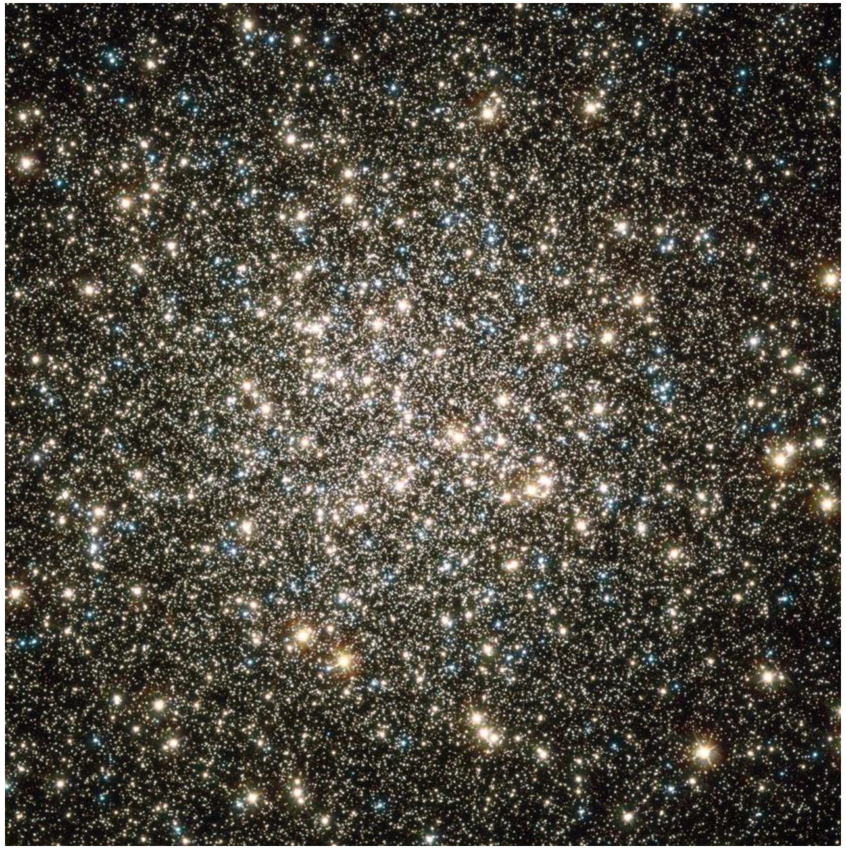
USING MACHINE LEARNING ALGORITHMS TO EXTRACT GLOBULAR CLUSTERS IN GAIA DR3



N.J. Baloyi[1], Z.M. Mguda[1], A. Faltenbacher[2]

[1]Unisa Centre for Astrophysics and Space Sciences (U-CASS), University of South Africa

[2]One Data München, Bavaria, Germany

baloynj@unisa.ac.za

UNISA | centre for astrophysics and space sciences

AfAS African Astronomical Society

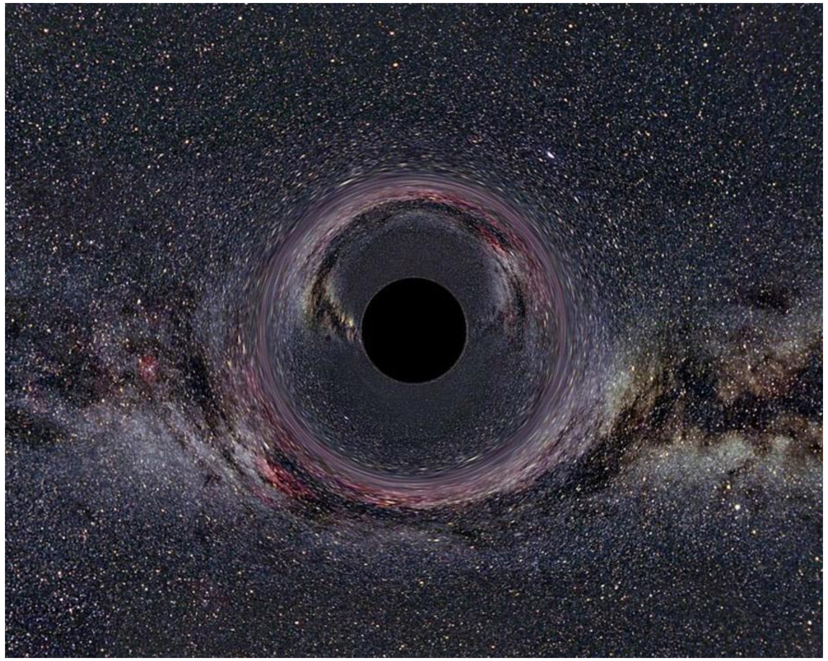
BIUST BOTSWANA INTERNATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

1. BACKGROUND ON GLOBULAR CLUSTERS

BRIEF HISTORY & PROPERTIES

- GCs are a spherical collection of stars bound by gravity, with a higher concentration of stars towards centre.
- Old (up to 13.5 Gyr), metal-poor, with masses up to $\sim 10^6\ M_\odot$
- Located in bulge, thick disk and stellar halo
- First known GC, Messier 22 (M22), was discovered in 1665 by Abraham Ihle
- Messier (1781) constructed the first-of-its-kind catalogue of “Nebulae and Star Clusters”, consisting of 28 known GCs
- Milky Way GC counts continued to increase over the next few centuries: ~80 in the 1910s to >160 by the 2010s.

RELEVANCE AND OPEN QUESTIONS



Relevance:

- Provide link between stellar halo of MW and its satellite dwarf systems
- Serve as probes for star and galaxy formation and evolution in the early, **metal-poor**, Universe
- Important for studying exotic phenomena such as blue stragglers and intermediate-mass black holes $\left(10^2 - 10^5\, M_\odot\right)$.

Ongoing research:

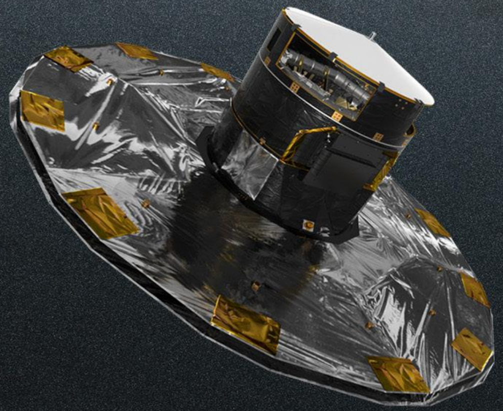
- Formation mechanisms? (e.g. core collapse)
- Connection between GCs and dwarfs (Taylor et al. 2025)
- Potential sites of supermassive black hole seeds? (Bañares-Hernández et al. 2025)

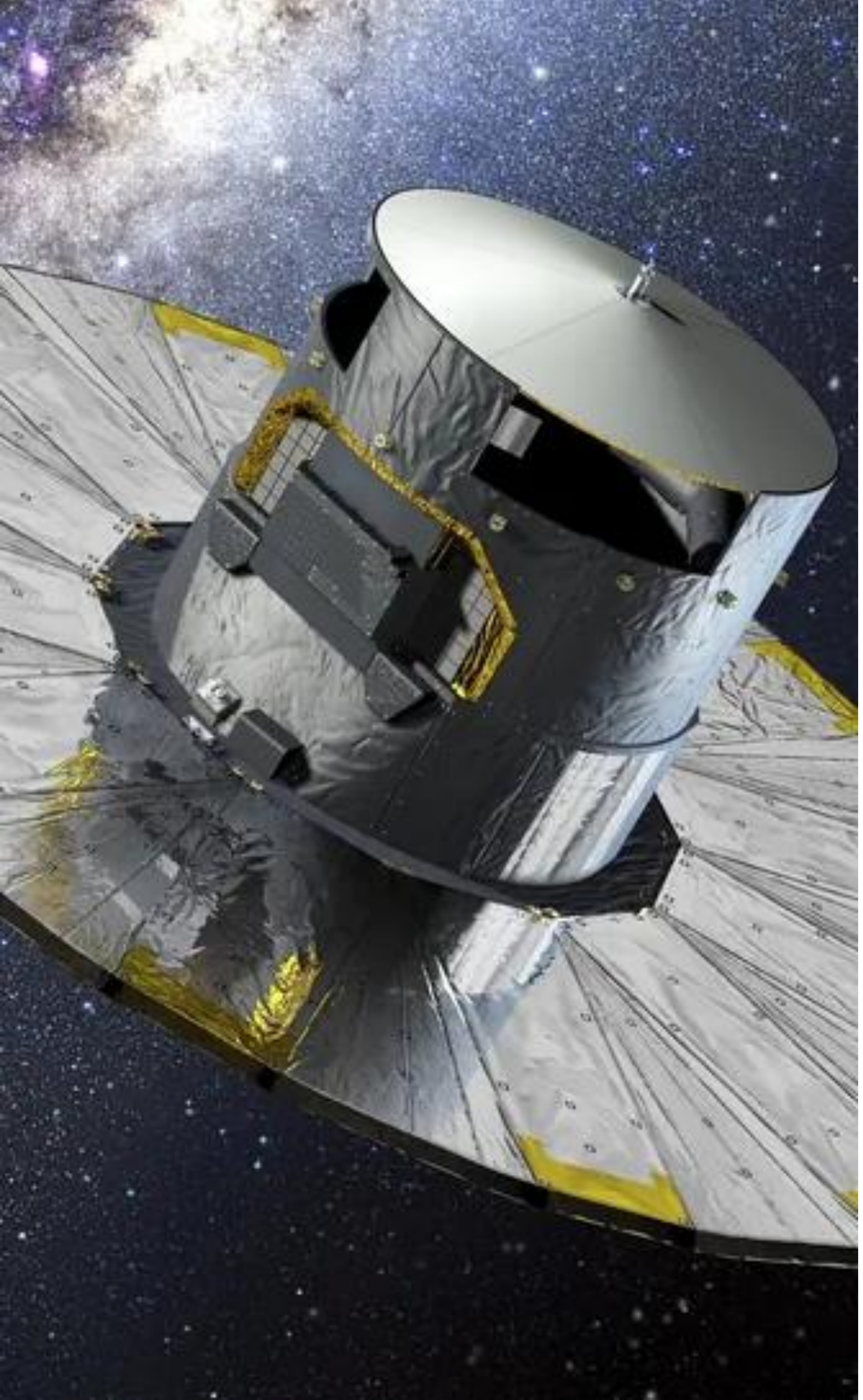
RESEARCH CONTEXT AND AIMS

- Modern astronomical surveys (e.g. *Gaia*, HST, JWST) observe millions of sources at unprecedented sensitivities
- BUT the resulting large data products make manual perusal impractical
- **Machine Learning** (ML) and **Artificial Intelligence** to “comb through” the data!
- **Main aim:** Carry out a blind, automated search for GCs in *Gaia* DR3, in regions beyond the thick Galactic disk…

2. DATA AND METHODS

Working with *Gaia* Data and Machine Learning techniques



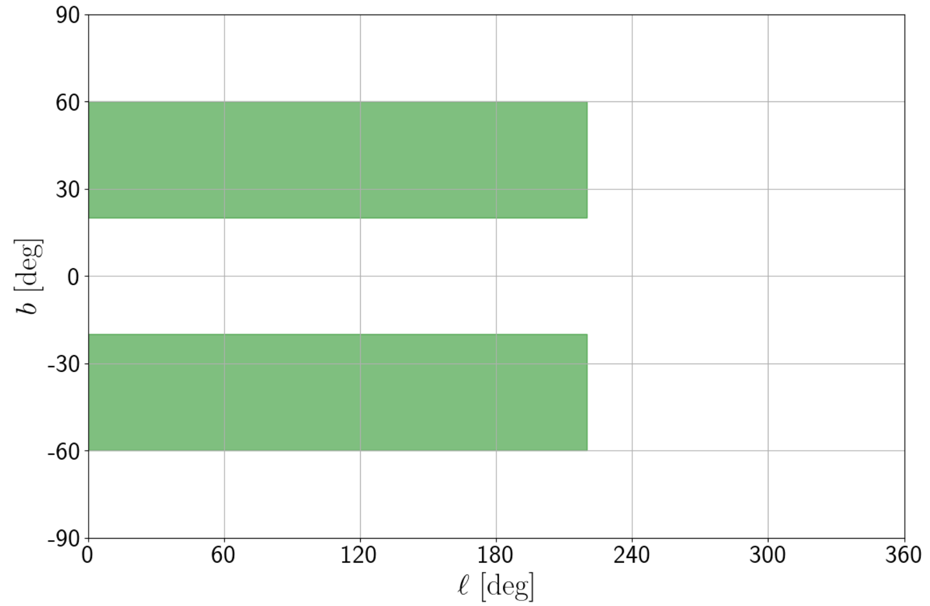


THE **Gaia** SURVEY

- *Gaia* is a space-based telescope mission of the European Space Agency
- Three main science functions: high-precision **astrometry, photometry and spectroscopy**
- Third data release (DR3) contains over 1.8 Billion sources (stars, galaxies and quasars)
- Unprecedented astrometric precision – down to 0.01 mas, equivalent to modern VLBI observations.

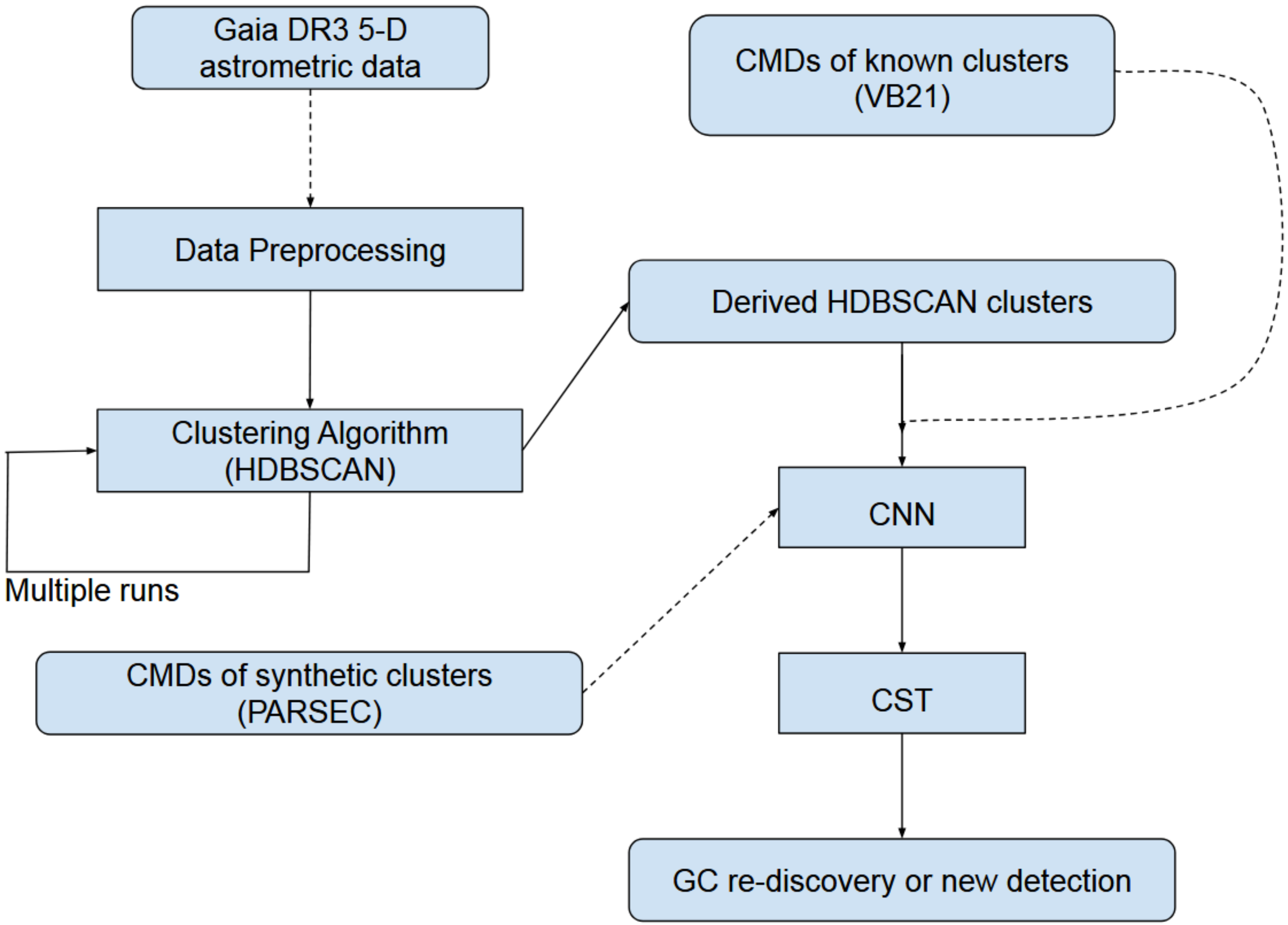
From Earth, *Gaia* would be able to see a 5 ZAR/BWP coin on the Moon!

DATA SELECTION

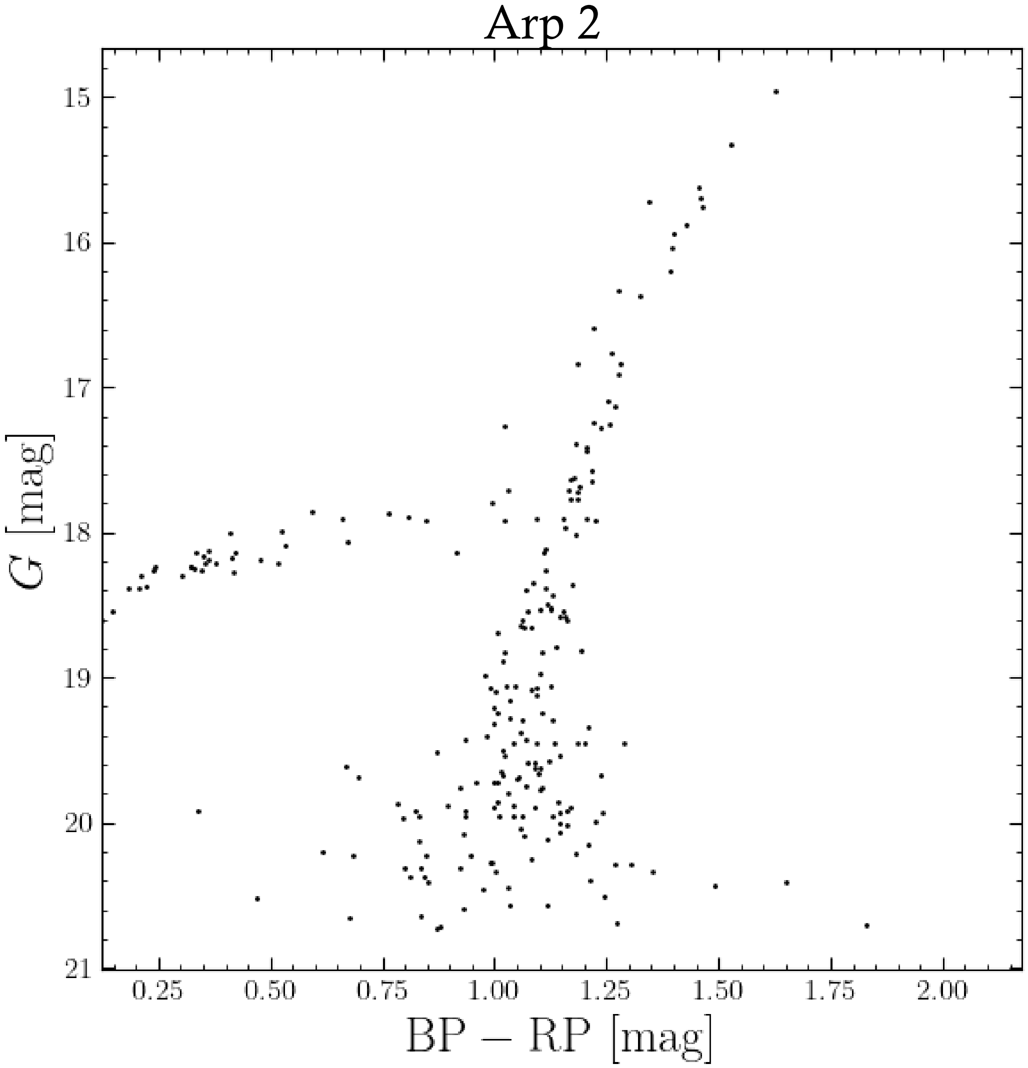


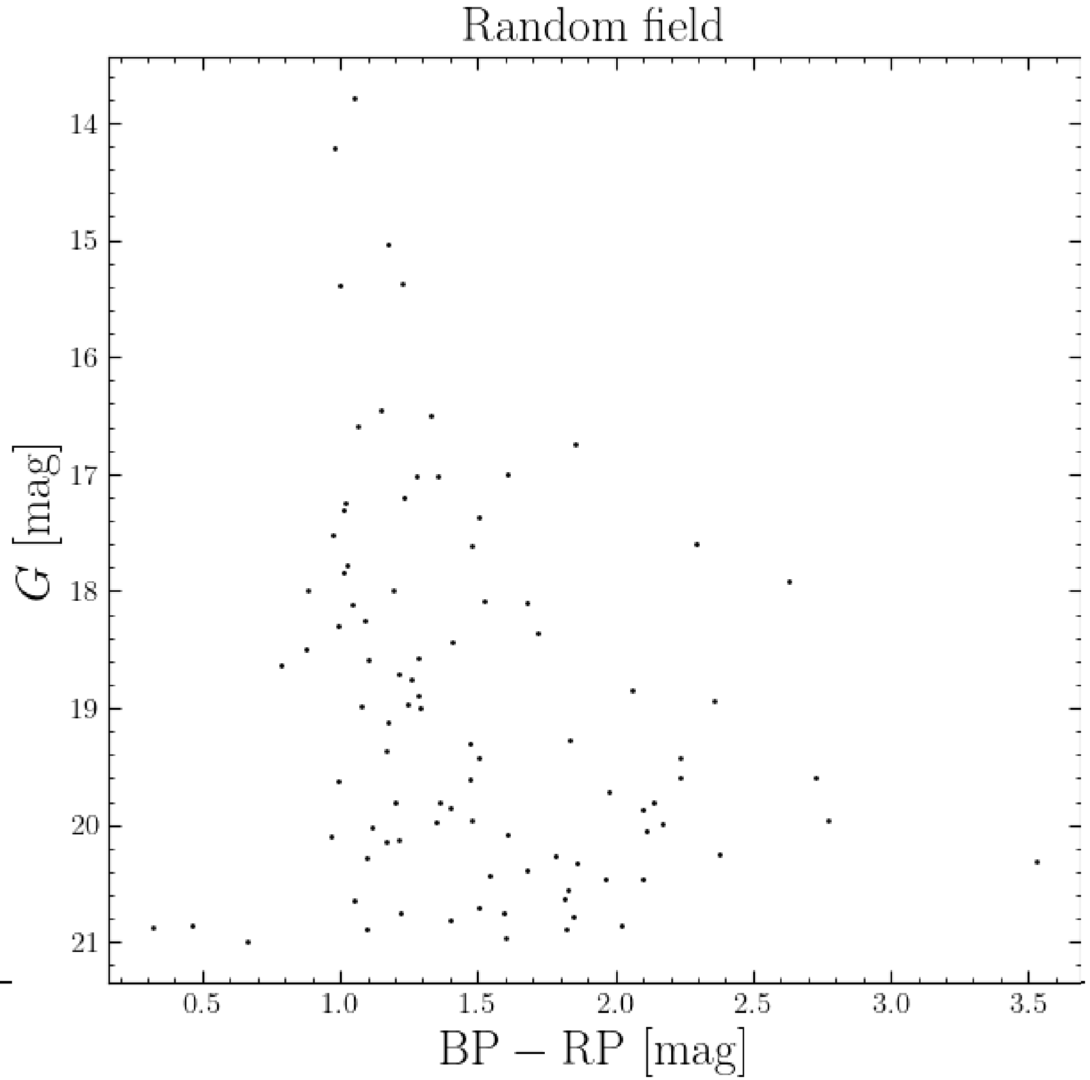
- 2 regions described by $0° < l < 220°$ and $20° < |b| < 60°$.
- **Some parameters used:**
  - 5-parameter astrometry, i.e. $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$,
  - Photometric G-band magnitude – denoted by **G**
  - **BP – RP**: colour difference between the blue and red photometers)
- **Some of the selection criteria:**
  - RUWE < 1.15 (Renormalised unit weight error)
  - Parallax/parallax error > –3
  - Visibility periods used ≥ 10
- Search regions contain over 109M sources, which after filtering, reduced to 65M (~60%)

PIPELINE SUMMARY

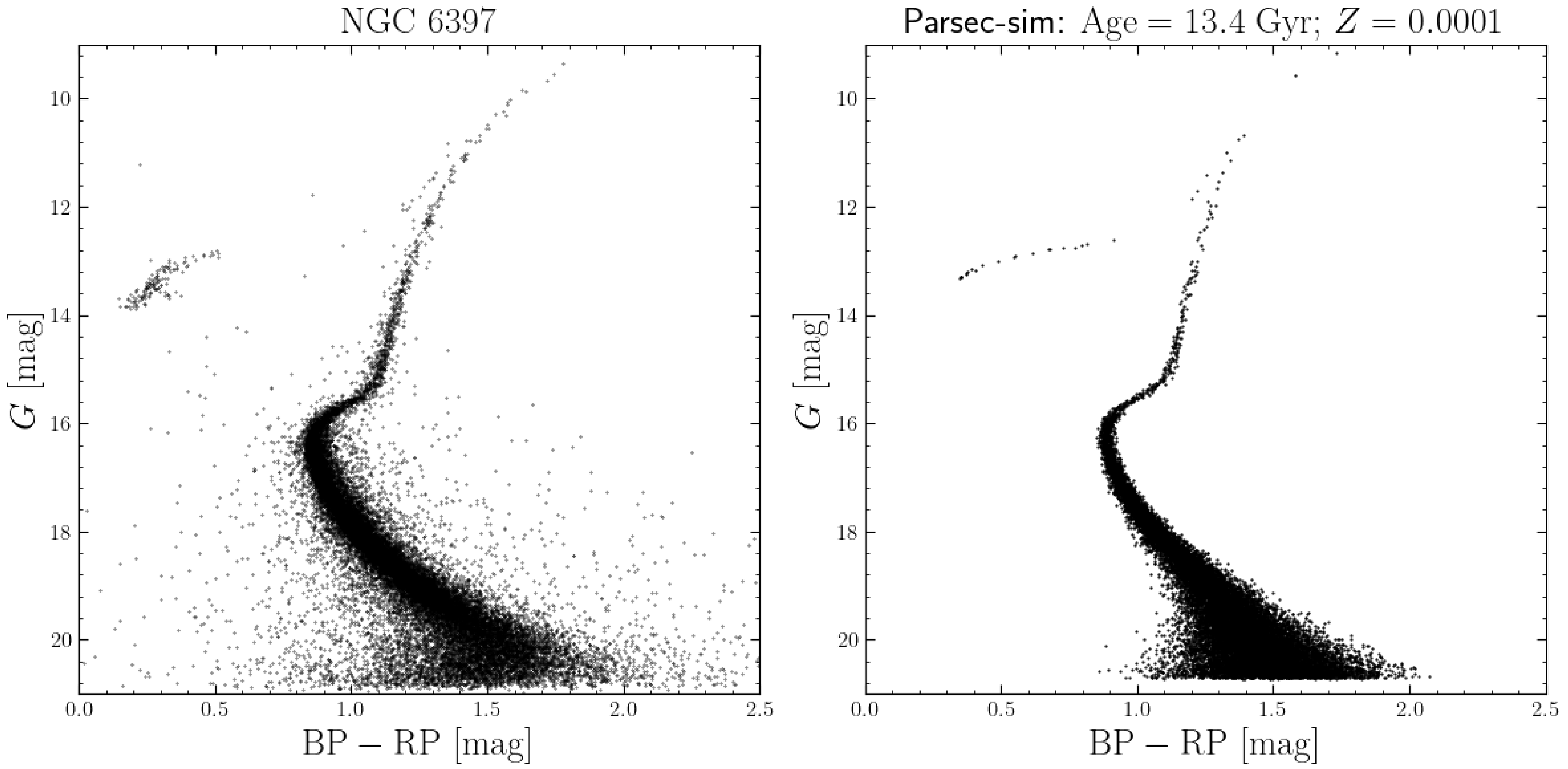


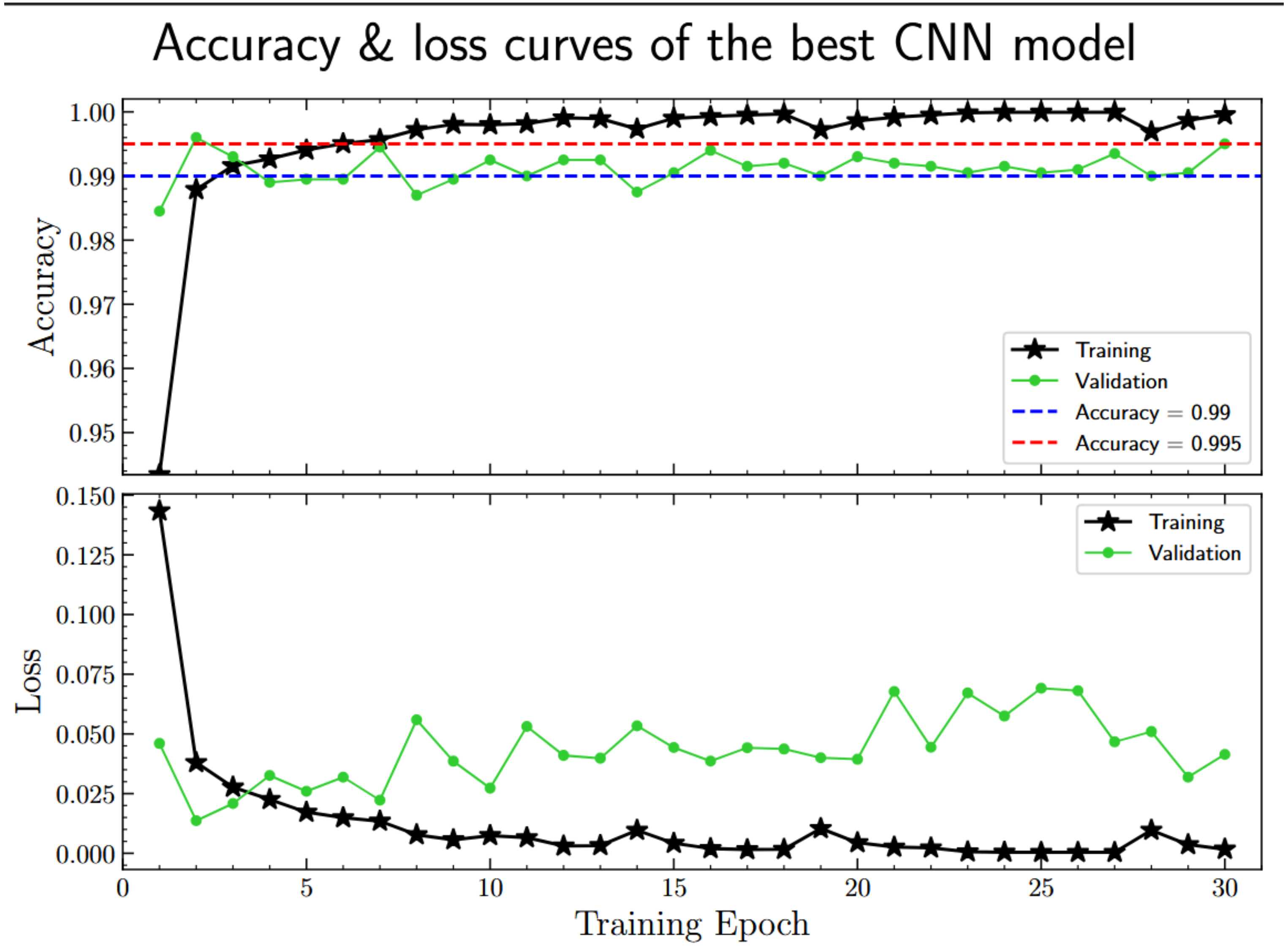
CNN TRAINING EXAMPLES





CNN TRAINING EXAMPLES





BEST CNN PERFORMANCE

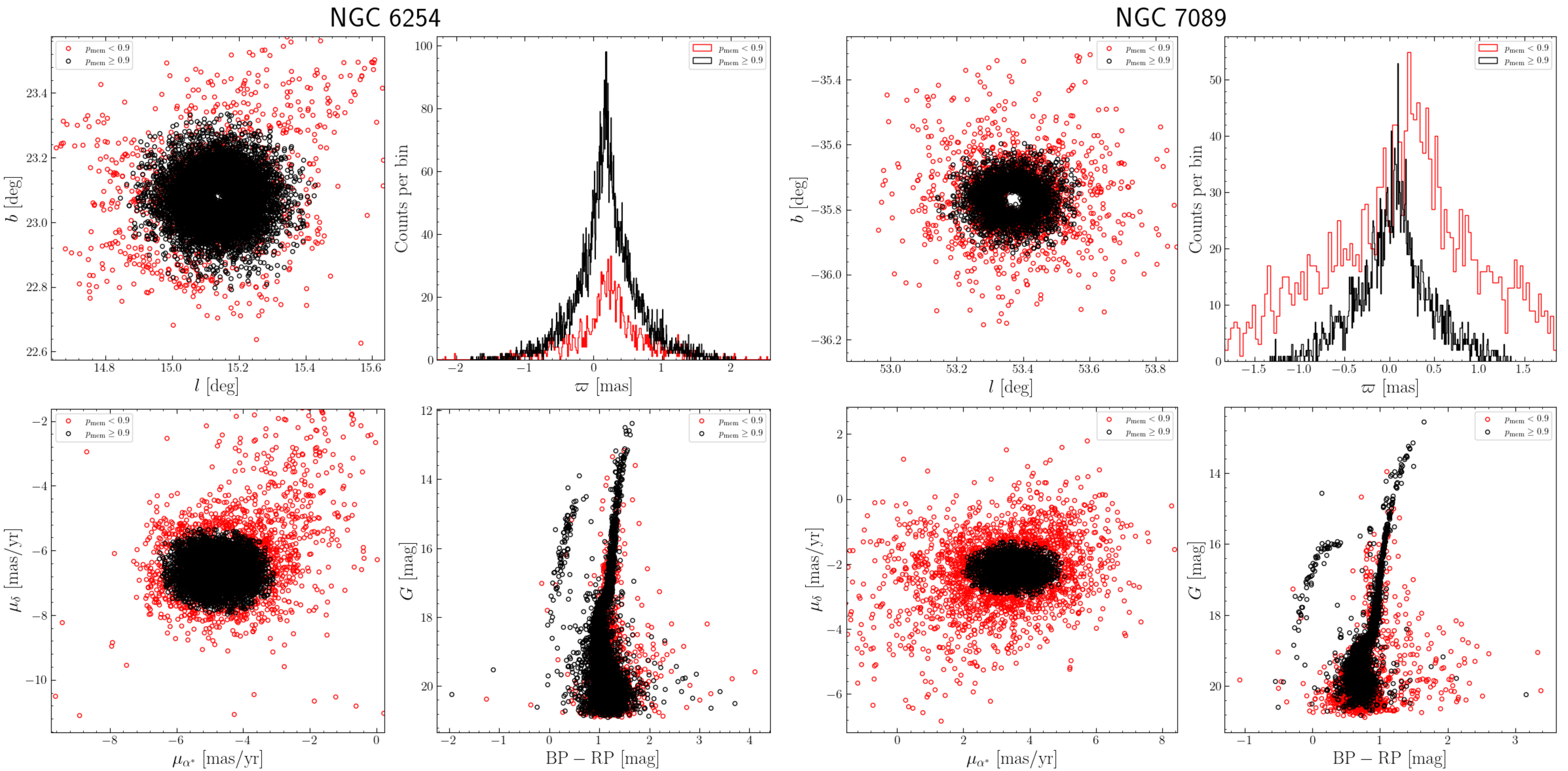
- **4 convolutional layers with $\mathbf{32 \times 32}$ neurons each**; each layer comes with a 3×3 kernel and a max pooling layer
- **2 Dense layers with 128 neurons each**
- **Image sizes of $\mathbf{128 \times 128}$**
- **30 training epochs**

3. KEY FINDINGS

Results from Clustering & CNN classification

DETECTION OF KNOWN GCs

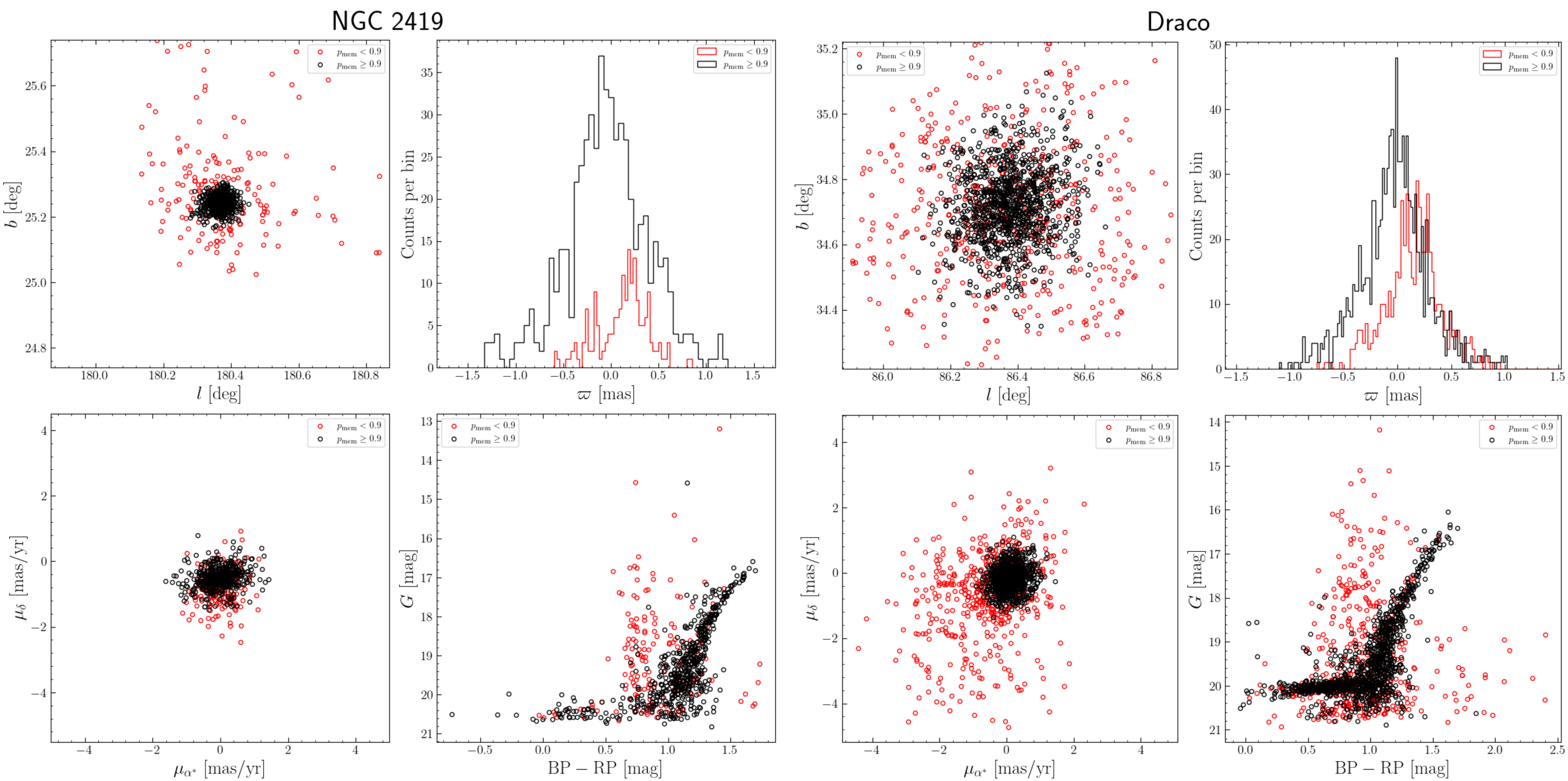
- Of the 28 known GCs in the search regions, we successfully detected and classified 23 (~82 %)
- 19 of the 23 detected GCs had mean position, parallax and proper motion (PM) values that were in good agreement with the literature – e.g. GC catalogue by Vasiliev & Baumgardt (2021)
- The 4 other clusters – **Eridanus, Pal 14, Sagittarius II & Terzan 8** – recorded astrometric values that deviated from literature. Reasons include:
  - Low number of observed of members (<100)
  - Foreground/Background contamination
  - Large distance (>50 kpc)
  - Too faint to be well-observed by Gaia (magnitude limit of ~21)
- We did not recover to detect 5 GCs – **Arp 2, Laevens 3, Munoz 1, Koposov 2 & Segue 3** – for similar reasons.



NGC 6254
NGC 7089
$p_{\rm mem} < 0.9$
$p_{\rm mem} \geq 0.9$
l [deg]
b [deg]
ϖ [mas]
Counts per bin
μ_{α^*} [mas/yr]
μ_δ [mas/yr]
BP − RP [mag]
G [mag]

OTHER OBJECTS OF INTEREST

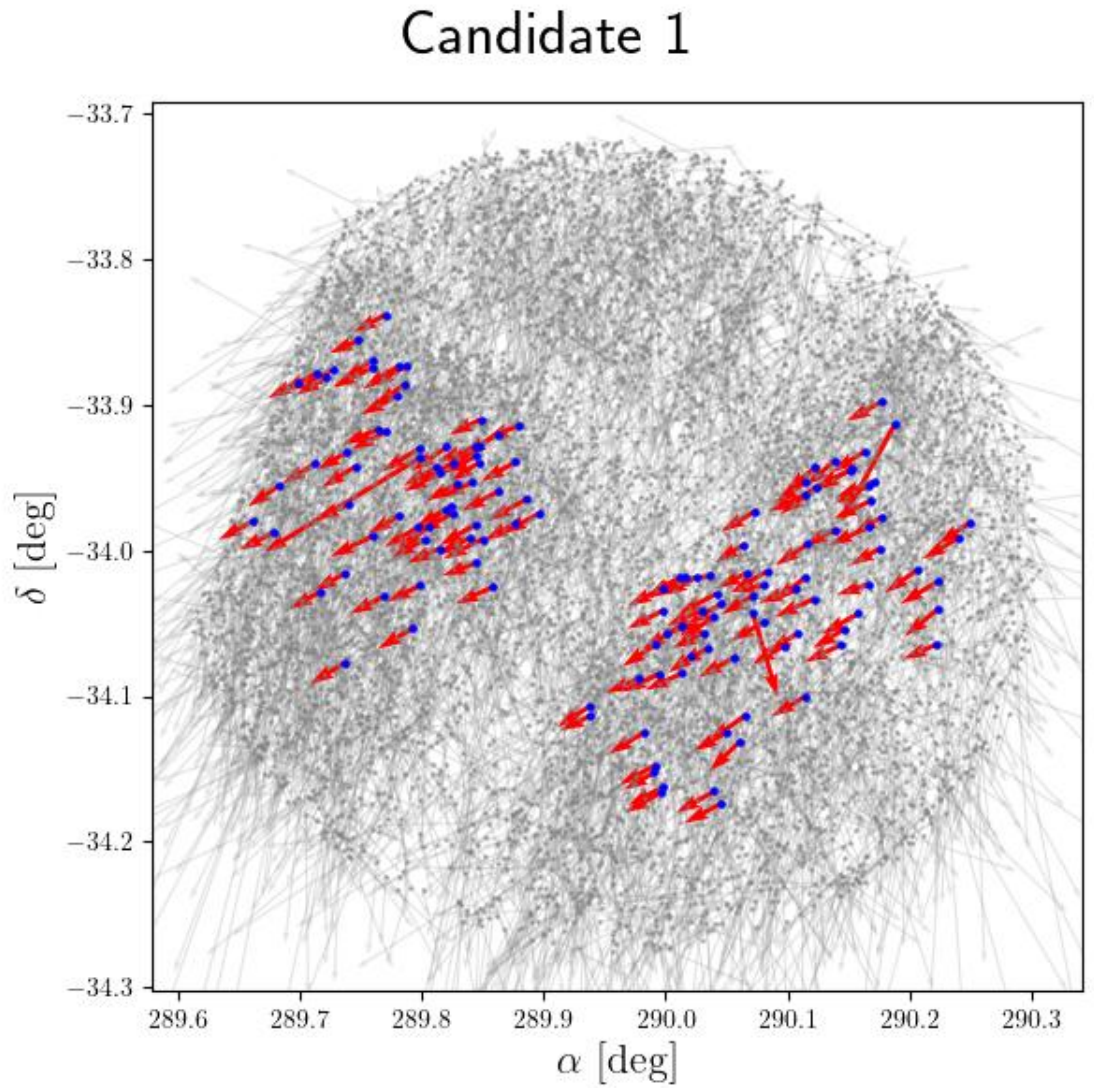
- We recovered 3 known dwarf spheroidal galaxies: **Draco, Ursa Major I, Ursa Minor**.
- Their CMDs are indistinguishable from those of bright, distant GCs, but their sizes and masses are a significant point of difference
- We identified 4 objects labelled HDBSCAN star clusters (HSCs) by Hunt & Reffert (2023):
  - 3 of the 4 HSCs were classified by the original authors as moving groups
  - 1 HSC was labelled “rejected” for failing to resemble a coherent stellar grouping
- We additionally identified 6 “candidates” with no matches to known star cluster/satellite galaxy catalogues.
- **Two of the candidates appear to be physically associated.**

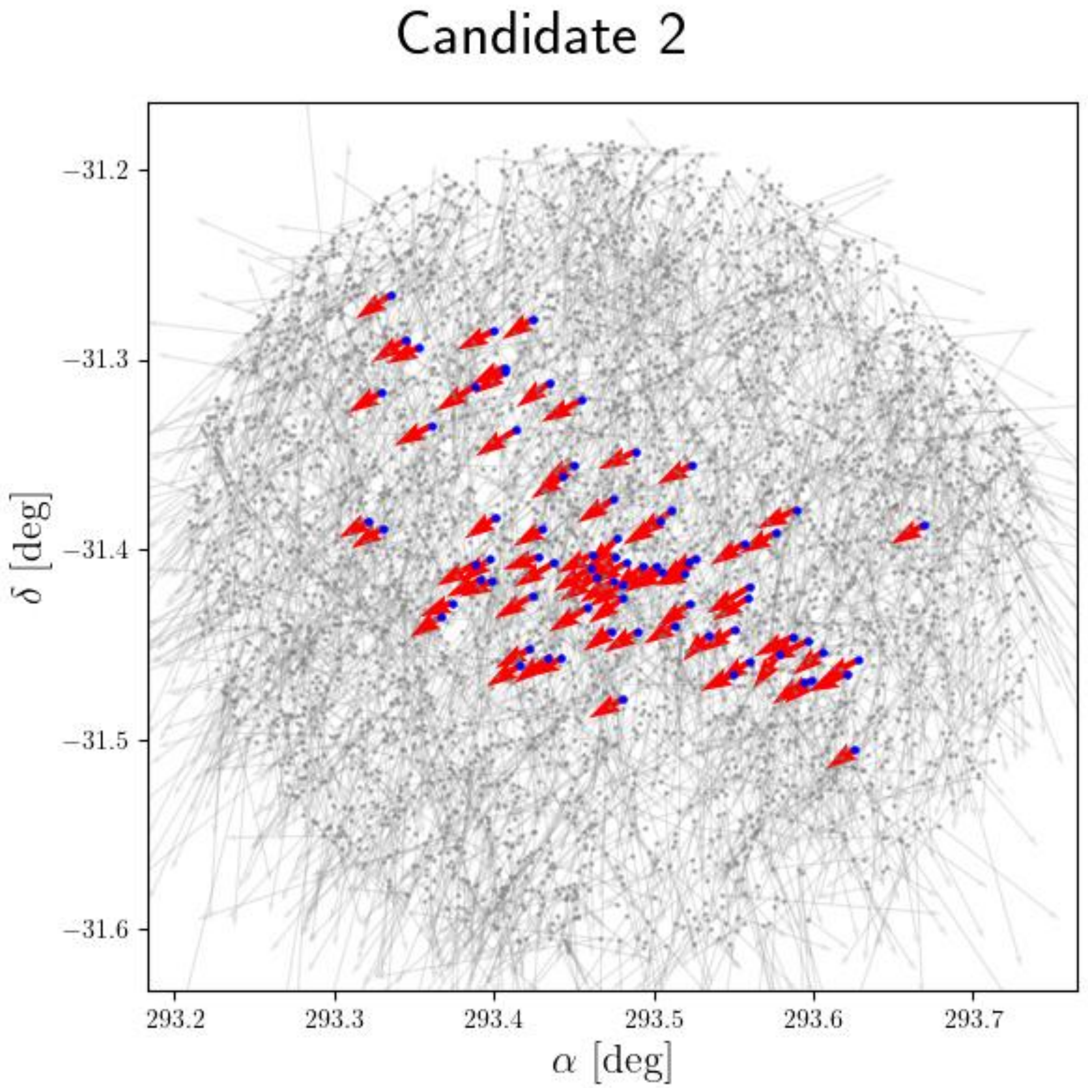


NGC 2419
Draco
b [deg]
l [deg]
Counts per bin
ϖ [mas]
μ_δ [mas/yr]
$\mu_{\alpha*}$ [mas/yr]
G [mag]
BP − RP [mag]
$p_{\rm mem} < 0.9$
$p_{\rm mem} \geq 0.9$

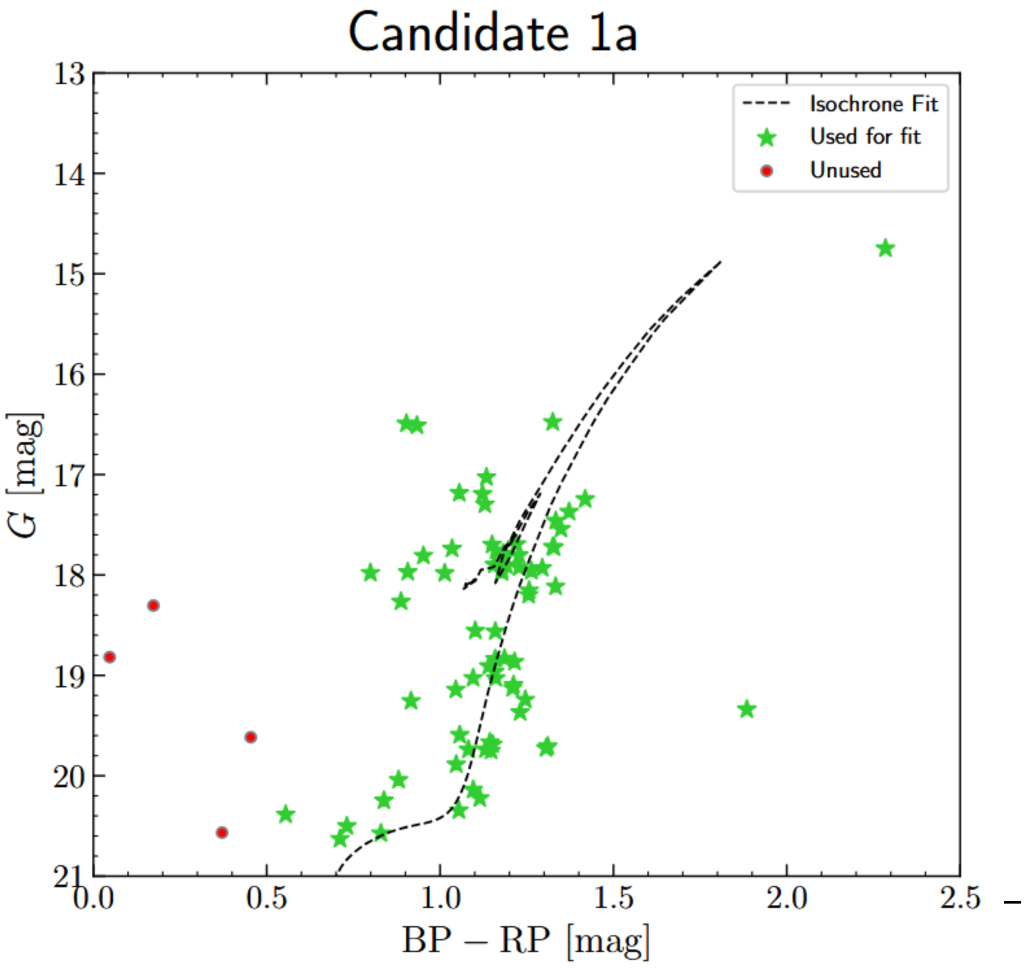
4. FURTHER INSPECTIONS

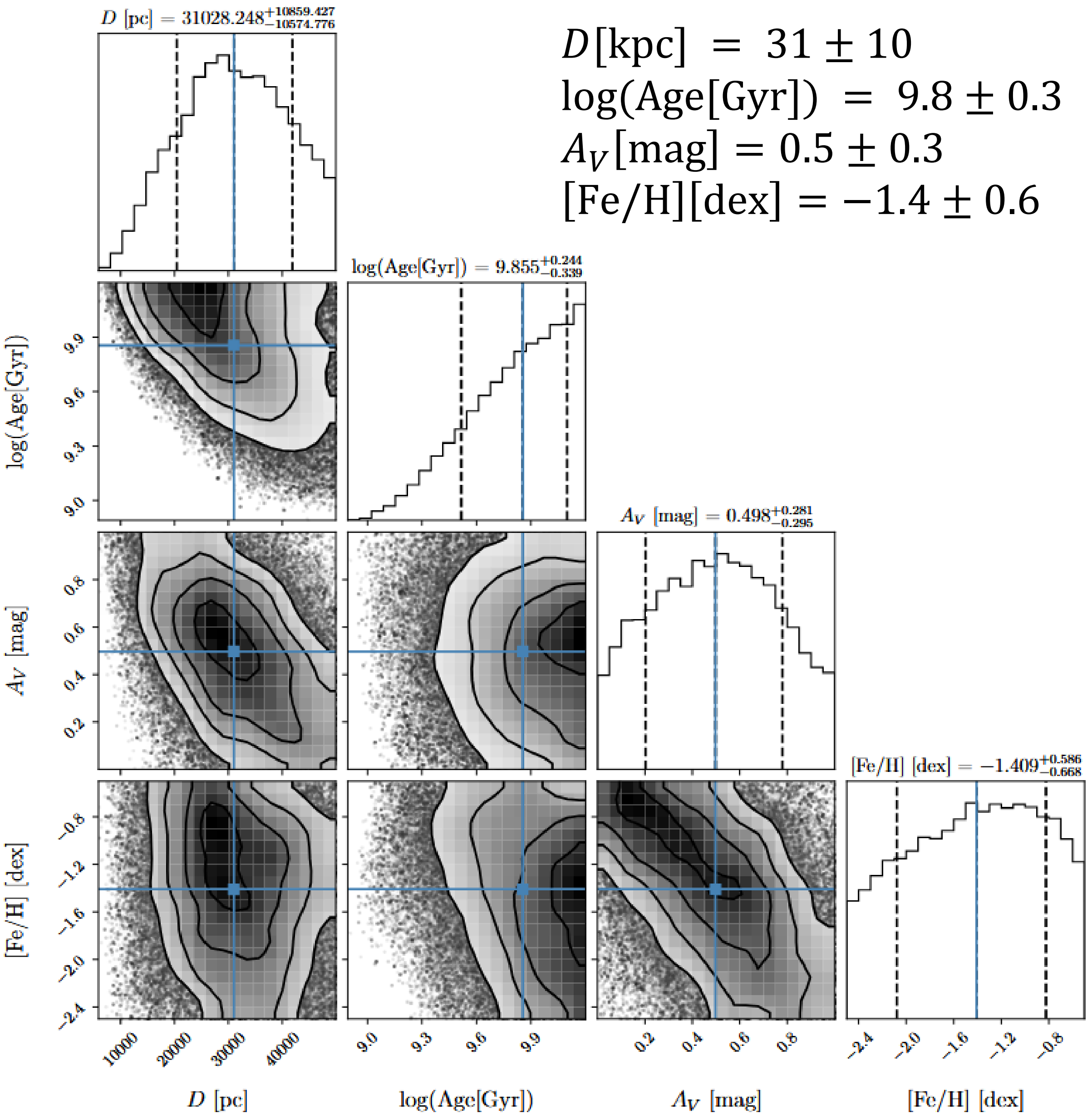
PROPER MOTION VECTOR PLOTS





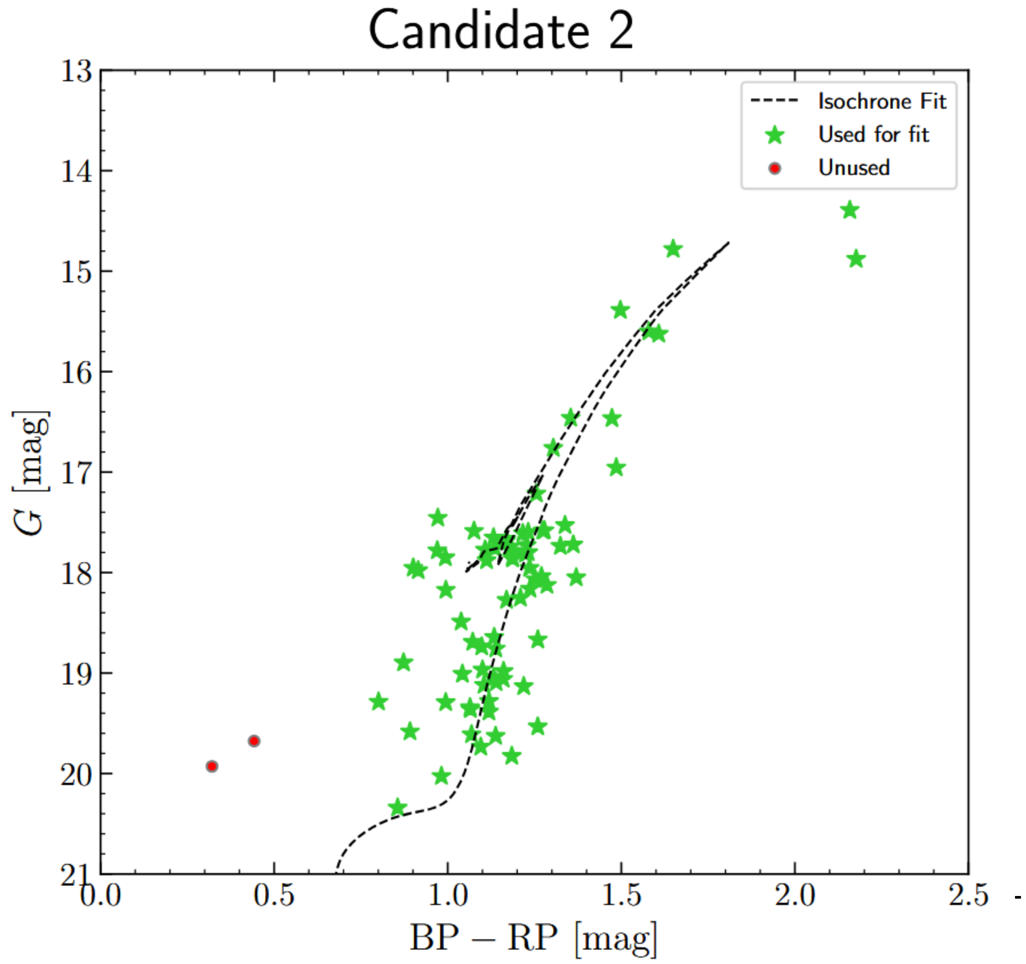
CMD FITTING

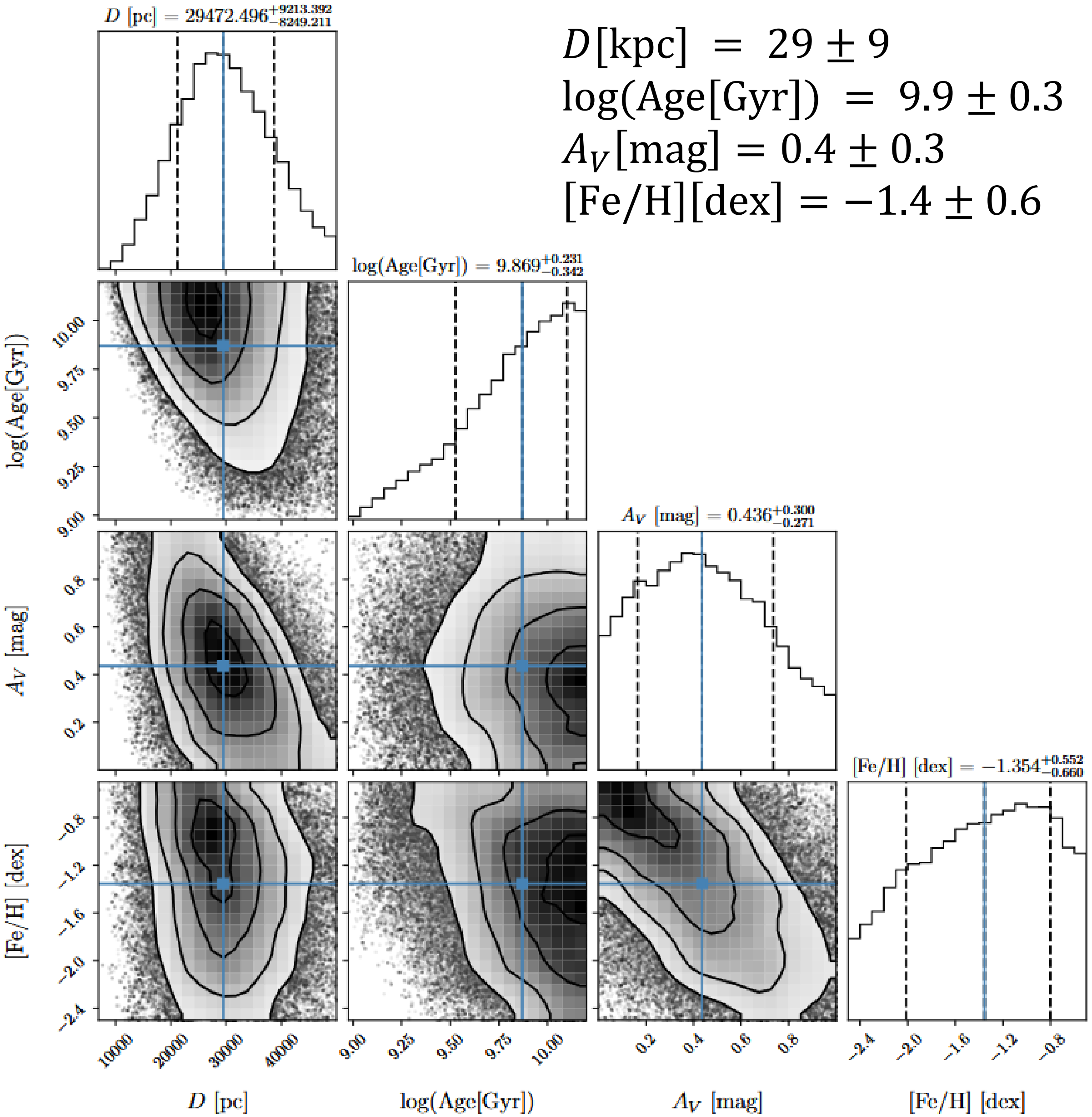




$$D[\text{kpc}] = 31 \pm 10$$
$$\log(\text{Age}[\text{Gyr}]) = 9.8 \pm 0.3$$
$$A_V[\text{mag}] = 0.5 \pm 0.3$$
$$[\text{Fe/H}][\text{dex}] = -1.4 \pm 0.6$$

CMD FITTING





$$D[\text{kpc}] = 29 \pm 9$$
$$\log(\text{Age[Gyr]}) = 9.9 \pm 0.3$$
$$A_V[\text{mag}] = 0.4 \pm 0.3$$
$$[\text{Fe/H}][\text{dex}] = -1.4 \pm 0.6$$

MASS ESTIMATION

- From Virial theorem, we made use of the following formula from Kulesh et al. (2024):

$$M_{vir} = 3\,\frac{\sigma_{V,\tan}^2 \bar{R}}{G}$$

- We first calculated the dispersions in proper motion, and converted them to velocity dispersions via CMD-derived distances
- C1 derived mass:

$$4.4 \pm 1.5 \times 10^6 M_\odot$$

- C2 derived mass:

$$4.6 \pm 1.4 \times 10^6 M_\odot$$

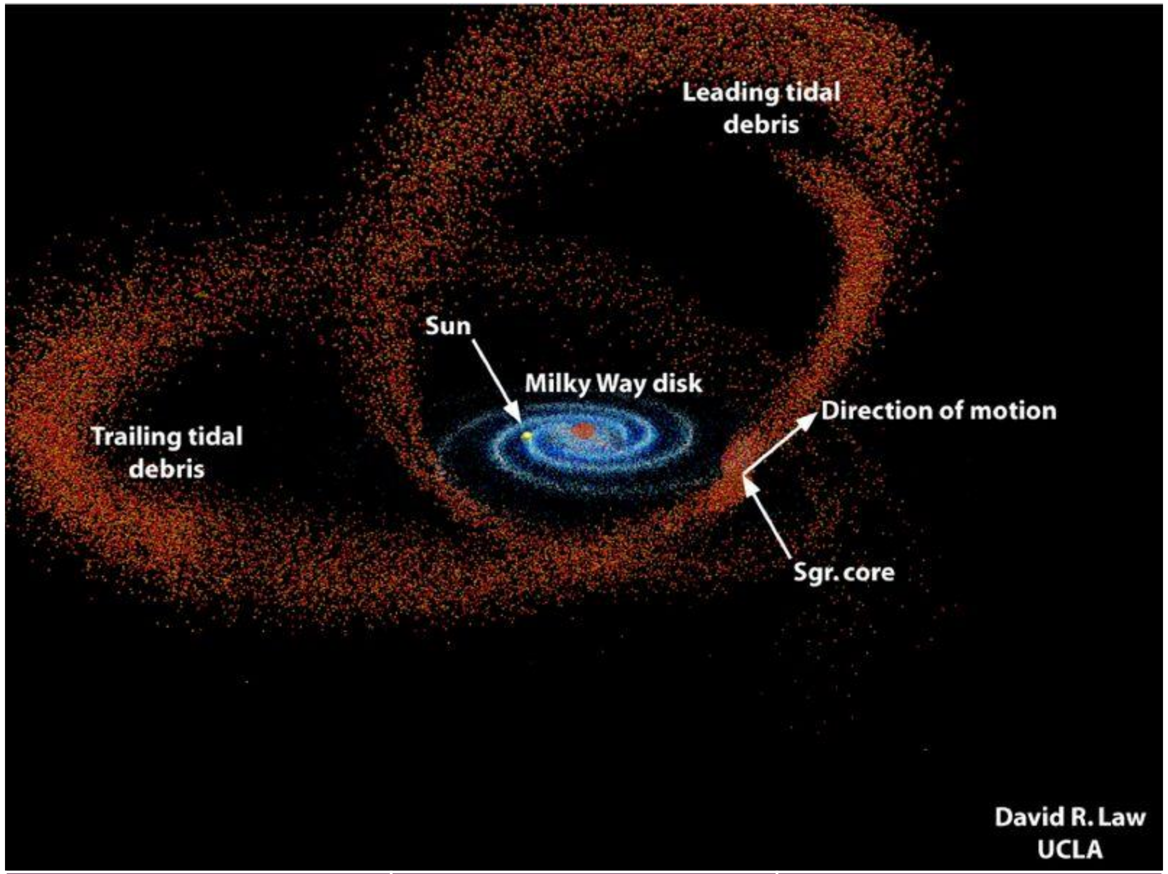
- This a very crude mass calculation; masses are highly inflated!
  For reference, we derived $M_{\mathrm{NGC\,2419}} = 2.5 \pm 0.3 \times 10^6\ M_\odot$; literature value is $\sim 8 \times 10^5 M_\odot$

Still, findings suggest some dynamical association between C1 and C2…

BIG REVEAL…

Candidates 1 and 2 are likely to be components of the Sagittarius system!

- C1 and C2 are ~6 and ~8 degrees away from Sgr dSph; both within $1.5R_{1/2}$ of the dwarf
- $D_{\mathrm{Sgr\ dSph}} = 26 \pm 2\ \mathrm{kpc}$ – agrees with both candidates.
- Members are tidally stripped stars from the Sagittarius Dwarf Elliptical galaxy.
- The result of merging with the MW over billions of years.
- Stream was originally proposed by Donald Lynden-Bell in 1995 after analysing globular cluster dynamics.



Object	μ_{α^*} [mas/yr]	μ_{δ} [mas/yr]
C1	-2.402 ± 0.030	-1.540 ± 0.030
C2	-2.589 ± 0.030	-1.651 ± 0.030
Sgr dSph	-2.679 ± 0.001	-1.394 ± 0.001

5. CLOSING REMARKS AND FUTURE CONSIDERATIONS

IMPORTANT NOTES AND CAVEATS

- We report a GC detection efficiency of 82%, on par with literature using similar techniques
- Our framework can be expanded to characterise a large variety of stellar groupings, e.g. open clusters, dwarf systems, moving groups and tidal streams.
- No new GC candidates were found – Gaia survey lacks the necessary depth to new GCs or faint dwarf systems

Caveats:

- The use of *Gaia* parallaxes in the clustering process requires revision: beyond 10 kpc, they are not useful distance measures.
- No consideration of astrometric error in HDBSCAN, or reddening in CMDs
- Method still needs to validated against mock data

POINTERS FOR THE FUTURE

- More work is required to understand GC formation and evolution via:
  - Deep, multiwavelength follow-ups
  - Dynamical simulations
  - Further study of tidal tails and multiple stellar populations
- Validation of method via mock data could help to constrain the **GC luminosity function** in *Gaia* or the properties of unknown, "to-be-discovered" GCs
- Further exploration of the link between halo GCs and dwarf satellites

SELECTED REFERENCES

baloynj@unisa.ac.za

- Bañares-Hernández, A, et al., 2025, A&A, 693, A104
- Castro-Ginard A., et al., 2022, A&A, 661, A118
- Hunt E. L., Reffert S., 2023, A&A, 673, A114
- Taylor E. D., et al., 2025, Nature, 645, pp. 327–331
- Vasiliev E., Baumgardt H., 2021, MNRAS, 505, 5978