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

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Abstract (different language)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgement

Even after eight years at university, it still feels *surprising* to see myself writing this PhD thesis and reaching this stage in my life. This thesis is for every LGBTQ+ child who grew up thinking the world is not for them.

I will forever be eternally indebted and grateful to the many, many people without whom I would never be here.

Dr. Michelle Cuthbert, your phenomenal commitment and passion for teaching had an impact on me that I will carry with me for the rest of my life. I still remember my first lesson of Year 12 physics, where you asked those of us intending to study physics at university to raise our hands. I think one or two hands were raised, of which I was not one. You proudly remarked that you would change our minds in the twelve months before we had to finalise our university and course applications. By the end of Year 13, a majority of our class went on to study physics – in no small part thanks to your teaching. I would never have studied physics – let alone go on to do a PhD – had I never been exposed to your infectiously passionate physics teaching.

Dr. Victoria Scowcroft, the summer project I did with you was a fantastic introduction to *Gaia*, statistics, and astronomy research that I cherish to this day. Your mentorship and guidance helped me to fall in love with doing research, and your humour helped me see how fun working in this field can be. I am eternally thankful for the sense of belonging in this field that you instilled in me.

Prof. Stijn Wuyts, thank you for your fantastic guidance and advice during my master's project. I had so much fun working on the project and I'm very thankful to you for introducing me to machine learning, something we worked together to discover more about. In particular, your scientific and methodical approach to solving problems is something I still aim for to this day. Your supervision and approach had a big impact on my abilities as a scientist.

To the PhD students at the University of Bath that I met during my final year – Abi, Anastasia, Bruno, Caroline, Charlotte, Eliot, and Nuria – thank you so much for adopting me into your lunch group and giving me so much insight. I learnt so much from you all about what a PhD entails and where I could go to do one. I would definitely have never applied abroad to do a PhD without you all.

To Loke and Laura, you are the best friends I could have ever wished for when moving abroad. I would never have kept going without your motivation, especially through the darkest times of the pandemic.

To my friends back in the UK – especially Claire, Jordan, Kate, and Mathy, along with everyone else I know from Backstage at the Bath Students' Union – thank you so much for always being amazing friends, and especially for all of the fun times we still managed to have even despite the pandemic. I miss you all constantly and I'm so happy I get to see you more regularly once again, even though it's still nowhere near often enough.

To the community at LSW and the PhD students in the IMPRS as a whole, thank you for being awesome. There are so many lovely people that I've been lucky enough to meet and learn from over the past four years. Thank you to all of you for the chats, the advice, and the fun. Especially thank you to everyone in the North building at LSW, who made me feel so welcome here right from my very first day.

From the bottom of my heart, thank you Zoe. You have been instrumental in motivating me to finish and brightening up every single day of my life since I met you here in Heidelberg. I'm so excited to see what the future holds, because I get to do it all with you.

Finally, thank you to Prof. Hans-Walter Rix for agreeing to be the second referee of this thesis, and thank you to Prof. Ralf Klessen and Prof. Matthias Bartelmann for agreeing to be the third and fourth examiners at my thesis defence.

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 CITEME, 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 and Lada 2003; Portegies Zwart et al. 2010). It is this property of the formation of OCs that makes them so useful: all stars in an OC will have the same age and initial composition, allowing parameters of the overall group of stars to 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 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 CITEME, as tracers of galactic structure (Cantat-Gaudin et al. 2020; Castro-Ginard et al. 2021), or even as calibrators of Cepheid variable stars (Medina et al. 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 et al. 2018), *Gaia* has revolutionised the study of our galaxy, including the study of OCs; with dozens of papers reporting thousands of new objects (e.g. Castro-Ginard et al. 2019, 2022, 2020; Liu and Pang 2019), and a number of works deriving dramatically improved parameters and members for OCs in the Milky Way (e.g. Cantat-Gaudin et al. 2018a; Tarricq et al. 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 et al. 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 (e.g. Castro-Ginard et al. 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 and Anders 2020; Piatti 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

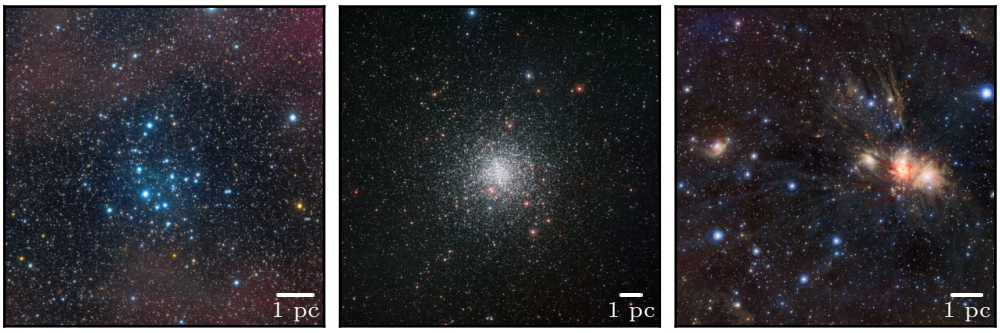

Fig. 1.1.: A visual comparison between the three main types of star cluster found in the Milky Way. *Left:* the open cluster NGC 2547. *Middle:* the globular cluster M 4. *Right:* the moving group/OB association Monoceros R2. All images contain a scale in the bottom right showing a length of 1 pc at the distance of each cluster. *Credit, left to right:* ESO / J. Pérez; ESO; ESO / J. Emerson / VISTA.

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. 1.3, I will discuss the history of OC observations up to before the release of *Gaia* DR2 in 2018, as well as briefly discussing the techniques and results from pre-*Gaia* observations. Section 1.4 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. Finally, Sect. 1.6 will briefly discuss some key pieces of theory surrounding the structure, dynamics, and lifetime of OCs, providing a good background on our theoretical knowledge of OCs 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 quickly discuss a 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 clusters, although I will also touch on globular clusters and moving groups. I differentiate between these three types of cluster approximately as follows, matching the observational definitions in Portegies Zwart et al. 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 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 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 will 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.

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,

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

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

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. Some MGs are also 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 1.1 and compared visually in Fig. 1.1. The figure shows three clusters; NGC 2547, M 4, 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 M 4 clearly contains significantly more stars. The stars in M 4 are older, with the cluster having a whiter, redder appearance. 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. 1.2). 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 adoptation of the reflecting telescope in the 17$^{\text{th}}$ and 18$^{\text{th}}$ centuries that led to catalogues of clusters like we use today.

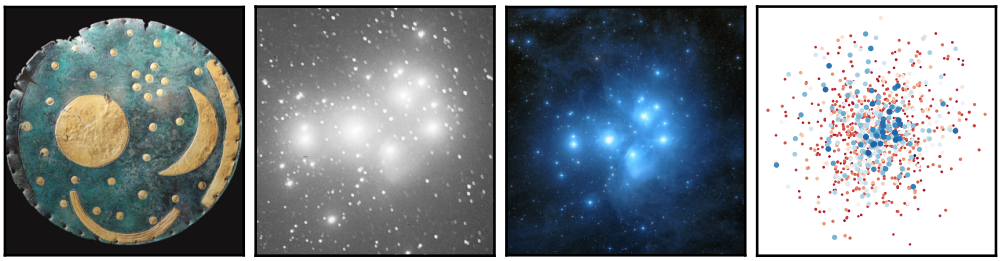

Fig. 1.2.: The Pleiades, as depicted throughout history and showing the clear improvements in astronomical data gathering over time. *Left:* the Nebra Sky Disc, depicting the Pleiades with its seven naked-eye visible stars in the upper center. The disc was discovered in 1999 in northern Germany and is 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 et al. 2018b. 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.

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 (e.g. Herschel 1786), with the number of known OCs jumping from a few dozen to around 700 in a little over a century. Figure 1.3 shows the evolution in size of OC catalogues over time, showing the peak of around 700 clusters by the turn of the 20$^{\text{th}}$ 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 20$^{\text{th}}$ century saw improvements to data gathering and techniques, with early photometric and spectroscopic methods allowing authors such as Rosenberg 1910 and Hertzsprung 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 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 used extensively in the present day as an essential tool to understand stellar evolution. Later, the differences in CMDs between different clusters were noticed and was interpreted as being a difference in age between the clusters, allowing for the ages of stars within star

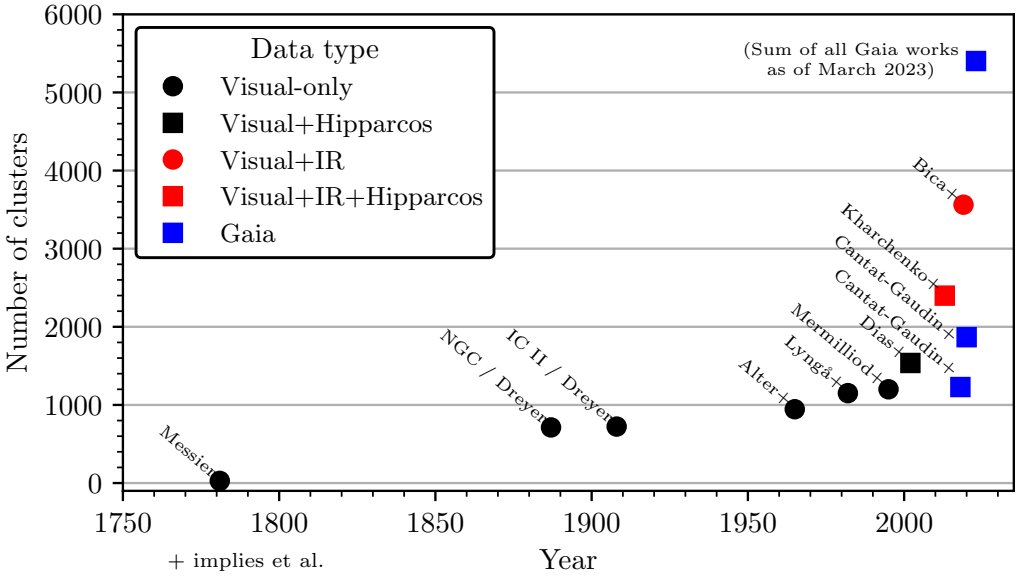

Fig. 1.3.: The size of OC catalogues over time. After the initial rise in the size of catalogues due to the advent of reflecting telescopes in the 18th and 19th centuries, it was not until the past 25 years and the advent of large-scale astrometric and IR datasets that the OC census significantly increased in size.
N.B.: this is not an exhaustive plot of all catalogues, and a number of old catalogues such as Herschel 1786 and Herschel 1864 without digitised versions are not included.

clusters to be estimated and beginning the foundation of our knowledge of stellar evolution 1.4.

While the 20th century saw huge strides in our understanding of stars and star clusters, the size of OC catalogues went relatively unchanged (Fig. 1.3). 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 et al. 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 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 et al. 1998 to discover a number of new OCs, with many being

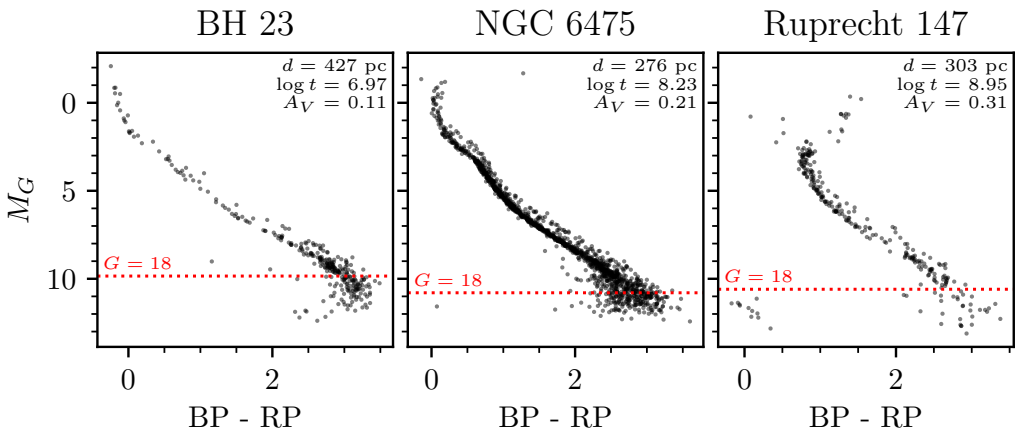

Fig. 1.4.: A comparison of the CMDs of a number of nearby OCs, using membership lists from later in this thesis in Sect. 3 and plotted with their absolute magnitude M_G against colour $BP - RP$. The OCs are plotted from left to right in order of increasing age, with their distance d, logarithmic age $\log t$ 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.

small objects near to the Sun that evaded detection with only two-dimensional visual observations.

This resulted in the catalogue of Dias et al. 2002 including over 300 more objects than the roughly ten years prior catalogue of Mermilliod 1995 (Fig. 1.3), 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.

The release of the Two Micron All Sky Survey (2MASS, Skrutskie et al. 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 Froebrich et al. 2007 to uncover over a thousand new OC candidates in the galactic disk, using IR data to peer through insterstellar 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 et al. 2007 using automated retrieval to extract cluster candidates. CITEME MORE REFERENCES HERE

Work predominantly with IR data culminated in the catalogue of Kharchenko et al. 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.

In around 20 years, the OC census more than doubled in size between the work of Mermilliod 1995 to the work of Kharchenko et al. 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 et al. 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 CITEME. 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 sample of stars, and also measures radial velocities and low-resolution spectra for a brighter subsample of sources (Gaia Collaboration et al. 2016). It is the incredible scale and precision of *Gaia* data that sets it apart from any previous datasets.

Figure 1.5 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

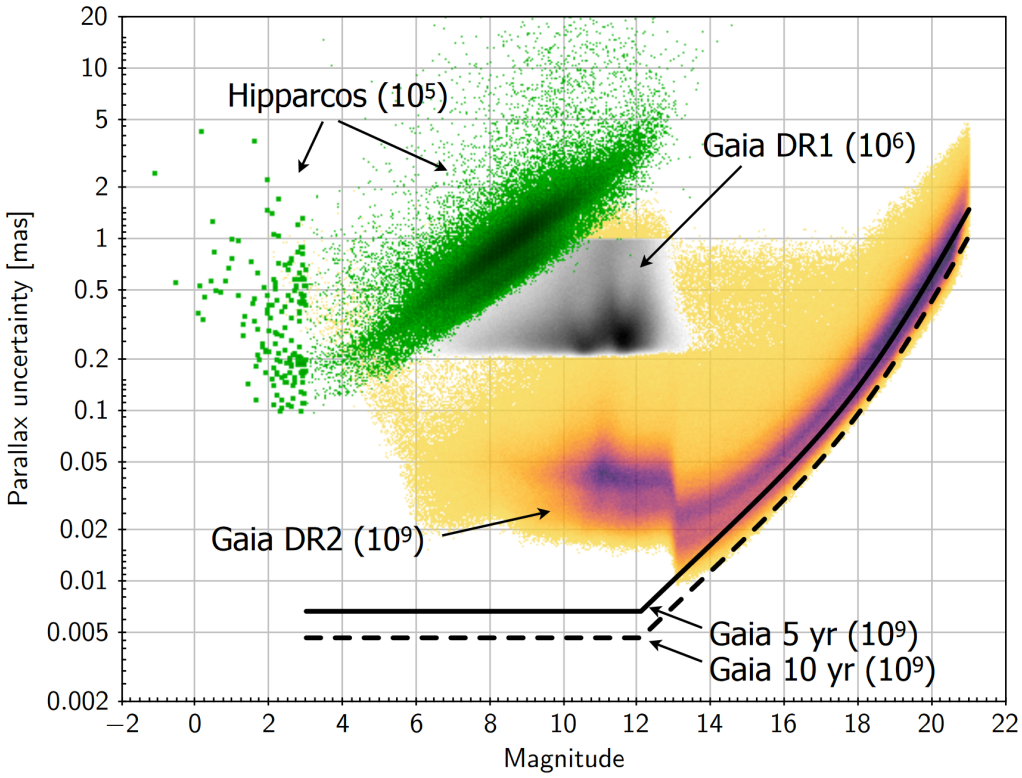

Fig. 1.5.: Comparison between the astrometric accuracy for all sources in the final data release of Hipparcos, *Gaia* DR1, and *Gaia* DR2. The predicted accuracy of future data releases using 5 and 10 years of data is shown by the solid and dashed lines respectively. *Credit*: *Gaia* DPAC.

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 and Tremaine 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 arcsecond 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 around 25 kpc CITEME, it is clear that measuring precise astrometry for even a small fraction of the stars in the galaxy requires an incredible level of precision.

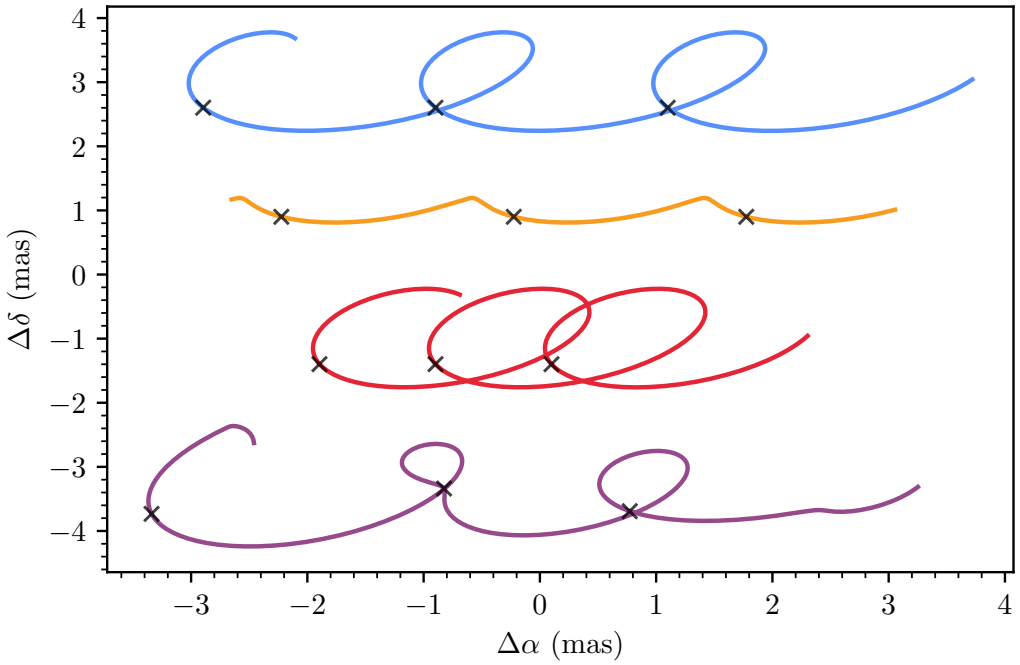

Fig. 1.6.: The predicted on-sky astrometric tracks of stars with different parameters, generated using astromet (Penoyre et al. 2022). All sources are at coordinates $\alpha, \beta = (0°, 45°)$, but are offset in the y direction for clarity of plotting. The first source has $\mu_{\alpha^*} = 2$ masyr^{-1}, $\mu_\delta = 0$, 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.

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 (particularly for binary stars) to be reconstructed to an exceptionally high level of precision for around 1 billion sources (see Fig. 1.6). *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 et al. 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 release of *Gaia* DR2. In many ways, *Gaia* is the perfect instrument for the study of OCs. Most OCs (such as NGC 2547 in Fig. 1.1) 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 et al. 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 (Cantat-Gaudin et al. 2018b).

Figure 1.7 shows the region around the high galactic latitude OC Blanco 1 in data from the Hipparcos-2 catalogue (van Leeuwen 2007), compared against the same region in data from *Gaia* DR3 (Gaia Collaboration et al. 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*ś $\sim 100\times$ 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 et al. 2018a produced an updated catalogue of OCs and OC membership lists, using pre-*Gaia* works such as Kharchenko et al. 2013 as input and trying to redetect their catalogued clusters in *Gaia* data. Cantat-Gaudin et al. 2018a were able to derive updated cluster membership lists with around twice as many members on average as in Kharchenko et al. 2013, as well as deriving cluster proper motions to around two orders of magnitude greater precision than in Kharchenko et al. 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*

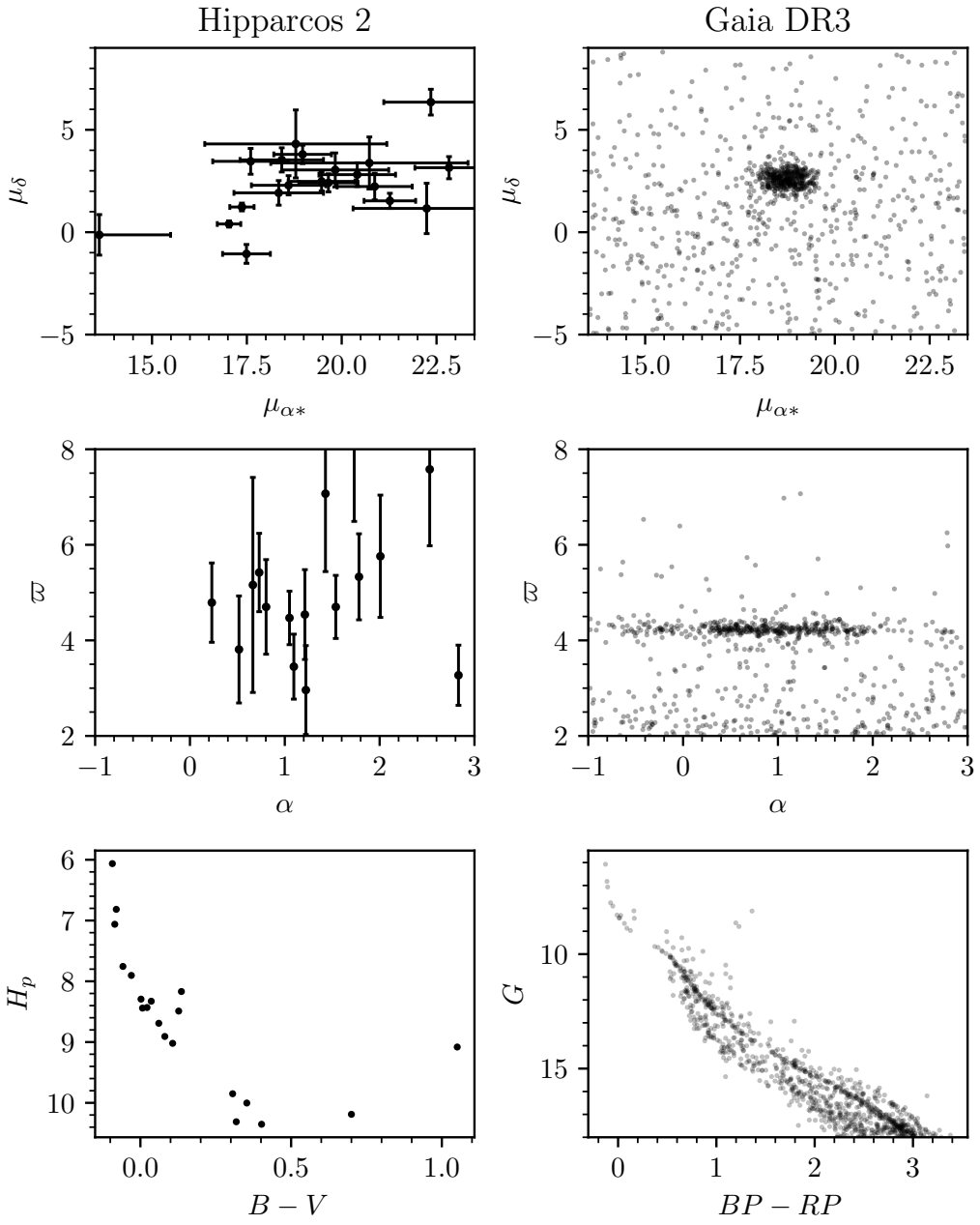

Fig. 1.7.: Comparison between the regions around the star cluster Blanco 1 in data from *Hipparcos* and *Gaia*. *Hipparcos*-2 data (van Leeuwen 2007) is shown on the left and *Gaia* DR3 data (Gaia Collaboration et al. 2022) is shown on the right. The top row shows proper motions, the middle row shows 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.

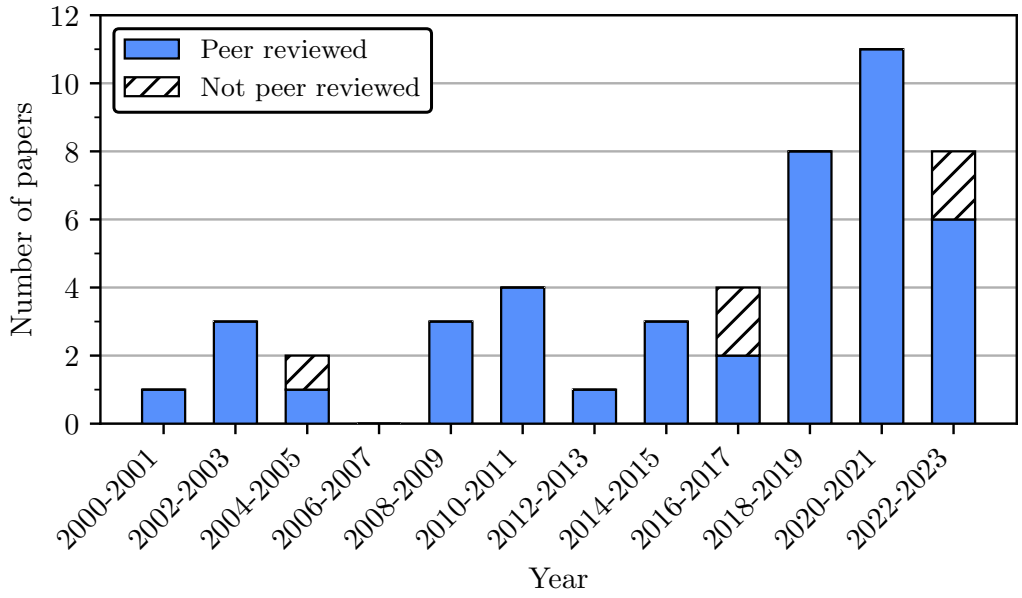

Fig. 1.8.: The approximate number of papers reporting new open clusters in the 21st century, shown as a stacked bar chart of peer reviewed and non-peer reviewed works. Data for 2022-2023 are incomplete.

data. This is clear in Fig. 1.3, with the catalogue of Cantat-Gaudin et al. 2018a containing around half as many clusters as Kharchenko et al. 2013. The reasons for the non-detection of such a large number of OCs remain unclear. How can so many objects catalogued before *Gaia* be undetectable?

Cantat-Gaudin and Anders 2020 provided some answers to this question, searching again in *Gaia* DR2 data for many 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 2022). Figure 1.8 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 et al. 2018) clearly corresponds with the number of papers reporting new OCs each year roughly doubling.

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 et al. 2007 refined this approach by using kernel 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 observational methods CITEME. 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 (e.g. Killestein et al. 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 and Reffert 2021), leveraging the latest computing hardware such as graphics processing units (GPUs) significantly more efficiently than previous methods.

While ML is not without caveats which will be discussed at length later in this thesis, ML has still been essential to the dramatic increase in newly reported OCs in the *Gaia* era. Castro-Ginard et al. 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,

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

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 et al. 2019 and a handful of works where small numbers of new OCs were noticed by mistake (e.g. Anders et al. 2022; Bastian 2019; Zari et al. 2018), 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 et al. 2019, 2022, 2020; Hao et al. 2022a; Hao et al. 2020; He et al. 2021; He et al. 2022a,b,c; Liu and Pang 2019; Qin et al. 2021; Qin et al. 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 et al. 2019 who used Gaussian mixture models (GMMs) and Jaehnig et al. 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 et al. 2013 believed that the OC census of 955 objects was largely complete within 1.8 kpc; 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 et al. 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 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.

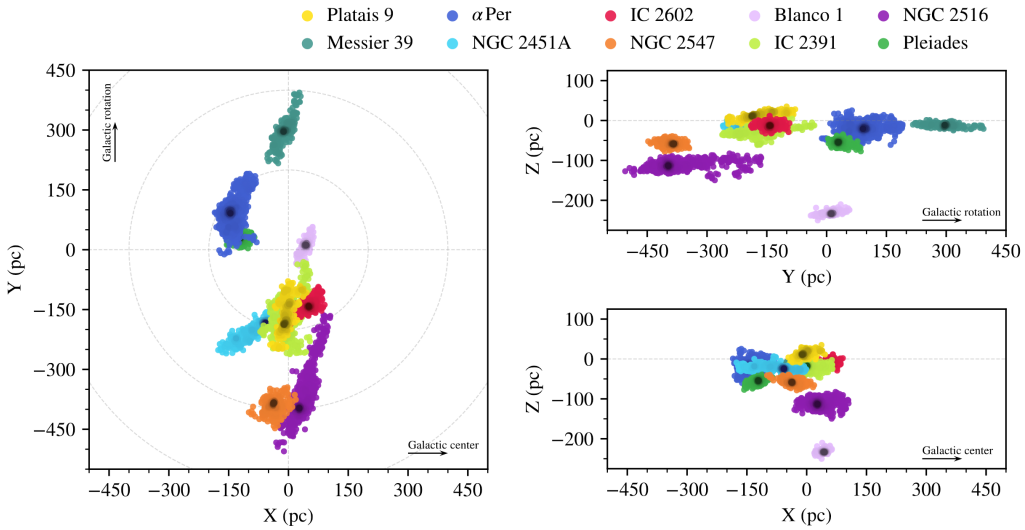

Fig. 1.9.: The detected tidal tails and comas of ten OCs near to the Sun. Clusters are shown as coloured density plots and plotted in heliocentric coordinates with the galactic centre to the right. *Credit:* Meingast et al. 2021

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 (Cantat-Gaudin 2022; Portegies Zwart et al. 2010), 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 et al. 2018 using *Gaia* DR1, and studied further by Lodieu et al. 2019, Röser et al. 2019, and Meingast and Alves 2019. Similar analyses have been performed on many more clusters, with Meingast et al. 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 from each cluster (Fig. 1.9). Tarricq et al. 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 et al. 2019, Oh and Evans 2020 and Pang et al. 2021 to study the dynamics of

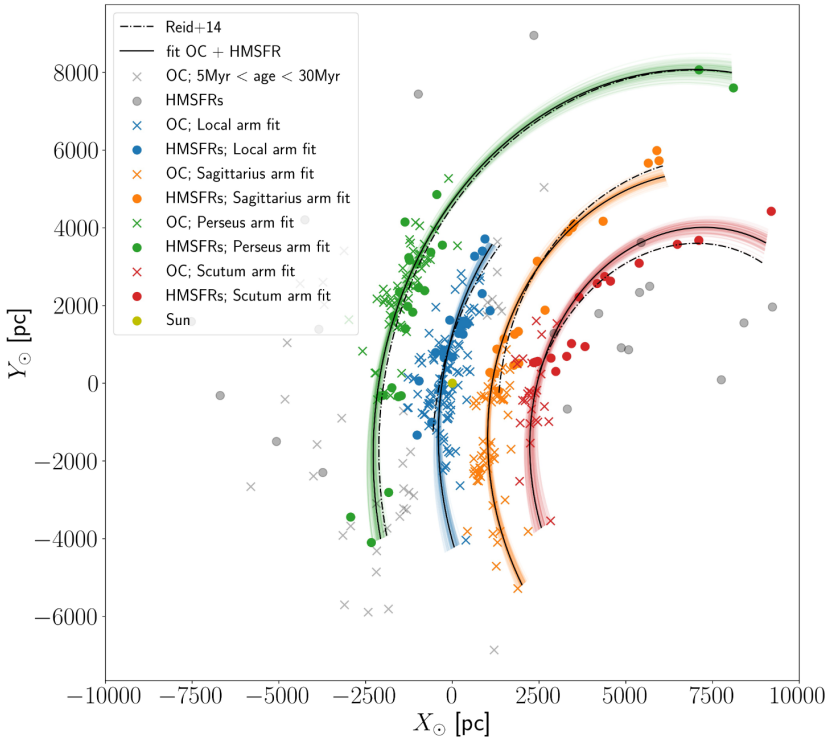

Fig. 1.10.: A model of the Milky Way's spiral arm structure as traced by OCs and high-mass star forming regions. *Credit:* Castro-Ginard et al. 2021

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 et al. 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 et al. 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 et al. 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. 1.10), finding that the addition of young OCs slightly changes the most likely spiral arm model relative to the fit of Reid et al. 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. 1.4). 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 et al. 2015). While BSSs have been extensively investigated in GCs, *Gaia* has allowed for many new investigations of BSSs in OCs (Cantat-Gaudin 2022), such as in Rain et al. 2020 who investigated BSSs in Trumpler 5, Trumpler 20, and NGC 2477, or Vaidya et al. 2020 who studied BSSs in a further seven OCs and found that BSSs are not mass-segregated in two of of the seven clusters they studied. Leiner and Geller 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 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 (e.g. Bastian and de Mink 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 et al. 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 (D'Antona et al. 2023; Milone and Marino 2022).

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 et al. 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 (Planck Collaboration et al. 2020) when using *Gaia* astrometry and Cepheid OC members. Works including Medina et al. 2021, Zhou and Chen 2021, and Hao et al. 2022b 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 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. 1.3 and 1.4, 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 et al. 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. 1.4, 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 and Tian 2015), despite the fact that publically available open-source implementations of these algorithms are often available and ready to use (e.g. Pedregosa et al. 2011).

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. 1.4.3), 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 (e.g. Castro-Ginard et al. 2018; He et al. 2021; Liu and Pang 2019). 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 et al. 2019, 2018, 2022, 2020 have 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 et al. 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 et al. 2022), meaning that current approaches will need to be ran on five to ten times as many sources[2].

[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 et al. 2016).

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 and Anders 2020; Cantat-Gaudin et al. 2018a). 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 and Anders 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 and Anders 2020).

While some studies have performed small investigations into clusters missing from *Gaia* on a case-by-case basis (e.g. Cantat-Gaudin and Anders 2020; Piatti 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 et al. 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 of these objects have not been independently verified Cantat-Gaudin 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 (e.g. Anders et al. 2020; Castro-Ginard et al. 2021). Given that so many clusters cannot be detected from recent works reporting new OCs before *Gaia*, with some works such as (Scholz et al. 2015) having as many as 100% of their clusters being impossible to redetect (Cantat-Gaudin et al. 2018a), 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 an 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. 1.8) can make the literature difficult to keep up with (Cantat-Gaudin 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 et al. 2023b (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. 1.4.3), 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 missing. Beyond the widespread disproving of the result in Kharchenko et al. 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 et al. 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 et al.

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 et al. 2019, 2018, 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. 1.5.1). 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 et al. 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 (Sect. 1.2 Portegies Zwart et al. 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 (e.g. Liu and Pang 2019), a neural network classifier (e.g. Castro-Ginard et al. 2018), or simple manual classification (e.g. He et al. 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 (e.g. Cantat-Gaudin et al. 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 (e.g. Liu and Pang 2019). Finally, all works adopt a different minimum size for an OC, ranging from as low as 8 stars (Castro-Ginard et al. 2018) to as high as 50 (Liu and Pang 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 and Anders 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 an analogous 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 50 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. 1.2 – requiring that an OC is an overdensity, a single stellar population, and gravitationally bound. However, no such way to transform observations into such a definition currently 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. 1.4. 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 and Tian 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* (E)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 (e.g. Cantat-Gaudin and Anders 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. 1.4.3), 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 et al. 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. 1.4). 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 et al. 2012) and are widely used in many areas of observational astronomy.

In practice, isochrones are difficult to fit, with age, extinction, distance and metallicity all being somewhat degenerate with one another. Figure 1.11 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.

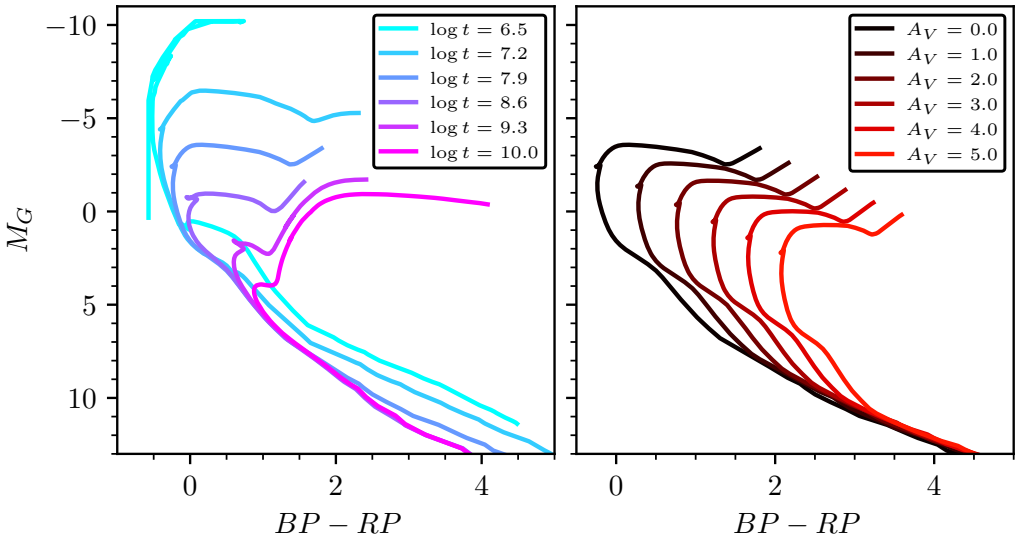

Fig. 1.11.: A comparison between stellar isochrones of various different parameters, derived from PARSEC stellar evolution models (Bressan et al. 2012) and shown in *Gaia* photometric bands. *Left:* isochrones of solar metallicity and zero extinction shown for six different ages. Most noticeably, as 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.

Isochrone fitting is further complicated by the presence of other cluster features, such as the presence of a binary sequence due to unresolved binaries (see binary sequences in Fig 1.4, 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, it was common to simply fit isochrones by hand (CITEME SOME EXAMPLES), which includes no robust uncertainty estimate and can open the door to human biases. The isochrones in Kharchenko et al. 2013 were fit using a hybrid method, with the authors fitting distances manually but then using χ^2 minimisation to fit cluster age and reddening values. Yen et al. 2018 developed this methodology further to perform χ^2 fitting of all cluster parameters. Finally, Hippel et al. 2006 created a full Bayesian methodology to fit isochrones to cluster CMDs, which is still used by works today in the *Gaia* era such as Bossini et al. 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 et al. 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 et al. 2020 and Kounkel et al. 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 with significantly less computational time, although this approach has the disadvantage of first requiring that isochrone fits are available to 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 et al. 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 (e.g. Cantat-Gaudin and Anders 2020; Cantat-Gaudin et al. 2018a). 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 et al. 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 1962) models are the most common models applied to star clusters. While originally derived for globular clusters, these models have also been shown to be a good fit to many OCs (e.g. in Piskunov et al. 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 $x_t = (r_t/r_c)^2$.

As an empirical model, the King 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 et al. 2013; Tarricq et al. 2022). King 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 1962 and King 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 et al. 2010). Due to the significantly simpler functional form of King 1962 models, they are used almost exclusively in the OC literature relative to King 1966 models (Cantat-Gaudin 2022; Portegies Zwart et al. 2010).

Finally, it is worth mentioning the model of Plummer 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 1962; King 1966 models, the Plummer 1911 model assumes star clusters do not have a physical limiting radius and extend to infinity, which is of course unrealistic. Nevertheless, the Plummer 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 which can be solved analytically (Dejonghe 1987). Plummer 1911 models are particularly popular in theoretical studies of star clusters due to this reason (Portegies Zwart et al. 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 et al. 2018 and Pang et al. 2021 have used similar methods on small scales to test if OCs are bound. The following useful definitions are all from Portegies Zwart et al. 2010.

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

$$T = \frac{3}{2} M \sigma_{1D}^2 \tag{1.3}$$

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

$$U = -\frac{GM^2}{2r_{vir}} \tag{1.4}$$

where G is the gravitational constant and r_{vir} is the theoretically definied virial radius of the cluster, a parameter that is difficult to calculate observationally as it requires three-dimensional positions. In practice, the virial radius can be defined as

$$r_{vir} = \frac{\eta}{6} r_{50} \tag{1.5}$$

where η is a constant that is model-dependent. For an ideal Plummer 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 1911 model (Portegies Zwart et al. 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

$$Q = \frac{T}{|U|} = \frac{\eta r_{50} \sigma_{1D}^2}{2GM} \approx \frac{1}{2} \quad \text{for a bound cluster.} \tag{1.6}$$

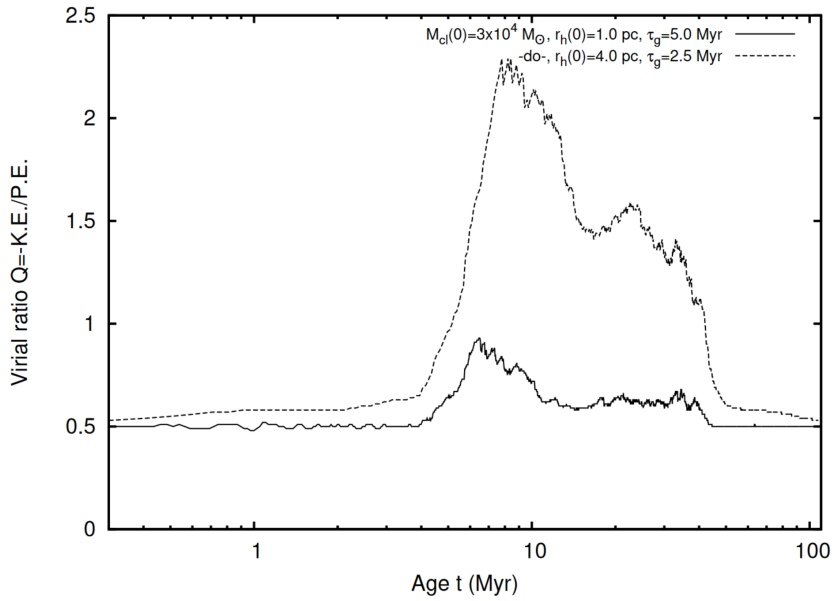

Fig. 1.12.: TODO

1.6.4 Timescales

1.6.5 Formation, evolution and destruction

1.7 The structure of this thesis

Chapter 1

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Chapter 1

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you

information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Chapter 1

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Chapter 1

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Chapter 1

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

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

> *Innovation distinguishes between a leader and a follower.*
>
> **— Steve Jobs**
> (CEO Apple Inc.)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

2.1 System Section 1

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there

no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Fig. 2.1.: Figure example: *(a)* example part one, *(c)* example part two; *(c)* example part three

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

2.2 System Section 2

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Fig. 2.2.: Another Figure example: *(a)* example part one, *(c)* example part two; *(c)* example part three

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look.

This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

2.3 System Section 3

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the

alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like

"Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

2.4 Conclusion

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

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 CITEME PAPER 2. 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 (Krause et al. 2020; Krumholz et al. 2019; Portegies Zwart et al. 2010).

Our view of OCs has always been complicated by their sparsity and their typical location in the galactic disk, making them challenging to isolate from field stars

along the line of sight (Cantat-Gaudin 2022). However, dramatically improved astrometric and photometric data from the *Gaia* satellite (Gaia Collaboration et al. 2016) are revolutionising our understanding of OCs and the overall Milky Way. Compared with the *Hipparcos* mission (Perryman et al. 1997), *Gaia* provides order of magnitude improvements in proper motion and parallax accuracy for around 10^4 times as many stars, with over 1 billion sources in total.

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

At the heart of all science with OCs, however, is the census of OCs itself. Particularly in the four years since *Gaia* Data Release 2 (DR2, Brown et al. 2018), many works have contributed major new insights into the census of OCs. Works such as Cantat-Gaudin et al. 2018b, Cantat-Gaudin and Anders 2020, and Jaehnig et al. 2021 provide new membership lists for OCs with a significantly higher number of stars and reduced outliers from the field when compared to pre-*Gaia* works. Thousands of new OCs have been reported using a range of unsupervised machine learning techniques, such as in Castro-Ginard et al. 2019, 2018, 2022, 2020, Cantat-Gaudin et al. 2019, or Liu and Pang 2019. The reliability of the census has also been improved, with works such as Cantat-Gaudin and Anders 2020 finding that a number of OCs discovered before *Gaia* are likely to be asterisms.

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

In Sect. 3.2, we describe the *Gaia* DR3 data used in this work and the quality cuts we adopted to filter out unreliable sources. In Sect. 3.3, we briefly recap our clustering method from Paper 1 and tweaks made to improve cluster recovery within 1 kpc. We then outline a method to validate cluster candidates using their photometry in Sect. 3.4, which we generalise to additionally infer ages, extinctions, and photometric distances to our clusters in Sect. 3.5. In Sect. 3.6, we crossmatch our catalogue against literature works. Section 3.7 presents an overview of our catalogue. We discuss the non-detections of some literature clusters in Sect. 3.8, and discuss required steps a future work will take to improve the reliability of our new cluster candidates in Sect. 3.9. Section 3.10 summarises this work.

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

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

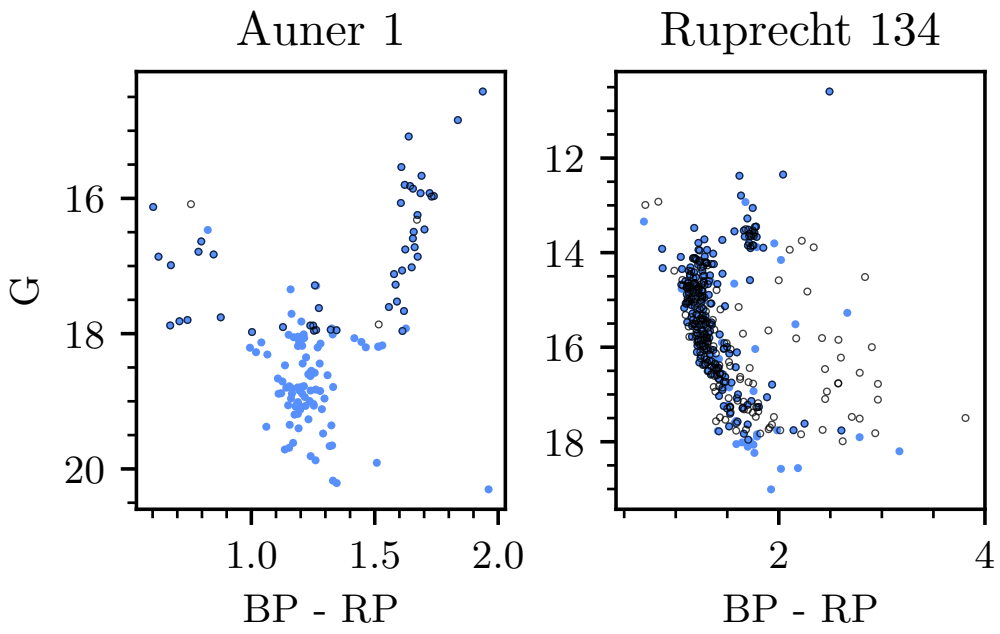

Fig. 3.1.: Comparison of cluster membership lists detected using *Gaia* DR3 data cut at $G < 18$ (black empty circles) and a Rybizki et al. 2022 v1 criterion greater than 0.5 (blue filled circles) using separate runs of HDBSCAN and our pipeline for each cut, shown for Auner 1 (left) and Ruprecht 134 (right).

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

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

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 (Lindegren et al. 2021a; Penoyre et al. 2022), as well as causing issues with photometric measurements due to blending (Golovin et al. 2023; Riello et al. 2021).

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

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

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

As a visual example, the CMDs of Auner 1 and Ruprecht 134 from clustering analyses using this cut and a $G < 18$ cut are compared in Fig. 3.1. Auner 1 is a distant and

difficult to detect cluster, for which only 51 stars are detected in the $G < 18$ trial for a cluster S/N of 10.8σ. 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 (including the nearby UFMG 88 reported by Ferreira et al. 2021) remained with a similar membership list to with no cut at all. In addition, the lack of a magnitude cut means that in sparse fields where faint sources have reliable astrometry, clusters such as the high galactic latitude Blanco 1 have membership lists down to fainter than $G \sim 20$, two magnitudes fainter than the membership list of Cantat-Gaudin and Anders 2020 for this cluster.

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

In total, 729.7 million sources in *Gaia* DR3 have a Rybizki et al. 2022 v1 quality flag of at least 0.5 and were selected for further clustering analysis in this work. This represents significantly more sources than the 301.7 million sources with $G < 18$, a cut adopted in works such as Castro-Ginard et al. 2022 or Cantat-Gaudin and Anders 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 et al. 2013 and Cantat-Gaudin and Anders 2020.

As in Paper 1, the HEALPix (Hierarchical Equal Area isoLatitude Pixelation) tessellation scheme was used to segment the entire *Gaia* dataset (Górski et al. 2005), with calculations performed by the Python package `Healpy` (Zonca et al. 2019). This has advantages over other methods to subdivide spheres into a finite number of regions, in that all regions at a given tessellation level have the same area, and spherical distortions are minimised. However, unlike in Paper 1, the origin of the HEALPix grid was set at the origin of galactic coordinates ($l, b = (0°, 0°)$), 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 et al. 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 (Campello et al. 2013; McInnes et al. 2017) is the most sensitive for recovering OCs in *Gaia* data.

Briefly, HDBSCAN is an updated version of the DBSCAN algorithm (Ester et al. 1996), for which only a minimum cluster size m_{clSize} and minimum number of points in the neighbourhood of a cluster core point m_{Pts} must be specified, unlike DBSCAN which instead uses m_{Pts} and a minimum, global distance between points in a cluster ϵ. DBSCAN has seen much use in the literature so far for OC recovery, such as in Castro-Ginard et al. 2019, 2018, 2022, 2020 or He et al. 2021; He et al. 2022a. 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 and Covey 2019 and Kounkel et al. 2020, as well as being used to find 41 new OCs in Paper 1, and being used by Tarricq et al. 2022 to reveal new tidal tails and comas of numerous OCs within 1.5 kpc. HDBSCAN has not yet been used to conduct a search through all *Gaia* data for OCs.

A major flaw of HDBSCAN, however, is its high false positive rate. In Paper 1, we show that this is due to the algorithm being overconfident, reporting dense random fluctuations of a given dataset as clusters. To mitigate this, we adopt the cluster significance test (CST) from Paper 1, which searches for field stars surrounding a cluster and compares the nearest neighbour distribution of cluster stars with that of field stars. This then produces a signal-to-noise ratio (S/N), with CST scores greater than 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 (e.g. as in He et al. 2022a; Kounkel et al. 2020), although a major issue is that converting *Gaia* parallaxes to distances is non-trivial and results in asymmetric errors and non-Gaussian parameter distributions (Bailer-Jones et al. 2021). Instead, we use the approach adopted in Paper 1, similar to that of works such as Castro-Ginard et al. 2018 and Liu and Pang 2019. We use *Gaia* positions, proper motions, and parallaxes directly, but with two preprocessing steps: firstly, recentring them into a coordinate frame with an origin at the centre of each respective field, which removes spherical distortions present at high declinations; secondly, rescaling all five axes of the dataset to have the same median and interquartile range, effectively removing the units of each axis of the data. Particularly for HDBSCAN, which can cope with varying density datasets, the choice to use these five simple recentred and rescaled features was found to have no 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 et al. 2021 were used to convert positions and parallaxes to a Cartesian coordinate frame, with proper motions converted to tangential velocities. At such small distances, the uncertainties in Bailer-Jones et al. 2021 are small and not prior-dominated, and so reliance on *Gaia*-derived distances for the single nearby field should not cause any issues.

3.3.2 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 3.2 shows histograms of the S/Ns of detected clusters, showing a clear spike in count for S/N < 0.5 and an increasing trend in S/N for S/N $\lesssim 3$ that deviates from the relatively straight log-linear relation in S/N present for S/N > 3, suggesting that an additional component of false positives is contributing to the otherwise log-linear component of reliable astrometric clusters at low S/Ns. This figure, our results from Paper 1, and the poor quality of the low-S/N clusters we detect strongly support that most low-S/N clusters are false positives; however, exactly where to set an S/N threshold is a non-trivial decision that has a large effect on the rest of the catalogue. A catalogue can choose to prioritise completeness, having a low threshold and including as many true positives as possible, but while inevitably including many false positives and sacrificing precision; or, a catalogue can do the opposite, having a lower completeness but also minimal false positives and maximised reliability of all objects in the catalogue.

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

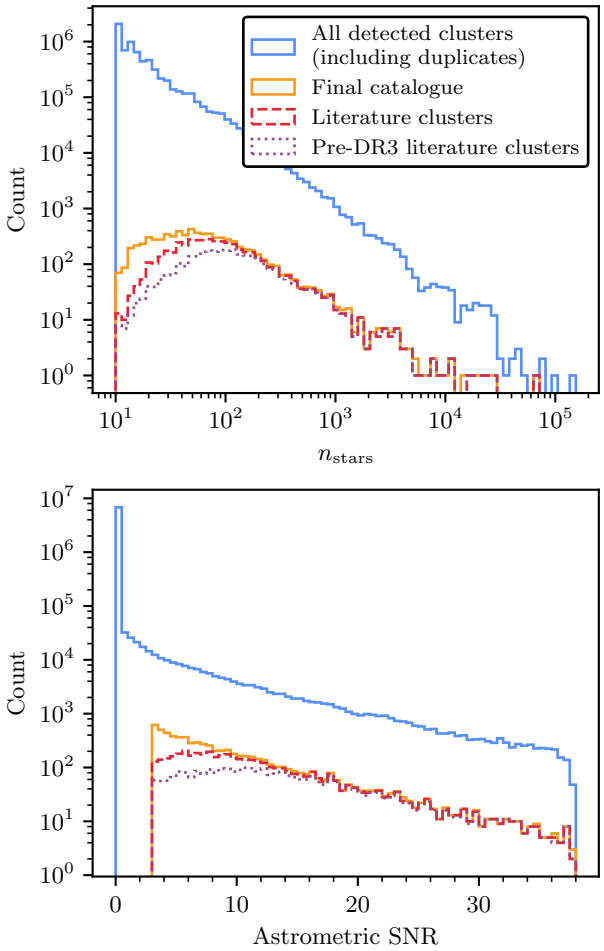

Fig. 3.2.: Statistics of all detected clusters compared against the final catalogue. *Top*: distribution of the number of member stars of detected clusters, n_{stars}, for all detected clusters in all fields before catalogue merging and duplicate removal (solid blue line), for the final catalogue (solid 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. 3.8.1, 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. 3.10.

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 et al. 2018b, 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 and Anders 2020 to be lost), it is nevertheless important to address to ensure the final catalogue is as complete as possible.

To mitigate this effect, clusters near to the edge of a central pixel were also kept. After extensive testing, it was found that cluster positions generally vary by no more than ~ 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 (see e.g. Anders et al. 2022; Kovaleva et al. 2020) 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 et al. 2022 attempted to derive structural parameters for 467 OCs within 1.5 kpc, but their method (based on fitting **king_structure_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_1962** definition of tidal radius as the radius at which a cluster's density begins to exceed that of the density of the field, but is model-independent and can be easily and efficiently computed for the entire catalogue by selecting the radius at which a cluster has the highest CST against field stars. For instance, for well-defined clusters such as the Pleiades and Blanco 1, this radius was found to exclude cluster tidal tails while corresponding well with literature tidal radius values in Kharchenko et al. 2013 (see Sect. 3.7 for a discussion of our cluster radii.)

Mean parameters such as mean proper motion and parallax were then calculated given the members within the cluster's estimated tidal radius, in addition to maximum likelihood cluster distances calculated using the method of Cantat-Gaudin et al. 2018b. To calculate more accurate distances, the parallax bias of member stars was corrected using the method in Lindegren et al. 2021b, which improved the accuracy of cluster distances particularly for distant clusters. As the Lindegren et al. 2021b parallax correction can only be applied for certain parameter ranges, for six clusters, too few sources (or no sources) had available corrections, and so we applied a simple global offset of $\varpi_0 = -17$ μas as derived in Lindegren et al. 2021b. These six clusters are flagged in the final catalogue as having less accurate distances. Overall, although the Cantat-Gaudin et al. 2018b 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 et al. 2011, who use Hubble Space Telescope parallaxes to a subset of Hyades member stars to derive its distance.

In addition, **king_structure_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_1962** (Eqn. 18) given $n(r_{50})$ and r_t. While approximate and less accurate than full Markov chain Monte-Carlo (MCMC) fits such as those performed in Tarricq et al. 2022, these

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.

core radii still provide a good approximation of a **king_structure_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 et al. 2019, 2018, 2022, 2020 successfully use neural networks to classify candidate clusters as real or false with their photometry, and while Cantat-Gaudin et al. 2020 and Kounkel et al. 2020 use neural networks to infer the ages, extinctions, and distances of their catalogued clusters, all of these works rely partially or entirely on existing examples of OCs detected in *Gaia*.

While such an approach mitigates issues with simulated training data, namely that stellar isochrones such as Bressan et al. 2012 are typically an imperfect fit to the observed CMDs of OCs (Cantat-Gaudin et al. 2020), it is difficult to guarantee that a small training dataset that relies mostly or entirely on examples of OCs from *Gaia* accurately covers a full range in parameters such as absolute extinction, differential

extinction, distance, metallicity, and age. In particular, due to the different cuts on *Gaia* data used in this work, we often detect significantly more member stars for many clusters and up to two magnitudes fainter than the membership lists of Cantat-Gaudin and Anders 2020; hence, particularly for more distant OCs, our membership differences have a significant impact on inferred parameters, making existing literature catalogues inappropriate to use as training data. Simulated data, if it can be simulated accurately enough, would offer an attractive way to quickly generate new training data applicable to new methodologies and new *Gaia* datasets or even other instruments, entirely based on a ground truth or 'best estimate' of how OCs should appear based on prior knowledge from stellar evolution models. Additionally, training data based on real clusters are biased towards an unknown selection effect of how a human defines a real cluster – whereas for simulated data, we are able to exactly state the distributions we assume real OCs are drawn from, hence giving more knowledge of any selection biases this may cause.

A key issue found in early experiments is that typical machine learning approaches are deterministic, and hence do not quantify the underlying uncertainties on their predictions. To aid with the use of simulated data, we adopt an approximate Bayesian neural network (BNN) framework using variational inference. In practice, true Bayesian machine learning is impractical to achieve with current methods; however, variational inference-based approaches offer an approximate and fast way to estimate the uncertainty of a neural network model by approximating parameters with simple probability distributions (Goan and Fookes 2020; Jospin et al. 2022), of which networks can then be sampled multiple times to produce a probability distribution for their output. The BNN approach we trialed had similar accuracy to a purely deterministic one except while also outputting uncertainties, allowing us to estimate the uncertainty of our classifier. We provide a broader overview of our adopted variational inference-based approach in Appendix **??**. 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 et al. 2020) to simulate single-population clusters from PARSEC evolution models (Marigo et al. 2017), with extinction calculated star-by-star using a Cardelli et al. 1989 extinction law with $R_V = 3.1$. Stars were sampled from these isochrones with SPISEA using a Kroupa 2001 IMF. In addition, SPISEA was used to supplement simulated OC CMDs with unresolved binary stars based on general relations derived in Lu et al. 2013 for

zero-age star clusters. The values in this work were found to correspond relatively well to *Gaia* observations, with a mass-dependent multiplicity frequency peaking at 100% for clusters of masses above $5\ M_{\odot}$. In practice, unresolved binary stars have negligible impact on the final cluster CMDs fed to the network, as typical binary sequences observed in *Gaia* photometry are smaller than the size of the pixels in input CMD images. SPISEA was also used to apply Gaussian-distributed differential reddening, with values up to a standard deviation of 0.6 in the highest cases, reflecting the most extreme examples of differentially reddened reliable clusters found in Cantat-Gaudin and Anders 2020.

Next, a random location on the galactic disk was selected for each cluster, which was used to simulate a realistic selection function and photometric errors. The magnitude-dependent selection function of *Gaia* DR3 at each given location was queried using the `selectionfunctions` package presented in Boubert and Everall 2020 and Boubert et al. 2020, which gives the basic probability that a source appears in *Gaia* as a function of position and G-band magnitude. We use the online version of their package updated for *Gaia* DR3. The `selectionfunctions` package is based on the `dustmaps` package from Green 2018. In addition, the selection function of every cluster was also corrected for the cuts to *Gaia* data applied in Sect. 3.2.2. During the preparation of this work, Cantat-Gaudin et al. 2023 released a new selection function for *Gaia* DR3 which suggested that the earlier work of Boubert and Everall 2020; Boubert et al. 2020 can be over-confident at the faint end; however, given that our cluster membership lists are overwhelmingly dominated by the selection function of our cuts on *Gaia* data at magnitudes $G > 18$, and not the pure selection function of *Gaia*, we found that it made too small of a difference to our simulated clusters to be worth updating our training data for, although we will adopt their work in future works. Realistic photometric uncertainties were added to sources based on the distribution of source uncertainties at the selected location, which are generally larger in crowded fields. We added systematic offsets in simulated BP and RP *Gaia* photometry for faint sources using relations in Riello et al. 2021.

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

10 000 examples of simulated real clusters were generated to use as one half of the simulated cluster dataset. Distributions of parameters such as age $\log t$, extinction A_V, differential extinction 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 and Anders 2020 (corrected for our deeper magnitude limit) was found to work well to produce an unbiased classifier, in other cases, such as for $\log t$ and $m - M$, the use of a uniform distribution (instead of one based on the expected distribution of clusters) was essential to avoid biasing the classifier towards certain ages or distances. These distributions are listed in Table 3.1.

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 *Gaia* dataset – $610\,000$ in the final version of the catalogue, or fewer than 0.1% – that it was not found to be necessary to first remove them from data used to generate false positives. This is similar to the false positive generation method used in Castro-Ginard et al. 2022.

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.

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 3.2.

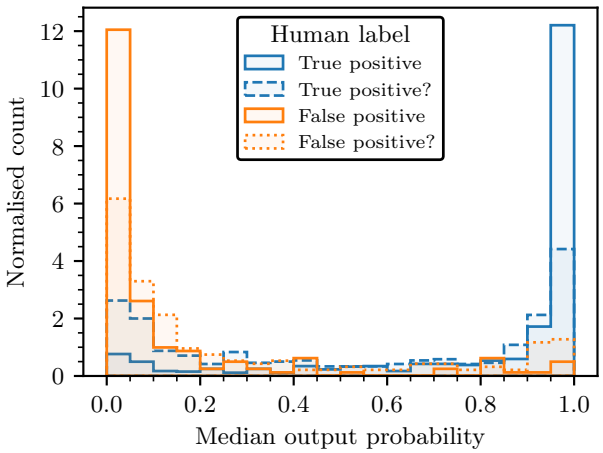

Fig. 3.3.: Performance of the CMD classifier on the independent test dataset of 2000 clusters detected by HDBSCAN in *Gaia* data and labelled by hand. Clusters are labelled as true positives or false positives, with clusters where the human classifier was less certain being additionally flagged.

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. 3.4.2), which can even fool a human classifier. This is particularly common in the halo and thick disk where most stars have a similar, old age. This is an important limitation of the human-classified test data to bear in mind, as a small fraction of clusters classified by hand as true positives will always in fact be false positives. Nevertheless, CMD classification is still a necessary validation tool to help ensure that detected cluster candidates are reliable, as many of the worst quality clusters can still be removed with this method.

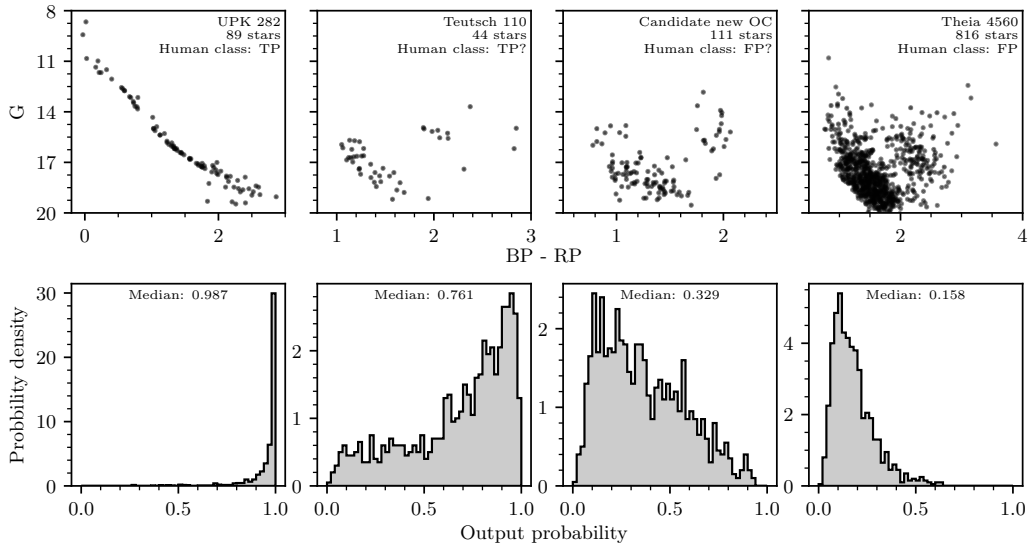

Fig. 3.4.: Four examples of classified cluster CMDs from the test dataset, with cluster CMDs on the top row and their PDFs of predicted probabilities on the bottom row. Cluster names and human-assigned labels are indicated on the figures. PDFs are generated by sampling the CMD classifier 1000 times for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

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

As input, the optimal network trialed used cluster CMDs converted to absolute magnitudes, with stars of absolute G magnitudes greater than 10 or lower than -2 cut away. Generally, this cuts certain very low mass M stars and bright O stars from cluster CMDs, which were found to be poorly simulated by PARSEC isochrones with

their inclusion only worsening network performance on real data. In practice, very few stars are cut due to this limitation, with O stars making up only a very small proportion of sources in young clusters and M dwarfs fainter than $M_G = 10$ only being brighter than $G = 20$ for clusters within 1 kpc, at which point the rest of the cluster CMD can be resolved well. In addition, $BP - RP$ colours were cut between -0.4 to 4, which in practice is a wide enough colour range to include almost all sources but while providing a good range to discretise cluster CMDs between. Sources with very low BP and RP fluxes that have overestimated BP or RP magnitudes were removed using cuts from Riello et al. 2021, as these also only confused the network, despite these systematics being simulated in the training data. Finally, in terms of structure, the optimal 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 120 filters respectively. Max pooling layers were placed between these convolutional layers to speed up training and inference. Convolution layer output was connected to a single densely connected layer of 128 nodes, with a final single node for output. The distance modulus of the cluster based on the parallax-derived cluster distances was also fed to the network as an auxiliary input into the 128 node dense layer, in a similar way to the network of Cantat-Gaudin et al. 2020 which also uses both photometric and astrometric input simultaneously. All layers used Rectified Linear Unit (ReLU) activation other than a sigmoid activation function applied to the final output to constrain network output in the range $[0, 1]$ as a probability distribution.

The final network had binary accuracies (the percentage of clusters given the correct true or false label) of 95% for both training and validation data, indicating that the network did not overfit to training samples when compared with other simulated data. Fig. 3.3 shows the performance of the network compared to the human-labelled test dataset of real clusters detected by HDBSCAN in *Gaia* after sampling the network 1000 times to generate PDFs for every object, with 85.5% of clusters labelled highly likely to be real and 91.3% of clusters labelled highly unlikely to be real having a median predicted probability greater or less than 0.5 respectively. Clusters where the human classifier was less certain have a much broader distribution, although this also reflects inherent uncertainties in the test dataset discussed in Sect. 3.4.3. Finally, only 4.3% and 2.5% of highly likely real and highly likely false clusters had predicted labels that disagree with human labels at more than the 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. 3.4.3. Instead, these data show that the CMD classifier can perform comparably well to human classification, except with the added bonuses of speed and reproducibility.

Fig. 3.4 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 et al. 2020, although the overdensity as detected in this work does not appear to contain a homogeneous population of stars and is hence classified weakly. CMD classifier median probabilities and confidence intervals for all clusters are listed in Table 3.4, 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 et al. 2020 use a neural network to infer $\log t$, A_V and $m - M$ for around 2000 OCs. In their work, a training dataset based on simulated OCs alone is not found to be sufficiently accurate to train a neural network. While simulated data were found to be accurate enough for the CMD classifier in Sect. 3.4, parameter inference is more challenging, as a network must learn to infer multiple parameters from a CMD alone and generalise this accurately to real data. However, our approach has a number of differences to theirs: firstly, we use a convolutional neural network, which may be better able to capture structure in CMDs due to its 2D approach, which may also reduce training data overfitting; secondly, our network is approximately Bayesian, and includes uncertainty estimates that quantify when it may have failed; finally, although Cantat-Gaudin et al. 2020

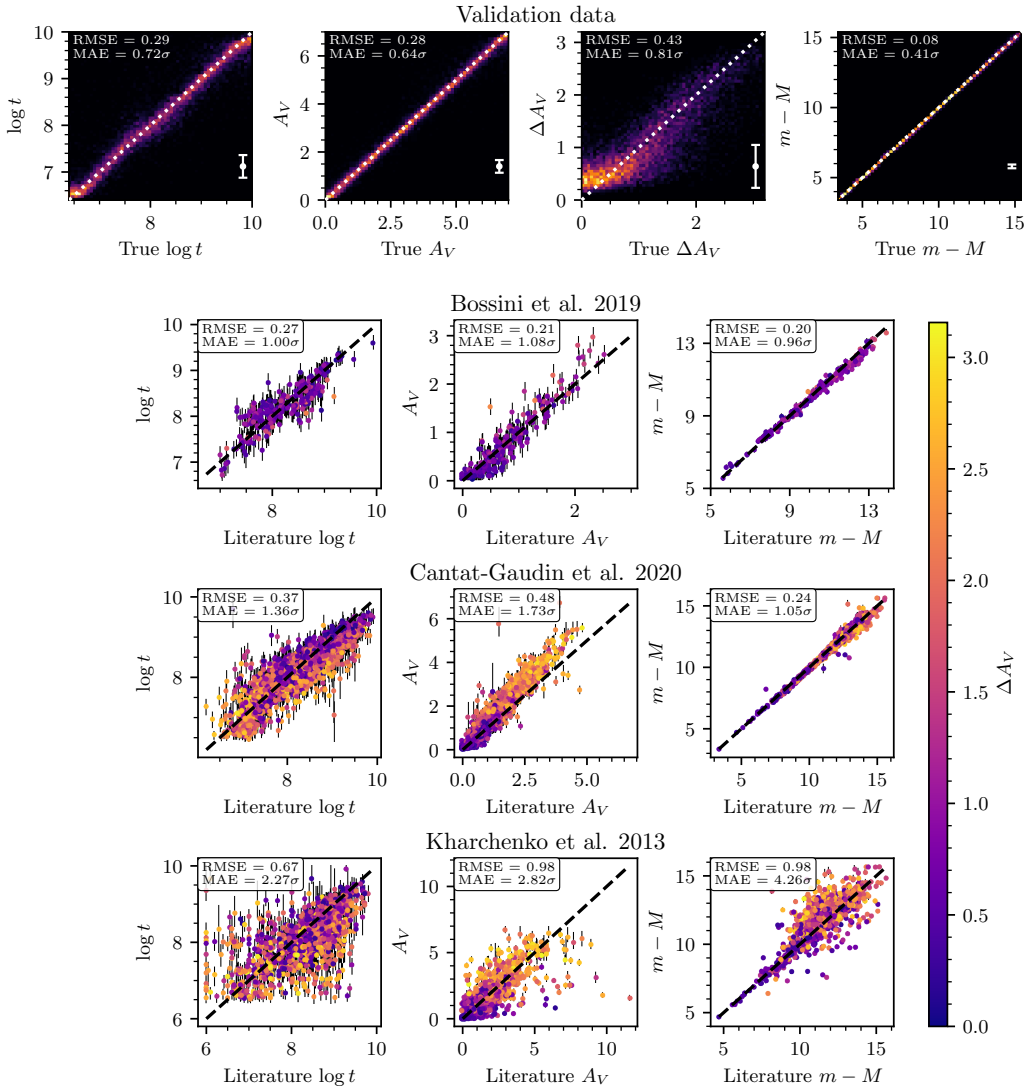

Fig. 3.5.: Photometric parameters derived in this work compared against test datasets. *Top row*: 2D histograms showing the performance of the trained photometric parameter inference network on all 10 000 clusters from the validation dataset. The mean output uncertainty is shown with white error bars. 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 et al. 2019, Cantat-Gaudin et al. 2020, and Kharchenko et al. 2013 respectively. Points are shaded based on the differential extinction we infer for each cluster.

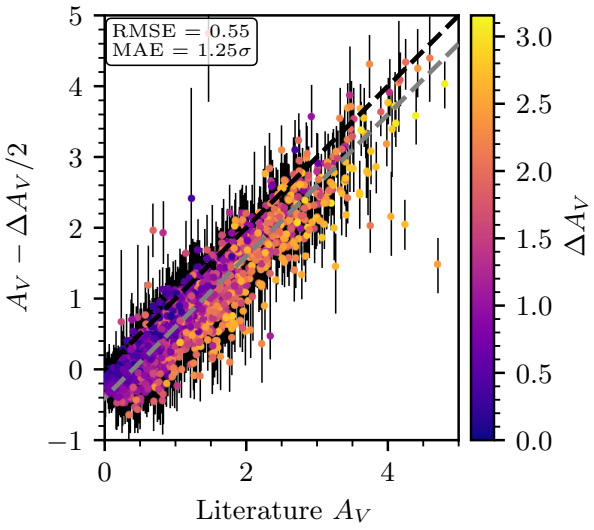

Fig. 3.6.: Extinction values from Cantat-Gaudin et al. 2020 compared against this work when corrected for differential extinction with an estimate of cluster differential extinction, plotted in the same style as Fig. 3.5. The dashed black line shows where y values equal x ones; the dashed grey line shows the same but offset by -0.4.

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. 3.4.

To create a parameter inference network, we used a similar network structure to that of Sect. 3.4.4, except with some tweaks to the network output to infer parameters. To better predict the aleatoric uncertainty of network output for this multiple-parameter network, network output was changed to a beta distribution for each parameter. These distributions can take any shape from a uniform (completely uncertain) distribution to a single point-like estimate. The output was then scaled to be within the minimum and maximum ranges of the training data. To train the network, 50 000 simulated clusters were created using the same methodology as in Sect. 3.4.1, changing the distribution of cluster extinctions A_V (as defined in Table 3.1) to simply be uniform between 0 and 7.

In initial comparisons with literature results, differential reddening was found to strongly correlate with disagreements in extinction (and to a lesser extent, age) between this work and others. A primary cause of this is that while many works (e.g. Bossini et al. 2019; Cantat-Gaudin et al. 2020) use the so-called 'blue edge'

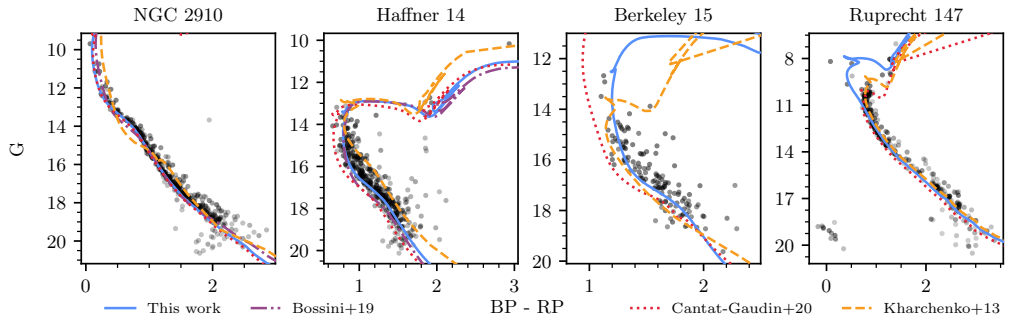

Fig. 3.7.: Predicted cluster isochrones from this work (solid blue line) compared with those from other works. Cluster members are plotted in black and shaded according to their membership probability.

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 and Anders 2020, which are constructed with the UPMASK algorithm (Krone-Martins and Moitinho 2014) and for many clusters only select stars in the core. This means that our CMDs are constructed from clusters with significantly larger angular extents on the sky and are hence often more strongly differentially reddened than in Cantat-Gaudin and Anders 2020, with many clusters having a blue edge at an extinction value up to 1 magnitude lower than in Cantat-Gaudin and Anders 2020. For instance, NGC 884 is an example of this, with our membership list being larger and more strongly differentially reddened. A blue-edge based definition of A_V means that different works 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_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 et al. 2013.

Against our validation dataset of $10\,000$ simulated clusters, the network performs well with no clear systematics in $\log t$, A_V or $m - M$. However, owing to the degeneracy between 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. 3.5.

Using the best trained network after a number of experiments, all clusters in our catalogue closer than a maximum distance of 15 kpc have ages, extinctions, differential extinctions, and distance moduli listed in Table 3.4. 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. 3.7 shows example predicted isochrones for four OCs in this work. In the first case, NGC 2910 is a cluster with a well-behaved isochrone where all works agree relatively well. On the other hand, Haffner 14 shows relatively strong differential reddening, and different definitions of differential reddening between different works cause isochrone fits to disagree. Berkeley 15 is a sparse cluster where both differential reddening and field star outliers affect different works in different ways, with our updated *Gaia* DR3 membership list having fewer outliers than that of Cantat-Gaudin et al. 2018b. 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. 3.5 compares all network predictions with values from four test datasets. An advantage of our simulated training approach is that network predictions can now be compared to other literature works, which act as independent test datasets which can verify the accuracy of our network. It is important to note that our results never agree perfectly, however, particularly since all works we compare to are based on *Gaia* DR2 or pre-*Gaia* OC membership lists that may be significantly less clean or have significantly fewer stars than our *Gaia* DR3 membership lists.

Bossini et al. 2019 provide a catalogue of precise OC parameters from Bayesian isochrone fitting using the BASE-9 algorithm (Hippel et al. 2006). A key difference is that their work uses metallicity estimates from the literature where available, whereas our approach is based entirely on *Gaia* DR3 parameters and assumes a given cluster can have any metallicity as drawn from a broad probability distribution based on literature values (Table 3.1). Nevertheless, our results still agree well with theirs in $\log t$, A_V and $m - M$. In cases where our $\log t$ estimates disagree most strongly, this is typically due to differences in OC membership list. There is however a possible minor systematic between our two works for OCs with extinctions below 0.6, many of which we infer smaller extinctions for than them; this may be as a result of A_V vs. metallicity degeneracies. However, their values are typically only 1 to 2σ from ours.

Our parameters agree less strongly with the results of Cantat-Gaudin et al. 2020, which are derived from a neural network trained on isochrone fits from a variety of works (including Bossini et al. 2019). This is to be expected to some extent, as while Bossini et al. 2019 only fit isochrones to a subset of OCs with clean membership lists and the least differential reddening, Cantat-Gaudin et al. 2020 fit isochrones to all known OCs at the time, including many sparse objects which may now have significantly different membership lists in our current *Gaia* DR3 work. However, some differences persist. A clear systematic in our and their A_V values is clear, although this is likely due to their different blue edge definition of extinction (whereas our network fits to the mean extinction in a cluster.) Figure 3.6 shows a crude conversion between our A_V values and their blue-edge A_V values. While this removes the systematic difference in gradient, our converted A_V values are still generally smaller than theirs by around 0.4 to 0.5 on average. This is likely due to two effects; firstly, as shown by the results on validation data, 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 et al. 2020, meaning that our clusters

are often larger and hence are more extremely differentially reddened between separate sides of the cluster; hence, a conversion between the works based on our 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. 3.7, see Ruprecht 147.) The training data we use for our photometric parameter inference are adapted from our CMD classifier in Sect. 3.4, for which blue straggler stars were not found to have a negative impact on the accuracy of our network and were hence not included. Future works using purely simulated data to train a photometric parameter inference neural network would benefit from inclusion of blue straggler stars in their training data, although in practice the origin of blue stragglers is still disputed, and these stars may hence be challenging to simulate accurate photometry for (Boffin et al. 2015; Cantat-Gaudin 2022).

Finally, our results have limited agreement with those of Kharchenko et al. 2013. While some clusters have similar values between their work and ours, particularly for A_V and particularly for the largest and most clearly defined clusters (Fig. 3.7), many sparse clusters that were difficult to detect before *Gaia* have very different photometric parameters. This typically appears to be caused by extremely different cluster membership lists. Before *Gaia*, OCs were often challenging to separate from field stars (Cantat-Gaudin 2022), requiring that suspected outliers be removed iteratively to improve CMD quality (Kharchenko et al. 2012). However, this process can also remove true cluster members, which can cause resulting cluster membership lists to be incorrect (Cantat-Gaudin and Anders 2020). This discrepancy with the results of Kharchenko et al. 2013 is also reported by Cantat-Gaudin et al. 2020, who also find that many photometric parameters derived before *Gaia* are strongly discrepant with current results. In addition, while the number of member stars reported in Kharchenko et al. 2013 is generally a poor predictor for whether or not a given cluster in their work has very different parameters to ours, there are some cases (such as clusters in their work with $A_V > 5$ that we derive much smaller values for) where the most discrepant clusters were also the smallest, with fewer than 20 member stars in reported in Kharchenko et al. 2013.

Although approximate, these results still agree well within the sample-limited but accurate Bayesian isochrone fits of Bossini et al. 2019 and agree relatively well (albeit with some caveats) with the machine learning derived parameters of Cantat-Gaudin and Anders 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. 3.4 and the third paper in this series, Hunt & Reffert, *in prep.*, where we will derive dynamical parameters for our census of star clusters.

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

Firstly, the standard errors on mean proper motions and parallaxes in *Gaia* DR2 can be as small as 5 to 10 μas for the largest clusters in catalogues such as Cantat-Gaudin and Anders 2020, although this is smaller than estimated upper limits on systematics in *Gaia* DR2 of 50 μas (Lindegren et al. 2018). Many reliable clusters are hence missed when treating DR2 positions exactly, as they have systematics significantly larger than their standard errors, with positions in DR3 that can deviate systematically from their DR2 positions by 50 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 et al. 2020 only used stars down to $G = 17$, whereas this work often has membership lists down to $G \sim 20$. Many clusters hence have significantly different membership lists that can result in different mean parameters, particularly for asymmetric clusters.

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

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

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

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

After an extensive search of the literature for recent catalogues, excluding works already listed entirely in other catalogues (such as Froebrich et al. 2007, which appears in its complete form within Bica et al. 2018), we crossmatch against 26 different works listed in Table 3.3. In addition, as our catalogue contains many moving groups, globular clusters, and a handful of clusters associated with the Magellanic clouds, we also crossmatch against the Kounkel et al. 2020 catalogue of predominantly moving groups, the Vasiliev and Baumgardt 2021 *Gaia* DR3 catalogue of globular clusters and the Bica et al. 2008 catalogue of star clusters in the Magellanic clouds. Names between catalogues were standardised as much as possible to facilitate easier comparison and remove duplicated clusters. One such example are ESO clusters, which are numbered based on their position in the form 'ESO XXX-XX' in the original work and Kharchenko et al. 2013, but with numbers that are separated by a space instead of a dash in Cantat-Gaudin and Anders 2020 and Dias et al. 2002, or often miss leading zeroes in Bica et al. 2018.

3.6.2 Recovery of clusters from prior works

Table 3.3 shows that this work has a high recovery rate of OCs from other works. As shown in Table 3.3, we recover 96.6% of clusters from Cantat-Gaudin and Anders 2020, higher than the 86.4% of clusters recovered in Paper 1. Generally, clusters not recovered in Paper 1 were sparse, barely-visible overdensities in *Gaia* DR2 which often now stand out strongly in *Gaia* DR3, including clusters such as Berkeley 91 and Auner 1, which we now detect reliably at S/Ns of 9.7σ and 12.5σ respectively. The fact that only Cantat-Gaudin and Anders 2020 was able to detect these clusters in DR2 is likely due to a difference in methodology – by starting with prior cluster positions, their search regions for these clusters are smaller and may help the clusters to stand out. However, the disadvantage of such an approach is that it may also introduce a handful of false positives, due to poor statistics inherent in such small search regions – in Paper 1, we comment that a handful of clusters in Cantat-Gaudin et al. 2020 may not exist, which may be the case for some of the 3.4% of clusters we are still not able to recover in *Gaia* DR3 despite the greatly improved astrometry and clear benefits to the S/N of other previously undetected clusters.

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

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

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

fundamental methodological difference (their works use the DBSCAN algorithm, see Paper 1 for a review) rather than a data one.

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

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

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

We discuss possible reasons for the non-detection of many literature OCs further in Sect. 3.8.

Finally, it is worth commenting on our detections of moving groups, globular clusters, and Magellanic cloud objects. We are only able to detect 18.1% of moving groups and clusters from the catalogue of Kounkel et al. 2020, despite this work using the same algorithm (HDBSCAN). Many of the groups reported in Kounkel et al. 2020 have large on-sky extents that are larger than the fields used in this work. However, although 2276 of their 8281 clusters are compact enough to be easily detectable in our fields, we only recover 622 (27.3%) of these compact groups, many of which correspond anyway to known nearby OCs. In Paper 1, we found that while HDBSCAN is the most sensitive clustering algorithm for application to *Gaia* data, it also reports a large number of false positives without additional postprocessing to remove clusters based on their statistical significance. It may be that these clusters are false positives, although this should be investigated further in detail (**zucker_disconnecting_2022**).

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

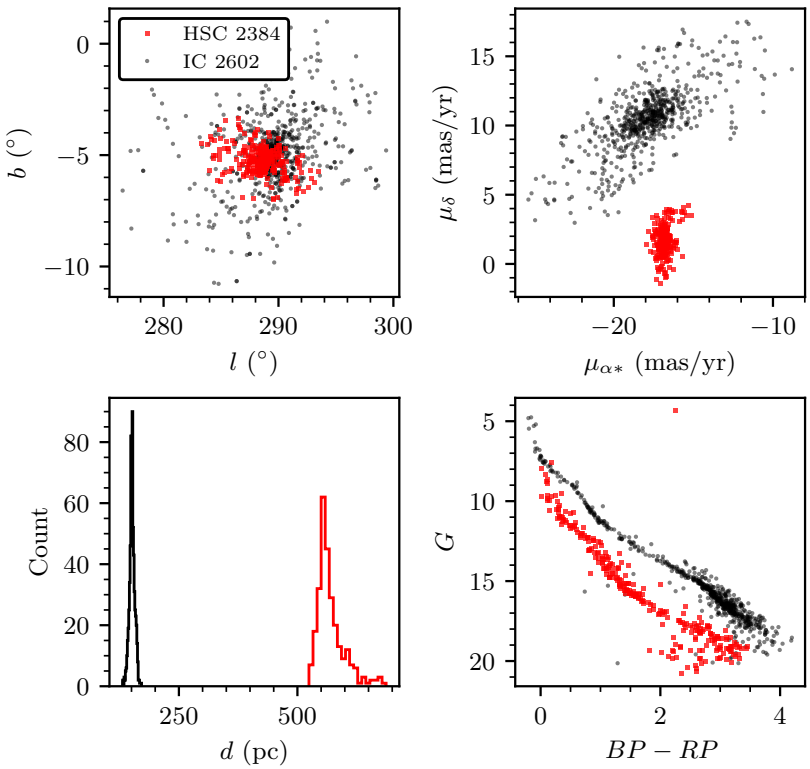

Fig. 3.8.: Member stars for the candidate new cluster HSC 2384 (red squares) compared against the nearby cluster IC 2602 (black circles). Four plots of are shown, comparing positions (top left), proper motions (top right) and photometry (bottom right). The bottom left plot shows a histogram of all distances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

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 et al. 2008 shows that *Gaia* data could be used to make limited inferences on existing Magellanic cloud clusters in a future work, although we do not appear to detect any new clusters in the Magellanic clouds as their distance is too high.

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, of which only 2749 matches are direct one-to-one matches where a single detected cluster can be easily assigned a single name.

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
							...		
HSC 1	1805	8.21	64	289.61	-38.03	3.32	-1.029 (0.054)	-8.941 (0.085)	2.09
HSC 2	1806	3.79	16	268.63	-29.53	0.13	1.680 (0.031)	-1.182 (0.032)	0.63
HSC 3	1807	3.89	24	273.73	-31.87	0.12	0.371 (0.019)	0.210 (0.025)	0.64
HSC 4	1808	3.32	17	269.07	-29.64	0.02	2.125 (0.067)	-11.895 (0.060)	0.11
HSC 5	1809	4.38	18	276.78	-33.09	0.12	0.150 (0.047)	-6.676 (0.049)	0.65
HSC 6	1810	4.57	21	267.71	-28.82	0.05	-0.292 (0.017)	-1.516 (0.023)	0.25
HSC 7	1811	3.12	18	261.40	-25.13	0.09	-5.033 (0.061)	-0.983 (0.060)	0.46
HSC 8	1812	3.33	28	267.67	-28.63	0.06	0.207 (0.014)	-0.211 (0.026)	0.34
HSC 9	1813	5.88	25	269.05	-29.33	0.16	2.120 (0.020)	-0.289 (0.021)	0.54
HSC 10	1814	4.56	12	268.23	-28.80	0.06	-0.200 (0.011)	-1.753 (0.013)	0.35
							...		

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 **??**.

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.

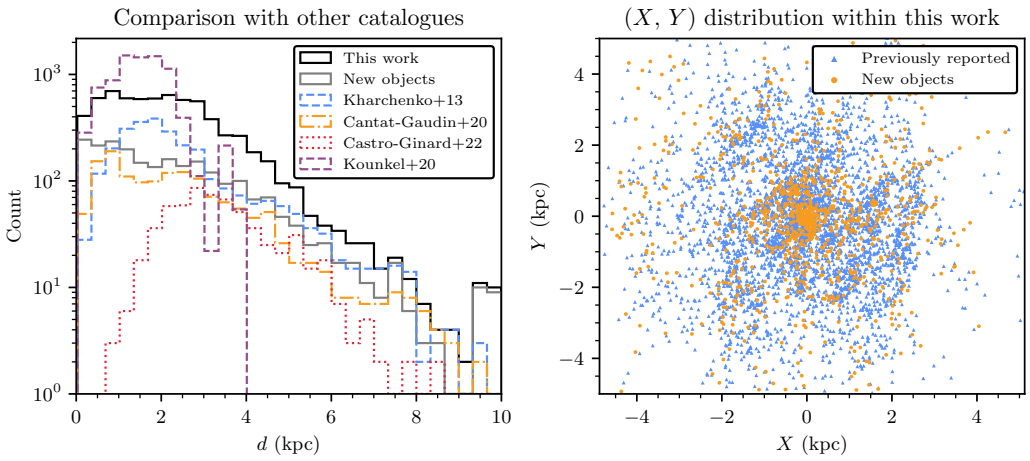

Fig. 3.9.: Distance and spatial distributions of clusters in this work. *Left*: the distance distribution of all clusters in this work that do not crossmatch to known GCs compared to other catalogues. *Right*: The distribution of clusters in this work in Cartesian coordinates centred on the Sun, cut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

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 3.4 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. 3.8.) However, many appear to be more consistent with unbound moving groups, and will require further classification based on their structure and dynamics. In addition, we provide a table of all crossmatches and non-crossmatches against the clusters in this work in Table **??**.

In the next sections, we discuss multiple aspects of the overall catalogue. Firstly, we discuss the overall catalogue of existing clusters in Sect. 3.7, including its distribution and the quality of its membership lists. Section 3.8 discusses why some literature clusters are undetected. Finally, Sect. 3.9 discusses why existing approaches to differentiate between moving groups and OCs are inadequate to classify the new

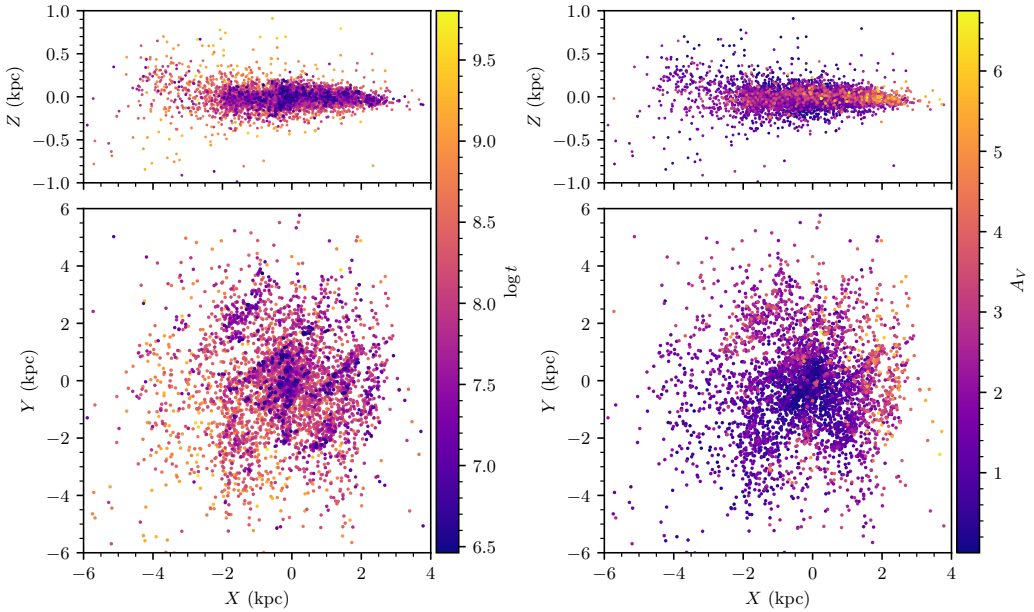

Fig. 3.10.: Spatial distributions of clusters detected in this work shaded on our derived $\log t$ and A_V values. *Left:* side-on and top-down distribution of clusters in heliocentric coordinates that do not crossmatch to known GCs. The galactic centre is to the right, with the Sun at $(0, 0)$. Only 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.

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. 3.4. Such

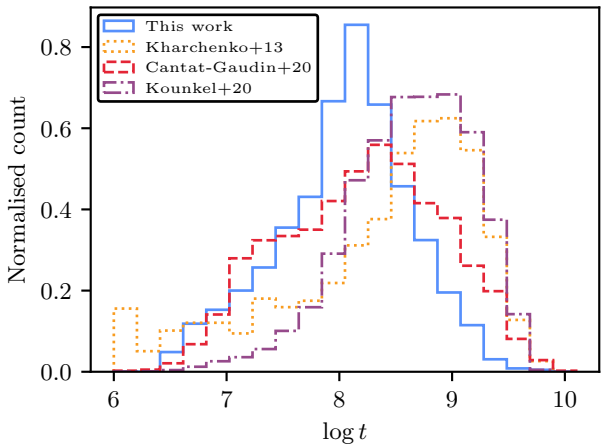

Fig. 3.11.: Histogram of ages of all clusters in this work with median CMD classes greater than 0.5 – specifically, all clusters with photometry that is compatible with a single population of stars. These are compared to the ages of all clusters in the catalogues of Kharchenko et al. 2013, Kounkel et al. 2020, and Cantat-Gaudin et al. 2020. Known GCs are excluded from the results of this work and the results of previous works for this plot.

clusters are included in our catalogue for completeness, as a low-quality CMD may be caused by a poor detection of a real OC by our cluster recovery method, and a cluster with a low CST that is not a guaranteed astrometric overdensity may still be a real cluster that could be validated by a future *Gaia* data release. However, these clusters are not particularly scientifically useful for studies of star clusters, as they cannot be validated as real within this work, or even with any currently available data.

Hence, in discussions of the overall structure of our results, we predominantly discuss the most reliable sample of 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.

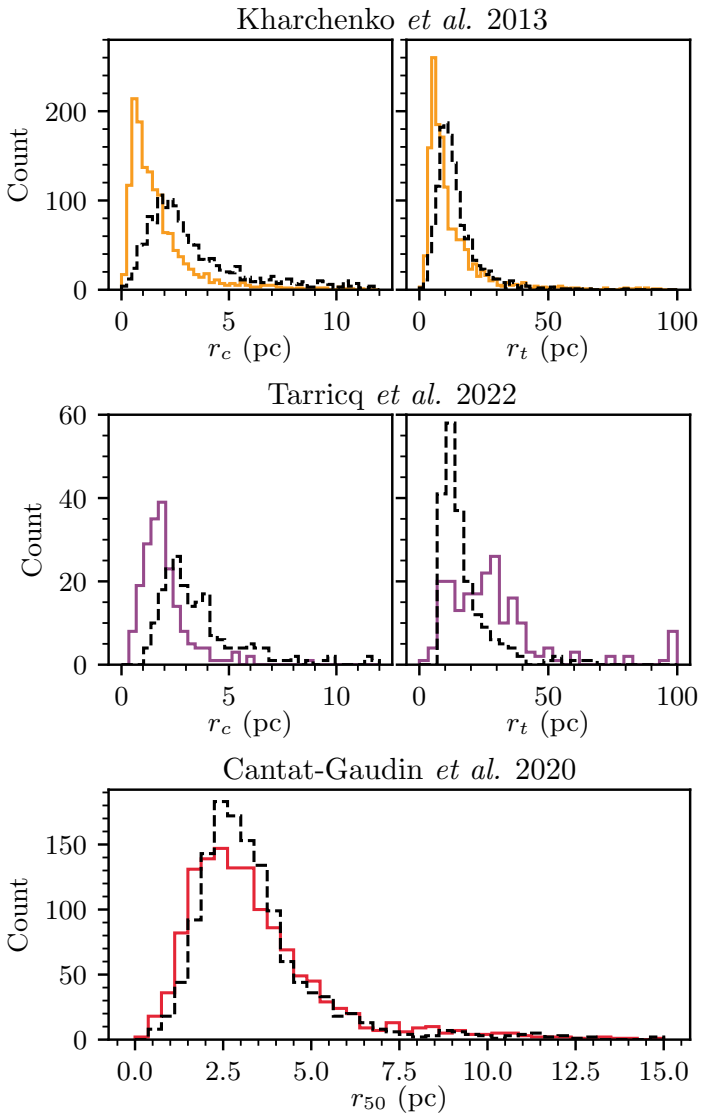

Fig. 3.12.: Cluster radii derived in this work (dashed black line) compared against the distributions of cluster radii in various literature works. *Top row:* r_c (top left) and r_t (top right) of 1446 clusters from Kharchenko et al. 2013 that we redetect in this work (solid orange curve) compared against our approximately estimated **king_structure_1962** radii for these 1446 clusters. *Middle row:* same as top, except for radii of 202 clusters from Tarricq et al. 2022 that have derived King radii (solid purple curve). *Bottom:* r_{50} measurements from Cantat-Gaudin and Anders 2020 compared against our r_{50} measurements for the 1343 clusters from their work that we redetect.

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 and Anders 2020, albeit with more stark differences when compared to those compiled before *Gaia*, such as Kharchenko et al. 2013. Comparisons are also useful to the catalogue of structures, moving groups, and star clusters of Kounkel et al. 2020 and papers based on *Gaia* DR3 data that report new clusters, such as Castro-Ginard et al. 2022.

Figure 3.9 shows the distance distribution of clusters in this work, as well as the X, Y distribution of clusters we re-detect and objects new to this work. Owing to the improved astrometry of *Gaia* DR3 and the clustering method we use (see Paper 1), our catalogue has a high total number of clusters in most distance bins relative to other catalogues. As expected from the results in Paper 1, HDBSCAN is a cluster recovery technique sensitive across all distance ranges. However, HDBSCAN is sensitive to all clusters within *Gaia* data, as it is unbiased on the shape of clusters it reports; hence, the catalogue contains a large number of moving groups, which are generally detected near to the Sun. The catalogue contains around 8x as many objects as the open cluster catalogue of Cantat-Gaudin and Anders 2020 within 500 pc, clearly visible as an overdensity of new objects and in the distance distribution of Fig. 3.9. These objects are often difficult to classify as being OCs or moving groups (see Sect. 3.9).

The age and extinction distribution of Fig. 3.10 is similar to that of Cantat-Gaudin et al. 2020. A number of structures stand out, including: the imprint of the galactic warp in X, Z plots for $X < -2$ kpc; the presence of spiral arm structure amongst young clusters very similar to that reported in works such as Castro-Ginard et al. 2021; and the general flatness of the distribution of compact star clusters in the Milky Way other than GCs, with few existing at heights of $|Z| > 250$ pc. Additionally, clusters towards the galactic centre generally have high A_V values of 5 or greater, suggesting that extinction may be a limiting factor in the detection of clusters in this direction.

Differences to pre-*Gaia* works are most apparent in the age histogram of Fig. 3.11, however. Our combined age distribution is relatively similar to that of Cantat-Gaudin et al. 2020, albeit with a slightly lower median age around $\log t \approx 8$ and no additional bump between $7 < \log t < 8$. However, the star cluster catalogue of Kharchenko et al. 2013 skews significantly older, with the most common (modal) age for clusters being around $\log t \approx 9$, an age range where we detect few clusters. A similar pattern is also visible for the catalogue of Kounkel et al. 2020, whose moving

group and star cluster catalogue contains many unbound, old structures. Many of these objects have similar ages to the typical ages of unclustered stars in the Milky Way disk. In Sect. 3.8, we elaborate on how some of these age differences may be caused by these catalogues containing a number of old false positive clusters.

Finally, Fig. 3.12 shows the distribution of cluster radii compared between this work and the works of Kharchenko et al. 2013, Tarricq et al. 2022, and Cantat-Gaudin and Anders 2020. Our cluster radii agree most strongly with those in Cantat-Gaudin and Anders 2020, with a similar distribution of cluster radii containing 50% of members r_{50}. The **king_structure_1962** core radii r_c that we derive, when compared against those in Kharchenko et al. 2013 and Tarricq et al. 2022, are generally larger. This may be due to our more populated membership lists, particularly for faint stars, due to our lack of a magnitude cut in our clustering analysis. Particularly for clusters with a high degree of mass segregation, this difference in memberships would cause our clusters to have larger observed cores. Our tidal radii r_t are slightly larger than those in Kharchenko et al. 2013, but much smaller than those in Tarricq et al. 2022. In the first case, the difference may be due to the improved precision of *Gaia* data compared to pre-*Gaia* works, causing us to detect more member stars at the outskirts of clusters and hence derive larger cluster tidal radii, with this effect again being stronger for mass segregated clusters. In the second case, since Tarricq et al. 2022 also explicitly searched for cluster tidal tails and comas in their work, it may be that their extended cluster membership lists mean that they report higher cluster tidal radii.

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. 3.13 compares our membership lists with those from Cantat-Gaudin and Anders 2020 for five clusters randomly selected from our catalogue. Our membership lists typically have a higher total number of stars, with virtually all new member stars being compatible with the existing cluster CMD. This is particularly the case for clusters in regions with minimal crowding, where *Gaia* has a high completeness of stars with 5-parameter astrometry down to $G \sim 20$, with our membership lists containing stars down to approximately this limit. For more distant clusters such as Kronberger 4, membership lists are comparable in quality to those of Cantat-Gaudin and Anders 2020, as *Gaia* DR3 data does not present a large improvement in the astrometric quality of these distant sources compared to

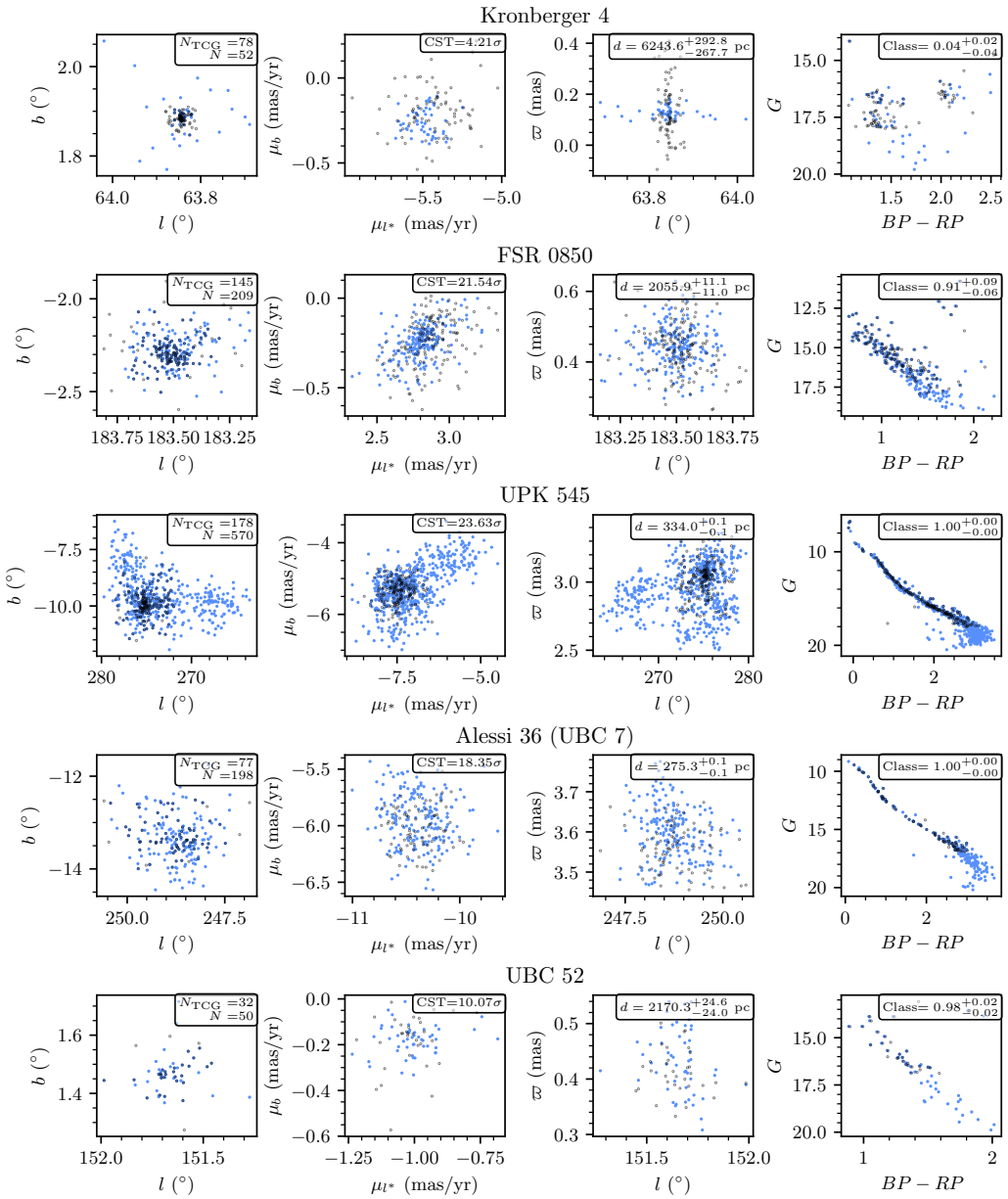

Fig. 3.13.: Membership list comparisons between this work and the catalogue of Cantat-Gaudin and Anders 2020, using three clusters selected at random (upper three) and two clusters selected at random that were detected in Castro-Ginard et al. 2018 using *Gaia* DR1 data. Stars assigned as members by this work are plotted with filled blue circles, while members reported by Cantat-Gaudin and Anders 2020 are plotted with empty black circles. The first three columns compare the astrometry of cluster members in galactic coordinates, proper motions, and parallax as a function of l. The final column compares colour-magnitude diagrams of each resulting membership list. For every cluster, various parameters are labelled on the plots: number of member stars in Cantat-Gaudin and Anders 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. 3.4.

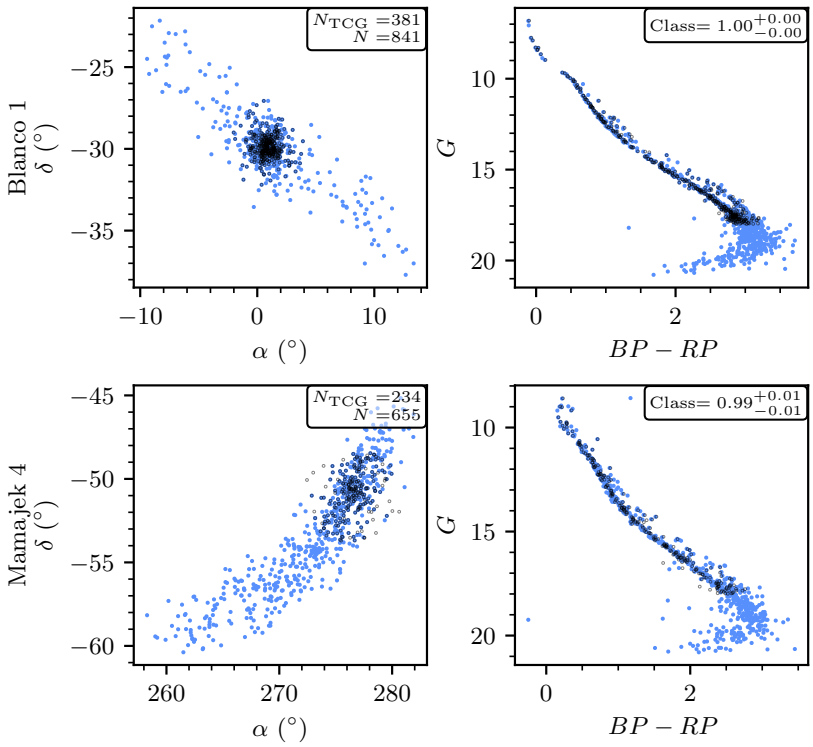

Fig. 3.14.: Two examples of clusters in the catalogue that have detected tidal structures. The spatial distribution of the clusters Blanco 1 (top row) and Mamajek 4 (bottom row) are plotted on the left, with member stars reported in this work shown as filled blue circles and compared against member stars from Cantat-Gaudin and Anders 2020 which are plotted as empty black circles. CMDs are shown in the two plots on the right for both clusters.

DR2. On average, our work contains 2.1 times as many member stars as the clusters we have in common with Cantat-Gaudin and Anders 2020, and 4.1 times as many member stars as the clusters we have in common with Kharchenko et al. 2013.

A second major advantage of our pipeline is that clusters are not forced to take a spherical shape, as with other methods such as Gaussian mixture models (Paper 1). Hence, we are able to detect tidal tails for many of the clusters in the catalogue, especially for those that are nearby and within $1 - 2$ kpc. Tarricq et al. 2022 use HDBSCAN to detect tidal tails for 71 nearby OCs, many of which we are also able to detect. Figure 3.14 shows two examples of nearby clusters with well-resolved tidal tails using our methodology, Blanco 1 and Mamajek 4, both of which have reported tidal tails stretching around 50 pc from the centre of the cluster. Virtually all stars within the tidal structures appear compatible with the isochrone of the cluster core, suggesting that they are stars with the same age, composition, and origin as the

stars in the cluster cores. Particularly for clusters within 1 kpc, many of the clusters in our catalogue have tidal tails or comas.

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

Firstly, for distant OCs, our method may return fewer members than some other approaches. At high distances ($d \gtrsim 5$ kpc), the errors on *Gaia* parallaxes and proper motions generally become much higher than the intrinsic dispersion of OCs, meaning that many members have low membership probabilities and can only be reliably assigned as members by incorporating error information. Our methodology does not use error information in the clustering analysis for reasons of speed and the fact that HDBSCAN does not directly include a way to consider errors on data in clustering analysis, although other methods such as UPMASK (Krone-Martins and Moitinho 2014) which do consider error information could return better membership lists for these distant clusters. This is visible for Kronberger 4 in Fig. 3.13, where the membership list of Cantat-Gaudin and Anders 2020 (which was compiled using UPMASK) has a slightly higher number of sources than our membership list, even though our list was compiled from a greater number of input sources due to our lack of a G-magnitude cut.

Secondly, HDBSCAN may sometimes return too many members, selecting regions larger than just an OC's core and tidal tails. This is particularly common for young clusters, which are often embedded in regions of high stellar density where recent hierarchical star formation has occurred (Portegies Zwart et al. 2010). These clusters can be difficult for HDBSCAN to isolate from other surrounding stars and sub-clusters. One particular example can be seen for UPK 545 in Fig. 3.13. Although the tail emerging from the cluster core in the upper-left of the (l, b) plot appears compatible with a tidal tail, the connected structure to the right of the cluster is not. It appears to have the same age and composition as the cluster core, with all members of the tail being photometrically consistent with it. However, this 'offshoot' from the cluster may be better described as a separate cluster, which may also be bound to the core of UPK 545 in a binary pair of clusters, due to their proximity. Edge cases such as these are impossible to deal with autonomously with our current methodology and HDBSCAN alone, and require manual selection and separation of certain clusters in the catalogue into multiple separate components.

On a whole, the primary advantage of our catalogue is its completeness, generally reporting more member stars than previous works in the literature and doing so

with a homogeneous methodology for a high number of total clusters. However, this is also the primary disadvantage of our catalogue: there are too many clusters and too many edge cases for all membership lists to be perfect, given only one clustering methodology. Hence, users of the catalogue who work with a small enough number of clusters are encouraged to manually check cluster membership lists and refine them depending on their application. To give one example, a user who wishes to only study cluster cores could refine our cluster membership lists by selecting a subset of them with Gaussian mixture models. With careful manual tweaking of the parameters of the mixture models, such a method could be used to remove tidal tails or possible other cluster components from our membership lists where necessary. Having discussed the general results of clusters in our catalogue, we next discuss the reasons why many clusters reported in the literature may not appear in our catalogue.

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

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

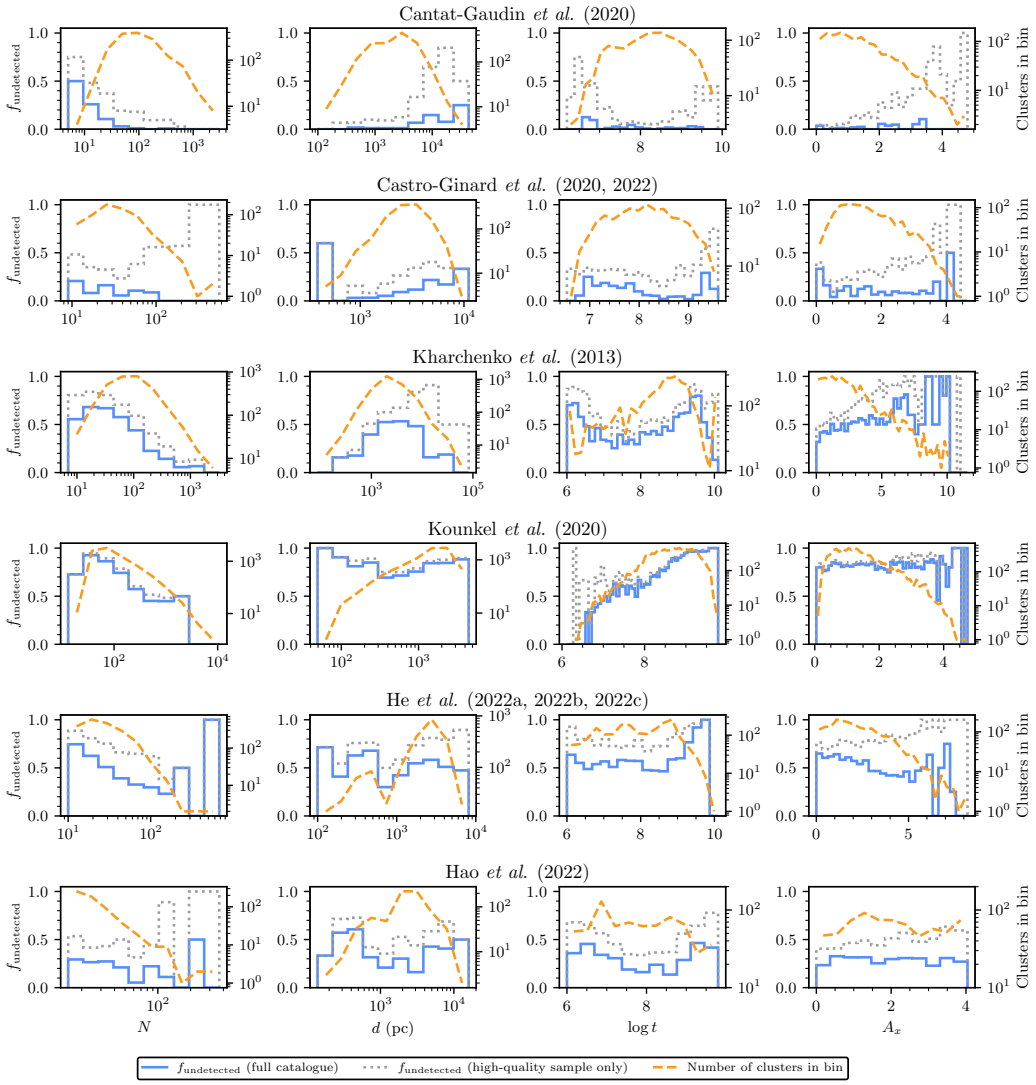

Fig. 3.15.: Plots showing the fraction of clusters undetected by this work when compared to various literature works or series of literature works, shown as a histogram of various parameters as a solid blue line for all clusters in the catalogue, and a dashed grey line for clusters in the high quality sample defined in Sect. 3.7.1. The dashed orange lines show the number of clusters in each bin. Optimum histogram bin widths were selected automatically using `numpy` (Harris et al. 2020). From left to right, each column shows the number of stars N, distance d, age $\log t$ and extinction A_x reported in each catalogue. For the top four groups of catalogues, extinctions were given in the V band. For the lower two, extinctions were given in *Gaia's* G band, which are generally slightly lower.

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 et al. 2020 and Castro-Ginard et al. 2022 respectively that we are unable to detect have reported parallaxes of less than 0.2 mas, suggesting that distance-related detection issues alone are not enough to explain why certain clusters from these works are not detected. Additionally, we note that Castro-Ginard et al. 2022 using *Gaia* EDR3 were only able to recover $\gtrsim$80% of clusters they found in DR2 in Castro-Ginard et al. 2020, and so DBSCAN itself between *Gaia* data releases is not able to reliably reconfirm all clusters it detected previously.

Nevertheless, Fig. 3.15 shows that our chance of recovering clusters at high distances can be lower for certain works. In particular, although we are unable to recover only 3.4% of clusters reported in Cantat-Gaudin and Anders 2020, most of the clusters from their work that we are unable to recover are small clusters at distances above 5 kpc, suggesting that an algorithmic limitation may contribute to why we are unable to recover remaining objects from Cantat-Gaudin and Anders 2020. A key difference between our work and Cantat-Gaudin and Anders 2020 is that their work used locations of clusters reported in the literature to narrow their search regions, which may in some cases be enough to make very distant clusters at the absolute limit of detectability in *Gaia* stand out. Future *Gaia* data releases with better data should provide additional clarity on whether or not such objects are real.

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 et al. 2022 to as high as 50 in Liu and Pang 2019, with most works coalescing around a value of between 10 and 12 (Krumholz et al. 2019). For the purposes of this work, we adopt a value of 10, and we should hence miss very few literature clusters due to this constraint alone.

Secondly, OCs generally have a population of stars with the same age and chemical composition, due to forming at the same time from the same molecular cloud (Cantat-Gaudin 2022). In practice, this is a difficult definition to constrain observationally, with the CMDs of OCs being broadened by effects such as differential extinction or outliers which are not true member stars, with these effects being worse with increasing distance and field star density. In addition, many OCs are not perfect single populations, with some hosting blue stragglers or having a clear second population in the form of an extended main-sequence turnoff (Cantat-Gaudin 2022). For the purposes of this work, we classify our clusters with our CMD classifier (see Sect. 3.4) and include all clusters in the final catalogue, instead leaving the task of removing clusters with poor photometry to the end user (recommending a minimum class value of 0.5). This means that no clusters are missing from the catalogue due to photometric reasons.

Finally, OCs must be distinguished from other types of single-population stellar overdensities. Star clusters can be divided into bound clusters (such as OCs and GCs) and unbound clusters (typically referred to as moving groups). Some works, such as Cantat-Gaudin and Anders 2020, use basic cuts on mean parameters to remove clear moving groups from their catalogue; we leave the classification of moving groups in our catalogue to a future work (Hunt & Reffert, *in prep.*) for reasons discussed in Sect. 3.9, 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 and Baumgardt 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 et al. 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 et al. 2021). Poor photometry and the bound or unbound status hence do not impact our recoverability of clusters in Fig. 3.15.

However, the sole quality cut applied to the catalogue that would affect its sensitivity is a cut on the astrometric S/N of detected clusters (derived using the CST) at 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. 3.3). Including such a quality cut dramatically improved the run merging process and hence our membership lists and completeness for reliable clusters, which is a more important scientific product than a list of low quality clusters that we cannot deem likely to be real clusters based on their S/N alone.

While we believe this is a fair trade-off to produce a catalogue that is as reliable as possible overall, it is likely that some real clusters are 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. 3.2. 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 et al. 2019, which was re-detected by Cantat-Gaudin and Anders 2020, but which HDBSCAN assigns as simply being a member of a tidal tail of a different and significantly larger nearby cluster, UPK 545, with no HDBSCAN m_{clSize} run separating the two objects. UPK 545 is shown in Fig. 3.13 on the third row. In this and other edge cases, our catalogue merges the two objects. An improved clustering algorithm that can separate edge-case binary clusters such as these autonomously would be helpful. However, only a small fraction of clusters (fewer than 1%) are affected by this issue.

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

While our crossmatching procedure as outlined in Sect. 3.6 aims to be as fair as possible, generally giving the benefit of the doubt to potential crossmatches, there are nevertheless cases where clusters reported in the literature still remain outside of our bounds for an accepted match. Generally, in all cases where this occurs, our detected cluster is significantly different to the literature object in at least one of the parameters considered for crossmatching, with these clusters representing ambiguous cases where it is not clear that the reported literature cluster is truly the same object.

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

This case is likely to have been repeated a few times, and appears particularly common with clusters detected in *Gaia* data using the DBSCAN algorithm (as in He et al. 2022a). In Paper 1, we commented that while DBSCAN has an excellent sensitivity and low false positive rate (depending strongly how the 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 et al. 2013 was compiled before *Gaia* and used infrared data from 2MASS (Skrutskie et al. 2006). Cantat-Gaudin and Anders 2020 are unable to recover a majority of the clusters from Kharchenko et al. 2013 using *Gaia* DR2 data, and we are unable to recover 48.4% of the clusters reported in their catalogue in *Gaia* DR3 data. Given that infrared light is significantly less affected by extinction than the visual light used to compile *Gaia* data, it begs the question of whether many clusters from Kharchenko et al. 2013 may still be missing from *Gaia*-based catalogues due to extinction limits.

However, Fig. 3.15 shows that extinction does not appear to play a major role in the non-detection of many clusters from Kharchenko et al. 2013. If extinction was a major contributor to why we are unable to detect so many of the clusters in their catalogue, then one would expect to see a linear trend in $f_{\text{undetected}}$; 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 et al. 2021), and given that our chance of detecting a cluster reported in Kharchenko et al. 2013 is uncorrelated with extinction for $A_V < 5$, limitations of *Gaia* data do not appear to be responsible for the bulk of non-detections of clusters from pre-*Gaia* works, despite assertions in recent works that *Gaia* data may be extinction-limited and unable to recover many highly reddened OCs from infrared datasets. Nevertheless, a handful of high-extinction clusters with $A_V > 5$ reported in the literature may still be challenging to recover in *Gaia* data.

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

In general, since most works in Fig. 3.15 use *Gaia* DR2 data or stronger cuts on *Gaia* data than our methodology, there are many cases where we should be able to detect their reported clusters easily and with a higher number of member stars and

statistical significance. The fact that we cannot suggests that some of these clusters may have been statistically insignificant associations of a small number of member stars.

The distance of undetected reported literature clusters is similarly revealing. In Sect. 3.8.1, we suggest that some clusters may be undetected in this work at high distances due to limitations of the HDBSCAN algorithm. However, given that HDBSCAN should be the most sensitive algorithm for recovery of nearby clusters (Paper 1), it makes little sense that we are unable to recover a number of nearby clusters within $1\ kpc$ for most of the works in Fig. 3.15. Many of these nearby and undetected objects may not be real, as there is no reason why we should not be able to detect them using the improved data of *Gaia* DR3 and the most sensitive algorithm for recovery of nearby OCs.

The age of undetected clusters paints a complicated picture. In principle, detecting an old cluster has two challenges. Firstly, as the cluster ages, the brightest stars in the cluster evolve into faint remnants, which reduces the number of stars visible in the cluster. This is a particular issue for distant old clusters, as the remaining fainter and longer-lived stars in a cluster may be below a survey's magnitude limit. In the case of *Gaia*, stars near 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 et al. 2010), reducing their signal-to-noise ratio relative to background field stars in positional data.

Although old clusters are likely to be harder to detect, in Paper 1, we found that the age of a reported cluster generally has the same effect on all algorithms: their lower number counts and sparsity affect all algorithms more or less equally in making them harder to detect. However, there are correlations between $f_{\text{undetected}}$ and $\log t$ for almost all papers in Fig. 3.15, despite all of them other than Kharchenko et al. 2013 being based on *Gaia* data and using methods found in Paper 1 to be equally affected by cluster age. Hence, these correlations may be more informative about the types of cluster in other catalogues that are false positives than on whether or not a given catalogue used a better method.

For all works other than Cantat-Gaudin and Anders 2020, clusters older than an age of around 1 Gyr ($\log t > 9$) are much less likely to be redetected. **zucker_disconnecting_2022** have recently investigated the nature of the groups reported in Kounkel et al. 2020, and find that many of them have ages $\sim$120 times larger than their dispersal times while being unbound and chemically homogeneous with their surrounding field

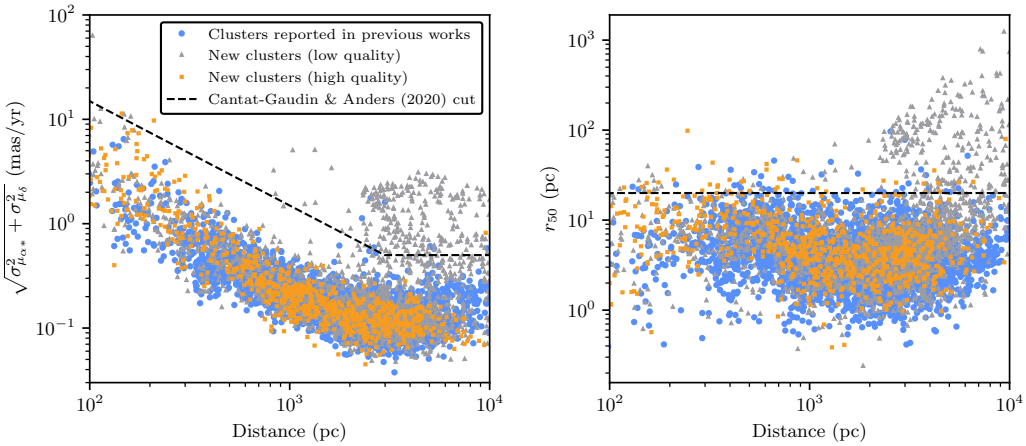

Fig. 3.16.: Geometric mean of the proper motion dispersion (left) and radius containing 50% of members (right) for the clusters reported in this work, as a function of distance. Clusters are split between those detected in previous works (blue circles) and those newly reported in this work, divided between the high quality (orange squares) and low quality (grey triangles) samples defined in Sect. 3.7.1. The cuts on cluster parameters to distinguish between bound OCs and unbound moving groups or associations proposed in Cantat-Gaudin and Anders 2020 are shown as a dashed black line.

stars – strongly suggesting that they are merely associations of field stars and not physical groupings. The fact that we are unable to redetect almost any of the groups older than 1 Gyr reported in Kounkel et al. 2020 supports this conclusion, with it being plausible that many of their oldest groups are instead associations of field stars, consistent with the mean ages of field stars in the galactic thin and thick disks of a few Gyr. The similar correlations with old clusters being undetected for other works may also suggest that a number of other old clusters reported in the literature are also associations of field stars with mean ages similar to that of the typical ages of unclustered field stars in the galactic disk.

The reasons for the non-detection of some young clusters are less clear, and are more surprising given that young clusters should be easier to detect. In the case of Cantat-Gaudin and Anders 2020, the handful of young clusters that we are unable to detect are also at high distances, which may mean that their non-detection is entirely a result of our own methodological limitations (see Sect. 3.8.1.) On the other hand, these distant, young clusters may have originally been detected by hand-searching for OB stars in pre-*Gaia* works and cataloguing them as OCs, but without a test of their physical nature, which could mean that they are associations. Similar reasoning could also be applied to the non-detected young clusters from Kharchenko et al. 2013. Both possibilities are plausible, and this should be investigated further in another work.

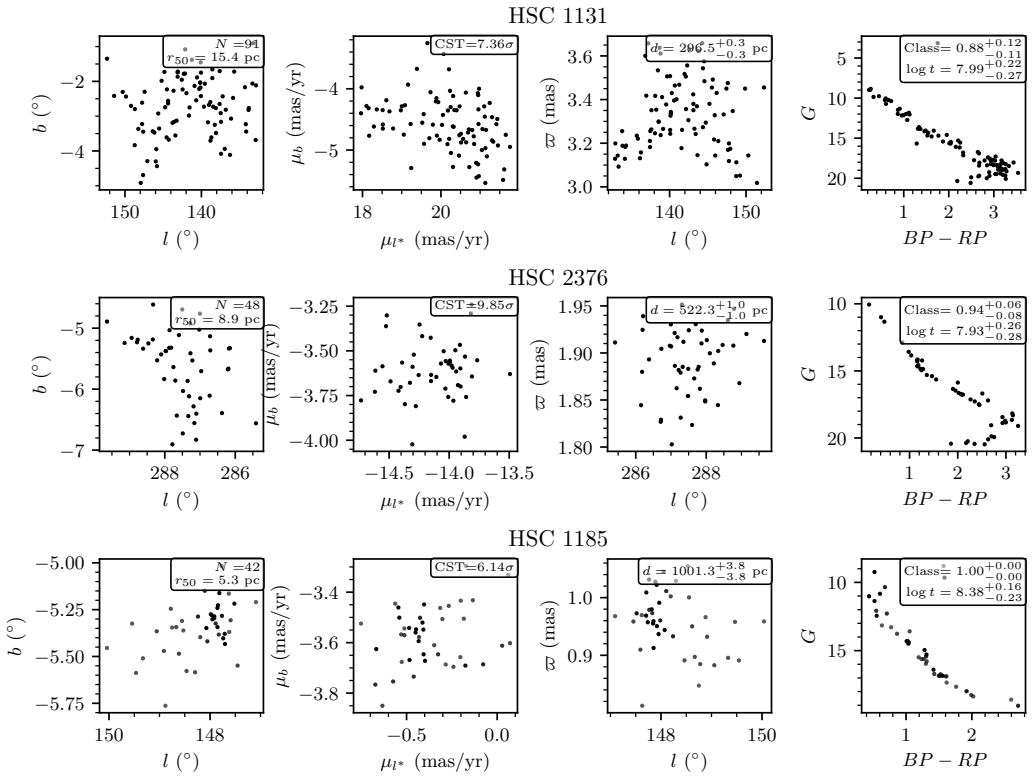

Fig. 3.17.: Three newly reported clusters randomly selected from the cluster catalogue and ordered by increasing distance, with member stars plotted as a function of their astrometric and photometric data as in Fig. 3.13. All clusters pass the cuts proposed in Cantat-Gaudin and Anders 2020, have good-quality CMDs passing the cuts from Sect. 3.4, and have astrometric significances of greater than 5σ, meaning they are almost certainly real overdensities in *Gaia* data.

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

For most works, extinction A_V does not predict the chance of redetecting a given cluster. In Sect. 3.8.1, 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 and Anders 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. 3.15 with respect to extinction, other than for a few dozen pre-*Gaia* clusters from the infra-red catalogue of Kharchenko et al. 2013 with $A_V \gtrsim 5$ that we are unable to redetect with *Gaia* data.

In summary, we find that there are many potential reasons for the non-detection of given clusters from the literature, all of which should be investigated in more depth in future works. Verifying that new clusters reported in the literature are real is arguably as important as reporting them. While we cannot provide conclusive reasons for the non-detection of given clusters, given the scope of this survey, the overall trends we have identified should still be helpful and suggestive in whether or not given objects are real. We provide a table of all clusters non-detected by this work in Table **??** 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_1962**) or Plummer-like (Plummer 1911) profiles. Figure 3.17 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 and Anders 2020 propose two tolerant cuts on cluster parameters, finding that requiring the geometric mean of proper motion dispersion to be less than a criterion (corresponding to ~ 5 kms^{-1}) and $r_{50} < 20pc$ removed objects highly unlikely to be OCs from their sample.

However, Fig. 3.16 shows that with the exception of some clusters that are clearly associated with stellar streams (based on their location, CMD, and sparsity at distances greater than ~ 3 kpc), most new clusters detected in this work are compatible with OCs given the tolerant cuts in Cantat-Gaudin and Anders 2020.

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

In the preparation of this work, much effort was put in to attempting to find a more stringent cut on basic astrometric parameters (or some combination of them) to distinguish OCs from moving groups. We found that whether or not a cluster is a bound OC cannot be decided accurately based on individual cuts on r_{50} or proper motion dispersions alone, and instead requires at least some modelling of the cluster's spatial profile, its velocity 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 et al. 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 et al. 2018, Kuhn et al. 2019, and Pang et al. 2021 to test the virial nature of OCs using *Gaia* data, albeit in limited studies of no more than 28 clusters in one work.

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

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

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 (e.g. Cantat-Gaudin et al. 2020; Castro-Ginard et al. 2021).

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

Our catalogue contains 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 and Anders 2020 being too lenient. In

an upcoming work, we will use the virial theorem to distinguish between bound and unbound clusters with a probabilistic methodology (Hunt & Reffert, *in prep.*).

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

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

4 The dynamics of Milky Way star clusters

> *TODO.*
>
> — **TODO**
> (TODO)

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.1 System Section 1

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information

about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Fig. 4.1.: Figure example: *(a)* example part one, *(c)* example part two; *(c)* example part three

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.2 System Section 2

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Fig. 4.2.: Another Figure example: *(a)* example part one, *(c)* example part two; *(c)* example part three

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A

blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.3 System Section 3

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the

alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there

no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.4 Conclusion

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5 Conclusion

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5.1 System Section 1

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text,

you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5.2 System Section 2

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A

blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5.3 Future Work

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Bibliography

Abadi, Martín, Ashish Agarwal, Paul Barham, et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* (cit. on p. 62).

Abadi, Martın, Paul Barham, Jianmin Chen, et al. (2016). "TensorFlow: A System for Large-Scale Machine Learning". In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, p. 21 (cit. on p. 62).

Anders, Friedrich, Tristan Cantat-Gaudin, Irene Quadrino-Lodoso, et al. (June 2020). "The Milky Way's Cluster Age Function in Light of Gaia DR2". In: *arXiv:2006.01690 [astro-ph]*. arXiv: 2006.01690 [astro-ph] (cit. on pp. 16, 22, 23, 27).

Anders, Friedrich, Alfred Castro-Ginard, Juan Casado, Carme Jordi, and Lola Balaguer-Núñez (Mar. 2022). "NGC 1605 Is Not a Binary Cluster". In: *Research Notes of the American Astronomical Society* 6, p. 58 (cit. on pp. 16, 54, 74).

Bailer-Jones, C. A. L., J. Rybizki, M. Fouesneau, M. Demleitner, and R. Andrae (Mar. 2021). "Estimating Distances from Parallaxes. V: Geometric and Photogeometric Distances to 1.47 Billion Stars in Gaia Early Data Release 3". In: *The Astronomical Journal* 161.3, p. 147 (cit. on pp. 47, 49).

Baratella, M., V. D'Orazi, G. Carraro, et al. (Feb. 2020). "The Gaia-ESO Survey: A New Approach to Chemically Characterising Young Open Clusters. I. Stellar Parameters, and Iron-Peak, α-, and Proton-Capture Elements". In: *Astronomy and Astrophysics* 634, A34 (cit. on p. 42).

Bastian, N. and S. E. de Mink (Sept. 2009). "The Effect of Stellar Rotation on Colour-Magnitude Diagrams: On the Apparent Presence of Multiple Populations in Intermediate Age Stellar Clusters". In: 398.1, pp. L11–L15. arXiv: 0906.1590 [astro-ph.GA] (cit. on p. 19).

Bastian, U. (Oct. 2019). "Gaia 8: Discovery of a Star Cluster Containing β Lyrae". In: *Astronomy & Astrophysics* 630, p. L8 (cit. on pp. 16, 74).

Becker, Burger, Mattia Vaccari, Matthew Prescott, and Trienko Lups Grobler (Feb. 2021). "CNN Architecture Comparison for Radio Galaxy Classification". In: *Monthly Notices of the Royal Astronomical Society* 503.2, pp. 1828–1846 (cit. on p. 62).

Bica, E., C. Bonatto, C. M. Dutra, and J. F. C. Santos (Sept. 2008). "A General Catalogue of Extended Objects in the Magellanic System". In: *Monthly Notices of the Royal Astronomical Society* 389, pp. 678–690 (cit. on pp. 73, 74, 76).

Bica, Eduardo, Daniela B. Pavani, Charles J. Bonatto, and Eliade F. Lima (Dec. 2018). "A Multi-band Catalog of 10978 Star Clusters, Associations, and Candidates in the Milky Way". In: *The Astronomical Journal* 157.1, p. 12 (cit. on pp. 73–75).

Binney, James and Scott Tremaine (1987). *Galactic Dynamics* (cit. on p. 10).

Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra (May 2015). "Weight Uncertainty in Neural Networks". In: *arXiv e-prints* arXiv:1505.05424 (cit. on pp. 134, 135).

Boffin, Henri M. J., Giovanni Carraro, and Giacomo Beccari (Jan. 2015). *Ecology of Blue Straggler Stars*. Vol. 413. Astrophysics and Space Science Library. Springer Berlin, Heidelberg (cit. on pp. 19, 70).

Bossini, D., A. Vallenari, A. Bragaglia, et al. (Mar. 2019). "Age Determination for 269 Gaia DR2 Open Clusters". In: *Astronomy & Astrophysics* 623, A108 (cit. on pp. 29, 42, 65, 66, 69, 70).

Boubert, Douglas and Andrew Everall (May 2020). "Completeness of the Gaia-verse II: What Are the Odds That a Star Is Missing from Gaia DR2?" In: *Monthly Notices of the Royal Astronomical Society* 497.4, pp. 4246–4261 (cit. on p. 58).

Boubert, Douglas, Andrew Everall, and Berry Holl (Apr. 2020). "Completeness of the Gaia-verse I: When and Where Were Gaia's Eyes on the Sky during DR2?" In: *Monthly Notices of the Royal Astronomical Society* 497.2, pp. 1826–1841 (cit. on p. 58).

Bravi, L., E. Zari, G. G. Sacco, et al. (July 2018). "The Gaia-ESO Survey: A Kinematical and Dynamical Study of Four Young Open Clusters". In: *Astronomy and Astrophysics* 615, A37 (cit. on pp. 32, 100).

Bressan, Alessandro, Paola Marigo, Léo. Girardi, et al. (Nov. 2012). "PARSEC: Stellar Tracks and Isochrones with the PAdova and TRieste Stellar Evolution Code". In: *Monthly Notices of the Royal Astronomical Society* 427, pp. 127–145 (cit. on pp. 28, 29, 56).

Breuval, Louise, Pierre Kervella, Richard I. Anderson, et al. (Sept. 2020). "The Milky Way Cepheid Leavitt Law Based on Gaia DR2 Parallaxes of Companion Stars and Host Open Cluster Populations". In: *arXiv:2006.08763 [astro-ph]*. arXiv: 2006.08763 [astro-ph] (cit. on p. 20).

Brown, A. G. A., A. Vallenari, T. Prusti, et al. (Aug. 2018). "Gaia Data Release 2 - Summary of the Contents and Survey Properties". In: *Astronomy & Astrophysics* 616, A1 (cit. on pp. 2, 9, 15, 42, 87).

Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander (2013). "HDBSCAN - Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining* 7819. Ed. by Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, pp. 160–172 (cit. on pp. 42, 45, 48).

Cantat-Gaudin, T. and F. Anders (Jan. 2020). "Clusters and Mirages: Cataloguing Stellar Aggregates in the Milky Way". In: *Astronomy and Astrophysics* 633, A99 (cit. on pp. 2, 14, 22, 25, 26, 30, 42, 45–47, 53, 57–59, 67, 70–75, 81–87, 89, 90, 92, 93, 95–97, 99–101, 132).

Cantat-Gaudin, T., F. Anders, A. Castro-Ginard, et al. (Aug. 2020). "Painting a Portrait of the Galactic Disc with Its Stellar Clusters". In: *Astronomy & Astrophysics* 640, A1 (cit. on pp. 2, 18, 30, 42, 56, 63–66, 69, 70, 73, 80, 82, 101).

Cantat-Gaudin, T., C. Jordi, A. Vallenari, et al. (Oct. 2018a). “A Gaia DR2 View of the Open Cluster Population in the Milky Way”. In: *Astronomy & Astrophysics* 618, A93 (cit. on pp. 2, 12, 14, 22, 30).

Cantat-Gaudin, T., A. Krone-Martins, N. Sedaghat, et al. (Apr. 2019). “Gaia DR2 Unravels Incompleteness of Nearby Cluster Population: New Open Clusters in the Direction of Perseus”. In: *Astronomy & Astrophysics* 624, A126 (cit. on pp. 16, 24, 42).

Cantat-Gaudin, T., A. Vallenari, R. Sordo, et al. (July 2018b). “Characterising Open Clusters in the Solar Neighbourhood with the Tycho-Gaia Astrometric Solution”. In: *Astronomy & Astrophysics* 615, A49 (cit. on pp. 6, 12, 42, 45, 52, 55, 68, 75).

Cantat-Gaudin, Tristan (Feb. 2022). “Milky Way Star Clusters and Gaia: A Review of the Ongoing Revolution”. In: *Universe* 8, p. 111 (cit. on pp. 15, 17, 19, 22, 23, 31, 42, 70, 90).

Cantat-Gaudin, Tristan and Timothy D. Brandt (Mar. 2021). “Characterizing and Correcting the Proper Motion Bias of the Bright Gaia EDR3 Sources”. In: *Astronomy & Astrophysics* 649, A124. arXiv: 2103.07432 (cit. on p. 44).

Cantat-Gaudin, Tristan, Morgan Fouesneau, Hans-Walter Rix, et al. (Jan. 2023). “An Empirical Model of the Gaia DR3 Selection Function”. In: *Astronomy & Astrophysics* 669, A55 (cit. on p. 58).

Cardelli, Jason A., Geoffrey C. Clayton, and John S. Mathis (Oct. 1989). “The Relationship between Infrared, Optical, and Ultraviolet Extinction”. In: *The Astrophysical Journal* 345, pp. 245–256 (cit. on p. 57).

Casado, Juan (2021). “New Open Clusters Found by Manual Mining of Data in Gaia DR2”. In: *Research in Astronomy and Astrophysics* 21.5, p. 117 (cit. on p. 74).

Casado, Juan and Yasser Hendy (Jan. 2023). “Discovery and Description of Two Young Open Clusters in the Primordial Group of NGC 6871”. In: *Monthly Notices of the Royal Astronomical Society*, stad071. arXiv: 2211.12843 [astro-ph] (cit. on p. 74).

Castro-Ginard, A., C. Jordi, X. Luri, T. Cantat-Gaudin, and L. Balaguer-Núñez (July 2019). “Hunting for Open Clusters in Gaia DR2: The Galactic Anticentre”. In: *Astronomy & Astrophysics* 627.A35 (cit. on pp. 2, 16, 21, 24, 42, 48, 56).

Castro-Ginard, A., C. Jordi, X. Luri, et al. (Oct. 2018). “A New Method for Unveiling Open Clusters in Gaia - New Nearby Open Clusters Confirmed by DR2”. In: *Astronomy & Astrophysics* 618, A59 (cit. on pp. 2, 15, 21, 24, 25, 42, 48, 49, 56, 84).

Castro-Ginard, A., C. Jordi, X. Luri, et al. (2022). “Hunting for Open Clusters in Gaia EDR3: 664 New Open Clusters Found with OCfinder”. In: *Astronomy & Astrophysics* 661, A118. arXiv: 2111.01819 (cit. on pp. 2, 16, 21, 42, 45, 46, 48, 56, 59, 62, 73, 74, 82, 89, 90, 97).

Castro-Ginard, A., C. Jordi, X. Luri, et al. (Jan. 2020). “Hunting for Open Clusters in \textit{Gaia} DR2: 582 New OCs in the Galactic Disc”. In: *Astronomy & Astrophysics* 635.A45 (cit. on pp. 2, 16, 21, 24, 42, 48, 56, 71–74, 89, 97).

Castro-Ginard, A., P. J. McMillan, X. Luri, et al. (May 2021). “On the Milky Way Spiral Arms from Open Clusters in Gaia EDR3”. In: *Astronomy & Astrophysics* 652, A162. arXiv: 2105.04590 (cit. on pp. 2, 18, 22, 42, 82, 101).

Chi, Huanbin, Feng Wang, and Zhongmu Li (Feb. 2023a). “LISC Catalog of Open Clusters.III. 83 Newly Found Galactic Disk Open Clusters Using Gaia EDR3”. In: *arXiv e-prints* arXiv:2302.08926 (cit. on p. 74).

Chi, Huanbin, Feng Wang, Wenting Wang, Hui Deng, and Zhongmu Li (Mar. 2023b). *Blind Search of The Solar Neighborhood Galactic Disk within 5kpc: 1,179 New Star Clusters Found in Gaia DR3*. arXiv: arXiv:2303.10380 (cit. on p. 23).

Chi, Huanbin, Shoulin Wei, Feng Wang, and Zhongmu Li (Dec. 2022). “Identify 46 New Open Clusters Candidates In Gaia EDR3 Using pyUPMASK and Random Forest Hybrid Method”. In: *arXiv e-prints* arXiv:2212.11569 (cit. on p. 74).

D’Antona, F, F Dell’Agli, M Tailo, et al. (Mar. 2023). “On the Role of Dust and Mass-Loss in the Extended Main Sequence Turnoff of Star Clusters: The Case of NGC 1783”. In: *Monthly Notices of the Royal Astronomical Society* 521.3, pp. 4462–4472 (cit. on p. 19).

Dejonghe, Herwig (Jan. 1987). “A Completely Analytical Family of Anisotropic Plummer Models”. In: 224, pp. 13–39 (cit. on p. 31).

Dias, W. S., B. S. Alessi, A. Moitinho, and J. R. D. Lépine (July 2002). “New Catalogue of Optically Visible Open Clusters and Candidates”. In: *Astronomy and Astrophysics* 389, pp. 871–873 (cit. on pp. 2, 8, 73–75).

Dillon, Joshua V., Ian Langmore, Dustin Tran, et al. (Nov. 2017). “TensorFlow Distributions”. In: *arXiv e-prints* arXiv:1711.10604 (cit. on pp. 62, 135).

Ester, Martin, Hans-Peter Kriegel, and Xiaowei Xu (1996). “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *KDD-96 Proceedings*, p. 6 (cit. on p. 48).

Ferreira, F. A., W. J. B. Corradi, F. F. S. Maia, M. S. Angelo, and J. F. C. Santos Jr. (Mar. 2021). “New Star Clusters Discovered towards the Galactic Bulge Direction Using Gaia DR2”. In: *Monthly Notices of the Royal Astronomical Society* 502, pp. L90–L94 (cit. on pp. 46, 74).

Ferreira, Filipe A., J. F. C. Santos, W. J. B. Corradi, F. F. S. Maia, and M. S. Angelo (Mar. 2019). “Three New Galactic Star Clusters Discovered in the Field of the Open Cluster NGC 5999 with Gaia DR2”. In: *Monthly Notices of the Royal Astronomical Society* 483, pp. 5508–5517 (cit. on p. 74).

Ferreira, Filipe A., J. F. C. Santos Jr., W. J. B. Corradi, F. F. S. Maia, and M. S. Angelo (June 2020). “Discovery and Astrophysical Properties of Galactic Open Clusters in Dense Stellar Fields Using Gaia DR2”. In: *Monthly Notices of the Royal Astronomical Society* 496.2, pp. 2021–2038. arXiv: 2006.05611 (cit. on p. 74).

Froebrich, D., A. Scholz, and C. L. Raftery (Jan. 2007). “A Systematic Survey for Infrared Star Clusters with $|b| < 20°$ Using 2MASS”. In: *Monthly Notices of the Royal Astronomical Society* 374, pp. 399–408 (cit. on pp. 8, 15, 72, 73).

Gaia Collaboration, A. G. A. Brown, A. Vallenari, et al. (2021). “Gaia Early Data Release 3: Summary of the Contents and Survey Properties”. In: *Astronomy & Astrophysics* 649, A1. arXiv: 2012.01533 (cit. on pp. 42, 43, 87, 91, 94, 102).

Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, et al. (Nov. 2016). “The Gaia Mission”. In: *Astronomy & Astrophysics* 595, A1 (cit. on pp. 9, 11, 21, 42, 43).

Gaia Collaboration, A. Vallenari, A.G.A. Brown, T. Prusti, and et al. (June 2022). “Gaia Data Release 3. Summary of the Content and Survey Properties”. In: *arXiv e-prints* arXiv:2208.00211 (cit. on pp. 12, 13, 20, 21, 43, 131).

Gal, Yarin and Zoubin Ghahramani (June 2015). “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *arXiv e-prints*, arXiv:1506.02142 (cit. on p. 134).

Goan, Ethan and Clinton Fookes (2020). “Bayesian Neural Networks: An Introduction and Survey”. In: *arXiv e-prints* arXiv:2006.12024 [cs, stat] (cit. on pp. 57, 134, 135).

Golovin, Alex, Sabine Reffert, Andreas Just, et al. (Feb. 2023). “The Fifth Catalogue of Nearby Stars (CNS5)”. In: *Astronomy & Astrophysics* 670, A19 (cit. on p. 45).

Górski, K. M., E. Hivon, A. J. Banday, et al. (Apr. 2005). “HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere”. In: *The Astrophysical Journal* 622.2, p. 759 (cit. on p. 47).

Green, Gregory M. (June 2018). “Dustmaps: A Python Interface for Maps of Interstellar Dust”. In: *Journal of Open Source Software* 3.26, p. 695 (cit. on p. 58).

Hao, C. J., Y. Xu, Z. Y. Wu, et al. (Apr. 2022a). “Newly Detected Open Clusters in the Galactic Disk Using Gaia EDR3”. In: *Astronomy & Astrophysics* 660, A4. arXiv: 2204.00196 (cit. on pp. 16, 74, 75, 97).

Hao, C. J., Y. Xu, Z. Y. Wu, et al. (Sept. 2022b). “Open Clusters Housing Classical Cepheids in Gaia DR3”. In: *Astronomy & Astrophysics*. arXiv: 2210.01521 [astro-ph] (cit. on p. 20).

Hao, ChaoJie, Ye Xu, ZhenYu Wu, ZhiHong He, and ShuaiBo Bian (Mar. 2020). “Sixteen Open Clusters Discovered with Sample-based Clustering Search of Gaia DR2”. In: *Publications of the Astronomical Society of the Pacific* 132, p. 034502 (cit. on pp. 16, 74, 75).

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, et al. (Sept. 2020). “Array Programming with NumPy”. In: *Nature* 585.7825, pp. 357–362 (cit. on p. 88).

He, Zhi-Hong, Ye Xu, Chao-Jie Hao, Zhen-Yu Wu, and Jing-Jing Li (2021). “A Catalogue of 74 New Open Clusters Found in Gaia Data-Release 2”. In: *Research in Astronomy and Astrophysics* 21.4, p. 093. arXiv: 2010.14870 (cit. on pp. 16, 21, 24, 48, 74, 75).

He, Zhihong, Chunyan Li, Jing Zhong, et al. (Mar. 2022a). “New Open Cluster Candidates Found in Galactic Disk Using Gaia DR2/EDR3 Data”. In: *The Astrophysical Journal Supplement Series* 260.1, p. 8. arXiv: 2203.05177 (cit. on pp. 16, 48, 49, 72, 74, 75, 93).

He, Zhihong, Xiaochen Liu, Yangping Luo, Kun Wang, and Qingquan Jiang (Sept. 2022b). “Unveiling Hidden Stellar Aggregates in the Milky Way: 1656 New Star Clusters Found in Gaia EDR3”. In: *The Astrophysical Journal Supplement Series* 264.1, p. 8 (cit. on pp. 16, 74, 75).

He, Zhihong, Kun Wang, Yangping Luo, et al. (June 2022c). “A Blind All-sky Search for Star Clusters in Gaia EDR3: 886 Clusters within 1.2 Kpc of the Sun”. In: *The Astrophysical Journal Supplement Series* 262.1, p. 7 (cit. on pp. 16, 74, 75).

Herschel, John Frederick William (Jan. 1864). “A General Catalogue of Nebulae and Clusters of Stars”. In: *Philosophical Transactions of the Royal Society of London Series I* 154, pp. 1–137 (cit. on p. 7).

Herschel, William (Jan. 1786). “Catalogue of One Thousand New Nebulae and Clusters of Stars. By William Herschel, LL.D. F. R. S.” In: *Philosophical Transactions of the Royal Society of London Series I* 76, pp. 457–499 (cit. on pp. 6, 7).

Hertzsprung, Ejnar (Jan. 1911). “Ueber Die Verwendung Photographischer Effektiver Wellenlaengen Zur Bestimmung von Farbenaequivalenten”. In: *Publikationen des Astrophysikalischen Observatoriums zu Potsdam* 63 (cit. on p. 6).

Hippel, Ted von, William H. Jefferys, James Scott, et al. (July 2006). “Inverting Color-Magnitude Diagrams to Access Precise Star Cluster Parameters: A Bayesian Approach*”. In: *The Astrophysical Journal* 645.2, p. 1436 (cit. on pp. 29, 69).

Hobbs, David, Erik Høg, Alcione Mora, et al. (Sept. 2016). “GaiaNIR: Combining Optical and near-Infra-Red (NIR) Capabilities with Time-Delay-Integration (TDI) Sensors for a Future Gaia-like Mission”. In: *arXiv e-prints*, arXiv:1609.07325. arXiv: 1609.07325 [astro-ph.IM] (cit. on pp. 22, 24).

Hosek Jr, M. W., J. R. Lu, C. Y. Lam, et al. (June 2020). “PyPopStar: A Python-Based Simple Stellar Population Synthesis Code for Star Clusters”. In: *The Astronomical Journal* 160.3, p. 143. arXiv: 2006.06691 (cit. on p. 57).

Hron, Jiri, Alexander G. de G. Matthews, and Zoubin Ghahramani (Nov. 2017). “Variational Gaussian Dropout Is Not Bayesian”. In: *arXiv e-prints* arXiv:1711.02989 (cit. on p. 134).

Huertas-Company, Marc, Vicente Rodriguez-Gomez, Dylan Nelson, et al. (Oct. 2019). “The Hubble Sequence at z $\sim$ 0 in the IllustrisTNG Simulation with Deep Learning”. In: *Monthly Notices of the Royal Astronomical Society* 489, pp. 1859–1879 (cit. on pp. 131, 134).

Hunt, Emily L. and Sabine Reffert (Feb. 2021). “Improving the Open Cluster Census: I. Comparison of Clustering Algorithms Applied to *Gaia* DR2 Data”. In: *Astronomy & Astrophysics* 646, A104 (cit. on pp. 15, 42, 74).

Jaehnig, Karl, Jonathan Bird, and Kelly Holley-Bockelmann (Aug. 2021). “Membership Lists for 431 Open Clusters in Gaia DR2 Using Extreme Deconvolution Gaussian Mixture Models”. In: *The Astrophysical Journal* 923.1, p. 129. arXiv: 2108.02783 (cit. on pp. 16, 42, 74).

Jospin, Laurent Valentin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun (Jan. 2022). “Hands-on Bayesian Neural Networks – a Tutorial for Deep Learning Users”. In: *arXiv e-prints* arXiv:2007.06823 (cit. on pp. 57, 134).

Kharchenko, N. V., A. E. Piskunov, E. Schilbach, S. Röser, and R.-D. Scholz (July 2012). “Global Survey of Star Clusters in the Milky Way - I. The Pipeline and Fundamental Parameters in the Second Quadrant”. In: *Astronomy & Astrophysics* 543, A156 (cit. on pp. 12, 70).

– (Oct. 2013). “Global Survey of Star Clusters in the Milky Way - II. The Catalogue of Basic Parameters”. In: *Astronomy & Astrophysics* 558, A53 (cit. on pp. 9, 12, 14, 16, 23, 29, 31, 47, 55, 65, 68, 70, 72–75, 80–83, 85, 91, 93–96, 98).

Killestein, T. L., J. Lyman, D. Steeghs, et al. (Feb. 2021). “Transient-Optimised Real-Bogus Classification with Bayesian Convolutional Neural Networks – Sifting the GOTO Candidate Stream”. In: *Monthly Notices of the Royal Astronomical Society* 503.4, pp. 4838–4854. arXiv: 2102.09892 (cit. on pp. 15, 62).

King, Ivan (1962). “The Structure of Star Clusters. I. an Empirical Density Law - NASA/ADS”. In: *The Astronomical Journal* 67, p. 471 (cit. on pp. 30, 31).

King, Ivan R. (Feb. 1966). “The Structure of Star Clusters. III. Some Simple Dynamical Models”. In: *The Astronomical Journal* 71, p. 64 (cit. on p. 31).

Kingma, Diederik P. and Jimmy Ba (Jan. 2017). “Adam: A Method for Stochastic Optimization”. In: *arXiv e-prints* arXiv:1412.6980 (cit. on p. 62).

Kounkel, Marina and Kevin Covey (Aug. 2019). “Untangling the Galaxy. I. Local Structure and Star Formation History of the Milky Way”. In: *The Astronomical Journal* 158.3, p. 122 (cit. on p. 48).

Kounkel, Marina, Kevin Covey, and Keivan G. Stassun (Oct. 2020). “Untangling the Galaxy. II. Structure within 3 Kpc”. In: *The Astronomical Journal* 160.6, p. 279. arXiv: 2004.07261 (cit. on pp. 30, 48, 49, 56, 64, 72–75, 80, 82, 95, 96).

Kovaleva, Dana, Marina Ishchenko, Ekaterina Postnikova, et al. (Sept. 2020). “Collinder 135 and UBC 7: A Physical Pair of Open Clusters”. In: *Astronomy & Astrophysics* 642, p. L4. arXiv: 2009.02223 (cit. on p. 54).

Krause, Martin G. H., Stella S. R. Offner, Corinne Charbonnel, et al. (May 2020). “The Physics of Star Cluster Formation and Evolution”. In: *Space Science Reviews* 216.4, p. 64. arXiv: 2005.00801 (cit. on p. 41).

Krone-Martins, A. and A. Moitinho (Jan. 2014). “UPMASK: Unsupervised Photometric Membership Assignment in Stellar Clusters”. In: *Astronomy and Astrophysics* 561, A57 (cit. on pp. 67, 86).

Kroupa, Pavel (Apr. 2001). “On the Variation of the Initial Mass Function”. In: *Monthly Notices of the Royal Astronomical Society* 322, pp. 231–246 (cit. on p. 57).

Krumholz, Mark R., Christopher F. McKee, and Joss Bland-Hawthorn (Aug. 2019). “Star Clusters Across Cosmic Time”. In: *Annual Review of Astronomy and Astrophysics* 57.1, pp. 227–303. arXiv: 1812.01615 (cit. on pp. 41, 90).

Kuhn, Michael A., Lynne A. Hillenbrand, Alison Sills, Eric D. Feigelson, and Konstantin V. Getman (Jan. 2019). “Kinematics in Young Star Clusters and Associations with Gaia DR2”. In: *The Astrophysical Journal* 870, p. 32 (cit. on p. 100).

Lada, Charles J. and Elizabeth A. Lada (2003). “Embedded Clusters in Molecular Clouds”. In: *Annual Review of Astronomy and Astrophysics* 41, p. 57 (cit. on p. 1).

Leiner, Emily M. and Aaron Geller (Feb. 2021). “A Census of Blue Stragglers in Gaia DR2 Open Clusters as a Test of Population Synthesis and Mass Transfer Physics”. In: 908.2, p. 229. arXiv: 2101.11047 [astro-ph.SR] (cit. on p. 19).

Leung, Henry W. and Jo Bovy (Mar. 2019). “Deep Learning of Multi-Element Abundances from High-Resolution Spectroscopic Data”. In: *Monthly Notices of the Royal Astronomical Society* 483, pp. 3255–3277 (cit. on p. 134).

Li, Zhongmu, Yangyang Deng, Huanbin Chi, et al. (Feb. 2022). “LISC Catalog of Star Clusters. I. Galactic Disk Clusters in Gaia EDR3”. In: *The Astrophysical Journal Supplement Series* 259.1, p. 19 (cit. on p. 74).

Li, Zhongmu and Caiyan Mao (Feb. 2023). “LISC Catalog of Star Clusters. II. High Galactic Latitude Open Clusters in Gaia EDR3”. In: *The Astrophysical Journal Supplement Series* 265.1, p. 3 (cit. on pp. 74, 75).

Lin, Yu-Chiung and Jiun-Huei Proty Wu (Mar. 2021). “Detection of Gravitational Waves Using Bayesian Neural Networks”. In: *Physical Review D* 103, p. 063034 (cit. on p. 134).

Lindegren, L., U. Bastian, M. Biermann, et al. (2021a). “Gaia Early Data Release 3: Parallax Bias versus Magnitude, Colour, and Position”. In: *Astronomy & Astrophysics* 649, A4. arXiv: 2012.01742 (cit. on p. 45).

Lindegren, L., J. Hernández, A. Bombrun, et al. (Aug. 2018). “Gaia Data Release 2. The Astrometric Solution”. In: *Astronomy and Astrophysics* 616, A2 (cit. on p. 71).

Lindegren, L., S. A. Klioner, J. Hernández, et al. (2021b). “Gaia Early Data Release 3 The Astrometric Solution”. In: *Astronomy & Astrophysics* 649, A2. arXiv: 2012.03380 (cit. on pp. 44, 55, 132).

Liu, Lei and Xiaoying Pang (Oct. 2019). “A Catalog of Newly Identified Star Clusters in GAIA DR2”. In: *The Astrophysical Journal Supplement Series* 245.2, p. 32. arXiv: 1910.12600 (cit. on pp. 2, 16, 21, 24, 25, 42, 49, 74, 90).

Lodieu, N., R. L. Smart, A. Pérez-Garrido, and R. Silvotti (Mar. 2019). “A 3D View of the Hyades Stellar and Sub-Stellar Population”. In: 623, A35. arXiv: 1901.07534 [astro-ph.SR] (cit. on p. 17).

Lu, J. R., T. Do, A. M. Ghez, et al. (Feb. 2013). “STELLAR POPULATIONS IN THE CENTRAL 0.5 Pc OF THE GALAXY. II. THE INITIAL MASS FUNCTION”. In: *The Astrophysical Journal* 764.2, p. 155 (cit. on p. 57).

Marigo, Paola, Léo Girardi, Alessandro Bressan, et al. (Jan. 2017). “A NEW GENERATION OFPARSEC-COLIBRISTELLAR ISOCHRONES INCLUDING THE TP-AGB PHASE”. In: *The Astrophysical Journal* 835.1, p. 77 (cit. on p. 57).

Marino, A. F., A. P. Milone, L. Casagrande, et al. (Aug. 2018). “Discovery of Extended Main Sequence Turnoffs in Galactic Open Clusters”. In: *The Astrophysical Journal* 863, p. L33 (cit. on p. 19).

McArthur, Barbara E., G. Fritz Benedict, Thomas E. Harrison, and William van Altena (May 2011). “Astrometry with the Hubble Space Telescope: Trigonometric Parallaxes of Selected Hyads”. In: *The Astronomical Journal* 141, p. 172 (cit. on p. 55).

McInnes, Leland, John Healy, and Steve Astels (Mar. 2017). “Hdbscan: Hierarchical Density Based Clustering”. In: *Journal of Open Source Software* 2.11, p. 205 (cit. on pp. 45, 48).

Medina, Gustavo E., Bertrand Lemasle, and Eva K. Grebel (Apr. 2021). “A Revisited Study of Cepheids in Open Clusters in the Gaia Era”. In: *arXiv:2104.14565 [astro-ph]*. arXiv: 2104.14565 [astro-ph] (cit. on pp. 2, 20).

Meingast, Stefan and João Alves (Jan. 2019). “Extended Stellar Systems in the Solar Neighborhood. I. The Tidal Tails of the Hyades”. In: 621, p. L3. arXiv: 1811.04931 [astro-ph.GA] (cit. on p. 17).

Meingast, Stefan, João Alves, and Alena Rottensteiner (2021). “Extended Stellar Systems in the Solar Neighborhood – V. Discovery of Coronae of Nearby Star Clusters”. In: *Astronomy & Astrophysics* 645, A84 (cit. on pp. 17, 42).

Mermilliod, J.-C. (1995). “The Database for Galactic Open Clusters (BDA)”. In: *Information \& On-Line Data in Astronomy*. Vol. 203. Springer Netherlands, pp. 127–138 (cit. on pp. 8, 9).

Milone, Antonino P. and Anna F. Marino (June 2022). *Multiple Populations in Star Clusters*. arXiv: arXiv:2206.10564 (cit. on p. 19).

Oh, Semyeong and Neil Wyn Evans (Oct. 2020). “Kinematic Modelling of Clusters with Gaia: The Death Throes of the Hyades”. In: 498.2, pp. 1920–1938. arXiv: 2007.02969 [astro-ph.SR] (cit. on p. 17).

Pang, Xiaoying, Yuqian Li, Zeqiu Yu, et al. (May 2021). “3D Morphology of Open Clusters in the Solar Neighborhood with Gaia EDR 3: Its Relation to Cluster Dynamics”. In: *The Astrophysical Journal* 912, p. 162 (cit. on pp. 17, 32, 100).

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. (2011). “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 21).

Penoyre, Zephyr, Vasily Belokurov, and N. Wyn Evans (June 2022). “Astrometric Identification of Nearby Binary Stars - I. Predicted Astrometric Signals”. In: *Monthly Notices of the Royal Astronomical Society* 513, pp. 2437–2456 (cit. on pp. 11, 45).

Perryman, M. a. C., L. Lindegren, J. Kovalevsky, et al. (July 1997). “The HIPPARCOS Catalogue”. In: *Astronomy and Astrophysics, Vol. 323, p.L49-L52* 323, p. L49 (cit. on pp. 7, 42).

Piatti, A. E., D. M. F. Illesca, A. A. Massara, et al. (2023). “Assessing the Physical Reality of Milky Way Open Cluster Candidates”. In: *Monthly Notices of the Royal Astronomical Society* 518.4, pp. 6216–6222. arXiv: 2211.15483 [astro-ph] (cit. on p. 87).

Piatti, Andrés E. (Jan. 2023). *Catching a Milky Way Open Cluster in Its Last Breath*. arXiv: arXiv:2301.04031 (cit. on pp. 2, 22).

Piskunov, A. E., E. Schilbach, N. V. Kharchenko, S. Röser, and R.-D. Scholz (June 2007). “Towards Absolute Scales for the Radii and Masses of Open Clusters”. In: *Astronomy & Astrophysics* 468.1, pp. 151–161 (cit. on p. 30).

Planck Collaboration, N. Aghanim, Y. Akrami, et al. (Sept. 2020). “Planck 2018 Results. VI. Cosmological Parameters”. In: 641, A6. arXiv: 1807.06209 [astro-ph.CO] (cit. on p. 20).

Platais, Imants, Vera Kozhurina-Platais, and Floor van Leeuwen (Nov. 1998). “A Search for Star Clusters from the HIPPARCOS Data”. In: *The Astronomical Journal* 116, pp. 2423–2430 (cit. on p. 7).

Plummer, H. C. (Mar. 1911). “On the Problem of Distribution in Globular Star Clusters”. In: *Monthly Notices of the Royal Astronomical Society* 71, pp. 460–470 (cit. on pp. 31, 32, 98).

Portegies Zwart, Simon F., Stephen L. W. McMillan, and Mark Gieles (Sept. 2010). “Young Massive Star Clusters”. In: *Annual Review of Astronomy and Astrophysics* 48, pp. 431–493 (cit. on pp. 1, 4, 17, 23, 24, 30–32, 41, 86, 95, 100).

Qin, Song-mei, Jing Li, Li Chen, and Jing Zhong (2021). “Discovery of Four New Clusters in the Cygnus Cloud”. In: *Research in Astronomy and Astrophysics* 21.2, p. 045. arXiv: 2008.07164 (cit. on pp. 16, 74).

Qin, Songmei, Jing Zhong, Tong Tang, and Li Chen (2023). “Hunting for Neighboring Open Clusters with Gaia DR3: 101 New Open Clusters within 500 Pc”. In: *The Astrophysical Journal Supplement Series* 265.1, p. 12. arXiv: 2212.11034 [astro-ph] (cit. on pp. 16, 74, 93).

Rain, M. J., G. Carraro, J. Ahumada, et al. (Oct. 2020). “The Blue Straggler Population of the Open Clusters Trumpler 5, Trumpler 20, and NGC 2477”. In: *arXiv:2010.06884 [astro-ph]*. arXiv: 2010.06884 [astro-ph] (cit. on p. 19).

Ratzenböck, Sebastian, Josefa E. Großschedl, Torsten Möller, et al. (Nov. 2022). “Significance Mode Analysis (SigMA) for Hierarchical Structures. An Application to the Sco-Cen OB Association”. In: *arXiv e-prints*, arXiv:2211.14225 (cit. on p. 102).

Reid, M. J., K. M. Menten, A. Brunthaler, et al. (Mar. 2014). “Trigonometric Parallaxes of High Mass Star Forming Regions: The Structure and Kinematics of the Milky Way”. In: 783.2, p. 130. arXiv: 1401.5377 [astro-ph.GA] (cit. on p. 18).

Reino, Stella, Jos de Bruijne, Eleonora Zari, Francesca d’Antona, and Paolo Ventura (July 2018). “A Gaia Study of the Hyades Open Cluster”. In: 477.3, pp. 3197–3216. arXiv: 1804.00759 [astro-ph.GA] (cit. on p. 17).

Riello, M., F. De Angeli, D. W. Evans, et al. (2021). “Gaia Early Data Release 3: Photometric Content and Validation”. In: *Astronomy & Astrophysics* 649, A3. arXiv: 2012.01916 (cit. on pp. 45, 58, 63).

Rosenberg, H. (Oct. 1910). “Über Den Zusammenhang von Helligkeit Und Spektraltypus in Den Plejaden”. In: *Astronomische Nachrichten* 186.5, p. 71 (cit. on p. 6).

Röser, Siegfried, Elena Schilbach, and Bertrand Goldman (Jan. 2019). “Hyades Tidal Tails Revealed by Gaia DR2”. In: 621, p. L2. arXiv: 1811.03845 [astro-ph.SR] (cit. on p. 17).

Russell, Henry Norris (May 1914). “Relations between the Spectra and Other Characteristics of the Stars”. In: *Popular Astronomy* 22, pp. 275–294 (cit. on p. 6).

Rybizki, Jan, Gregory Green, Hans-Walter Rix, et al. (2022). “A Classifier for Spurious Astrometric Solutions in Gaia EDR3”. In: *Monthly Notices of the Royal Astronomical Society* 510.2, pp. 2597–2616 (cit. on pp. 44–46).

Santos-Silva, T., H. D. Perottoni, F. Almeida-Fernandes, et al. (Nov. 2021). “Canis Major OB1 Stellar Group Contents Revealed by Gaia”. In: *Monthly Notices of the Royal Astronomical Society* 508, pp. 1033–1055 (cit. on p. 74).

Schmeja, S., N. V. Kharchenko, A. E. Piskunov, et al. (Aug. 2014). “Global Survey of Star Clusters in the Milky Way - III. 139 New Open Clusters at High Galactic Latitudes”. In: *Astronomy & Astrophysics* 568, A51 (cit. on p. 75).

Scholz, R.-D., N. V. Kharchenko, A. E. Piskunov, S. Röser, and E. Schilbach (Sept. 2015). “Global Survey of Star Clusters in the Milky Way - IV. 63 New Open Clusters Detected by Proper Motions”. In: *Astronomy & Astrophysics* 581, A39 (cit. on p. 22).

Sim, Gyuheon, Sang Hyun Lee, Hong Bae Ann, and Seunghyeon Kim (Oct. 2019). “207 New Open Star Clusters within 1 Kpc from Gaia Data Release 2”. In: *Journal of the Korean Astronomical Society* 52, pp. 145–158. arXiv: 1907.06872 (cit. on pp. 16, 74, 92).

Skrutskie, M. F., R. M. Cutri, R. Stiening, et al. (Feb. 2006). “The Two Micron All Sky Survey (2MASS)”. In: *The Astronomical Journal* 131, pp. 1163–1183 (cit. on pp. 8, 93).

Tarricq, Y., C. Soubiran, L. Casamiquela, et al. (Dec. 2020). “3D Kinematics and Age Distribution of the Open Cluster Population”. In: *arXiv:2012.04017 [astro-ph]*. arXiv: 2012.04017 [astro-ph] (cit. on p. 2).

Tarricq, Y., C. Soubiran, L. Casamiquela, et al. (2022). “Structural Parameters of 389 Local Open Clusters”. In: *Astronomy & Astrophysics* 659, A59. arXiv: 2111.05291 (cit. on pp. 17, 30, 31, 42, 48, 54, 55, 81, 83, 85).

Tian, Hai-Jun (Sept. 2020). “Discovery of a Young Stellar "Snake" with Two Dissolving Cores in the Solar Neighborhood”. In: *The Astrophysical Journal* 904.2, p. 196. arXiv: 2005.12265 (cit. on p. 74).

Vaidya, Kaushar, Khushboo K. Rao, Manan Agarwal, and Souradeep Bhattacharya (June 2020). “Blue Straggler Populations of Seven Open Clusters with Gaia DR2”. In: *arXiv:2006.05189 [astro-ph]*. arXiv: 2006.05189 [astro-ph] (cit. on p. 19).

van Leeuwen, Floor (2007). *Hipparcos, the New Reduction of the Raw Data*. Vol. 350 (cit. on pp. 12, 13).

Vasiliev, Eugene and Holger Baumgardt (Feb. 2021). “Gaia EDR3 View on Galactic Globular Clusters”. In: *Monthly Notices of the Royal Astronomical Society* 505.4, pp. 5978–6002. arXiv: 2102.09568 (cit. on pp. 73–75, 90, 132).

Wen, Yeming, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse (Apr. 2018). “Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches”. In: *arXiv e-prints* arXiv:1803.04386 (cit. on p. 134).

Xu, Dongkuan and Yingjie Tian (June 2015). “A Comprehensive Survey of Clustering Algorithms”. In: *Annals of Data Science* 2.2, pp. 165–193 (cit. on pp. 21, 26).

Yeh, Fu Chi, Giovanni Carraro, Marco Montalto, and Anton F. Seleznev (Feb. 2019). “Ruprecht 147: A Paradigm of Dissolving Star Cluster”. In: *The Astronomical Journal* 157.3, p. 115 (cit. on p. 17).

Yen, Steffi X., Sabine Reffert, Elena Schilbach, et al. (July 2018). “Reanalysis of Nearby Open Clusters Using Gaia DR1/TGAS and HSOY”. In: *Astronomy & Astrophysics* 615, A12 (cit. on pp. 29, 30).

Zari, E., H. Hashemi, A. G. A. Brown, K. Jardine, and P. T. de Zeeuw (Dec. 2018). “3D Mapping of Young Stars in the Solar Neighbourhood with Gaia DR2”. In: *Astronomy & Astrophysics* 620, A172 (cit. on pp. 16, 74).

Zhou, Xiaoyue and Xiaodian Chen (Apr. 2021). “Galactic Open Cluster Cepheids – a Census Based on Gaia EDR3”. In: *arXiv:2104.11929 [astro-ph]*. arXiv: `2104.11929 [astro-ph]` (cit. on p. 20).

Zonca, Andrea, Leo P. Singer, Daniel Lenz, et al. (Mar. 2019). “Healpy: Equal Area Pixelization and Spherical Harmonics Transforms for Data on the Sphere in Python”. In: *Journal of Open Source Software* 4.35, p. 1298 (cit. on p. 47).

List of Figures

List of Tables

List of Listings

Appendix A

A.1 Appendices for Chapter 3

A.1.1 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 et al. 2022), except while also having columns referencing the cluster name and cluster ID we assign them to, the cluster membership probability, and a flag for if the star is a member within our estimated tidal radius r_t.

A.1.2 Table of crossmatch results

Here we provide a table of all crossmatches to all literature clusters that meet our adopted crossmatch criteria from Sect. 3.6 in Table A.2. 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 3.4 that corresponds to this object. For clusters that we do not redetect, only a blank row with the cluster name, source paper, and type of crossmatch is shown.

A.1.3 Bayesian neural networks

Given that Bayesian neural networks (BNNs) are only just beginning to see use in the astronomical literature (e.g. Huertas-Company et al. 2019), here we provide a brief background overview of the advantages and caveats of the approximate BNN methodology we adopted in Sect. 3.4 and Sect. 3.5.

BNNs are a somewhat elusive area of open research in machine learning. Their appeal is clear: unlike a deterministic approach or an approach based on simply

Tab. A.1.: Description of the columns in the tables of detected clusters.

Col.	Label	Unit	Description
1	Name	–	Designation
2	Internal ID	–	Internal designation
3	All names	–	All literature names
4	Kind	–	Estimated object type[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 and Baumgardt 2021 GC catalogue, otherwise `o` (OC) or `m` (moving group) for clusters according to the empirical cuts in Cantat-Gaudin and Anders 2020. [(d)] Flag indicating six clusters for which parallax bias correction using the method of Lindegren et al. 2021b was not possible, and a global offset was used instead (see Sect. 3.3.3). [(e)] Corrected using cluster distances to be relative to cluster centre. [(f)] Cluster CMD classes derived using the neural network in Sect. 3.4. [(g)] Indicates 25 clusters merged by hand (see Sect. 3.3). [(h)] Indicates nine clusters with members from an additional Gaussian mixture model clustering step. [(i)] Method used to assign name to cluster (see Sect. 3.6.3.)

Tab. A.2.: 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
					...						
176	Basel 1	Cantat-Gaudin+20	gaia dr2	0.01	0.04	0.03	0.00	0.01	0.00	0.01	0.00
176	Basel 1	Dias+02	position	0.03	0.12	-	-	-	-	-	-
176	Basel 1	Kharchenko+13	hipparcos	0.01	0.04	0.40	0.09	1.04	0.24	0.06	0.17
179	Basel 10	Bica+18	position	0.01	0.04	-	-	-	-	-	-
179	Basel 10	Dias+02	position	0.01	0.04	-	-	-	-	-	-
179	Basel 10	Cantat-Gaudin+20	gaia dr2	0.01	0.07	0.03	0.00	0.05	0.01	0.01	0.00
179	Basel 10	Kharchenko+13	hipparcos	0.01	0.07	0.30	0.05	2.49	0.51	0.02	0.00
179	Basel 10	Kharchenko+13	position	0.01	0.07	-	-	-	-	-	-
183	Basel 11A	Cantat-Gaudin+20	gaia dr2	0.01	0.01	0.02	0.00	0.05	0.00	0.02	0.00
183	Basel 11A	Kharchenko+13	hipparcos	0.01	0.04	0.52	0.12	1.66	0.42	0.11	0.81
183	Basel 11A	Dias+02	position	0.02	0.06	-	-	-	-	-	-
183	Basel 11A	Bica+18	position	0.03	0.06	-	-	-	-	-	-
183	Basel 11A	Kharchenko+13	position	0.01	0.04	-	-	-	-	-	-
3003	Basel 11B	Kharchenko+13	position	0.11	0.25	-	-	-	-	-	-
184	Basel 11B	Kharchenko+13	hipparcos	0.02	0.06	1.28	0.37	0.24	0.06	0.17	1.40
184	Basel 11B	Kharchenko+13	position	0.02	0.06	-	-	-	-	-	-
184	Basel 11B	Dias+02	position	0.01	0.02	-	-	-	-	-	-
184	Basel 11B	Cantat-Gaudin+20	gaia dr2	0.01	0.03	0.02	0.00	0.01	0.00	0.03	0.00
184	Basel 11B	Bica+18	position	0.00	0.01	-	-	-	-	-	-
6363	Basel 11B	Kharchenko+13	hipparcos	0.11	0.39	2.15	0.64	1.99	0.59	0.22	1.98
6363	Basel 11B	Kharchenko+13	position	0.11	0.39	-	-	-	-	-	-
					...						

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. 3.6, 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, $\theta_r = \theta / \max(r_t,\, r_{t,\mathrm{lit}})$

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

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

Many approaches for variational inference exist in the literature, with a common approach being dropout regularisation as an approximation of a BNN (Gal and Ghahramani 2015), having also been used within astronomy (e.g. Huertas-Company et al. 2019, Leung and Bovy 2019). However, this approximation is not inherently Bayesian (Hron et al. 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 et al. 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 and Fookes 2020). Due to the time-consuming nature of repeated samples of all kernel and bias posterior distributions, we also apply an approximation known as Flipout to more efficiently sample them with a lower runtime while preserving good training characteristics (Wen et al. 2018). Similar approaches using Bayes by backpropagation and Flipout have seen some use in the astronomy literature (e.g. Lin and Wu 2021). We use the implementations of `DenseFlipout`

and `Convolution2DFlipout` layers in TensorFlow Probability (Dillon et al. 2017), minimising the evidence lower bound (ELBO) loss (Blundell et al. 2015).

In initial tests, these approximations produced network outputs with reliable uncertainty estimates that correspond well to the uncertainty inherent to classifying star cluster CMDs. It is worth noting from the literature that variational-inference based approaches are still more overconfident than a true BNN when applied to unseen data (Goan and Fookes 2020), and that this approach is still an imperfect estimator of the true uncertainty of our model; nevertheless, our adopted method was found to be as accurate as a traditional deterministic network architecture of the same configuration when applied to our training data, but while providing an estimate of its uncertainty and without dramatically increasing runtime during training or sampling.

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/`.

Declaration

I hereby declare that this thesis is my own work and that I have used no other than the stated sources and aids.

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, May 2nd, 2023

Emily Lauren Hunt