Astronomy & Astrophysics manuscript no. output
July 13, 2026

Milky Way open cluster age determination with spectroscopic priors and machine learning-accelerated inference

I.N. Kallimanis[1, 2], A. Vallenari[1], D. Bossini[2], A. Bragaglia[3], M. Pasquato[4], R. Carrera[3], C. Boeche[1], D. Hatzidimitriou[5]

[1] INAF - Osservatorio Astronomico di Padova, vicolo dell'Osservatorio 5, 35122 Padova, Italy
[2] Università di Padova, Dipartimento di Fisica e Astronomia G. Galilei, Padova, Italy
[3] INAF-Osservatorio di Astrofisica e Fisica dello Spazio, via P. Gobetti 93/3, 40127 Bologna, Italy
[4] Istituto Nazionale di Astrofisica, IASF-Milano, via Alfonso Corti 12, I-20133 Milano, Italy
[5] National and Kapodistrian University of Athens, Panepistimiopolis, 15784 Zografos, Athens, Greece

July 13, 2026

ABSTRACT

Context. Open clusters are excellent tracers of the Galactic thin disk, providing insights into its formation and evolution. Precise cluster age determination is essential for studying cluster evolution as well as the evolution of the disk as a whole.
Aims. With this work, our aim was to produce an updated catalog of cluster parameters (age, extinction, distance) by constraining cluster ages via newly available spectroscopic measurements of metallicity.
Methods. Using high-quality astrometry from Gaia Data Release 3, we derived new cluster memberships via a machine learning-based clustering algorithm (AstroLink). We compiled metallicity measurements from multiple ground-based spectroscopic surveys (Gaia-ESO, GALAH DR3, APOGEE-OCCAM) and Gaia GSP-Spec into a uniform scale. The ages of 320 clusters were determined using a machine learning-accelerated approximate Bayesian isochrone-fitting code (based on ASteCA), adopting a multiband approach that incorporates Gaia XP and synthetic photometry.
Results. This work resulted in a robust sample of open clusters with precisely determined parameters and uncertainties. In addition, this catalog may provide a foundation for machine learning models to infer parameters for a larger sample of objects without spectroscopic data, which is to be explored in future work.

Key words. Galaxy: open clusters and associations, Stars: fundamental parameters, Catalogs, Methods: Data Analysis

1. Introduction

Open clusters (OCs) are key tracers of the Galactic thin disk. Formed from the same interstellar medium (ISM), their member stars share a common age and chemical composition, making OCs ideal laboratories for studying stellar evolution and Galactic structure. As OCs evolve, their own internal dynamics and relaxation, amplified by perturbations due to external forces, lead to their gradual dissolution into the field star population (Binney & Tremaine 2008; Gieles et al. 2006). Therefore, understanding their formation and dissolution offers valuable insights into the history of the thin disk, the star formation process, and the assembly of the Milky Way.

A comprehensive and precise census of OCs, along with accurate determination of their fundamental parameters, is essential for studies of Galactic evolution. The *Gaia* mission (Gaia Collaboration et al. 2016) has transformed the field of stellar cluster studies by providing high-precision astrometric, photometric and spectroscopic data, enabling the discovery and characterization of thousands of new OCs. Several catalogs of cluster parameters have been published in recent years (e.g., Bossini et al. 2019; Tarricq et al. 2022; Dias et al. 2021), and various machine learning (ML) techniques have been applied to identify cluster candidates and members (e.g., Castro-Ginard et al. 2022; Perren et al. 2023; Hunt & Reffert 2023).

In particular, determining cluster ages is often approached using Bayesian inference methods (e.g., Bossini et al. 2019; Perren et al. 2022; Li et al. 2025), as well as ML-based techniques (e.g., Cantat-Gaudin et al. 2020; Hunt & Reffert 2024; Cavallo et al. 2024a). However, photometric data alone are often insufficient for precise age estimation due to strong degeneracies between distance modulus and extinction and, most notably, between age and metallicity. Independent spectroscopic metallicity measurements are crucial to alleviating these degeneracies.

Our aim was to improve the accuracy of OC parameter estimation by incorporating distance information from Gaia Data Release 3 (DR3) data (Gaia Collaboration et al. 2023c), as well as metallicity information from recent high-resolution spectroscopic surveys. We derived the membership for our cluster sample, and we performed isochrone fitting using multiple photometric bands to mitigate degeneracies. Additionally, we implemented a novel ML-based surrogate model for the likelihood function, enabling faster and more efficient inference without compromising accuracy.

In section 2 we present the data. In section 3 we derive the cluster membership. In section 4 we describe our method for parameter estimation. In section 5 we present the results and in section 6 we give the conclusions.

2. Data

Gaia DR3 includes very precise (often to the mmag level) $G/BP/RP$ photometry for stars down to magnitude 20 (Montegriffo et al. 2023). Moreover, its very precise astrometry makes

Article published by EDP Sciences, to be cited as https://doi.org/10.1051/0004-6361/202558766

membership determination more straightforward than in the past.

Moreover, the precise synthetic photometry derived by Gaia Collaboration et al. (2023a) also provides a homogeneous photometric catalog in other photometric systems by convolving existing passsbands with XP spectra. We opted to fit the G/BP/RP Gaia Bands, but we utilized a multi-band approach by simultaneously fitting the synthetic g, i, z SDSS bands as well, available in Gaia Collaboration et al. (2023a). A multiband approach is often adopted to alleviate the degeneracies between age, metallicity, distance, extinction, since different passbands have different sensitivity to those parameters (see among others Vanderbeke et al. 2014; Ivezić et al. 2008; Smolinski et al. 2011; Anders et al. 2022; Cavallo et al. 2024b; Kordopatis et al. 2023). More details on selecting the appropriate photometric bands and their compatibility with the isochrones can be found in Appendix A.

Color-magnitude diagrams (CMDs), even when constructed using multiple photometric bands, do not always adequately constrain the metallicity. Thus, the choice of metallicity prior can significantly impact the convergence of the isochrone fitting procedures. Therefore, careful selection and homogenization of metallicity information is crucial.

Bossini et al. (2019) includes hard metallicity priors derived from photometry, and also from spectroscopy for 88 clusters. Since this work, several new spectroscopic surveys have become available, substantially increasing the number of OCs with well-constrained metallicities. However, metallicity estimates from different surveys and instruments often differ due to different analysis methods, assumptions and instrumentation used.

In our analysis, we made use of a few high-resolution surveys providing metallicity information for a number of OCs. GALAH data are obtained through the HERMES spectrograph ($R \sim 28000$) in the visual range (Buder et al. 2021). OCCAM (Donor et al. 2020) provides cluster metallicities from combined observations of the infrared APOGEE survey ($R \sim 22500$). Gaia-ESO (Gilmore et al. 2022; Randich et al. 2022) utilizes the FLAMES instrument of the VLT and typically targets a large amount of stars for each OC included in the survey. Thus, cluster metallicities were obtained by the combined observations of more than 100 stars for most systems (Randich et al. 2022), of which tens are of very high-resolution ($R \sim 47000$) spectra obtained through FLAMES-UVES(Magrini et al. 2023), with the rest being from FLAMES-GIRAFFE ($R \sim 17000 - 31750$). When ground-based high-resolution spectroscopy is unavailable, metallicity estimates from the *Gaia* GSP-Spec can be utilized. These are obtained through spectra of the *Gaia* RVS instrument ($R \sim 11500$) in a small spectral window around the Ca triplet region. These metallicities have greater uncertainties than the ground-based ones, but they are available for a much greater number of stars (Recio-Blanco et al. 2023).

The catalogs used in this work refer to solar reference values of Grevesse et al. (2007) and are generally consistent, although systematic offsets remain. To account for these, we applied corrections to bring all metallicity measurements onto a common scale, as follows.

Open clusters typically exhibit solar α-element abundances, with small variations of $\sim \pm 0.1$ *dex* for some objects (Casamiquela et al. 2019; Magrini et al. 2023). Thus, we adopt $[M/H] = [Fe/H]$ throughout this work and do not attempt to homogenize the more variable and sometimes inconsistent α-abundance measurements across surveys.

GALAH DR3 metallicities have already been homogenized to the APOGEE scale (Spina et al. 2021). Thus, we utilized GALAH DR3 (Buder et al. 2021) instead of the newer GALAH DR4 (Buder et al. 2025). GSP-Spec metallicities were calibrated using the latest coefficients from the *Gaia* Archive Recio-Blanco et al. (2023).

Comparisons of OC metallicities from different catalogs indicate that average cluster metallicities from GALAH, OCCAM, and GSP-Spec differ from one another primarily by zero-point offsets, which are comparable and often smaller than the typical intrinsic metallicity spread within each cluster. Accordingly, we applied zero point corrections to align the catalogs to the GALAH scale: Median $[Fe/H]_{OCCAM} - [Fe/H]_{GALAH} = -0.0165$, $\sigma = 0.0410$ for OCCAM, $[M/H]_{GSP-Spec} - [Fe/H]_{GALAH} = -0.086$, $\sigma = 0.049$) for GSP-Spec. The Gaia-ESO Survey (Gaia-ESO) metallicities show a small linear trend relative to GALAH and OCCAM; we corrected for this trend as well ($[Fe/H]^*_{GES} = \frac{[Fe/H]_{GES} - 0.087}{0.83}$). After this correction, the spread between GaiaESO and OCCAM is $\sigma = 0.047$. The high statistics of OC members in Gaia-ESO result in a typical dispersion of 0.05 *dex*. A comparison of the common objects between the catalogs, after corrections have been applied can be seen in Fig. 1.

For clusters with multiple available metallicity sources, we adopt a priority ordering of: Gaia-ESO > GALAH > OCCAM > GSP-Spec. OCCAM is preferred over GALAH when the number of OCCAM stars exceeds twice the number in GALAH ($N_{OCCAM} > 2 \times N_{GALAH}$). For clusters with fewer than two stars in GALAH and OCCAM, or fewer than five stars in GSP-Spec, we adopted metallicities in the order GALAH > OCCAM > GSP-Spec, and assumed a typical spread consistent with the corresponding catalog. The metallicity and corresponding internal cluster standard deviation are used as the center and σ of a Gaussian prior for each cluster. In Table G.1, we report the source of the metallicity prior, the number of spectroscopic observations and the metallicity prior for each cluster.

There are 480 clusters with well-defined spectroscopic metallicity in these catalogs. According to estimations from (Cantat-Gaudin et al. 2020), these objects are in the $\log Age$ range between 6.6 to 9.86 (median 8.64). Their extinctions A_V vary from 0 to 4.4 (median 0.9) and their distance spans from very local (0.1 kpc) to very distant (12.6 kpc), with a median value of 1.7 kpc. For those objects, we proceeded to derive membership, as described in the next section. Moreover, in Table 1 we summarize the characteristics of the metallicity priors for the 320 objects in our final sample (see Section 5).

Table 1. Metallicity prior summary.

Source	Clusters	$< N_{\rm spec} >$	Typical $\sigma_{\rm [M/H]}$
GaiaESO	45	448	0.04
GALAH DR3 ($N \geq 2$)	27	15	0.04
GALAH DR3 (N=1)	19	1	0.04
OCCAM	31	23	0.03
GSP-Spec ($N \geq 5$)	27	10	0.05
GSP-Spec (N<5)	171	2	0.05

Notes. This table summarizes the prior for the final cluster sample. $< N_{\rm spec} >$ is the average number of spectroscopic observations per cluster.

3. Member stars

In order to derive cluster parameters, high probability membership is required. The membership catalog of

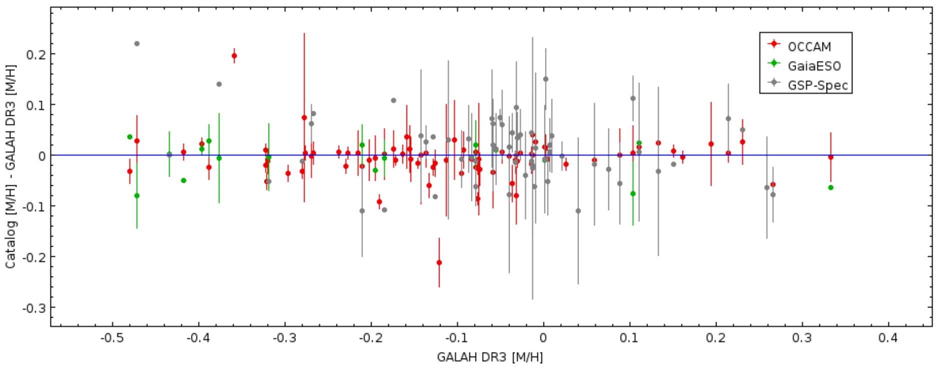


Fig. 1. Comparison of catalogs to the Spina et al. (2021) GALAH DR3 scale, after linear corrections are applied.

Cantat-Gaudin et al. 2020) relies on *Gaia* DR2, which included significantly fewer stars than the more recent DR3 and less precise astrometry. Hunt & Reffert (2023) uses *Gaia* DR3 astrometry to derive a census of astrometric overdensities across the whole *Gaia* DR3 dataset using the clustering algorithm HBDSCAN (Campello et al. 2013) which does not include a treatment of the parameter uncertainties.

In order to refine the memberships of our cluster sample, we derived new memberships from *Gaia* DR3 astrometric data while also taking into account the astrometric uncertainties. We applied a magnitude cut of $G < 19$, to discard the dimmest stars with uncertain photometry and astrometry. We utilized an unsupervized ML clustering algorithm, AstroLink (Oliver et al. 2024). Similar to other ML clustering algorithms, AstroLink detects overdensities in the arbitrary parameter space of the input data. It was developed to study galactic halo structure in simulations and is conceptually similar to the OPTICS algorithm (Ankerst et al. 1999). It offers a clustering performance similar to HDBSCAN (Campello et al. 2013), but it is faster and requires less hyperparameter tuning. AstroLink implicitly relies on prominence of a cluster over a fitted background distribution (Oliver et al. 2024). The prominence threshold for an overdensity to be considered significant is the only tunable hyperparameter of the algorithm. It is most often decided automatically by the AstroLink package without any configuration. In cases where the algorithm failed to find the cluster with the automatic tuning, we progressively lowered the threshold until the cluster was detected. Like most other clustering algorithms, AstroLink does not incorporate measurement uncertainties in the clustering process. However, we addresed these uncertainties through the following Monte Carlo (MC) scheme: we perturbed the astrometric parameters according to the observational uncertainties, assuming Gaussian errors. The clustering process was repeated 100 times (as was done by Tarricq et al. 2022 with HDBSCAN). Each star was then assigned a detection rate based on the fraction of runs in which it was identified as a cluster member.

By default, clustering was performed in a five-dimensional parameter space: $(\mu_\alpha, \mu_\delta, \varpi, \alpha, \delta)$, corresponding to proper motion, parallax, and sky position. Parallax ϖ was always included. Even in far away objects where it is minimally informative, it does not degrade the performace of the algorithm. The universal inclusion of the parallax is standard practice in similar work such as Hunt & Reffert (2023). Hunt & Reffert (2021) showed that clusters with very small parallaxes appear as distinct Gaussian overdensities in parallax space due to *Gaia* observational errors, making them detectable with clustering algorithms. In our tests, AstroLink performed significantly better when parallax was included, failing to detect clusters in about 40% of cases when it was omitted. Given our focus on high-quality membership for a relatively small sample, the clustering results were visually inspected to ensure well-defined CMDs.

Clustering was performed within a circular region of radius $5 \times r_{50}$, where r_{50} is the sky cone radius within which 50% of the member stars are enclosed. We utilized the r_{50} values determined by Cantat-Gaudin et al. (2020). This reliably selects the central part of the clusters, which is most critical for isochrone fitting. In the cases where $r_{50} < 1°$, we performed clustering within a radius of 1°.

Clustering algorithm performance often depends on the scaling of the parameter space. We investigated the use of different scalers (RobustScaler and StandardScaler), as suggested by Hunt & Reffert (2021). AstroLink memberships do not strongly depend on the chosen scaler, provided that a cluster can be detected. However, we found that in some cases, clusters could only be reliably detected using StandardScaler. Usually multiple clusters are identified within each search region. We retained the one whose mean proper motion and parallax is closest to the value reported by Cantat-Gaudin et al. (2020) or Hunt & Reffert (2023), if the first is unavailable.

The probable members were selected from the upper quantiles of the detection rate distribution. In fields without significant crowding, a lot of stars reach a detection rate of 1, and we adopted a threshold of 0.5. In crowded fields, the maximum detection rate may be lower, because the cluster is not found in every MC iteration.

In particularly contaminated CMDs, we discarded outliers by selecting stars with Mahalanobis distance (Mahalanobis 2018) less than 3 from the cluster center in the 5D astrometric space. In the case of a 5D Gaussian distribution, this is equivalent to $\chi^2_{k=5} < 3^2$ cut, which would effectively discard 11% of the points that are furthest away from the center. This ensured that sparse points and outliers are rejected. This was implemented for 16 clusters where visual inspection revealed heavy contamination of the CMD. An example is shown in Fig. C.5 [1]. The central part of the clusters, which is the most informative for age determination, was recovered very well in the majority of cases. The derived membership is generally more complete than that of Cantat-Gaudin et al. (2020), owing to the use of Gaia DR3, a more sensitive algorithm, and a deeper magnitude limit ($G < 19$ vs. $G < 18$). Our results are broadly consistent with Hunt & Reffert (2023) when adopting the same search radius and magnitude cut, with 87% of our members also identified in their catalog. About 15% of their members are absent from ours, primarily low-probability sources (median probability < 0.5). Conversely, 13% of our members are not listed by Hunt & Reffert (2023); these tend to have lower detection rates (median 0.69)

[1] In Table G.1 , the application of a Mahalanobis distance cut is indicated by a flag, described also in Table G.3.

compared to the shared subset (median 0.99). Overall, the two catalogs show strong overlap and agreement. Appendix C provides more information about the comparison of our membership with Cantat-Gaudin et al. 2020; Hunt & Reffert 2023).

After deriving and verifying the memberships for our initial cluster sample, we visually inspected the resulting CMDs for overall quality. Approximately 410 of the 480 OCs were deemed good enough to attempt isochrone fitting on.

4. Isochrone fitting

We aimed to obtain ages and other cluster parameters through isochrone fitting. Since OCs are simple stellar populations, their photometry should be reproduced by a single isochrone.

4.1. Isochrone set

The PARSEC 1.2S isochrones (Chen et al. 2014) are well suited for this study, as they have been specifically calibrated to reproduce Gaia photometry. A caveat is needed here regarding rotation; in high and intermediate mass stars, stellar rotation affects both the luminosity and the color of a star. Rotationally induced mixing increases luminosities and lifetimes. Centrifugal forces introduce a latitudinal dependence on the effective surface gravity and on the effective surface temperature of the star. This implies that the observed color of a rotating star changes with the stellar axis inclination angle and random inclination angles result in a spread in color and magnitudes. All this affects the age determination, but since information about rotation is usually not available, we adopt the nonrotating isochrones for the whole catalog. The implications of this choice are further discussed in 5.1 where we report the results by Boeche et al. (2026) who use the method developed here to derive the age of a clusters using isochrones that include rotation. For the whole catalog, we used a grid of PARSEC 1.2S isochrones [2] with a step size of 0.01 dex in both $[M/H]$ and log(Age).

4.2. Bayesian inference procedure

We utilized the ASteCA package, presented in (Perren, G. I. et al. 2015), due to its relative simplicity and capacity for optimization to our specific needs. ASteCA provides tools for generating synthetic clusters corresponding to given parameter combinations and comparing them to the observed data via already implemented likelihood functions, which we employ to perform Bayesian inference.

A likelihood derived from a simple distance metric from the isochrone, such as χ^2, is usually not adequate. Not all stellar evolution phases of the isochrone are equally likely to correspond to observed stars, as stars move through the different evolutionary stages at different rates, depending on their mass. Thus, isochrone fitting procedures incorporate an assumed initial mass function (IMF) of the cluster. Moreover, some of the cluster stars are unresolved binaries, which also follow a statistical distribution that needs to be accounted for.

We adopt the Chabrier et al. (2014) IMF and assume the Duchêne & Kraus (2013) distribution of binary mass ratios across the isochrone. We assume that the fraction of synthetic binaries to single stars is the same across the masses. This binary fraction is an additional free parameter for each CMD. However, this parameter does not directly correspond to the true binary

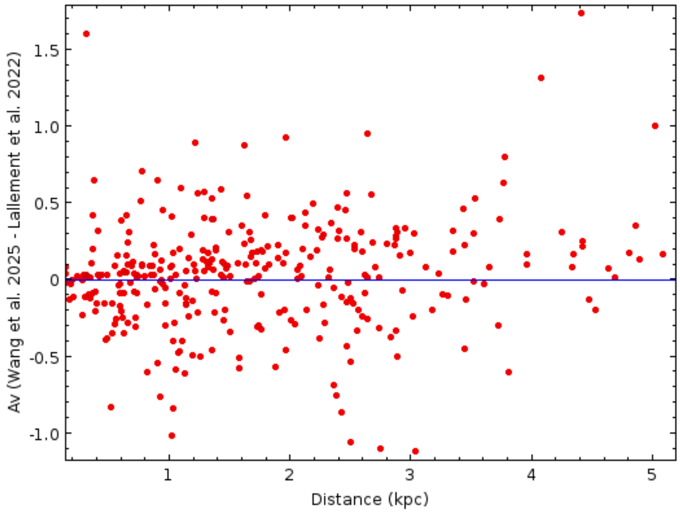


Fig. 2. Comparison of extinction values from the Wang et al. (2025) and Lallement et al. (2022) extinction maps.

fraction in the cluster. Determining the exact binary fraction of the clusters is beyond the scope of this work. Moreover, the real distribution of binary fractions across masses is not exactly uniform as assumed here, and also depends on the mass of the stars (Offner et al. 2023). A provisional binary fraction was nonetheless included as a nuisance parameter, so as to avoid introducing bias through an arbitrary fixed choice. For these reasons, we do not further discuss the results concerning the binary fraction, which should be regarded as an internal parameter.

We fitted the CMDs $(G, BP{-}RP)$ and synthetic SDSS $(i, g{-}z)$ simultaneously, with the parameters $[M/H]$, $\log Age, \mu, A_V$ varying together for both CMDs, and using the provisional binary fractions $a_{G,BP,RP}, \alpha_{g,i,z}$ for each corresponding CMD, making a total of 6 parameters. The parameter μ corresponds to the intrinsic, unreddened distance modulus, and A_V is equivalent to A_0.

For each sampled point of the parameter space, the isochrone was moved to the corresponding (unreddened) distance modulus μ. Then, extinction A_V was applied. For the *Gaia* bands, the Danielski et al. (2018) color-dependent extinction approximation was used with the Fitzpatrick et al. (2019) coefficients available from the *Gaia* Archive. For the SDSS bands, the standard Cardelli et al. (1989) extinction law was assumed, with $R_V = 3.1$, as already implemented in the code. The Cardelli et al. (1989) law has a negligible difference from the Fitzpatrick et al. (2019) law in the SDSS i, used in one of the CMDs, and hence the impact of the extinction law in the age determination is minimal. The $g - z$ color exhibits a slightly larger difference. At $A_V = 2$ mag, the median difference between the two laws is ~ 0.01 mag in color, which is comparable to the observed photometric scatter. At $A_V = 4$ mag, the difference is ~ 0.07 mag in color. The majority of our objects are at much lower extinctions (90% of the objects converge to $A_V < 2.25$), thus this difference is not significant. We chose to directly fit the extinction without an informative prior from available sources, because even the most recent extinction maps predict significantly different values, even for quite local clusters. This difference is not accounted for by the stated uncertainties of the maps, which are typically < 0.05 mag in A_V. As shown in Fig. 2, the difference of extinction values from the Wang et al. (2025) and Lallement et al. (2022) maps can vary by up to > 1 mag with a standard deviation of $\sigma = 0.36\ mag$.

[2] The isochrones can be obtained at `https://stev.oapd.inaf.it/cgi-bin/cmd`.

After moving the isochrone to the appropriate distance and extinction, we then sampled the IMF and binarity distributions, generating a synthetic cluster CMD with the same number of stars as the observed cluster. The synthetic CMDs were then compared to the real cluster CMDs. The likelihood of the points in the parameter space was obtained by comparing the Hess diagrams (binned 2D histograms, Gaposchkin 1948) of the real and synthetic CMDs.

The Hess diagrams were compared bin by bin (see below for details on the binning), so that $\log L_{CMD} = \Sigma \log l_i$, where l_i is the likelihood that the corresponding bin of the real CMD is reproduced by the synthetic CMD. Thus, multiple CMDs can be fit together by summing their corresponding likelihoods: $\log L = \log L_{G,BP,RP} + \log L_{g,i,z}$.

We adopt the Tremmel et al. (2013) Poisson likelihood ratio, because it is statistically appropriate for comparing the binned counts of Hess diagrams. The counts of each bin of the real cluster can be considered Poissonian, as the observation/detection of each star in a bin of the Hess diagram has a probability related to the observational uncertainties. Likewise, the counts in synthetic Hess diagrams are also Poissonian, since the synthetic stars are generated by stochastically sampling the IMF. In Appendix B we elaborate on the implementation of the likelihood function.

Determining the optimal binning of the CMDs is crucial for the inferred parameters and the derived uncertainties. Too small bins make the fit more susceptible to convergence issues due to noise, while too large bins dilute details and artificially inflate the uncertainties. In the majority of clusters, uniform binning as described by Knuth (2006), works adequately well. In clusters where the density of CMD varies significantly, Bayesian Blocks (Scargle et al. 2013) with varying bin size works better. Both methods of binning aim to maximize the information content of each bin through Bayesian optimization.

We assumed Gaussian priors for the metallicity and distance modulus. The metallicity prior was obtained as described in Section 2.

For the distance modulus, we utilized a prior derived from the parallaxes of the cluster stars:

$$< \mu >= -5 \times \log_{10} \varpi_{median} - 5,$$

$$\sigma_\mu = \max(| -5 \times \log_{10}(\varpi_{median} \pm \sigma_\varpi) - 5 - \mu|)$$

The Gaussian prior is interpretable and computationally convenient. However, the true shape of the distribution of parallax-derived μ of the cluster stars is not symmetrically Gaussian. That is why we used the maximum value of σ_μ.

The parameter of $\log Age$ was left to vary freely within a flat prior within the range of the isochrone grid (i.e. $\log Age = 6.6 - 10.13\ dex$. Extinction parameter A_V was sampled from a flat prior between values of 0 and 3, and when a fit failed due to a cluster being highly reddened, between 0 and 7. [3] The provisional binary fractions of the two CMDs, $\alpha_{G,BP,RP}, \alpha_{g,i,z}$ were sampled from a flat prior between 0 and 1.

For valid Bayesian inference, it is necessary to acquire enough unbiased samples (θ_i) of the posterior distribution $\log P(\theta_i) = \log p(\theta_i) + \log L(\theta_i)$, where $\log p(\theta_i)$ is the prior probability. This can be done by sampling the posterior directly, as in Markov chain Monter Carlo (MCMC) methods (e.g. emcee, Foreman-Mackey et al. 2013), or by sampling the prior and weighting its samples by the likelihood value, as done in nested sampling approaches, as in MultiNest/nestle (Feroz et al. 2009), UltraNest (Buchner 2021), nautilus (Lange 2023) and others. We choose to utilize the nautilus importance nested sampler as it is very fast, robust and provides unbiased posterior estimates.

[3] The smaller prior range aided the performance of the fit in the majority of cases.

4.3. Modifications to ASteCA

By default, ASteCA presamples the IMF and photometric uncertainty model distributions once per isochrone grid point ($[M/H], \log Age$ pair) and reuses the same set of samples each time a synthetic cluster is generated for the same parameters. This approach has the advantage of improving computational speed and making the process deterministic.

However, generating synthetic CMDs is inherently stochastic, since the location of stars on the CMD is derived from sampling statistical distributions. Using fixed IMF samples and errors and can introduce small biases at certain grid points, due to stochastic effects. This effect is significant enough that multiple runs of the fit with different initial seeds would need to be performed and the outputs combined.

For clusters with few member stars (and thus low-information CMDs) or when wide priors are used, the impact of this bias may be negligible: statistical uncertainties are sufficiently large that the choice between neighboring grid points is insignificant. In contrast, our analysis benefits from highly informative metallicity constraints, which narrow uncertainties to the scale of the (already fine) grid resolution. Under these conditions, the deterministic IMF sampling can introduce a nonnegligible bias.

To mitigate this, we modified ASteCA (v.0.5.8) so that the IMF is resampled independently in each synthetic cluster generation step. This modification also reduces RAM usage, enabling the use of very fine isochrone grids. Additionally, we incorporated perturbations to the colors and magnitudes of the synthetic stars, sampled from the error distributions of the observed photometry. The modified version of the code resamples these perturbations for each synthetic cluster realization.

Since the priors and data constrain the fit to very small statistical uncertainties, nonphysical artifacts caused by the interpolation between isochrones influence the calculated posterior and slow down convergence. To mitigate this, we rounded ages the continuosly sampled ages and metallicities to the grid resolution (the 2nd decimal). Thus, plus or minus one grid point is the smallest uncertainty possible to obtain.

We used the ASteCA implementation of the Tremmel Poisson likelihood ratio, and other ASteCA modules. Slight modifications were applied to support the proper use of Bayesian blocks binning where required.

4.4. Tempered likelihood

In our code, the IMF and observational error bars are sampled in every synthetic cluster generation step, which makes the likelihood unbiased but also noisy.

A rough or noisy likelihood surface with multiple sharp local maxima poses challenges for efficient exploration by standard sampling methods, as the amplitude of local fluctuations can exceed the overall gradient of the likelihood function. Consequently, many commonly used Bayesian sampling techniques fail to converge reliably.

We can formally define the computed log-likelihood as $\log L^*(\theta) = \log L(\theta) + \epsilon(t, [\theta])$, where $\epsilon(t, [\theta]) \sim N(0, \sigma^2)$ is a random fluctuation, which in the general case can also depend on the location on the parameter space.

As a sampler explores the parameter space, it should progressively obtain more samples from the regions of higher likelihood. If we consider a new proposed point θ' of the parameter space, to replace θ, the proposal depends on the sign of $\Delta(\theta, \theta') = \log L(\theta') - \log L(\theta)$. With the noisy realization, the computed quantity becomes $\Delta^* = \Delta + \epsilon(t') - \epsilon(t)$. Due to the two noisy computations, it exhibits random variance $\mathrm{Var}[\Delta^*] = 2 \times \sigma^2$. The proposal will often fail (it will be incorrectly accepted or rejected) if $\Delta(\theta, \theta')$ is not significantly larger than $\sqrt{2} \times \sigma$. This effect is stronger in MCMC-based methods, where each new proposed point is by construction close to the previous one. Thus, we opt for Nested Sampling approaches which are less affected, since each new point is drawn directly from the prior distribution. However, efficiency is still impacted gravely. Reducing the stochastic noise is required for convergence. In principle, this could be done by simply averaging multiple evaluations of the log-likelihood at each point. However, this is prohibitive due to the computational cost of multiple likelihood evaluations.

Instead, to mitigate the noise, we applied tempering: dividing the $\log L^*$ with a value T is equivalent to exponentiating the likelihood as $L^{*1/T}$. This introduces a nonlinear transformation of the true likelihood that increases the effect of the prior, while also decreasing variance: $\mathrm{Var}[\frac{\log L^*}{T}] = \frac{\sigma^2}{T^2}$. This procedure is analogous to increasing the statistical temperature of a random distribution, and is often done to aid exploration of complex parameter spaces, as in simulated annealing (Kirkpatrick et al. 1983) or parallel-tempered MCMC (Earl & Deem 2005).

We selected a value of $T \sim \sigma$, by sampling the prior at multiple points and evaluating the likelihood there multiple times, effectively measuring the noise of the likelihood function at random points of the prior. This temperature sets the absolute variance of the likelihood function equal to unity, which in our tests was enough to help most samplers explore the parameter space efficiently. Tempering does not affect the signal to noise ratio of the likelihood in any way.

The Tempered Likelihood was sampled through the `nautilus` importance nested sampler (Lange 2023). This sampler is very efficient, as it uses neural networks to accelerate the proposal of new sampled points. As a result, it requires much fewer likelihood evaluations than other samplers to reach convergence. We reach an effective sample size of more than 20000 for each cluster. The tempered posterior is typically broader than the actual one, with its modes shifted toward the prior distribution.

4.5. Surrogate likelihood

The credible intervals obtained by sampling the tempered likelihood cannot, in the general case, be transformed into the untempered ones. This is because tempering does not result in a mere rescaling of the posterior, it changes its shape. When a parameter has a noninformative prior and follows a Gaussian posterior distribution, the tempered posterior is also Gaussian, but broadened by a factor of $\sqrt{T}$. Since the priors on metallicity and parallax are informative and the parameters are correlated, no analytical transformation can be applied.

In nested sampling, each sample drawn from the prior is assigned a weight, in order to represent the posterior:

$$\log w = \log p + \log l,$$

where p and l are the prior and likelihood values, respectively. When sampling the tempered likelihood,

$$\log l = \frac{\log L^*}{T}.$$

Reweighting the samples by the true likelihood at each point,

$$\log w' = \log p + T \times \log l,$$

should in principle transform the tempered posterior into the untempered one. The effective sample size of the weighted samples is defined as:

$$N_{\mathrm{eff}} = \frac{(\sum w)^2}{\sum w^2}.$$

Since the tempered posterior is broader, this procedure should reduce N_{eff} even in the ideal case, since fewer samples will be present at the statistically significant parts of the true posterior. However, in the likelihood-dominated parts of the posterior, $w \sim (L^*)^{1/T}$, and hence $w' \sim L^*$. Since $N_{\mathrm{eff}} \propto 1/\mathrm{Var}[w^2]$, reweighting increases the variance of the weights drastically, which in turn catastrophically reduces N_{eff}.

Since it is impractical to use a version of the actual likelihood for inference, we can employ an approximate Bayesian approach. What is needed is an unbiased, smooth, and deterministic estimate of the true average likelihood at each point in the parameter space. This can be obtained by fitting the samples obtained during the tempered run with an appropriate ML model. The model must capture the true structure of the likelihood without overfitting the noise.

Gaussian processes (Williams 1997) can approximate an arbitrary function with heteroscedastic noise. Although they are the most formally correct option, they are computationally very expensive and prohibitive for our sample sizes.

An alternative is Bayesian additive regression trees (BART), which approximates arbitrary functions with homoscedastic Gaussian noise (Chipman et al. 2010). The assumption of homoscedasticity is only approximately valid across the entire parameter space, but it holds near the peaks of the posterior, which are typically very narrow. The surrogate model needs to be precise and unbiased only in these most statistically important regions.

XBART (He et al. 2019) is a heuristic but very fast implementation of BART, which retains the Bayesian characteristics of the original algorithm while making it practical as a surrogate likelihood. Its performance is similar to that of random forests or gradient boosting algorithms, but it requires minimal hyperparameter tuning.

XBART by definition approximates a function $f(x)$ when trained with noisy values of the function, as

$$f_i' = f(x_i) + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2).$$

The theoretical assumption of noisy training data and the Bayesian treatment of uncertainty make XBART especially better suited to this task than similar but non Bayesian models such as random forest, as it is by construction (instead of implicitly) resistant to overfitting the noise in the training data: In fact, it can provide a posterior prediction at each point instead of a single value, effectively also giving an estimate of the noise at each point, instead of a just the mean value.

In Figure 3, we demonstrate the performance of XBART. In the toy example of two blended Gaussian peaks, and apply 10% additive noise (which corresponds to a much lower signal to noise ratio than that of the real likelihood function). We then fit the noisy values with XBART to obtain a very good approximation of the original function.

The surrogate likelihood was then sampled using the `nautilus` sampler like before. Thus, the final posterior estimate was obtained.

The posterior obtained through the surrogate model had to be validated. Since the XBART model is theoretically robust against overfitting (estimator variance), we had to ensure that the bias of the estimator is small as well in the bias-variance tradeoff (Geman et al. 1992). The main evaluation metric is the bias-to-variance ratio near the peak of the final posterior. This can be defined as

$$B/V = \frac{< L_s - L >}{\sqrt{Var[L_s - L]}},$$

where L is the real and L_s the surrogate likelihood. If the surrogate model is successful, this ratio should be less than 1, so that any introduced bias is subdominant to the noise. Similar metrics have been applied for validating the performance of Gaussian process emulators (Bastos & O'Hagan 2009).

The ratio can be calculated by selecting the most highly weighted points of the posterior, evaluating the true (noisy) likelihood there, and comparing it with the surrogate prediction. The total residuals should be roughly Gaussian and centered at zero. Figure 4 is an actual example of successful validation. This procedure can be interpreted as a posterior predictive check of the surrogate model, since it evaluates whether the surrogate reproduces the true likelihood in the regions where the posterior density is highest. In Appendix F we demonstrate the performance of the surrogate for the case of a real OC in the catalog.

Unsuccessful surrogates can result from incomplete training. In cases of high T, the tempered samples may be spread over a large region around the true peak due to the broadening effect of tempering. Moreover, if significant mode shifting has occurred, the statistically significant regions of the untempered likelihood surface may be far from those of the tempered surface and thus underrepresented in training. These issues can be alleviated by augmenting the training set with additional samples from important areas of parameter space. However, in practice, this problem is very uncommon. We visually inspected the failed cases, and in most of them, the cluster CMDs were poorly defined, in spite of our initial visual check of the membership. In these cases, the failure of the surrogate model reflects the highly degenerate and uninformative nature of the likelihood surface itself. Consequently, these clusters would not have converged to a meaningful or well-constrained solution regardless of the sampling method used. Therefore, we discarded these clusters ($\sim$ 60 OCs) from the final catalog.

4.6. Performance of the algorithm

In summary, the algorithm we developed is an amortized approximate bayesian inference procedure based on a surrogate model (Cranmer et al. 2020). The algorithm steps are summarized below:

1. Sample the prior at N points, calculate $\log L$ m times at each point. Estimate the tempering factor $T = \sigma$ as the median or higher quantile of the calculated noise at each point.
2. Run `nautilus` on the tempered $\log L_T = \log L/T$ until $N_{\text{eff}} \geq 20000$, and obtain $\theta_i, \log L(\theta_i)$
3. Train $XBART(\theta)$ on $\theta_i, T \times \log L(\theta_i)$
4. Sample the surrogate likelihood $\log L_s(\theta) = XBART(\theta)$ with `nautilus` until $N_{\text{eff}} = 10000$
5. Validate the surrogate through a posterior predictive check
6. Obtain credible intervals from the surrogate posterior

The surrogate model massively boosts the performance of the code[4].

5. Results

We obtained a sample of 350 OCs for which the fitting procedure converged. For all the clusters with converging results, we visually checked the quality of the isochrone fit, as well as the synthetic CMDs of the best solution (generated as described in Section 4.2), to discard clusters with poor fits or unrealistic stellar distributions. Of those clusters, 320 are well fitted by the isochrone according to extensive visual verification. An example is shown in Appendix E. Of course, the results presented here depend on the isochrone set used (see for instance Nguyen et al. (2022) for a discussion of the effect of different stellar isochrones on the age determination). The log Ages span between 6.5 and 10.0, with a median value of 8.73. Their distribution is shown in Fig. 5. Young objects are underrepresented in the sample, because they are not well represented in the spectroscopic surveys. Extinctions A_V range from 0.04 to 4.51 mag with a median value of 0.92 mag (Fig. 6). Distance modulus μ ranges from 5.7 to 14.4 mag, with median value of 10.7 (Fig. 7). The fitted parameters are presented in Table G.1. Correlations between the parameters are investigated in Appendix D.

5.1. $\log Age$ uncertainties

In the approximate Bayesian framework we employed, uncertainties are represented by the 16th and 84th quantiles of the posterior probability distribution for each parameter. If the posterior is Gaussian, this would correspond to the 1σ confidence intervals for the parameter. In general, clusters with sparse CMDs and vague features have wider error bars. In a few cases, bimodal distributions are obtained. Then, the credible intervals can no longer be interpreted as standard error bars. To provide physically meaningful uncertainties that represent the spread around the most likely fit, we investigated the objects exhibiting the widest uncertainties, for bimodality. For 17 such OCs (flagged in Table G.1), we report intervals derived from the dominant posterior mode. The relative uncertainties of the $\log Age$ are shown in Fig. 8. The following trends in the uncertainties can be roughly identified. Old nearby clusters with CMDs well populated in all phases tend to have smaller uncertainties, close to the grid resolution of 0.01 dex. Some young clusters can also have low uncertainties, when the pre-main sequence turn-on is well fitted. Clusters that are more sparse and far away have larger uncertainties, especially when the red clump (RC) is not well populated. For such objects, we typically obtain uncertainties of > 0.05 dex.

As a more general trend, the nominal uncertainties are always smaller if a cluster has more stars, because the statistics are more robust. This results in some clusters having nominal uncertainties smaller than 0.01 dex, which is the isochrone grid resolution. For these cases, the grid resolution should be considered as the true formal uncertainty.

[4] The runtime of this procedure is $\sim$ 20 minutes per cluster, using 6 threads on a modern processor and ~20 GB of RAM. More instances can be run in parallel, as resources allow. In fact, the majority of the time is spent on step 2, while training and sampling the surrogate usually requires less than 5 minutes.

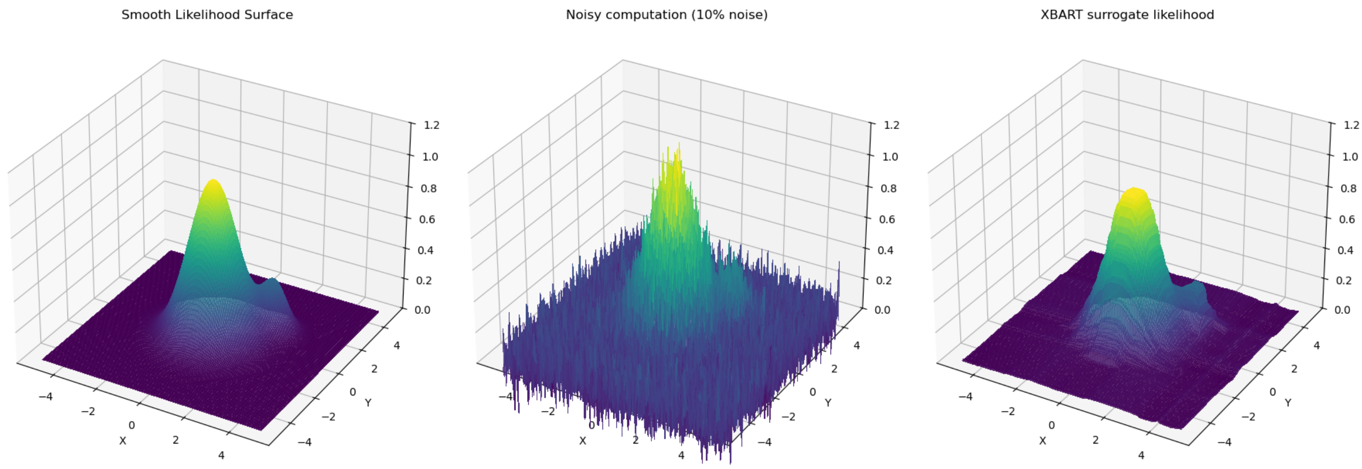


Fig. 3. Smoothing effect of the surrogate model, shown in a toy two Gaussian peaks example. On the left is the ideal deterministic likelihood surface. On the center is the noisy realisation (10% additive noise), that is difficult to explore. On the right is the XBART approximation. Structure is retained without overfitting the noise. The approximation is conservative, as the peak is slightly flatter than the original, as the noisy evaluations have obscured some information on the peak.

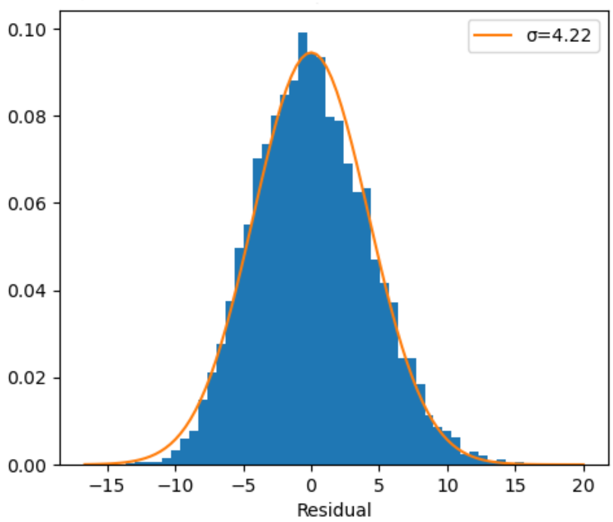


Fig. 4. Validation of the Surrogate likelihood for the fit of ASCC 111. The distribution of residuals ($\log L_{surrogate} - \log L$) is Gaussian, as per model assumption. The orange line is the best fit Gaussian centered at zero.

The reported uncertainties are internal and depend on the assumption that the models can completely accurately reproduce the data. However, a multitude of unmodeled effects can also affect inference.

For instance, stellar rotation can be significant in objects of young and intermediate age. For example, in NGC 2509, neglecting rotation, we derive $\log Age = 9.20 \pm 0.01$. As discussed in Boeche et al. (2026), stars at the main sequence turn-off (MSTO) in NGC 2509 are affected by rotation with typical $v \sin i$ ranging from 80 to 140 km/s. Boeche et al. (2026) use the same membership in the same photometric filters, and derive the cluster parameters with exactly the same procedure described here. The authors make use of the PARSEC 2.0 isochrones (Nguyen et al. 2022) including rotation. Assuming Ω/Ω_{crit} in the range 0.1-0.9, they find that the best fit is obtained for $\Omega/\Omega_{crit} = 0.9$, where $\log Age = 9.08^{+0.02}_{-0.01}$ was derived. The fact that the rotation parameter Ω/Ω_{crit} is unknown introduces an additional age uncertainty of 382 Myr or 24%.

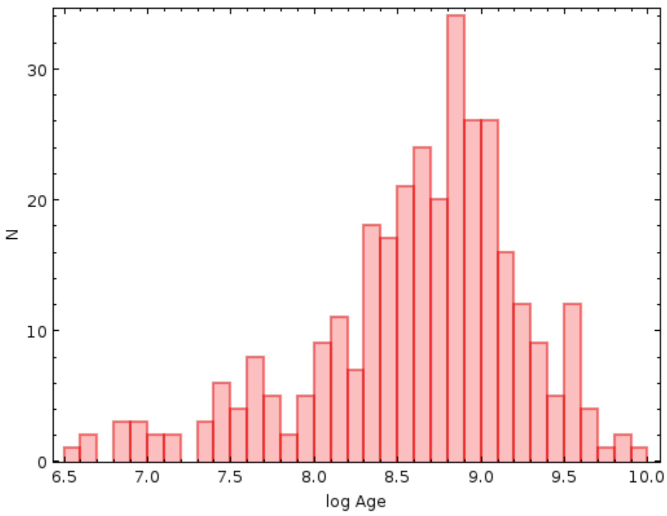


Fig. 5. Distribution of cluster ages in the sample. Median $\log Age = 8.73$, $\sigma_{\log Age} = 0.63$

Another model-dependent effect is the calibration of the lower MS of the isochrones. The Mass-Radius relation used in the isochrones tends to systematically underestimate the radii and luminosities, at the low mass regime, i.e. $M < 0.5 M_{\odot}$ (Torres et al. 2010). Chen et al. (2014) have calibrated the PARSEC 1.2S isochrones (used here) for this effect. However, some small discrepancies may still remain. Imperfect calibration of the lower MS should be more important for nearby objects. In order to better quantify the effect, we applied a magnitude cut at $G = 18$ for the 9 closest objects in our catalog. These are at distances of $200 - 300$ pc. We found that applying the magnitude cut results in younger age estimation by 0.1 in $\log Age$. This effect becomes less significant the further away the cluster is. At ~ 900 pc, no

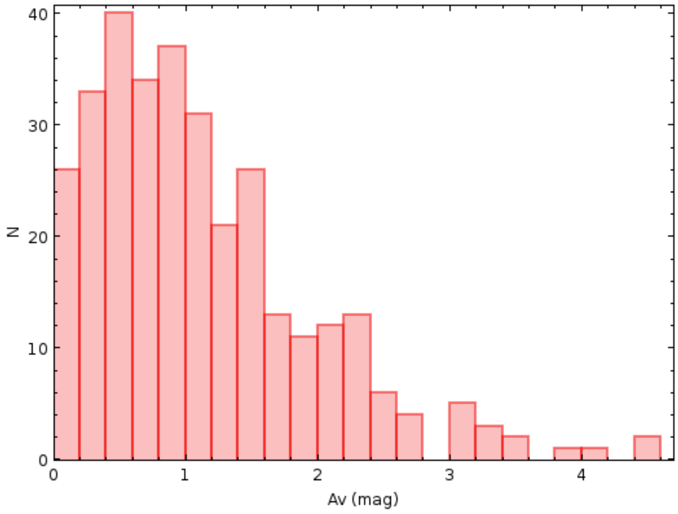


Fig. 6. Distribution of cluster extinction parameter A_V in the sample. Median $A_V = 0.92$, $\sigma_{A_V} = 0.84$

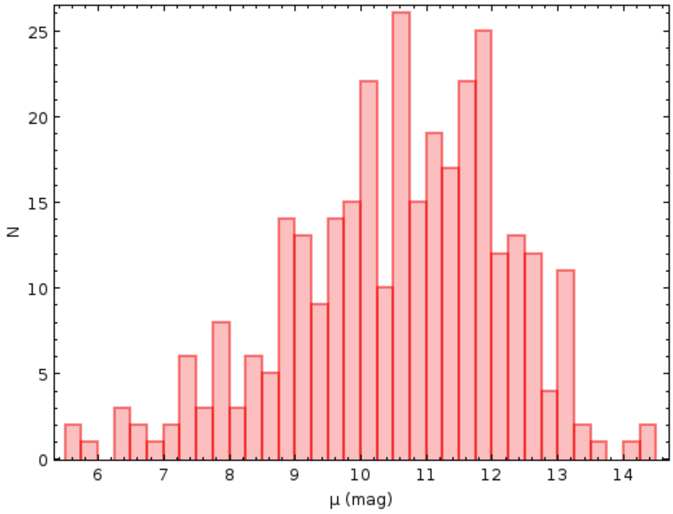


Fig. 7. Distribution of cluster distance modulus μ in the sample. Median $\mu = 10.7$, $\sigma_\mu = 1.6$

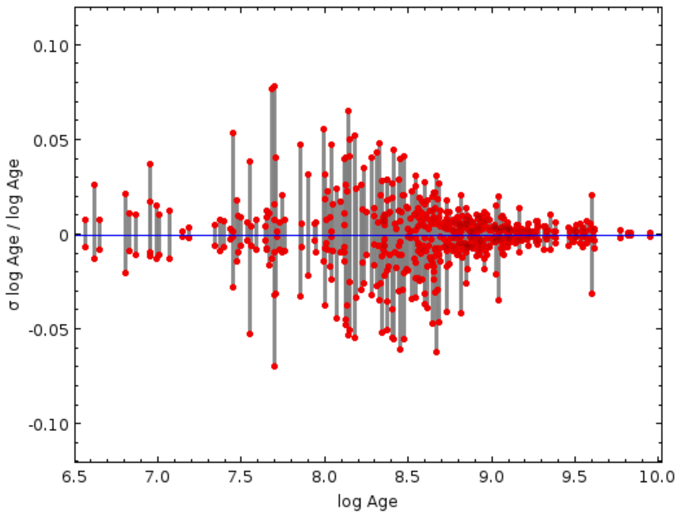


Fig. 8. Asymmetric relative uncertainties (relative to the median of the posterior, which is the expected value) for the studied OCs. Median spread is $\sigma_{\log Age} = \pm 0.057$.

star with $M < 0.5 M_\odot$ can be observed within our magnitude limit of $G < 19$.

5.2. Posterior metallicity

The metallicity parameter was sampled within the informative spectroscopic prior. Thus, the uncertainty of the metallicity could propagate to the other inferred parameters. Moreover, the fit was allowed to converge to different values, even far away from the center of the prior, if the data is informative enough.

In the majority of the objects, a metallicity value consistent with the prior was recovered. The median difference of posterior-prior is not significant (< 0.01 dex in $[M/H]$ regardless of catalog). This indicates that the observed metallicity scale is generally consistent with the scale of the isochrones, after the minimal homogenization we applied [5]. However, the median scatter is $\sim 0.03 - 0.04$ dex depending on the survey, which is comparable to the uncertainties on the prior values.

As a consistency check, we calculated the $z-$score of the posterior metallicity values with respect to the prior:

$$z = \frac{[M/H]_{post} - [M/H]_{prior}}{\sigma_{[M/H],prior}}$$

In our whole sample, 38 clusters (13%) have shifted more than 1σ away from the center of the metallicity prior ($|z| > 1$, see Fig. 9). Moreover, for some ($\sim$ 40%) of these objects, the prior was derived by very few GSP-Spec stars (N<5). In our determination, we assumed the nominal uncertainties on the GSP-Spec $[M/H]$, which are known to be underestimated Babusiaux et al. (2023). This may have over-restricted the prior range for those particular objects, partially explaining the results.

Visual inspection of the diverging objects reveals successful CMD fits. Moreover, the posterior metallicity values stay consistent between different runs of the algorithm, ruling out numerical instability or stochastic effects. As further validation, in Appendix D we show that the diverging objects do not exhibit outlier behavior in any other fitted parameter, nor do they correlate with low membership counts. The values of metallicity determined by the fit are indeed those that allow for a good isochrone fit with these particular constraints.

The behavior of the metallicity posterior validates our choice to sample the metallicity parameter instead of fixing it to a single value. In the majority of cases, the prior dominates, but the fit can converge to a slightly different value of metallicity if required.

5.3. Posterior distance

One of the fitted parameters is the distance modulus, which corresponds to the photometric distance of the object. It is correlated with other parameters, especially the extinction parameter A_V. We chose to crudely constrain it with a prior derived from the parallaxes of the cluster members, as described in 4.2.

If we compare the inferred distance to the geometric distance (inversion of the median cluster parallax), we notice that they are consistent up to 1 kpc, with median difference of ~ -3 pc,

[5] According to Bressan et al. (2012), the PARSEC isochrones utilize solar abundances from Grevesse & Sauval (1998) and Caffau et al. (2011), unlike our metallicity catalogs which use Grevesse et al. (2007). However, as also noted in `https://stev.oapd.inaf.it/cmd_3.9/faq.html`, this difference is not significant compared to other approximations included in the models.

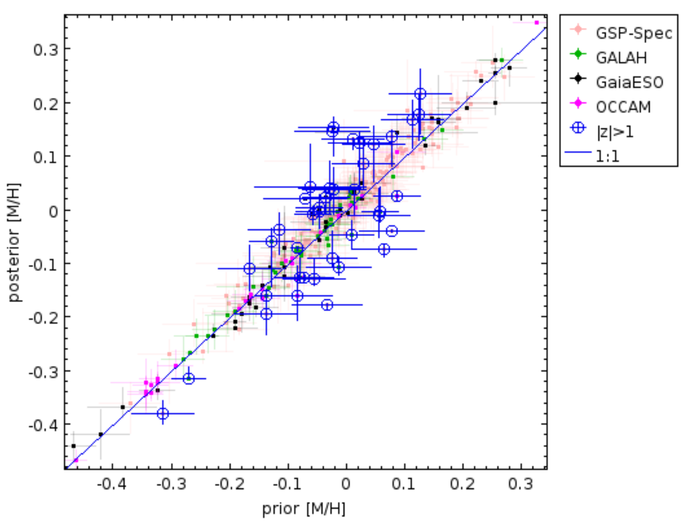


Fig. 9. Metallicity posterior vs prior. Circled blue points are the objects with $|z| > 1$, indicating metallicity posterior shift from the prior greater than 1σ.

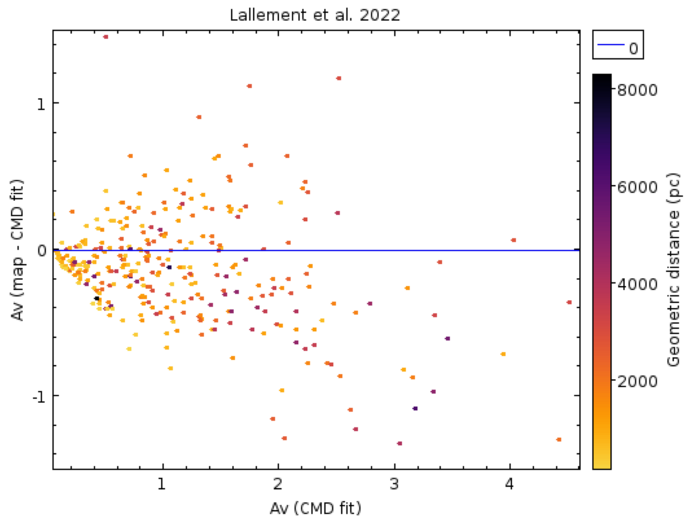


Fig. 11. Comparison of fitted extinction to the map derived by Lallement et al. (2022).

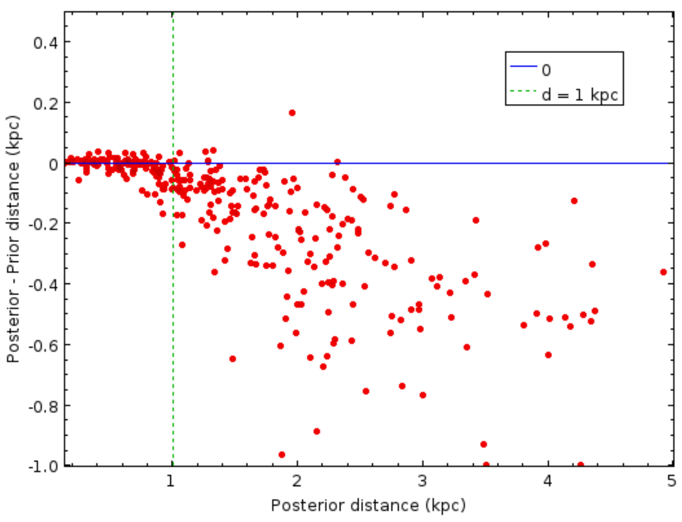


Fig. 10. Inferred photometric distances compared to inverted median parallax distance, assumed as prior.

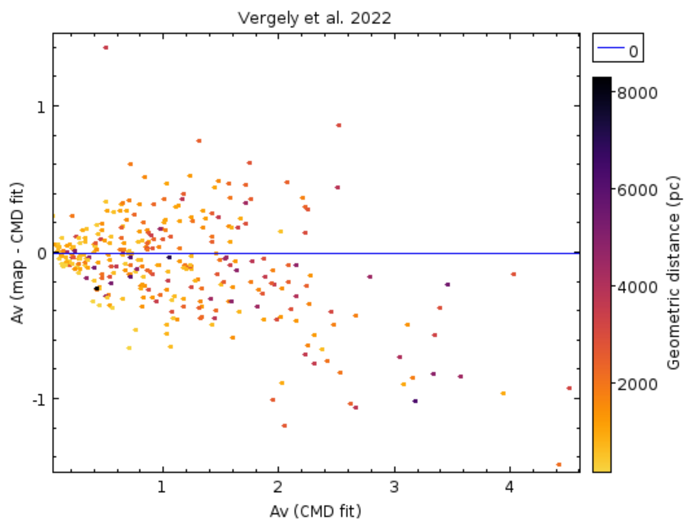


Fig. 12. Comparison of fitted extinction to the map derived by Vergely et al. (2022).

$\sigma = 34$ pc, as shown in Fig. 10. At larger distances, the photometric distance is systematically lower than the prior, due to the known and complex systematics of inverting parallaxes of faraway objects. In Luri et al. (2018) it is shown that distances inferred from parallax inversion tend to be overestimated, especially for greater distances. In that paper it is also suggested that parallax information should be combined with other data in a Bayesian approach, for a more precise and accurate inference of the distance, as done here. The fact that we do not recover the prior in all cases actually demonstrates that our prior choice does not bias the fit. A correction of the parallax, utilizing color and position-dependent zero point correction as presented in detail in Groenewegen (2021) or Lindegren et al. (2021) would not significantly alter the results of our analysis.

5.4. Extinction and maps

Deriving the interstellar extinction with different methods can lead to systematically different results, due to the complex caveats involved with each specific method. As shown in Fig. 2, extinction maps in particular can often significantly disagree with each other.

We compare our extinction determination with the two recent 3D maps of interstellar extinction, by Lallement et al. (2022) and Vergely et al. (2022) in Figures 11 and 12, respectively. Figure 13 presents a comparison to the more recent Wang et al. (2025) extinction map.

The systematic difference between fit and the Vergely et al. (2022) map (Median $A_{V,map} - A_{V,fit} = -0.07$ mag, $\sigma = 0.35$ mag), which was calibrated with spectroscopic data, is of the same magnitude as the difference with the Lallement et al. (2022) map (Median $A_{V,map} - A_{V,fit} = -0.13$ mag, $\sigma = 0.39$ mag). In both cases, extreme outliers are clusters at greater distance, but in general, random differences appear at all distances. It should be noted that the geometric distances derived from the parallaxes (and used as a prior in our fit) provide better agreement with the maps than the fitted photometric distances. This is due to the fact these maps use parallax inversion in their construction. When we calculate the integrated extinctions from the maps using our fitted distances, the systematic offset increases in the Lallement et al. (2022) map, and is doubled in the case of

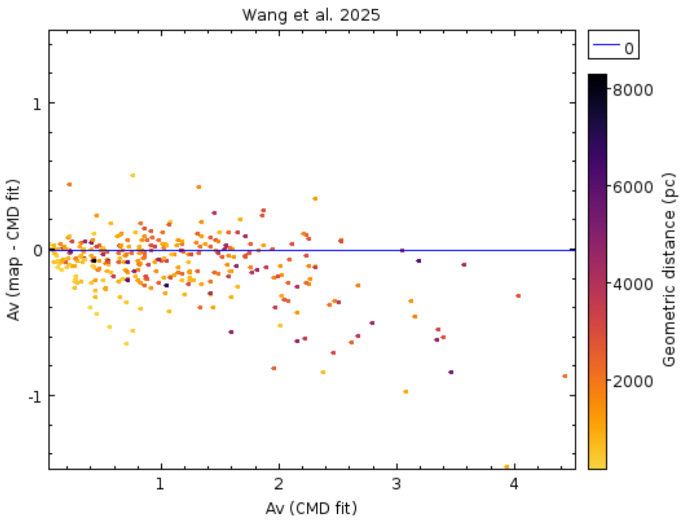


Fig. 13. Comparison of fitted extinction to the map derived by Wang et al. (2025).

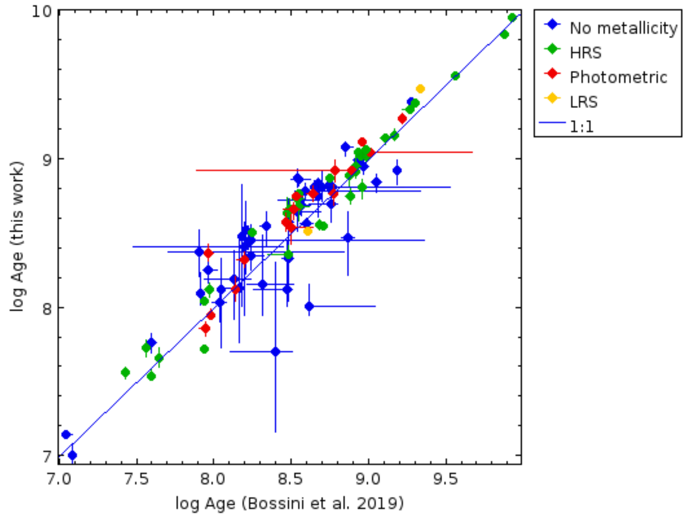


Fig. 14. Ages compared to Bossini et al. (2019). Blue points represent clusters without spectroscopic metallicity priors in the previous work. Red points utilized a photometric metallicity prior. Yellow points utilized a low resolution spectroscopic (LRS) prior. For green points, a High Resolution Spectroscopic (HRS) prior was used.

the Vergely et al. (2022) map. In the Wang et al. (2025) map the systematic offset remains the same in both distance estimations (Median $A_{V,map} - A_{V,fit} = -0.08$ mag, $\sigma = 0.23$ mag), with the scatter slightly increasing when using photometric distances. Of the three maps, Wang et al. (2025) is more compatible with our estimates. All maps tend to predict lower extinctions than the fitted ones, especially for the majority of clusters with $A_V < 1$ mag. The effect at very low extinctions ($A_V \sim 0.1$ mag) can be attributed to the flat prior imposed on extinction, starting at $A_V = 0$ mag. Clusters compatible with zero extinction can converge to a slightly positive value. Moreover, a known tendency of our fitting procedure is to fit the center of the main sequence instead of the blue edge, when there is a significant broadening due to a number of reasons, such as differential reddening, presence of binaries, photometric errors in crowded regions or stellar rotation. This can partially explain this systematic, but visual examination of our catalog reveals that significant differences exist even in objects where the blue edge is fitted. At all distances the scatter seems to dominate any systematic.

6. Discussion

We compare our results to established catalogs of cluster parameters derived with similar methodology. We also compare to parameter determinations done with methods not based on isochrone fitting. Finally, we present an example of scientific exploitation of our catalog of OCs.

6.1. Comparison to CMD fitting methods

6.1.1. Comparison to Bossini et al. (2019)

Only 95 objects in our sample are also included in Bossini et al. (2019). Most clusters with spectroscopic priors in that previous work are present in our catalog, and our derived ages are in agreement with previous estimates. The majority of the clusters in Bossini et al. (2019) employ non-spectroscopic priors. In some cases no prior is available and the metallicity is assumed to be exactly solar. In some of these objects we obtain significantly different age values (blue points in Fig. 14). When no prior exists in Bossini et al. (2019), the standard deviation of the difference in $\log Age$ is 0.24 dex. When a prior is utilized, the standard deviation decreases to 0.12 dex. Moreover, in the whole common sample, our estimates are slightly older, with median difference of 0.07 dex. For the subset of objects with a high-resolution spectroscopic prior, the median difference is 0.03 dex. For objects with non-spectroscopic priors it increases to 0.08 dex. This highlights the importance of the metallicity prior for the accuracy of OC age determination.

6.1.2. Comparison to Cantat-Gaudin et al. (2020)

All of our OCs are also present in the Cantat-Gaudin et al. (2020) catalog. The approach adopted in that work differs from ours, as it relied on a pure ML approach without a Bayesian uncertainty framework. However, the neural network regressor used was directly trained on real data and the results of Bossini et al. (2019), so similar behavior is expected. Since metallicity is not predicted in that work, any effects of the age–metallicity degeneracy present in the cluster parameters of the training set should also propagate into the derived ages of that catalog.

Our results are very compatible with Cantat-Gaudin et al. (2020), as seen in Fig. 15. There is a small systematic offset of 0.05 dex in $\log Age$. The standard deviation of the difference between the two catalogs is 0.2 dex. This is comparable to the age uncertainty claimed by Cantat-Gaudin et al. (2020). Our method tends to obtain slightly higher ages for a number of clusters. These clusters usually have poorly populated red clumps, but their MSTO is routinely well-fitted.

One systematic difference is found in the treatment of extinction A_V and distance modulus μ. We derive systematically lower distances (Median $\mu_{CG2020} - \mu_{fit} = 0.12$ mag, $\sigma = 0.20$ mag) and higher extinctions (Median $A_{V,CG2020} - A_{V,fit} = -0.16$ mag, $\sigma = 0.26$ mag) than Cantat-Gaudin et al. (2020), especially in more distant objects, as seen in (Fig. 16, 17). The systematics between our estimates and those of Cantat-Gaudin et al. (2020) are further explored in Appendix D.

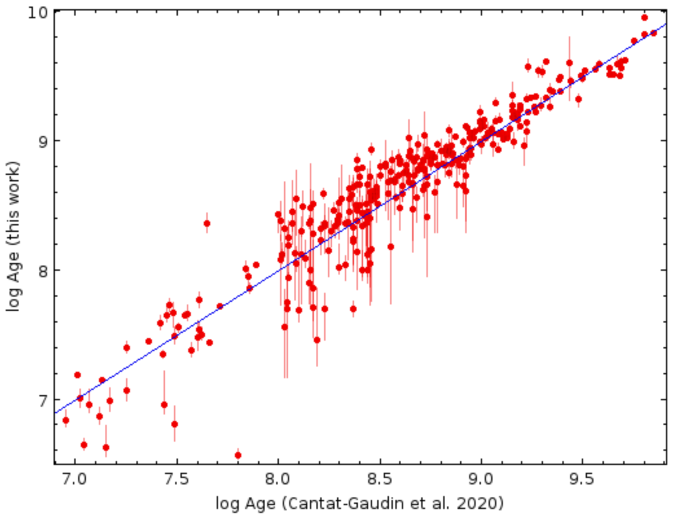


Fig. 15. Comparison of age estimation to Cantat-Gaudin et al. (2020). The uncertainty of the Cantat-Gaudin et al. (2020) log *Age* is ∼ 0.2 dex according to the validation of the algorithm. Blue line indicates 1 : 1 agreement.

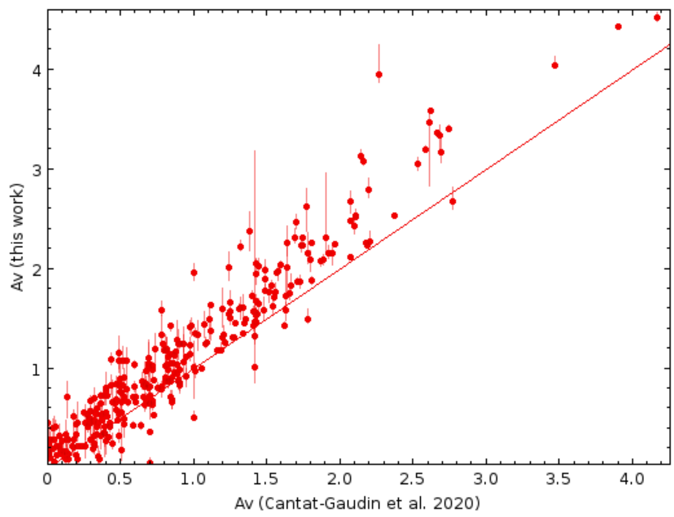


Fig. 17. Extinction A_V compared to Cantat-Gaudin et al. (2020). Blue line indicates 1 : 1 agreement. We systematically predict higher reddening in faraway clusters.

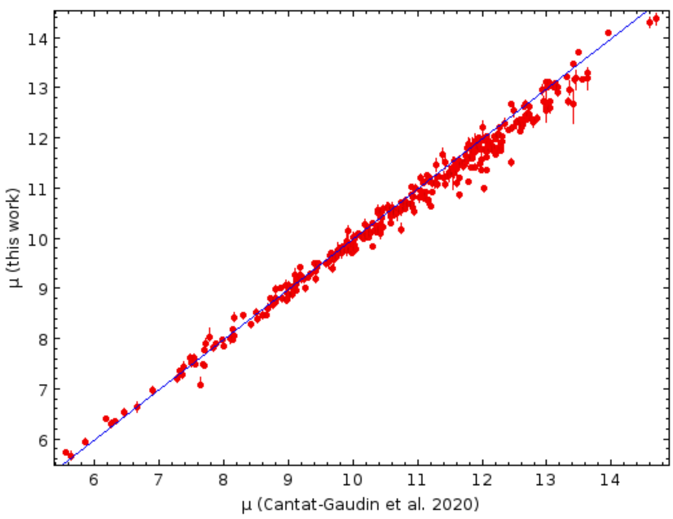


Fig. 16. Distance modulus μ compared to Cantat-Gaudin et al. (2020). Blue line indicates 1 : 1 agreement. Notice our systematic difference at higher μ.

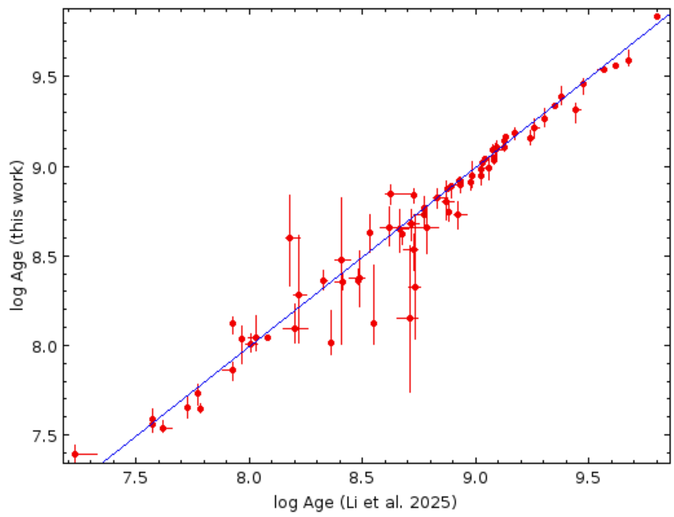


Fig. 18. Comparison of age estimation to Li et al. (2025). Only the higher quality subset of the catalog is plotted.

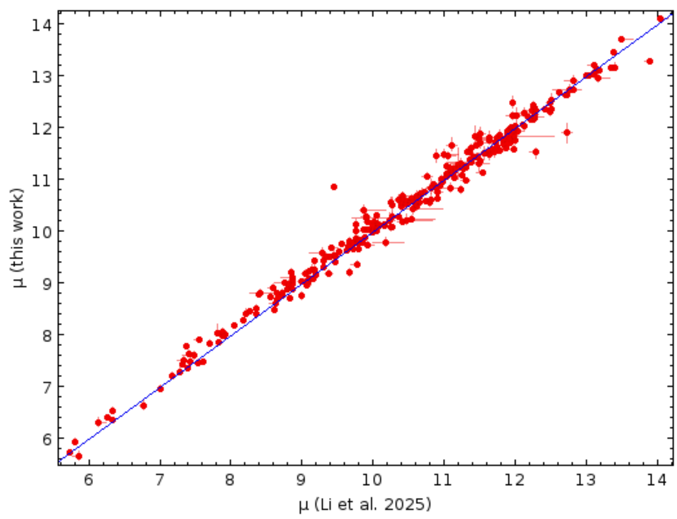


Fig. 19. Distance modulus μ compared to Li et al. (2025). Blue line indicates 1 : 1 agreement.

6.1.3. Comparison to Li et al. (2025)

Li et al. (2025) utilize a Bayesian method by Li & Shao (2022) to derive parameters for a large number of clusters using *Gaia* DR3 G/BP/RP photometry. There are 284 objects in common between their catalog and ours. Of those, 74 are flagged as belonging to their high quality subset.

Our results match very well for clusters of both young and old ages, with more scatter in the intermediate ages. In In the common sample of 284 objects, median $\log Age_{Li25} - \log Age = -0.02$, $\sigma = 0.24$. If we compare to the high quality subset of the catalog (plotted in Fig. 18), the spread reduces to $\sigma = 0.13$.

The fitted values of extinction and distance also agree very well. The systematic differences between our estimates of A_V (median $A_{V,Li25} - A_V = 0.01$, $\sigma = 0.20$) and μ (median $\mu_{Li25} - \mu = 0.02$, $\sigma = 0.21$) are very small, as seen in Figs. 19, 20.

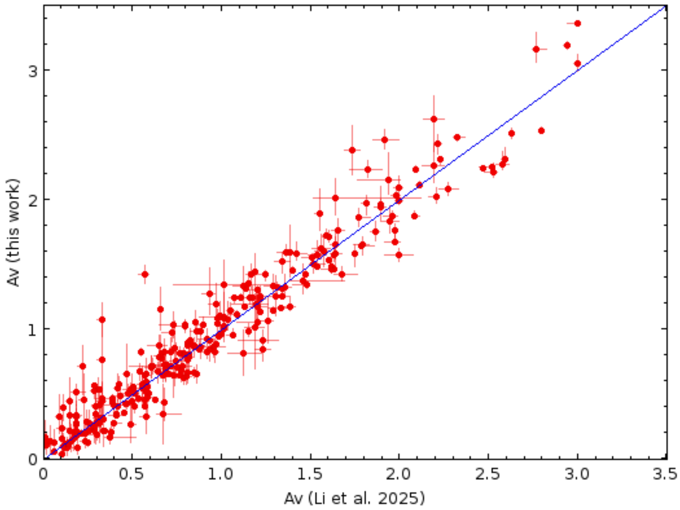


Fig. 20. Extinction A_V compared to Li et al. (2025). Blue line indicates 1 : 1 agreement.

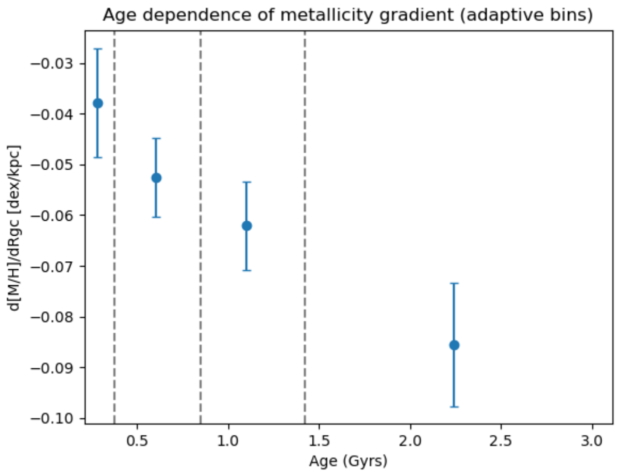


Fig. 21. The dependence of metallicity gradient on age. Clusters with $\log Age < 8.3$ are excluded.

6.2. Comparison with asteroseismology, gyrochronology and eclipsing binaries

Beyond isochrone fitting, highly precise age estimates can be achieved through asteroseismology. Studies of eclipsing binary systems also offer reliable age constraints by providing insights into stellar masses. Additionally, gyrochronology serves as another useful method for estimating ages. Although such detailed information is available for only a limited number of objects, these cases remain valuable benchmarks for understanding the uncertainties and biases involved in age determinations. In the following, we discuss a few OCs whose age is derived with the above methods.

6.2.1. NGC 6791

Brogaard et al. (2012) estimate the age of NGC 6791 as $\log Age = 9.92^{+0.035}_{-0.026}$, after accounting for multiple sources of uncertainty, including prolonged star formation and different values of CNO abundances. Our isochrone-derived value of $\log Age = 9.95 \pm 0.01$ is compatible within the errors of the eclipsing binary method.

6.2.2. NGC 6866

Brogaard et al. (2023) have used asteroseismology to estimate the age of this cluster as $\log Age = 8.63 \pm 0.05$, which is not consistent with our estimation of $\log Age = 8.92 \pm 0.02$ within the errors. The authors make use of the MESA models (Choi et al. 2016) and utilize a convective overshooting parameter much lower than the one of the PARSEC models we utilize. They note that this choice of convective overshooting is necessary to obtain the best possible fit for this particular cluster. They point out that this explains the difference of their age estimate to other values based on the PARSEC models, such as the one by Bossini et al. (2019) ($\log Age = 8.89^{+0.05}_{-1.00}$) which is compatible to our own estimation.

6.2.3. NGC 6811

Sandquist et al. (2016) analysed an eclipsing binary in this cluster and derived $\log Age = 9.0$ for the primary star and $\log Age = 9.07$ for the companion star, which, they note, are mutually inconsistent within their uncertainties. Our determination of $\log Age = 9.04 \pm 0.01$, is in between those values. Our determination is also compatible with the gyrochronology of G-type dwarfs, resulting in $\log Age = 9.02 \pm 0.03$ (Curtis et al. 2019). However, some asteroseismic ages are much lower. For instance, Çelik Orhan (2021) derived $\log Age$ in the range between 8.85 − 8.91 dex.

6.2.4. NGC 6819

Rodrigues et al. (2017) derived $\log Age = 9.35 \pm 0.03$ through asteroseismology. Brewer et al. (2016) determined $\log Age = 9.38 \pm 0.05$ using eclipsing binaries. Both values are compatible with our estimation of $\log Age = 9.38 \pm 0.01$.

6.2.5. NGC 2506

Knudstrup et al. (2020) determine the age of this cluster as $\log Age = 9.30 \pm 0.02$ using eclipsing binaries. We estimate $\log Age = 9.33 \pm 0.01$, which is compatible within the errors.[6]

6.3. The variable metallicity gradient

The variations of the metallicity gradient of the OCs can provide insights into the formation and chemical evolution of the Milky Way disk and to processes such as radial migration and OC disruption (see, e.g. Gaia Collaboration et al. 2023b). Since all of our clusters have reliable and homogenous metallicity determinations, we can use our updated sample to rederive the metallic-

[6] This cluster is adequately well fitted by the isochrone, but we could not obtain a successful fit of the surrogate likelihood. In exception to our general methodology and because this cluster is important, we report values derived from the tempered posterior. At tempering factor $T = 9.1$, the posterior distribution of the parameters is very narrow and Gaussian. Since multimodality is not present and the distribution is not skewed, we can recover the internal uncertainties by approximately rescaling the tempered posterior by $1/\sqrt{T}$. This yields $\delta \log Age < 0.01$. Thus, we report as internal uncertainty the grid resolution, $\delta \log Age = 0.01$. This uncertainty is most probably underestimated. The rest of the fitted parameters are $\mu = 12.49 \pm 0.03$, $A_V = 0.26 \pm 0.02$. These results are not included in Table G.1, due to the different methodology.

ity gradient of the thin disk, as probed by OCs. We restrict galactocentric radius R_{GC} to $6 - 12$ kpc, within which the majority of our clusters lie. We also exclude clusters with $\log Age < 8.3$, as they are not evenly distributed along R_{GC}. There are 233 OCs within this range. We perform adaptive binning in $\log Age$, as follows: Starting from 40 quantile bins, we required each bin to contain at least 20 clusters, span an adequate R_{GC} range, and yield a statistically significant linear fit of the metallicity gradient ($p < 0.05$). We also discarded bins whose R_{GC} distribution was grossly discrepant from the full sample (evaluated via a two-sample KS test, requiring $p \geq 0.05$ to pass [7]). Failing bins were iteratively merged with the adjacent bin with the closest median $\log Age$, until all criteria are met.

As a result, we obtained four age bins, containing between 46 and 93 clusters. The slope for the four age bins is shown in Fig. 21. We obtain a slope of -0.062 ± 0.009 dex/kpc at an age of 1 Gyr and -0.086 ± 0.012 dex/kpc at 2.2 Gyrs. Fig. 21 indicates that the metallicity gradient is steeper for older clusters, possibly due to the disruptive effect of the galactic potential on clusters migrating inward. Our result is compatible with Gaia Collaboration et al. (2023b). Carbajo-Hijarrubia et al. (2024) also find that the metallicity gradient becomes steeper for older clusters.

7. Summary and future work

We derived reference parameters for 320 OCs, for which reliable spectroscopic priors exist. We compiled a homogeneous metallicity catalog from ground-based spectroscopic surveys and GSP-Spec data. We derived updated membership using *Gaia* DR3 astrometry in order to improve upon the completeness and purity of previous determinations. Our updated approximate Bayesian isochrone fitting code leverages ML and is fast, reliable and precise. Our novel approximate Bayesian method for handling stochastically computed likelihoods is robust and utilizes Bayesian principles wherever possible, guaranteeing robust and interpretable uncertainty estimation. For this study, we utilized a very high quality and homogeneous photometric dataset, that includes the Gaia bands and synthetic photometry derived from the Gaia XP spectra in a multiband approach that minimizes fit degeneracies. As a result, we obtained a sample of OCs, with well-defined ages, thus improving on the accuracy and completeness of previous works.

This sample of clusters can serve as a reliable basis for training ML models to derive cluster parameters, a possibility to be explored in future work. Moreover, as more metallicity determinations become available with compatible and homogenized scales, the catalog can be expanded to more objects. Many clusters with already available metallicities present significantly distorted CMDs due to a range of effects that are difficult to accurately model. These objects cannot be reliably tackled with standard isochrone fitting procedures. For these clusters and those that lack a reliable determination of the metallicity, a ML approach is to be explored in the future.

[7] For each bin, we tested the null hypothesis that its R_{GC} distribution and that of the full sample are drawn from the same parent distribution. We retained the bin only if this hypothesis could not be rejected at the 5% level ($p \geq 0.05$). Bins with $p < 0.05$ were considered to have grossly discrepant radial coverage due to selection effects, and were merged with their neighbors. Not rejecting the null hypothesis does not prove that the distributions are identical. It is a lenient criterion that efficiently rejects the most severely skewed bins.

8. Data availability

The catalog of 320 OCs with their parameters and validation statistics, as well as the catalog of member stars are available at CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via `http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/??/??`

Acknowledgements. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101072454, as part of the Milky Way-*Gaia* Doctoral Network (`https://www.mwgaiadn.eu/`). We acknowledge funding from Bando Astrofisica Fondamentale INAF 2022 (High resolution spectroscopy of open clusters, PI Bragaglia) and 2023 (open clusters and stellar structures in the local Galactic disk, PI Vallenari). This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This work has made use of public auxiliary data provided by ESA/Gaia/DPAC/CU5 and prepared by Carine Babusiaux. This research has made use of the VizieR catalog access tool (Ochsenbein et al. 2000), CDS, Strasbourg Astronomical Observatory, France (DOI : 10.26093/cds/vizier). We thank professor Carine Babusiaux (IPAG Grenoble) for her useful clarifications on interstellar extinction. We would also like to thank the anonymous referee for their useful comments and suggestions.

References

Anders, F., Khalatyan, A., Queiroz, A. B. A., et al. 2022, A&A, 658, A91
Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. 1999, SIGMOD Rec., 28, 49–60
Babusiaux, C., Fabricius, C., Khanna, S., et al. 2023, A&A, 674, A32
Bastos, L. S. & O'Hagan, A. 2009, Technometrics, 51, 425
Binney, J. & Tremaine, S. 2008, Galactic Dynamics: Second Edition
Boeche, C., Alonso-Santiago, J., Bragaglia, A., et al. 2026, A&A, 710, A285
Bossini, D., Vallenari, A., Bragaglia, A., et al. 2019, A&A, 623, A108
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Brewer, L. N., Sandquist, E. L., Mathieu, R. D., et al. 2016, AJ, 151, 66
Brogaard, K., Arentoft, T., Miglio, A., et al. 2023, A&A, 679, A23
Brogaard, K., VandenBerg, D. A., Bruntt, H., et al. 2012, A&A, 543, A106
Buchner, J. 2021, The Journal of Open Source Software, 6, 3001
Buder, S., Kos, J., Wang, X. E., et al. 2025, PASA, 42, e051
Buder, S., Sharma, S., Kos, J., et al. 2021, MNRAS, 506, 150
Caffau, E., Ludwig, H.-G., Steffen, M., Freytag, B., & Bonifacio, P. 2011, Sol. Phys., 268, 255
Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, in Advances in Knowledge Discovery and Data Mining, ed. J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Berlin, Heidelberg: Springer Berlin Heidelberg), 160–172
Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. 2020, A&A, 640, A1
Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, A&A, 618, A93
Carbajo-Hijarrubia, J., Casamiquela, L., Carrera, R., et al. 2024, A&A, 687, A239
Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245
Casamiquela, L., Blanco-Cuaresma, S., Carrera, R., et al. 2019, MNRAS, 490, 1821
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2022, A&A, 661, A118
Cavallo, L., Spina, L., Carraro, G., et al. 2024a, AJ, 167, 12
Cavallo, L., Spina, L., Carraro, G., et al. 2024b, AJ, 167, 12
Çelik Orhan, Z. 2021, MNRAS, 506, 4413
Chabrier, G., Hennebelle, P., & Charlot, S. 2014, ApJ, 796, 75
Chen, Y., Girardi, L., Bressan, A., et al. 2014, MNRAS, 444, 2525
Chipman, H. A., George, E. I., & McCulloch, R. E. 2010, The Annals of Applied Statistics, 4, 266
Choi, J., Dotter, A., Conroy, C., et al. 2016, ApJ, 823, 102
Cranmer, K., Brehmer, J., & Louppe, G. 2020, Proceedings of the National Academy of Sciences, 117, 30055
Curtis, J. L., Agüeros, M. A., Douglas, S. T., & Meibom, S. 2019, ApJ, 879, 49
Danielski, C., Babusiaux, C., Ruiz-Dern, L., Sartoretti, P., & Arenou, F. 2018, A&A, 614, A19
Dias, W. S., Monteiro, H., Moitinho, A., et al. 2021, MNRAS, 504, 356
Donor, J., Frinchaboy, P. M., Cunha, K., et al. 2020, AJ, 159, 199
Duchêne, G. & Kraus, A. 2013, ARA&A, 51, 269
Earl, D. J. & Deem, M. W. 2005, Physical Chemistry Chemical Physics (Incorporating Faraday Transactions), 7, 3910
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601

Fitzpatrick, E. L., Massa, D., Gordon, K. D., Bohlin, R., & Clayton, G. C. 2019, ApJ, 886, 108
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Gaia Collaboration, Montegriffo, P., Bellazzini, M., et al. 2023a, A&A, 674, A33
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
Gaia Collaboration, Recio-Blanco, A., Kordopatis, G., et al. 2023b, A&A, 674, A38
Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023c, A&A, 674, A1
Gaposchkin, C. P. 1948, AJ, 53, 193
Geman, S., Bienenstock, E., & Doursat, R. 1992, Neural Computation, 4, 1
Gieles, M., Larsen, S. S., Scheepmaker, R. A., et al. 2006, A&A, 446, L9
Gilmore, G., Randich, S., Worley, C. C., et al. 2022, A&A, 666, A120
Grevesse, N., Asplund, M., & Sauval, A. J. 2007, Space Sci. Rev., 130, 105
Grevesse, N. & Sauval, A. J. 1998, Space Sci. Rev., 85, 161
Groenewegen, M. A. T. 2021, A&A, 654, A20
He, J., Yalov, S., & Hahn, P. R. 2019, in Proceedings of Machine Learning Research, Vol. 89, Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, ed. K. Chaudhuri & M. Sugiyama (PMLR), 1130–1138
Hunt, E. L. & Reffert, S. 2021, A&A, 646, A104
Hunt, E. L. & Reffert, S. 2023, A&A, 673, A114
Hunt, E. L. & Reffert, S. 2024, A&A, 686, A42
Ivezić, Ž., Sesar, B., Jurić, M., et al. 2008, ApJ, 684, 287
Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. 1983, Science, 220, 671
Knudstrup, E., Grundahl, F., Brogaard, K., et al. 2020, MNRAS, 499, 1312
Knuth, K. H. 2006, arXiv e-prints, physics/0605197
Kordopatis, G., Schultheis, M., McMillan, P. J., et al. 2023, A&A, 669, A104
Lallement, R., Vergely, J. L., Babusiaux, C., & Cox, N. L. J. 2022, A&A, 661, A147
Lange, J. U. 2023, MNRAS, 525, 3181
Li, L. & Shao, Z. 2022, ApJ, 930, 44
Li, L., Shao, Z., Li, Z., & Fu, X. 2025, AJ, 170, 288
Lindegren, L., Bastian, U., Biermann, M., et al. 2021, A&A, 649, A4
Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9
Magrini, L., Viscasillas Vázquez, C., Spina, L., et al. 2023, A&A, 669, A119
Mahalanobis, P. 2018, Sankhyā: The Indian Journal of Statistics Series B, 80, 1
Montegriffo, P., De Angeli, F., Andrae, R., et al. 2023, A&A, 674, A3
Nguyen, C. T., Costa, G., Girardi, L., et al. 2022, A&A, 665, A126
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, Astronomy and Astrophysics Supplement Series, 143, 23
Offner, S. S. R., Moe, M., Kratter, K. M., et al. 2023, in Astronomical Society of the Pacific Conference Series, Vol. 534, Protostars and Planets VII, ed. S. Inutsuka, Y. Aikawa, T. Muto, K. Tomida, & M. Tamura, 275
Oliver, W. H., Elahi, P. J., Lewis, G. F., & Buck, T. 2024, MNRAS, 530, 2637
Perren, G. I., Pera, M. S., Navone, H. D., & Vázquez, R. A. 2022, A&A, 663, A131
Perren, G. I., Pera, M. S., Navone, H. D., & Vázquez, R. A. 2023, MNRAS, 526, 4107
Perren, G. I., Vázquez, R. A., & Piatti, A. E. 2015, A&A, 576, A6
Randich, S., Gilmore, G., Magrini, L., et al. 2022, A&A, 666, A121
Recio-Blanco, A., de Laverny, P., Palicio, P. A., et al. 2023, A&A, 674, A29
Rodrigues, T. S., Bossini, D., Miglio, A., et al. 2017, MNRAS, 467, 1433
Sandquist, E. L., Jessen-Hansen, J., Shetrone, M. D., et al. 2016, ApJ, 831, 11
Scargle, J. D., Norris, J. P., Jackson, B., & Chiang, J. 2013, ApJ, 764, 167
Smolinski, J. P., Lee, Y. S., Beers, T. C., et al. 2011, AJ, 141, 89
Spina, L., Ting, Y. S., De Silva, G. M., et al. 2021, MNRAS, 503, 3279
Tarricq, Y., Soubiran, C., Casamiquela, L., et al. 2022, A&A, 659, A59
Torres, G., Andersen, J., & Giménez, A. 2010, A&A Rev., 18, 67
Tremmel, M., Fragos, T., Lehmer, B. D., et al. 2013, ApJ, 766, 19
Vanderbeke, J., West, M. J., De Propris, R., et al. 2014, MNRAS, 437, 1734
Vergely, J. L., Lallement, R., & Cox, N. L. J. 2022, A&A, 664, A174
Wang, T., Yuan, H., Chen, B., et al. 2025, ApJS, 280, 15
Williams, C. K. I. 1997, Regression with Gaussian Processes, ed. S. W. Ellacott, J. C. Mason, & I. J. Anderson (Boston, MA: Springer US), 378–382

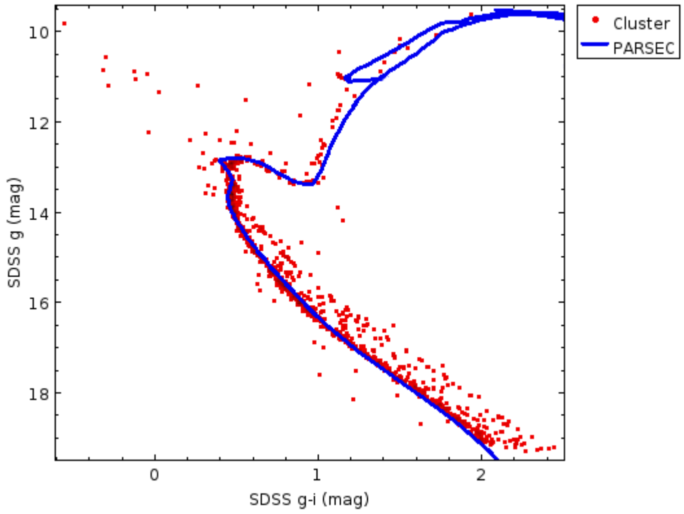
Appendix A: Photometric band choice

In order to better constrain our fit, we adopted a multi-band approach by including the *Gaia* G, BP, RP bands and in addition the synthetic SDSS[8] g, i, z photometry, as derived by Gaia Collaboration et al. (2023a). This synthetic photometry, readily available through the *Gaia* Archive, is very precise ($< 0.01 mag$) and homogeneous.

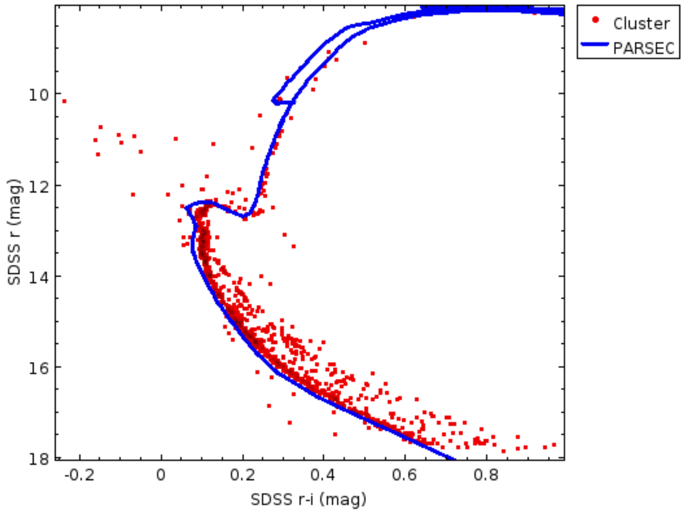
Unfortunately, this dataset does not include good photometry toward the blue part of the spectrum that is more sensitive to extinction, and the synthetic u band is only valid for a very small percentage of high signal to noise stars, because it is poorly covered by the Blue Photometer (BP) (Gaia Collaboration et al. 2023a). We do not make use of the SDSS r band, as it does not fit the PARSEC isochrones as consistently as the other bands. In particular, it is incompatible with the best-fit isochrones of the *Gaia* Bands, as seen in Fig. A.1 for the case of NGC 2682. This is most likely due to the filter response curve adopted in the PARSEC isochrones, as in Gaia Collaboration et al. (2023a) it is shown that the synthetic r band matches the photometry from ground-based observations as well as the other bands which we utilize.

Although the spectral windows of the SDSS bands is covered by the Gaia bands, the SDSS bands are much narrower. Thus, they contain information that has been integrated away in the Gaia bands. Hence, utilizing these bands together does not amount to "using the same data twice", which would lead to invalid statistical inference. For this reason, we do not use filters of comparable bandwidth in equivalent spectral windows, such as the Johnson V together with SDSS g, although they are both very consistent with the Gaia Bands and our chosen isochrone sets.

All stars in the published catalog have $flux/flux\_error > 30$, which, according to Gaia Collaboration et al. (2023a) amounts to valid photometry.



(a)



(b)

Fig. A.1. Best fit isochrone of NGC 2682 synthetic SDSS bands. The r band is inconsistent with the other bands. The offset ~ 0.02 in $r-i$ color is very visible partially due to the smaller color range compared to $g-i$. However, there is also offset in magnitude as well.

Appendix B: Implementation of Tremmel et al. (2013) Poisson likelihood ratio

The full expression for the Tremmel et al. (2013) likelihood in the context of Hess diagram fitting, is the sum over all bins of:

$$\log L = \sum_{i}^{B}[\log\Gamma(n_i+m_i+\frac{1}{2})-\log\Gamma(m_i+\frac{1}{2})-m_i\log 2-n_i\log 2]+C \tag{B.1}$$

where n_i is the number of stars in a bin of the observed CMD and m_i the number of stars in the bin of the synthetic one. The term C incorporates every all constant terms, which can be ignored as they have no effect on inference.

As discussed in Tremmel et al. (2013), this formulation of the Poisson Likelihood Ratio is well-defined even to the limit of zero bin counts. Thus, this likelihood is robust even for the case of clusters with low number of stars per CMD bin. This enables us to calculate a statistically meaningful likelihood even for clusters with a small number of members (see Table G.1).

Since the number of observed stars is always the same for a given cluster, the Σn_i term can also be considered constant. The Σm_i term effectively penalizes models that predict stars at bins where there are no observed stars.

By default, ASteCA removes the bins where no stars are observed. This is a sensible approach in the context of incomplete coverage of CMD features of the observed clusters; missing parts of the observed CMD would lead to an "unfair" penalization of an otherwise correct set of parameters. This default likelihood differs from the one defined as Eq. B.1 by:

$$\log L_{trunc} = \log L - \sum_{i:n_i=0}(-m_i\log 2) \tag{B.2}$$

When the likelihood is evaluated for a well-fitting model, synthetic stars are not in any of the discarded bins, so $\log L_{trunc} = \log L$. However, when the model is not fitting some features of the data, synthetic stars are predicted inside bins that are discarded, so $\log L_{trunc} > \log L$. The likelihood is less peaked around the best-fitting value. In clusters with very well defined CMDs and many counts per bin, the $\log\Gamma$ terms of the likelihood strongly dominate and the two likelihoods are numerically equivalent. Moreover, in every case, Eq. B.2 is a conservative (broader and less peaked) approximation of the true likelihood, yielding wider credible intervals. However, in cases of clusters

[8] Sloan Digital Sky Survey

with few counts per bin, Eq. B.1 is the more statistically correct option.

In order to address this, we modified ASteCA to keep all the bins of the Hess diagrams, and utilize this more formally correct version of the likelihood in order to better fit some objects. This proved necessary for 43 OCs (marked with a flag in Table G.1). We tested the modified likelihood in clusters well-fitted using the original likelihood and obtained identical results. It should be noted that, as of the time of writing, the latest version of ASteCA (0.6.3) has adopted a formulation of the likelihood that is mathematically equivalent to this modification.

The Poisson likelihood ratio depends on discrete Poissonian statistics of the bin counts. Thus, there is no straightforward way to impose statistical weights on more important parts of the CMD. In cases where this was necessary (for instance, the RC of clusters with a small amount of stars), we inflate the uncertainties of the fainter main-sequence stars. Since the synthetic clusters are generated by sampling those uncertainties, these less informative parts are less likely to drive the fit, but the Poissonian characteristics of the bin counts are retained and the likelihood remains formally correct. It is indicated in the catalog when this is done, as a flag in Table G.1. This modification of the original ASteCA algorithm is implemented only where the best-fit isochrone was consistently missing few high-probability RC stars. This was implemented only for ten OCs.

Appendix C: Member stars

In this section, we provide a few examples of comparison of our membership determination with literature catalogs. We typically detect more stars across all luminosities, compared to Cantat-Gaudin et al. (2020). This is only in part due to the more sensitive clustering algorithm. This comparison is especially interesting, since our clustering algorithm has certain similarities with UPMASK (Cantat-Gaudin et al. 2018), used in this prior work. In particular, both AstroLink and UPMASK detect the prominence of an overdensity over the background, while UPMASK also utilizes a MC outlier rejection, albeit implicitly.

In the case of NGC 2682 (see Fig. C.1), we recovered more than twice as many stars as the Cantat-Gaudin et al. (2020) catalog. More stars were recovered at every magnitude, as an effect of the detection algorithm. At fainter magnitudes, the higher completeness and precision of *Gaia* DR3 data play a relevant role. Indeed, due to *Gaia* DR2 data quality, Cantat-Gaudin et al. (2020) utilize a magnitude cutoff of $G = 18$, while our cutoff is at $G = 19$.

However, NGC 2682 represents something close to the ideal case of a well defined cluster. Sparser objects can be more challenging, but we still recovered more stars than Cantat-Gaudin et al. (2020) within the same radius in most cases, as seen in Figure C.2.

In Fig.s C.3 and C.4 we compare our results with the DR3-based member selection of Hunt & Reffert (2023) (constrained to the same sky cone and magnitude limit as our search). In objects with well defined CMD features (as in the case of NGC 188 in Fig. C.3), and in sparser objects such as ASCC 11 (Fig. C.4) the membership determinations were equivalent.

Finally, Fig.C.5 shows an example of outlier rejection using the Mahalanobis distance cut for NGC 6404. This was implemented for 16 OCs, marked by a corresponding flag as described in Appendix G.

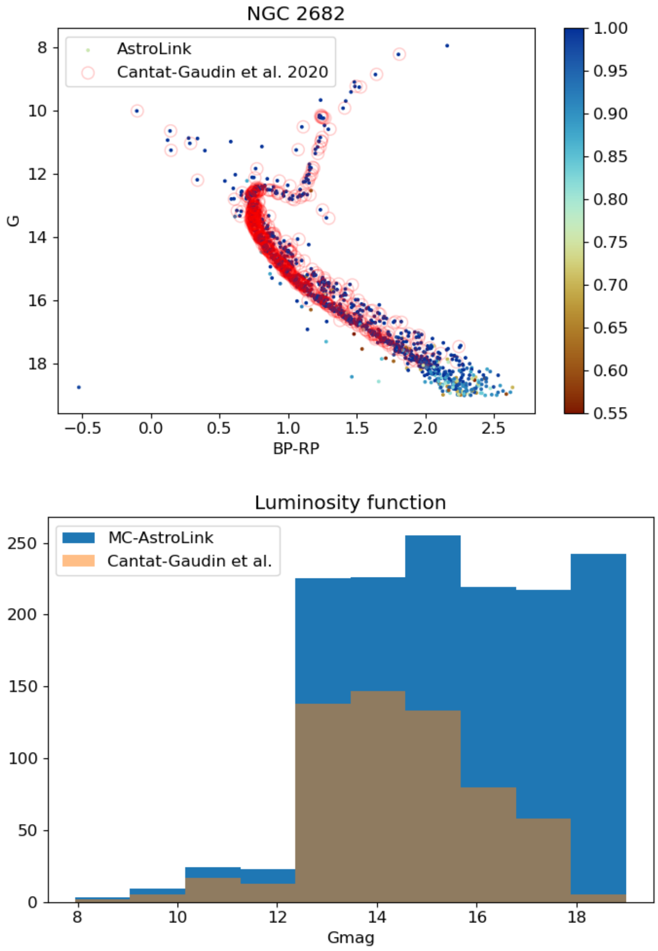


Fig. C.1. Comparison of the membership of NGC 2682 to Cantat-Gaudin et al. (2020). There are far more stars at every luminosity, more dramatically at the faint end until cutoff at $G = 19$ mag. Colorbar corresponds to the detection rate of the MC outlier rejection scheme. On the CMD (upper panel), the circled stars are the ones detected by Cantat-Gaudin et al. (2020).

Appendix D: Parameter correlations

In 5.2 we showed that for a number of objects, the Bayesian update caused the expected value of the posterior metallicity to move away from the prior. Fig. D.1 compares age, extinction and distance modulus to the fitted values from Cantat-Gaudin et al. (2020). We also compare the deviation of the fitted parameters to the membership count of the clusters, and we calculate the Pearson correlation coefficient ρ. This coefficient ranges from -1 to 1 (perfect anticorrelation to perfect correlation). The detected correlations are considered statistically significant, if the $p-$value is $p < 0.05$. Only some moderate anti-correlation between age and extinction ($\rho = -0.32$, $p < 0.001$) and age and distance modulus ($\rho = -0.23$, $p < 0.001$) can be detected. As shown in Fig. D.1, outliers in metallicity (red points) do not exhibit outlier behavior in the other fitted parameters and in general they do not correlate with low membership counts.

Appendix E: An example best-fit synthetic cluster

In Fig. E.1 we present the best fit synthetic CMD for OC NGC 188. The stochastic nature of the cluster simulation entails that each realization of a cluster with the same parameters

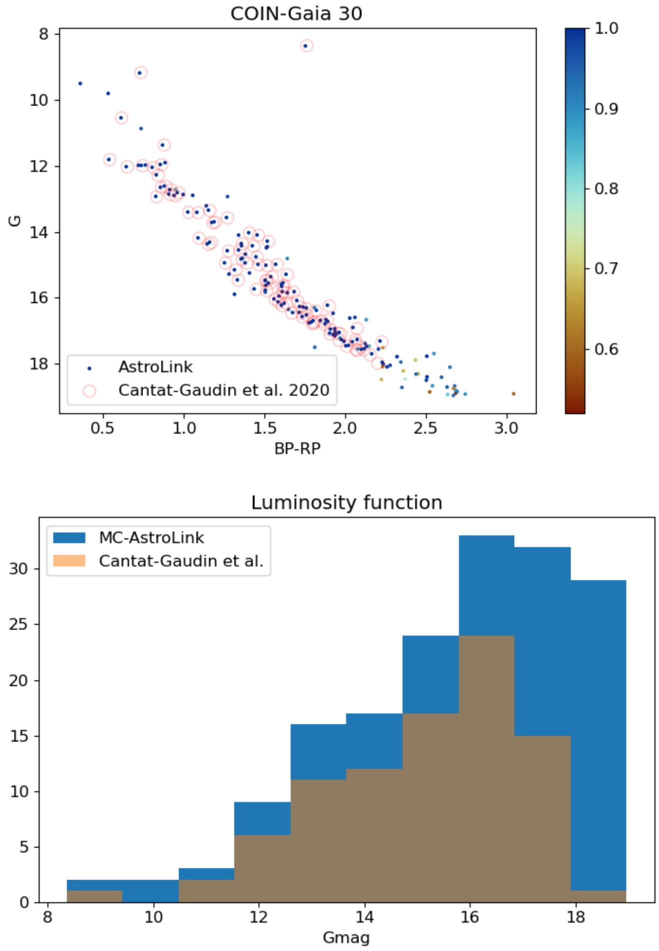


Fig. C.2. Comparison of the membership of COIN-Gaia 30 to Cantat-Gaudin et al. (2020). We recover more stars at every luminosity bin. Colorbar corresponds to the detection rate of the MC outlier rejection scheme. On the CMD (upper panel), the circled stars are the ones detected by Cantat-Gaudin et al. (2020).

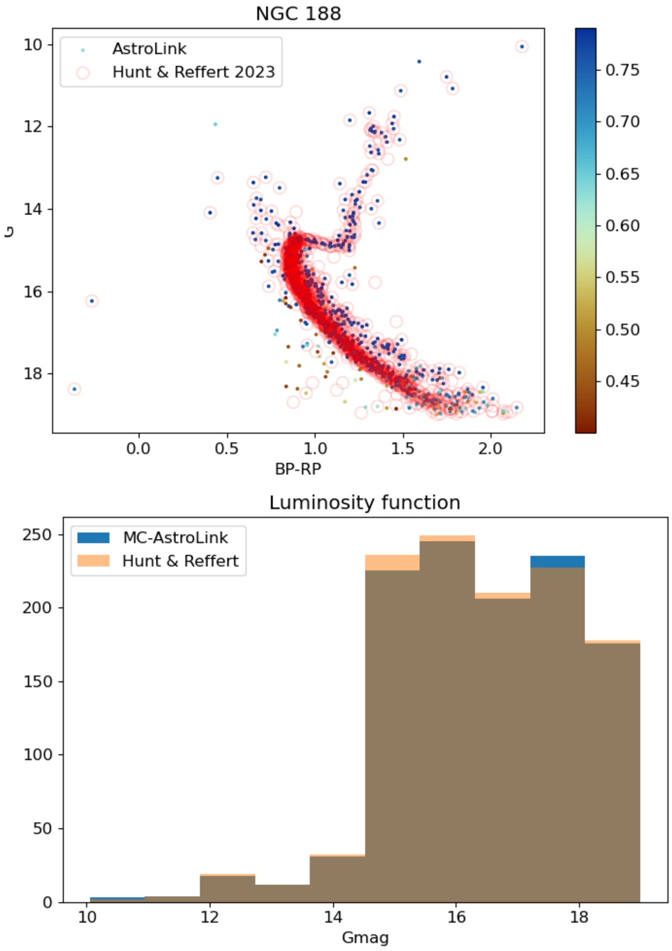


Fig. C.3. Comparison of the membership of NGC 188 to Hunt & Reffert (2023). Colorbar corresponds to the detection rate of the MC outlier rejection scheme. On the CMD (upper panel), the circled stars are the ones detected by Hunt & Reffert (2023).

will be slightly different. The true likelihood of each point in the parameter space corresponds to the average of multiple realizations, which is implicitly derived by our surrogate model, as described in 4.2.

Appendix F: The true posterior of ASCC 111

In order to visually demonstrate the ability of our surrogate model to recover the true shape of the posterior, we give the example of OC ASCC 111. This OC has relatively few member stars (143), and hence its likelihood has a smaller numerical range and relatively low noise. This enables us to also perform the fit by averaging multiple likelihood evaluations. Here, averaging eight evaluations of the likelihood was enough for convergence, but in the majority of clusters, many more evaluations would be required, rendering such analysis extremely computationally expensive.

In Fig. F.1, we show an estimation of the posterior by averaging the likelihood 8 times at each point. Even then, the sampler was run with 250 live points (instead of the 2000 we usually utilize for best accuracy) and only up to 1500 effective samples, instead of the 10000 effective samples obtained through 2000 live points we use a standard. This is enough for inferring the inter quantile range of the 1σ credible intervals, but is not ideal for recovering the shape of the true posterior.

In Figure F.2, we show the tempered posterior at $T = 5.35$. It is noticeable that metallicity $[M/H]$ and distance modulus μ, which are strongly constrained by the prior are not very affected by the tempering. On the other hand, $\log Age$, which shows significant bimodality, is greatly affected. The two modes are merged together and the structure is distorted. Moreover, in both A_V and $\log Age$, the credible intervals are much wider.

In Figure F.3, it can be seen that the XBART surrogate likelihood satisfactorily reverts the effects of tempering. The positions of the modes of $\log Age$ is identical that of the averaged case, and the credible intervals of all the other parameters are identical. The small difference in the median value of $\log Age$ between the averaged likelihood and the surrogate is not a cause for alarm, since the exact position of the median is very sensitive to the relative weight of the modes. Moreover, the posterior obtained through averaging cannot be considered very accurate in itself, due to the low effective sample size and the sampler settings necessary to acquire it. The true validation of the surrogate likelihood is done with the posterior-predictive check, shown in Figure 4. There it is seen that, in the statistically significant parts of the likelihood, there is no bias present in the surrogate model and the only variation is due to the stochastic noise.

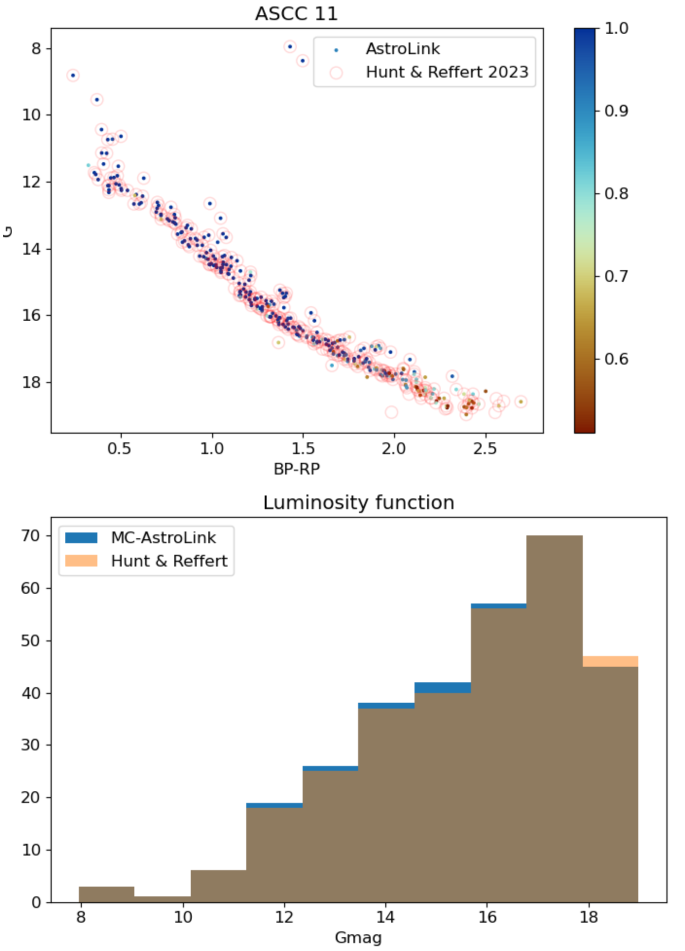


Fig. C.4. Comparison of the membership of ASCC 11 to Hunt & Reffert (2023). Colorbar corresponds to the detection rate of the MC outlier rejection scheme. On the CMD (upper panel), the circled stars are the ones detected by Hunt & Reffert (2023).

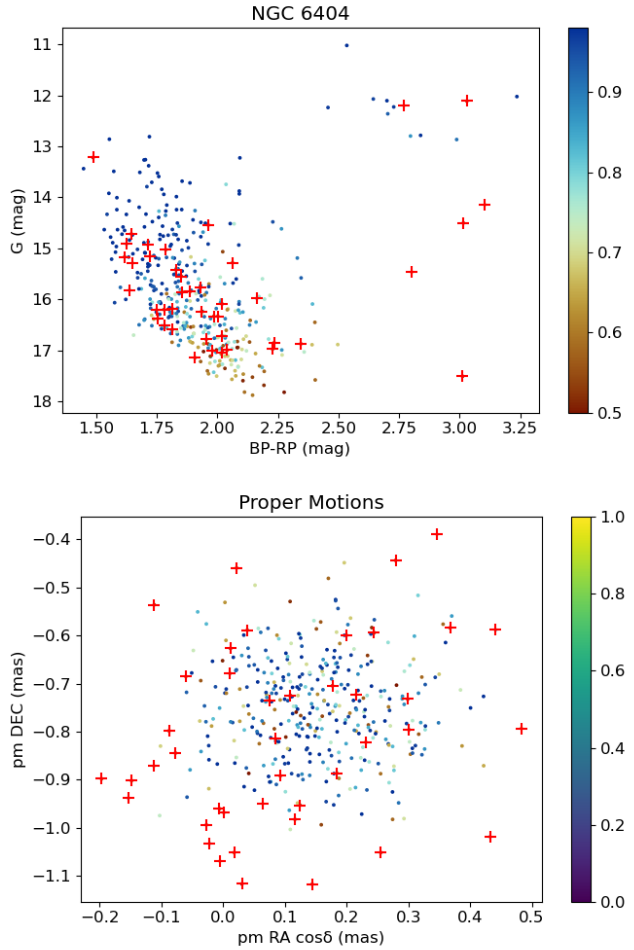


Fig. C.5. Outlier rejection through a Mahalanobis distance cut. Crossed stars have a Mahalanobis distance > 3 from the center of the cluster.

Appendix G: Summary of online tables

In Table G.1 the main results of this work are presented. This table includes basic cluster parameters, the spectroscopic priors utilized, and the posterior parameters inferred. The uncertainties are represented by the 16th and 84th percentiles of the posterior distribution of the surrogate likelihood. The online version of the table also includes several diagnostics of the fit quality, the tempering factor and B/V diagnostic of the surrogate model, which are described in Table G.3 and the README file of the online tables. Flags also provide information about the binning scheme used and and the variant of the PLR likelihood used, as described in Appendix B.

Table G.2 contains the member stars for the clusters in this work, derived through the AstroLink algorithm, as described in Section 3.

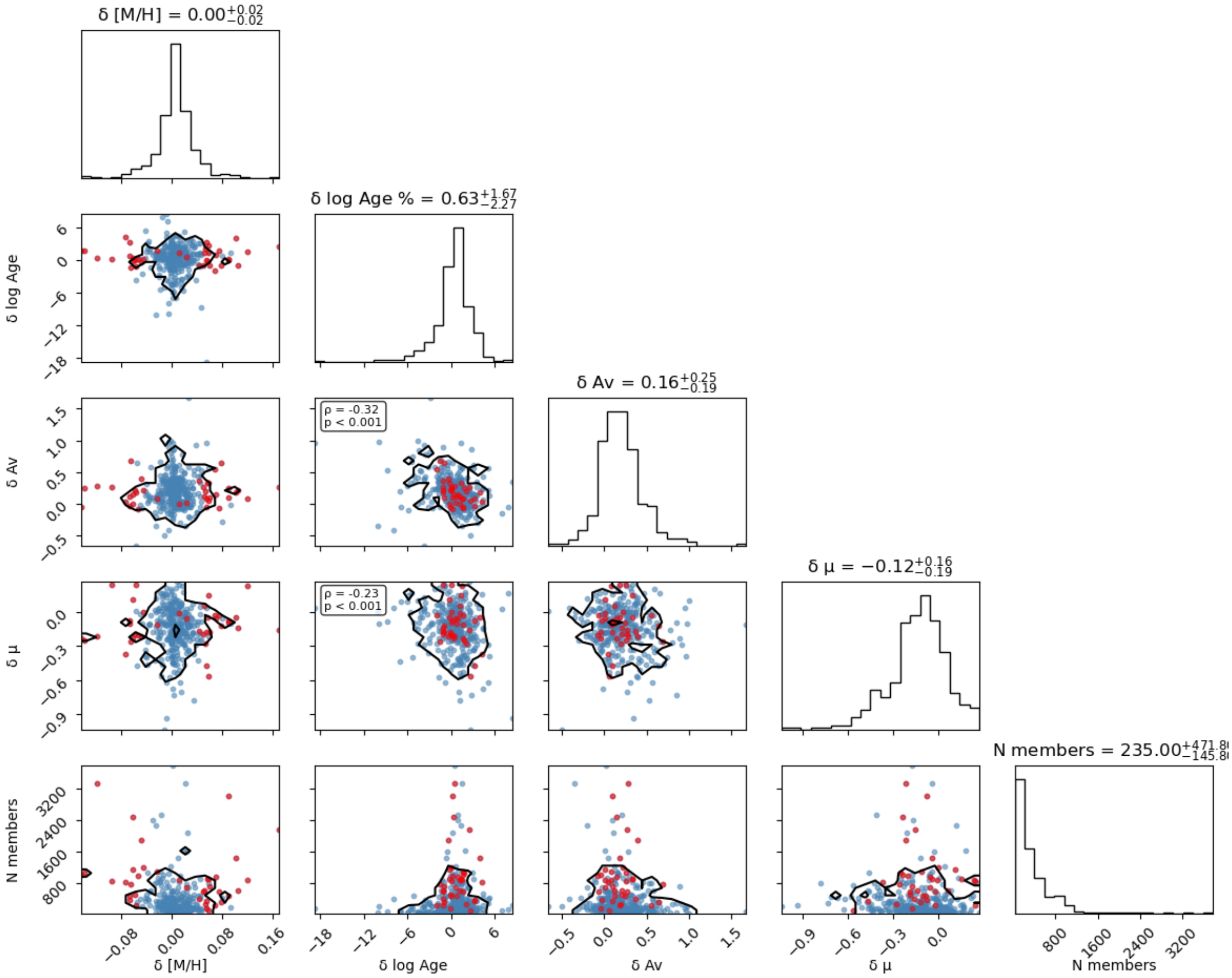


Fig. D.1. Residual correlation plot of fitted vs literature values. For the metallicity, we compare to the spectroscopic priors. For the other values we compare to Cantat-Gaudin et al. (2020). For the log *Age* parameter we show the percentage of difference, as more physically meaningful. We also include the number of high probability member stars from our catalog. Red points indicate OCs for which the fit converged to a metallicity value > 1σ away from the spectroscopic prior. The ρ Pearson correlation value is displayed when significant ($p < 0.05$).

Table G.1. First 10 rows of the main results table.

Cluster	R.A.	Dec.	$N_{\rm stars}$	$[M/H]_{prior}$	$N_{\rm spec}$	Spec source	$[M/H]_{post}$	log *Age*	A_V	μ
ASCC 11	53.04	44.84	307	-0.34 ± 0.06	17	OCCAM	$-0.32^{+0.04}_{-0.06}$	$8.85^{+0.05}_{-0.06}$	$0.81^{+0.05}_{-0.07}$	9.41 ± 0.07
ASCC 111	302.94	37.52	143	0.13 ± 0.05	1	GSPSPEC	$0.15^{+0.05}_{-0.06}$	$8.00^{+0.44}_{-0.11}$	$0.72^{+0.08}_{-0.10}$	$9.65^{+0.07}_{-0.08}$
ASCC 16	81.42	1.63	976	-0.07 ± 0.05	62	OCCAM	-0.13 ± 0.01	$7.14^{+0.00}_{-0.01}$	$0.13^{+0.02}_{-0.04}$	7.49 ± 0.00
ASCC 19	81.99	-1.93	130	-0.09 ± 0.05	183	OCCAM	-0.10 ± 0.05	$7.00^{+0.07}_{-0.08}$	$0.33^{+0.14}_{-0.13}$	7.77 ± 0.05
ASCC 23	95.00	46.72	172	-0.03 ± 0.05	1	GSPSPEC	0.00 ± 0.05	$8.32^{+0.40}_{-0.29}$	$0.33^{+0.11}_{-0.09}$	8.88 ± 0.08
ASCC 41	116.69	0.14	201	-0.08 ± 0.05	1	GSPSPEC	$-0.07^{+0.05}_{-0.07}$	$8.03^{+0.07}_{-0.14}$	$0.14^{+0.12}_{-0.10}$	$7.43^{+0.09}_{-0.07}$
ASCC 71	185.10	-67.52	105	0.01 ± 0.05	3	GSPSPEC	-0.00 ± 0.05	$8.12^{+0.21}_{-0.39}$	$0.91^{+0.09}_{-0.10}$	10.61 ± 0.09
ASCC 88	256.85	-35.58	79	-0.01 ± 0.05	1	GSPSPEC	0.02 ± 0.05	$8.13^{+0.19}_{-0.30}$	$2.02^{+0.07}_{-0.06}$	$9.72^{+0.07}_{-0.06}$
ASCC 90	264.78	-34.82	79	0.05 ± 0.05	4	GSPSPEC	$0.07^{+0.05}_{-0.04}$	$8.80^{+0.08}_{-0.09}$	0.70 ± 0.10	$8.80^{+0.03}_{-0.04}$
Alessi 1	13.30	49.54	96	0.03 ± 0.05	2	GSPSPEC	0.05 ± 0.05	$8.99^{+0.07}_{-0.06}$	$0.32^{+0.12}_{-0.11}$	$9.26^{+0.09}_{-0.08}$

Notes. The online version of this table contains more columns, excluded from this sample for the sake of brevity.

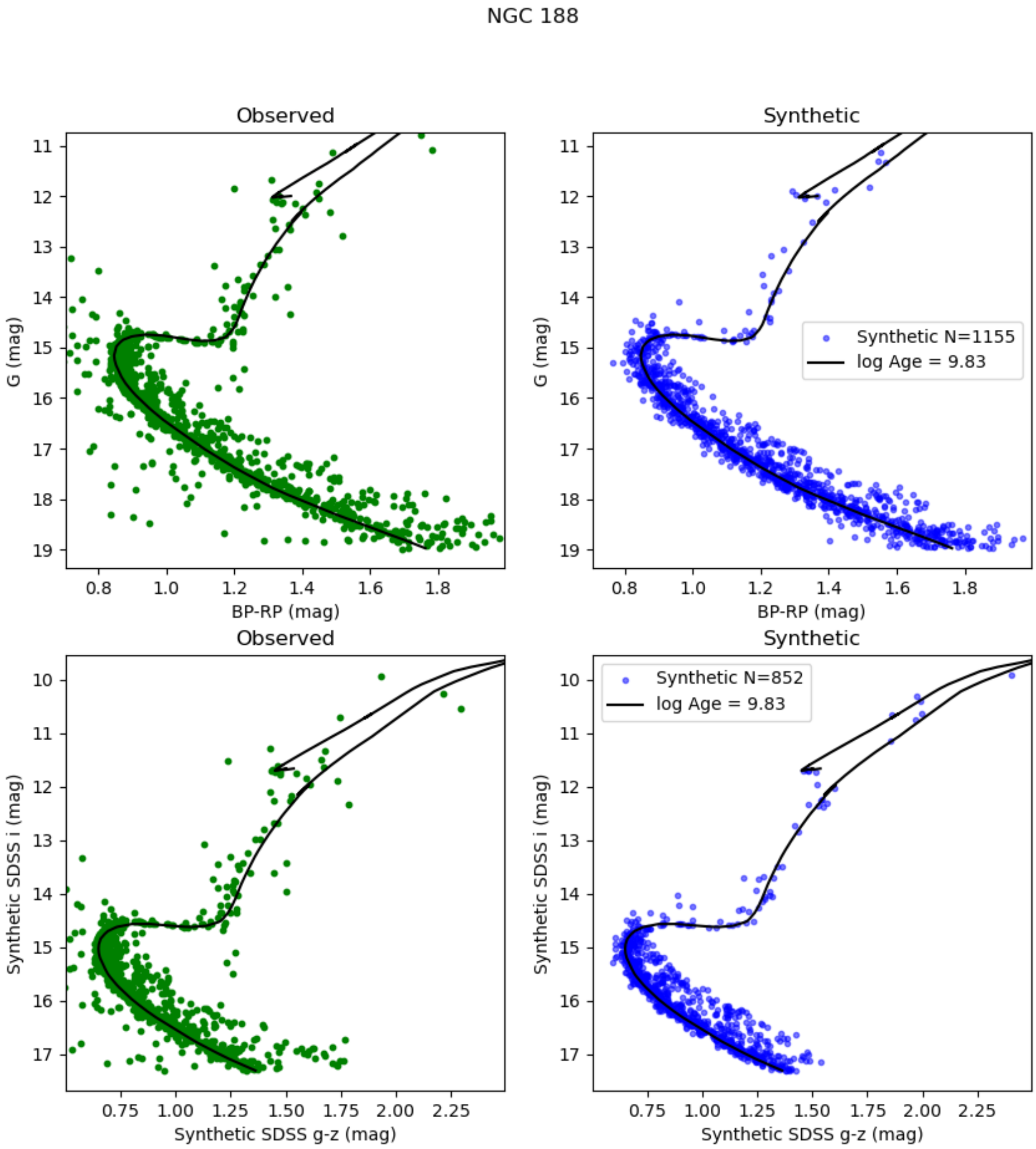


Fig. E.1. Observed OC NGC 188 vs best fit synthetic cluster

Table G.2. First 10 rows of the membership catalog.

source_id	R.A.	Dec.	ϖ	μ_α	μ_δ	G	$BP-RP$	detection rate	cluster
238643333751692000	53.231	44.211	1.215	1.012	-3.053	14.304	0.969	0.91	ASCC 11
238644948659379000	53.270	44.258	1.248	0.742	-3.078	15.787	1.260	0.92	ASCC 11
238645803354866000	53.262	44.293	1.271	0.767	-3.260	11.509	0.325	0.82	ASCC 11
238647800518504000	53.572	44.329	1.195	0.876	-3.140	15.529	1.209	0.81	ASCC 11
238650961613529000	53.601	44.444	1.207	0.854	-3.057	11.539	0.489	1	ASCC 11
238652744022830000	53.399	44.429	1.308	0.792	-3.146	17.545	1.799	0.62	ASCC 11
238655050422388000	53.533	44.512	1.262	0.676	-3.226	15.699	1.316	0.87	ASCC 11
238680888945616000	53.699	44.654	1.149	0.878	-3.105	13.417	0.888	1	ASCC 11
238681709280394000	53.742	44.732	1.111	0.858	-3.131	16.273	1.418	0.98	ASCC 11
238681915438859000	53.751	44.770	1.050	0.863	-3.160	18.237	2.127	0.52	ASCC 11

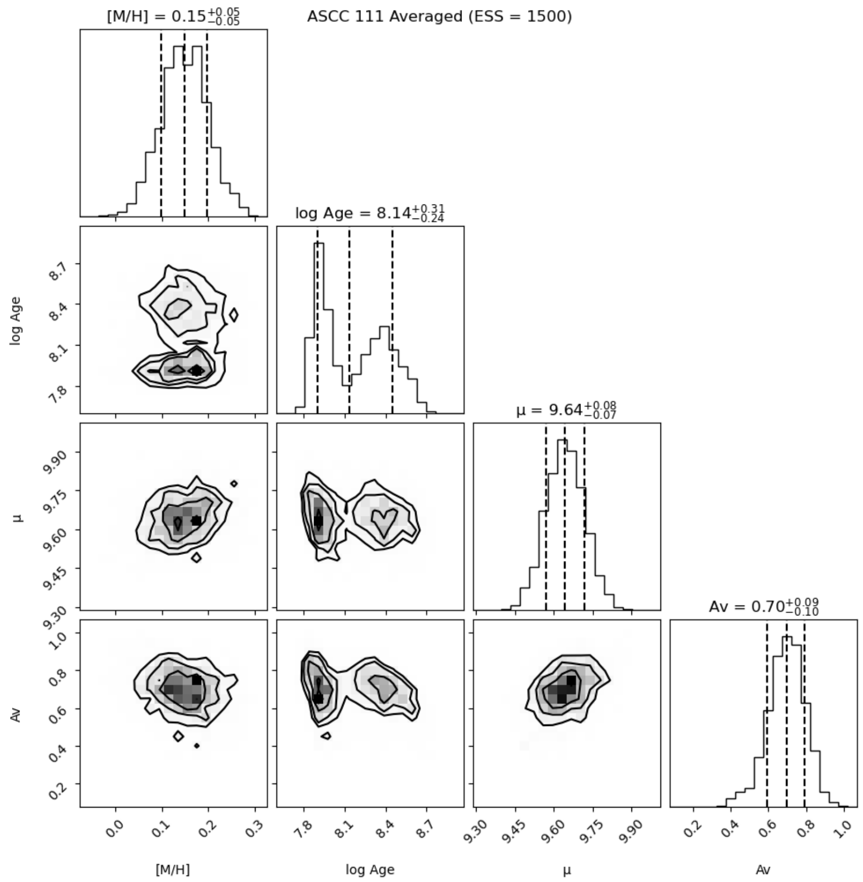


Fig. F.1. Estimation of the posterior by averaging the likelihood 8 times.

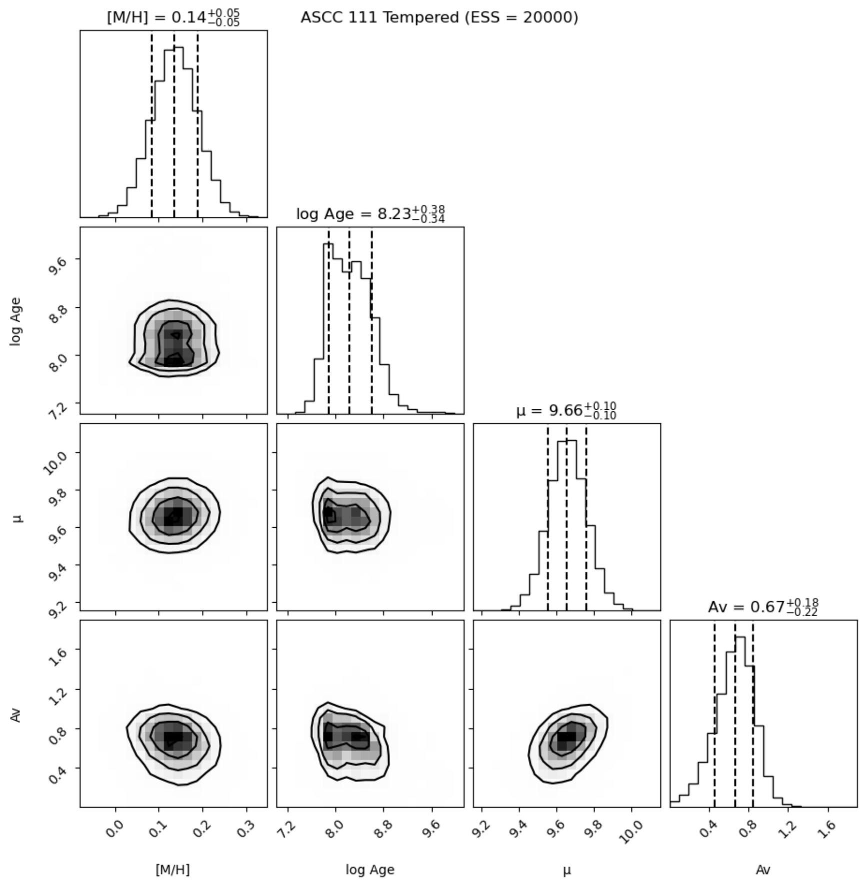


Fig. F.2. The tempered posterior at $T = 5.35$

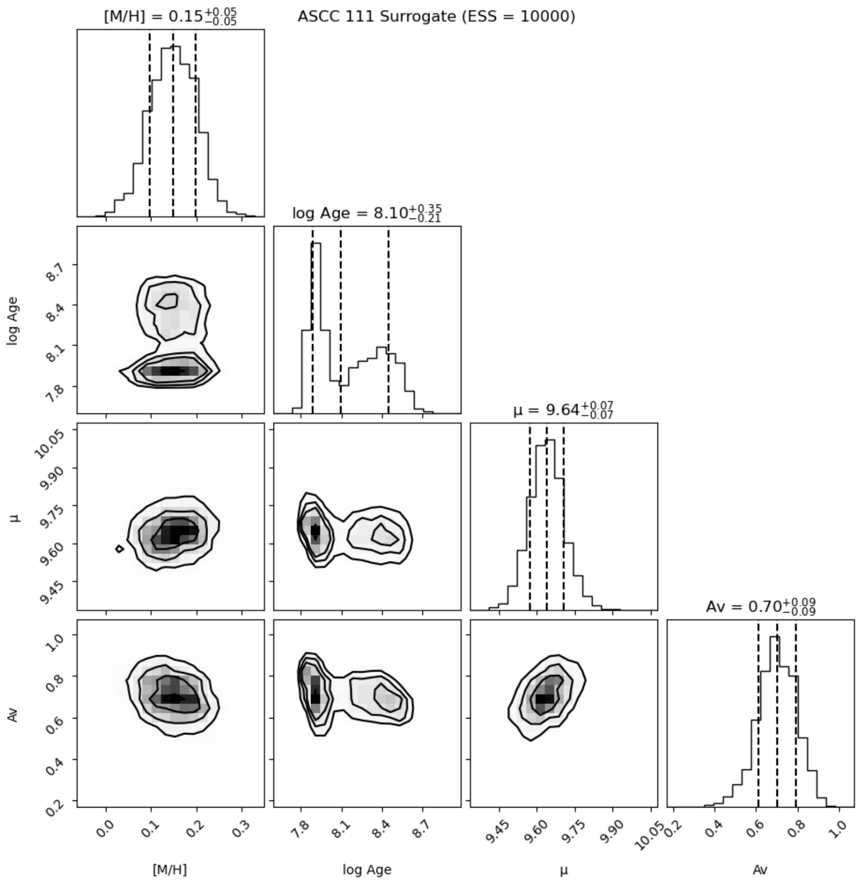


Fig. F.3. Posterior obtained through XBART surrogate likelihood trained on the samples from the tempered posterior shown at Figure F.2

Table G.3. Description of auxiliary information of Table G.1.

flag	range	description
N_{eff}	>= 10000	Effective Sample Size of surrogate
T	>= 1	Tempering factor
N_{train}	> 20000	Effective Sample Size of training
B/V	$[-1, 1]$	Bias to Variance diagnostic
Posterior Mass	$[0.5, 1]$	Relative weight of dominant mode
BINNING	K/B	Knuth or Bayesian Blocks binning
DISCARD_BINS	T/F	Discarded empty bins
INFLATE	T/F	Inflated photometric uncertainties
MAHALANOBIS_CUT	T/F	$MD > 3$ astrometric cut
BIMODAL	T/F	Significant bimodality in $\log Age$