Faculty 1 – Physics and Electrical Engineering

Bachelor thesis in Physics

Membership determination on stellar clusters using HDBSCAN and GMM

Posted by:	Lennart Rathjen
Mat.-Number:	6187492
Submitted on:	7. October 2025

First examiner:	Dr. Marco Scharringhausen [1]
Second examiner:	Prof. Dr. Claus Laemmerzahl [2]

Supervisor:	Research Prof. Dr. Jinyoung Serena Kim [3]

[1] German Aerospace Center (DLR), Bremen, Germany
[2] Center of Applied Space Technology and Microgravity (ZARM), University of Bremen, Germany
[3] Steward Observatory/ Department of Astronomy, University of Arizona, USA

Abstract

Stellar clusters of different ages and properties are essential for understanding stellar evolution and the mechanisms governing the universe. This thesis applies and validates a relatively new methodological approach for identifying cluster members by combining five-parameter astrometry from Gaia DR3 with the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm, further refined through a Gaussian Mixture Model (GMM). The method was tested on the Pleiades, M67, and λ Ori, which are at distinct stellar evolutionary stages. This enables a comprehensive and robust validation yielding 1188, 1092, and 653 cluster members respectively. Mean positions and proper motions derived for the clusters agree well with recent studies, supporting the validity of the identified members. Isochrone fitting to Color-Magnitude Diagrams of the respective cluster members was performed using four different evolutionary grids (MIST, PARSEC, BHAC15, Dartmouth). The clusters yield typical ages of $125 - 143$ Myr for the Pleiades, $3.8 - 5.2$ Gyr for M67, and $3.5 - 6$ Myr for λ Ori, with metallicities near solar and extinction values, consistent with literature. The derived distances are ~ 135 pc for the Pleiades, ~ 866 for M67, and ~ 400 for λ Ori. Systematic uncertainties remain, especially in the isochrone fitting procedure and age determination. Nevertheless, the results demonstrate that HDBSCAN and GMM reliably identify cluster members and provide astrophysical parameters consistent with recent literature, while simultaneously offering new insights into membership determination. These findings establish the method's potential for future studies of stellar clusters and stellar populations.

Contents

1. Introduction

Most stars typically form in clusters within turbulent and clumpy giant molecular clouds (C. J. Lada and E. A. Lada, 2003; Kuhn et al., 2019). As these molecular clouds collapse due to locally induced gravitational fluctuations, stars form and bound together by their mutual gravitational pull, while stellar feedback mechanisms transfer momentum back into the cloud. Simultaneously, residual gas and dust from the parent molecular cloud remain bound, embedding stars during their earliest evolutionary stages and obscuring them in the visible spectrum (Hao et al., 2023).

Stellar feedback mechanisms such as outflows, stellar winds, radiation pressure, and supernovae (Sills et al., 2018), together with tidal forces with surrounding clusters and clouds, however, expel these remnant materials from the clusters vicinity (Kuhn et al., 2019). This expulsion of gas intensifies with age and typically clears the cluster within 2-3 Myrs (Rangwal et al., 2017; C. J. Lada and E. A. Lada, 2003), which makes the cluster optically visible. Since most of the initial cluster mass resides in the remnant material, its expulsion weakens the internal gravitational pull, often leading to dispersal and the formation of unbound associations (Kuhn et al., 2019). Consequently, only about $4 - 7\%$ of embedded clusters reach the bound open cluster stage (Hao et al., 2023). These open clusters typically contain up to a few hundred stars, occupying the same region of space and share a common origin within the same giant molecular cloud (ESA/Hubble, n.d.). This makes them physically related and valuable for astrophysical research, particularly for testing stellar evolutionary models and enhancing our understanding of galactic structure and dynamics.

With its ~ 1.8 billion sources in its Data Release 3 (DR3), the Gaia mission has significantly advanced the study of stellar clusters by providing precise astrometric and photometric measurements, including position, proper motion, parallax, and more (Gaia Collaboration et al., 2023). Among the nearby stellar open clusters, λ Ori (Collinder 69) is especially interesting due to its young age $\sim 3 - 6\,[\text{Myrs}]$ and a distance of $\sim 400\,[\text{pc}]$ (Murdin and Penston, 1977; Cao et al., 2022). Located north of the Orion complex, it has been suggested to be affected by a past supernova event, which cleared away much of its remnant materials (Kounkel et al., 2018), making it accessible to Gaia observations and an excellent case study for investigating early cluster evolution. In addition to λ Ori, this thesis also examines the well-characterized stellar cluster Pleiades (M45) and M67 (NGC 2628). The Pleiades, located in Taurus at a distance of $\sim 135\,\text{pc}$ and an age of $\sim 110\,\text{Myr}$ (Liu et al., 2025; Alfonso and García-Varela, 2023; Gossage et al., 2018), is one

of the most extensively studied young clusters. Its proximity and well-defined stellar population make it a prime benchmark, and its well-defined main sequence makes it ideal for isochrone fitting to membership. In contrast, M67 located within the constellation Cancer, is one of the oldest known stellar clusters, with an age of $\sim 4 - 4.5\,\mathrm{Gyrs}$ and a distance of $\sim 800 - 900\,\mathrm{pc}$ (Reyes et al., 2024; Ghosh et al., 2022). Due to its well-developed main sequence, turn-off point, and red giant branch, it is an excellent benchmark for testing stellar evolution models and complements both the Pleiades and λ Ori. These two benchmark clusters serve as reference cases to assess the robustness of the adopted methodology, while λ Ori serves as a more complex and less constrained case study in early evolutionary phases.

Although many stellar clusters have been studied, membership determination remains challenging, due to significant field star contamination (Santos-Silva and Gregorio-Hetem, 2012) and restricted precise and reliable data. Building on previous work, this thesis introduces a fairly new approach to refine membership selection. To achieve this, filtered astrometric and photometric measurements from Gaia DR3 are used to identify likely members of the Pleiades, M67, and λ Ori, through the combination of clustering algorithms, including Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN, Campello, Moulavi, and Sander (2013) and McInnes, Healy, and Astels (2017)) and the Gaussian Mixture Model (GMM) (described in detail in Section 4). Unlike classical methods, relying primarily on proper motion and parallax cuts, validated through Color-Magnitude Diagrams (CMDs) or radial velocities, this approach makes use of unsupervised clustering. Here, HDBSCAN is especially effective in identifying members in large datasets with varying density and strong field star contamination, while GMM provides probabilistic membership estimates, making the two methods complementary for robust membership determination. To further validate the reliability of this approach, several stellar evolutionary model grids (MIST, PARSEC v2.0, BHAC15, and Dartmouth) are employed to derive key astrophysical parameters such as age, distance, extinction, and metallicity with associated Student-t-distribution uncertainties.

The structure of the thesis is as follows. Section 2 reviews the methodological approaches and data usage adopted by previous studies, which serve as the reference framework for this work. Section 3 outlines the data used in this work and the applied selection criteria. Section 4 describes the operational principles of HDBSCAN and GMM for membership determination, followed by a description of the theoretical isochrone fitting process used to derive key astrophysical parameters. Section 5 presents the results of the analysis, which are interpreted in Section 6. Finally, Section 7 summarizes the main findings and provides an outlook for future research.

1.1 Goal of the Thesis

The key goal of this thesis is to implement the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) in combination with a Gaussian Mixture model (GMM) in a multi-dimensional parameter space (position, proper motion, parallax, radial velocities), to obtain membership of the three evolutionary distinct clusters: the Pleiades, M67, and λ Ori. This thesis therefore attempts to answer whether this approach can successfully and reliably identify the membership of these stellar clusters. In this context, HDBSCAN is used for removing field stars contamination (hard-labeling), while GMM provides probabilistic membership assignment (soft-labeling), yielding a refined list of probable candidate members. To evaluate the reliability of the membership, astrophysical parameters (age, distance, extinction, metallicity) derived through theoretical isochrone fitting are compared against reported values in literature. The three clusters were chosen purposefully for their well-established astrophysical properties and different evolutionary stages, making them ideal benchmarks for validating this method. Hence, this thesis serves as a methodological test for the used unsupervised clustering algorithms, with the broader aim of contributing to future studies of stellar clusters and stellar populations.

Research question: How effective is the combined use of HDBSCAN and GMM in identifying reliable memberships compared to established methods?

2. State of the Art

This thesis compares the memberships and derived astrophysical parameters to Gaia-based catalogs of T. Cantat-Gaudin et al. (2020b) and Emily L. Hunt and Sabine Reffert (2024), (available through VizieR (T. e. a. Cantat-Gaudin, 2020; E. L. Hunt and S. Reffert, 2024)) to validate the identified memberships of the Pleiades, M67, and λ Ori. These catalogs provide comprehensive membership lists for all three clusters, including both astrometric and photometric measurements. This makes them especially suitable for testing and validating the methodology, whereby their comparison is limited to overall membership consistency and agreement in derived astrometric parameters.

In addition, cluster parameters derived in this work through isochrone fitting are compared with literature values. These selected reference studies include a range of methodologies for membership and parameter determination, therefore providing a more robust and unbiased basis for comparison. For this purpose, each study is briefly introduced, outlining the data, methodology, and reported mean cluster parameters such as distance and age. A full overview of the reported values, including position, proper motion (in RA and Dec), distance, age, extinction, and metallicity is summarized in Table 7.1 (Appendix A).

2.1 General Catalogs

Cantat-Gaudin 2020

The catalog presented by T. Cantat-Gaudin et al. (2020b) is a revised and expanded version of the initial catalog by T. Cantat-Gaudin et al. (2020a), based on Gaia DR2 astrometric and photometric data down to $G = 18\,\text{mag}$. Probable members with a membership probability of $p > 0.7$ were assigned using unsupervised photometric membership assignment in stellar clusters (UPMASK). The catalog provides position, proper motion, parallax, Gaia photometry for CMD validations, and includes mean literature estimates of cluster age, distance, and extinction. These parameters were derived by Bossini et al. (2019) through isochrone fitting using PARSEC stellar evolutionary tracks with the BASE-9 algorithm.

For λ Ori, they report 620 candidate members ($p > 0.7$), a distance of $\sim 406\,\text{pc}$ and an age of $\sim 12.6\,\text{Myr}$, while for the Pleiades (Melotte 22), they report 952 candidate members. The derived distance is $\sim 136\,\text{pc}$, and an age of $\sim 78\,\text{Myr}$. M67 (NGC 2682) counts 598 members at a distance of $\sim 881\,\text{pc}$, and an age of $\sim 4.2\,\text{Gyr}$.

Hunt 2024

The catalog presented by Emily L. Hunt and Sabine Reffert (2024) (Census I - III) is a homogeneous all-sky collection of galactic open clusters based on Gaia DR3 astrometric and photometric data down to $G \sim 20\,\text{mag}$. Candidate clusters were identified using the density-based clustering algorithms HDBSCAN, followed by validation with density tests and Bayesian convolution neutral network trained to recognize clusters sequences in Gaia CMD's. The catalog provides refined cluster membership, positions, parallaxes, proper motion, and Gaia photometry, as well as basic astrometric parameters. These were combined with mean literature values of age, distance, and extinction derived by Bossini et al. (2019) and T. Cantat-Gaudin et al. (2020b), yielding 1247 members for λ Ori, a distance of $\sim 400\,\text{pc}$ and an age of $\sim 5.2\,\text{Myr}$. Pleiades counts 1721 members, a distance of $\sim 134\,\text{pc}$, and an age of $\sim 121.6\,\text{Myr}$, while for M67 they report 1844 members. The derived distance is $\sim 837\,\text{pc}$, and an age of $\sim 1.6\,\text{Gyr}$.

2.2 λ Ori

Armstrong 2024

To refine astrometric precision of identified λ Ori members by T. Cantat-Gaudin et al. (2020b), Armstrong and Tan (2024) cross-matched members with Gaia DR3 and filtered for RUWE > 1.4, yielding 563 members. In addition, they crossed-matches cluster members to radial velocities (RV) provided by the Survey of Survey (Tsantaki et al., 2022), removing source with RV inconsistencies, reducing their membership for kinematic analysis to 347. This slightly shifts the mean astrometric values compared to T. Cantat-Gaudin et al. (2020b). Based on the 563 stars, they report a mean distance of $\sim 400\,\text{pc}$, with a derived age via the minimum area traceback method of $\sim 4.1\,\text{Gyr}$.

Cao 2022

The study of Cao et al. (2022) constructed a dataset for λ Ori, collecting optical (Gaia DR2, APASS DR9, Pan-STARRS DR1), near and mid infrared (IR) (2MASS, WISE W1-W4) photometry. Gaia photometries were cut at magnitudes of $> 18.5\,\text{mag}$, while sources with magnitudes of $< 14\,\text{mag}$ were discarded for Pan-STARRS DR1 sources. Candidate members were obtained by literature lists collected in the YSO Corral database and APOGEE Net, further enhanced with a two-component Gaussian Mixture Model and quality cuts, yielding 875 probable members ($p > 0.95$) and 486 possible contaminants. An additional kinematic fit in RA / Dec and proper motion was performed to separate the cluster center from nearby subregions, leaving 357 stars as suitable members of λ Ori. For age estimation, Cao et al. (2022) employed different stellar isochrones (SPOTS, PARSEC, DSEP, Feiden magnetic DSEP, MIST) to account for variations in input and physics such as convection, magnetic activity, and diffusion, yielding an age range between $2 - 4\,\text{Mry}$.

2.3 M67

M67 is a well-studied old open cluster, making it suitable for testing clustering methods, since it hosts various stellar types and masses with stellar population at the turn-off point and older. This turn-off point is important to test age of the cluster.Ghosh et al. (2022) applied HDBSCAN followed by a two-component GMM to Gaia EDR3 data "within [a] 150 arcmin [radius] from the cluster centre " (Ghosh et al. (2022), p.3), yielding 1269 members at a probability threshold of $p \geq 0.6$. Through PARSEC isochrone fitting with the Automated Stellar Cluster Analysis tool (ASteCA), they derived a mean age of $\sim 4.26\,\mathrm{Gyr}$ and a distance of $\sim 870\,\mathrm{pc}$.

In contrast, Reyes et al. (2024) analyzed Gaia DR3 sources "within 2 degrees of the centre of the cluster „(Reyes et al. (2024), p. 2), selecting candidate members by cuts in proper motions, parallax, and radial velocities, identifying 488 members. They refined the sample for MIST isochrone fitting, by excluding the Renormalized Unit Weighted Error (RUWE) > 1.2, yielding 369 stars with an age of $\sim 3.95\,\mathrm{Gyr}$ and a distance of $\sim 836\,\mathrm{pc}$.

2.4 Pleiades

The Pleiades cluster is a rather young cluster with all its stars on the main sequence. Its close proximity makes it an ideal study ground for testing stellar evolutionary models. Liu et al. (2025) implemented HDBSCAN to projection-effect corrected and filtered Gaia DR3 sources within a 70 pc radius of Pleiades center, derived by Emily L. Hunt and Sabine Reffert (2023) (Census II), identifying 1763 members. Through PARSEC v1.2s isochrone fitting, they derived an age of $\sim 123\,\mathrm{Myr}$.

Gossage et al. (2018) used Hipparcos photometric measurements of members identified by Madsen, Dravins, and Lindegren (2002) and Lindegren, Madsen, and Dravins (2000), and 2MASS photometry of members identified by Stauffer et al. (2007), to generate stellar evolution models using MESA, focusing on rotational effects. The models were fitted to the photometric membership in CMDs using MATCH, to derive ages and metallicites, yielding an age of $\sim 110 - 160\,\mathrm{Myr}$.

In contrast, Alfonso and García-Varela (2023) cross matched a classical Bayesian Model on proper motion and DBSCAN to Gaia DR3 sources within 10 degree of Pleiades center, to obtain a refined membership of 958 members. Astrometric parameters were derived using BASE-9 Bayesian software, yielding an age of $\sim 98\,\mathrm{Myr}$ at a distance of $\sim 136\,\mathrm{pc}$.

3. Data

This work makes use of Gaia Data Release 3 (DR3) sources for the Pleiades, M67, and λ Ori, with corresponding search radii of 4°, 1°, and 5° based on the cluster's center. The extracted sources provide position in right ascension (α) and declination (β), proper motions (μ_α, μ_β), parallax (ω), and three-band photometry (G, G_B, G_R). However, G_B and G_R bands are combined to form the color index $G_B - G_R$, which serves as a proxy for stellar temperature and mass, and is widely used in Color-Magnitude Diagrams (CMDs). To ensure data quality and minimize faulty entries, only sources complete in all astrometric and photometric parameters and with non-negative parallax values are selected. A parallax signal-to-noise ratio greater than 5 was required, and the G-band photometric error had to be less than 0.005 mag for the Pleiades and M67, and less than 0.001 mag for λ Ori. The G-band cut was implemented, since low G-band errors generally correlate with reliable parallaxes and proper motions, indicating good overall astrometric quality (Ghosh et al., 2022). Additional photometric quality filters include relative errors of below 0.05 for G-band photometry and the $G_B - G_R$ color index. After applying these criteria, the final samples, hereafter refereed to as *Source Data*, yield 102561 sources for the Pleiades, 21407 for M67, and 315964 for λ Ori. To account for potential contamination from unresolved binaries or other systematics, sources with Gaia Renormalized Unit Weight Error (RUWE) above 1.4 are flagged (Gaia Collaboration et al., 2023). For the isochrone fitting procedure, apparent Gaia G magnitudes were converted to absolute (M_G), using:

$$M_G = G - 5 \cdot \log_{10}\Big(\frac{1000}{\omega}\Big) + 5 \tag{3.1}$$

Figure 3.1 displays unfiltered Gaia DR3 sources against the Source Data for all three clusters in position, proper motion, and corresponding CMD. This visualizes the effects and importance of applying filters prior to analyzing the data.

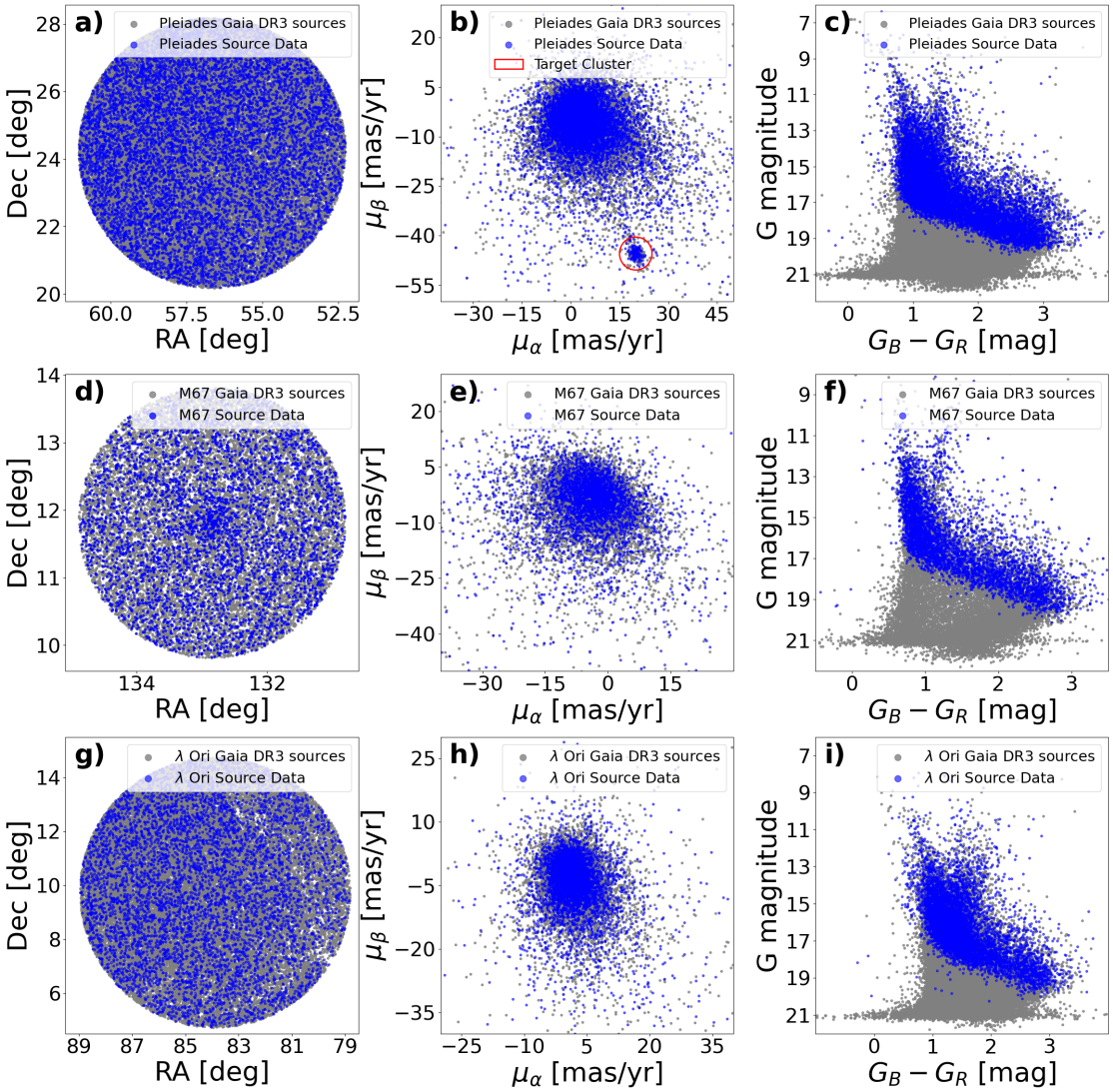

Figure 3.1. Visualization of Gaia DR3 data for the Pleiades (a - c), M67 (d - f), and λ Ori (g - i) in spatial distribution (left), proper motion (middle), and Color-Magnitude Diagram (right). Pre-filtered sources (Source Data) are highlighted in blue. For visual clarity, each point represents approximately 10, 5, and 50 individual sources for the respective clusters.

4. Methodology

This section outlines the methodological approach employed to identify probable members of all three clusters and to derive their astrophysical parameters. As the first step in identifying cluster members, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is implemented. This unsupervised density-based algorithm assigns hard-labels to the data and discards most of the field star contamination. To further refine the membership, a Gaussian Mixture Model (GMM) is applied, assigning probabilistic membership (soft assignment) by fitting Gaussian components to the given data. For deriving astrophysical parameters such as age, distance, extinction, and metallicity, this thesis implements CMD-driven isochrone fitting, using various stellar evolutionary tracks, including PARSEC v2.0 MIST, BHAC+15, and Dartmouth. The respective code used for this thesis is available on `https://github.com/SummerSunny-bit/Employment-of-HDSCBAN-and-GMM`

4.1 HDBSCAN: Density-Based Clustering

HDBSCAN introduced by Campello, Moulavi, and Sander (2013) is a robust unsupervised density-based algorithm, designed to identify clusters of varying densities and shapes by hard-labeling a source point to an overdensity. HDBSCAN improves upon the original DBSCAN algorithm proposed by Ester et al. (1996) to overcome the fixed global threshold of the two hyperparameters to identify overdensities. These are the linking length ϵ within a particular neighborhood, which describes the maximum distance between two sources, and the minimum number of points (minPts.) to define a cluster. Since they are globally fixed thresholds, DBSCAN is less effective in local neighborhoods, with varying overdensities in close proximity, which can lead to fusions of different overdensities(Ester et al., 1996). HDBSCAN overcomes this limitation by replacing the hard global ϵ threshold with a hierarchical density-based approach. Instead of employing a fixed neighborhood size, HDBSCAN builds a minimum spanning tree (MST) of the respective dataset, based on mutual reachability distances, representing the connectivity of data points based on local density estimates. This method builds a hierarchy, from which clusters are then extracted based on their relative persistence using a stability-based selection criterion. This allows for unsupervised identification of outliers as noise, and recovers clusters that vary in shape and density. The general operational principle is described below and illustrated in Figure 7.1 (Appendix B).

4.1.1 Theoretical Foundations

HDBSCAN constructs a hierarchy of clusters based on mutual reachability distances, extracting the most persistent clusters using a formal stability measure. The following definitions are adapted from the foundational framework of Campello, Moulavi, and Sander (2013), which formalizes this approach.

Core Distance For a given element x, the *core distance* $d_{\text{core}}(x)$ is defined as the distance to its k-th nearest neighbor, where $k = \text{minPts}$. If the distance is less or equal to a certain threshold τ, which defines the radius of a local neighborhood, x qualifies as a *coreobject* located in a sufficiently dense area of the dataset. Thus, a small core distance indicates a dense neighborhood, while a large core distance suggest a more sparse region, serving as a proxy for local density.

Mutual Reachability Distance To smooth out density gradients and avoid chaining effects, HDBSCAN defines the *mutual reachability distance* between two elements x and y as

$$d_{\text{mreach}}(x, y) = \max(d_{\text{core}}(x), d_{\text{core}}(y), d(x, y))$$

This replaces the raw Euclidean distance used in DBSCAN and ensures that distances respect local density estimates, leading to better clustering in non-uniform data.

Hierarchical Tree Construction The mutual reachability graph forms a complete weighted graph, where each data point is a vertex and edges weights correspond to the mutual reachability distance. HDBSCAN constructs a *minimum spanning tree* (MST) from this graph, and extends it by adding self-loops to each vertex, weighted by the corresponding core distance. Figure 4.1 illustrates the MST for λ Ori sources identified by HDBSCAN alongside the central λ Orionis star. By progressively removing the heaviest weighted edges in decreasing order, HDBSCAN separates the hierarchy at different density levels, revealing increasingly finer structures. The resulting hierarchy can be visualized as a *condensed cluster tree*, where each branch represents a cluster that exists across a specific range of distance thresholds (McInnes, Healy, and Astels, 2017).

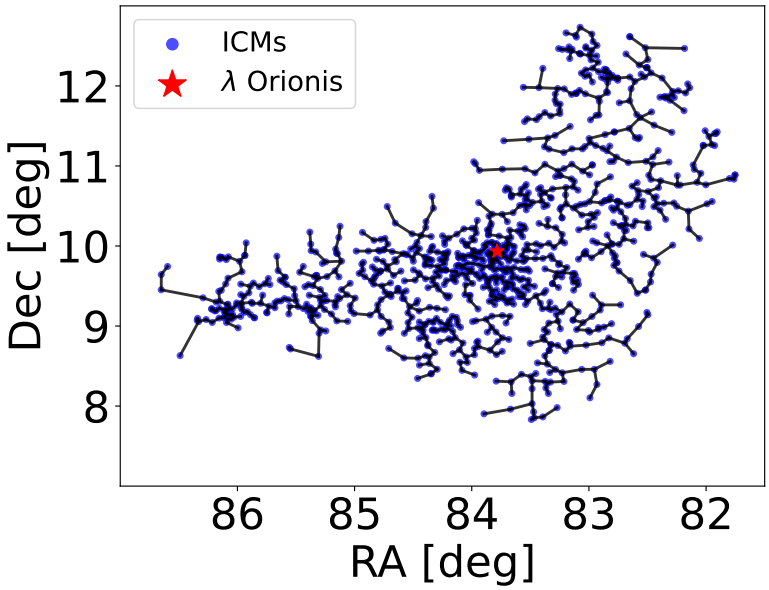

Figure 4.1. Minimum Spanning Tree (MST) of λ Ori constructed from the initial candidate members (ICMs) in positional space, with branches shown in black. The MST is derived as part of the HDBSCAN clustering process and visualizes internal density structures, along the λ Orionis star.

4.1.2 Cluster Selection via Stability

One of HDBSCAN's innovations is the introduction of a *cluster stability* criterion to extract meaningful clusters from the hierarchy. The stability $S(C)$ of a cluster C is defined as

$$S(C) = \sum_{x \in C} (\lambda_{\max}(x) - \lambda_{\min}(x)),$$

where $\lambda = d_{\text{mreach}}^{-1}$ represents the inverse scale of mutual reachability distance, acting as a density-based scale. Clusters that persist over a wide range of λ, are considered more stable, indicating compact and well-separated structures. Points that do not belong to any stable cluster are classified as noise, allowing HDBSCAN to effectively separate cluster members from field stars.

4.1.3 Implementation in framework

In this thesis, HDBSCAN is employed on *Source Data* using the official library (McInnes, Healy, and Astels, 2017), described in McInnes and Healy (2017). It is used as a first-pass filtering step to identify primary cluster candidates and reduce field star contamination. The clustering is performed in a five-dimensional astrometric parameter space of position in right ascension (RA, α) and declination (Dec, β), proper motion in both RA (μ_α) and Dec (μ_δ), and parallax (ω). Prior to clustering, the data is standardized using the z-score to ensure each feature contributes equally to clustering process, since HDBSCAN is sensitive to the scale of input features. The primary hyperparameter *minimum cluster size*, denoted as *min_cluster_size*, determines the smallest group of points considered a cluster. In line with the methodology in Ghosh et al. (2022), this thesis adopts a low value for *min_cluster_size* to avoid missing small substructures in the astrometric space. This choice reflects the initial goal of employing HDBSCAN, not as to finalize membership selection, but to retaining a liberal set of cluster candidate.

Definition of Initial Candidate members

Subsequently to the clustering, the largest identified group in the feature space is assumed to correspond to the target cluster. This cluster is selected and its stars are referred to as *initial candidate members* (ICM). The ICMs count 1089 stars for λ Ori, 1432 for the Pleiades and 1393 for M67. These groups are thought to be slightly larger than the expected number of true members, as the intentions are to exclude obvious field stars, while retaining all potential cluster members for probabilistic refinement. A more restrict HDBSCAN threshold would risk removing true members and degrade the performance of the subsequent Gaussian Mixture Model (GMM), which allows for a refine cluster membership and address the limitations of hard cluster assignment.

4.2 Gaussian Mixture Model

Gaussian Mixture Models (GMMs) are probabilistic-based clustering algorithms, assuming a given population is composed of a finite number of multivariate Gaussian distributions (Fraley and Raftery, 2002; Reynolds, 2009). While algorithms such as HDBSCAN employ hard-labeling approaches, GMMs adopt a soft assignment strategy by assigning probabilities to each point, indicating their likelihood of belonging to each Gaussian component. This probabilistic framework is especially advantageous when dealing with datasets, characterized by overlapping components between clusters and field contamination, as sharp separations are often not well defined. GMMs address this complexity by modeling the intrinsic substructure

within the data and by quantifying membership uncertainty on a per-object basis. In this thesis, GMMs allow for statistical refinement of any ICMs, especially for borderline sources, allowing for a more flexible and statistically robust classification. The general operational principle of GMM is displayed 7.2 (Appendix B) and is outlined subsequently.

4.2.1 Theoretical Foundations

Gaussian Mixture Models assume that the given dataset is generated from a weighted sum of K multivariate normal distributions (Reynolds, 2009; Ghosh et al., 2022). The probability density function for a point x in d-dimensional space is defined as

$$p(x) = \sum_{i=1}^{K} \omega_i \cdot G(x|\mu_k, \Sigma_i),$$

where ω_i is the mixing coefficient for the i-th component, such that $p(x) = \Sigma_i^K \omega_i = 1$. The parameters μ_i and Σ_i are the mean vector and the full covariance matrix of the i-th Gaussian component. $G(x|\mu_i, \Sigma_i)$ is the multivariate Gaussian density function of the form

$$G(x, \mu_i, \Sigma_i) = \frac{\exp\left[-0.5(x-\mu_i)^T \sum_i^{-1}(x-\mu_i)\right]}{(2\pi)^{d/s}\sqrt{|\Sigma_i|}}$$

The parameters ω_i, μ_i and Σ_i are estimates using the expectation-maximization (EM) algorithm. This iterative algorithm alternates between (1) the expectation (E) step, computing the posterior probability (Pr), that the component k generated the data point x_i

$$Pr(i|x_t, \mu_i, \sigma_i) = \frac{\omega_i \cdot G(x_t|\mu_i, \Sigma_i)}{\sum_{j=1}^{K} \omega_j \cdot G(x_t|\mu_j, \Sigma_j)}$$

and (2) the maximization (M) step. The M-step updates the parameters of each Gaussian component (specifically their location (mean), normalization (weight), and shape (covariance)), to maximize the expected log-likelihood given these probabilities. The new updated parameters are then considered the new initial parameters for the next iteration, until it reaches convergence. After convergence, each data point is assigned a membership probability, which is the highest posterior $Pr(i|x_t)$ over all components k.

4.2.2 Implementation in framework

To refine the ICMs identified by HDBSCAN, a Gaussian Mixture Model is employed using the *sklearn.mixture.GaussainMixture* class from the scikit-learn library (Pedregosa et al., 2011). It is performed in a five-dimensional astrometric parameter space consisting of position (α, β), proper motion (μ_α, μ_δ), and parallax (ω). Prior to clustering, the features are standardized using the z-score, ensuring that each feature contributes equally to the fit. The number of Gaussian components, denoted as $n_{components}$, was fixed at two, assuming the data consists of one cluster and one field star population. This constraint offers a more direct interpretation, while maintaining computational efficiency. By increasing the number of components, it potentially could refine the membership, however, as the complexity of evaluating which component accounts for the cluster increases as well, underestimating or overestimating the membership could become a risk. Figure 4.2 shows the two components identified by GMM with their respective mean centers in position. The ellipses indicate the Gaussian fit at $1\,\sigma$, $2\,\sigma$, and $3\,\sigma$ confidence levels. However, this is a simplification of the method, as the GMM performs the clustering in a 5D parameter space, therefore applying ellipsoids onto the data instead of 2D ellipses.

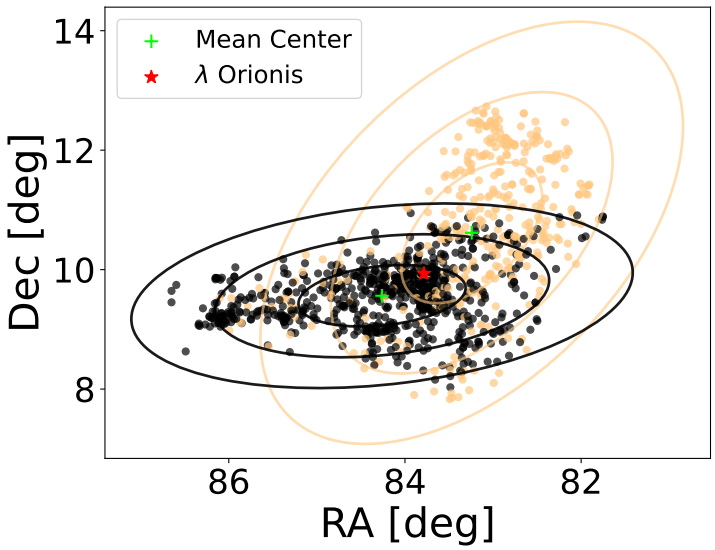

Figure 4.2. Visualization of the two components identified by the Gaussian Mixture Model, along with their respective mean centers and the λ Orionis star in simplified 2D positional space. Actual GMM derived these components in 5D parameters space by fitting different ellipsoids. Ellipses represent the Gaussian fit at $1\,\sigma$, $2\,\sigma$, and $3\,\sigma$ confidence levels. The black component corresponds to the main cluster members.

Probabilistic Membership Assignment

After the GMM assigned membership probabilities based on two components, this thesis establishes certain probability thresholds to obtain a high-confidence membership. These thresholds help to distinguish high-confidence members against likely field stars and contamination. Therefore, a threshold of $p \geq 0.7$ is adopted for λ Ori, following the methodology of T. Cantat-Gaudin et al. (2020b). Even though the Cantat-Gaudin 2020 catalog also includes membership lists ($p > 0.7$) for the Pleiades and M67, this thesis adopts a probability threshold of $p \geq 0.6$, in line with the methodology of Ghosh et al. (2022). This choice ensures consistency with the reference framework for cluster parameter comparison.

While this work focuses on high-probability members for consistency with reference catalogs, it is acknowledged, that lower-probability sources may still contain genuine cluster members. Therefore, borderline cases, likely excluded from the core analysis, could influence completeness and substructure detection, and may justify further investigations in future work. However, this selection balances purity and completeness, ensuring that the core of the cluster is preserved, while retaining some stars that may reside in the tails of the cluster's distribution. The total number of the high-probability sources, denoted as *members* in the subsequent analysis, are summarized in Table 4.1. Additionally, mean cluster parameters in position (α, β), proper motion (μ_α, μ_β), and parallax (ω) were calculated with uncertainties corresponding to $1\,\sigma$ standard deviation.

Table 4.1. Counts of Initial candidate members (ICMs) identified by HDBSCAN and the high-confidence membership assigned by GMM.

	ICMs	Members	Threshold p
Pleiades	1432	1188	≥ 0.6
M67	1393	1092	≥ 0.6
λ Ori	1089	653	≥ 0.7

4.3 Advantages and Limits of HDBSCAN and GMM

The methodological approach of employing HDBSCAN and GMM is well-suited for detecting overdensities in noisy datasets, and to refine a probabilistic membership. However, while HDBSCAN is effective in identifying clusters, as (1) it does not require a predefined number of clusters, (2) while identifying clusters with arbitrary shapes and densities, (3) is robust to noise, and capable of explicitly labeling

outliers, since (4) it adapts to local density variations, it remains sensitive to the choice of the *minimum cluster size* and *minPts.*. Additionally, it may struggle with overlapping populations and its non-probabilistic nature produces hard assignments, making it more sensitive to borderline cases and uncertainty estimates. These limitations however, are partly compensated by GMM, due to its probabilistic nature, performing well in dense regions with overlapping populations, while being less sensitive to borderline cases. Still, it assumes Gaussian distribution in limited shapes, thus it can over- or undershoot membership probability if the cluster's distribution does not follow a Gaussian distribution. Therefore, by combining both methods, the strengths of each are adopted, enabling this hybrid method to determine membership assignment more accurately, especially in complex regions, where the individual methods would not be sufficient. Nevertheless, the combined methodology still shows limitations from both algorithms, such as parameter sensitivity and the underlying requirement for a Gaussian data distribution.

4.4 Isochrone Fitting

To estimate cluster parameters such as age, distance, metallicity ([Fe/H]), and extinction (A_V), I perform theoretical isochrone fitting on CMDs, using various stellar evolutionary models. For all clusters, MIST and PARSEC (v1.2s and v2.0) are employed, while for λ Ori the BHAC15 and Dartmouth models are supplemented via the MADYS package, to evaluate the consistency of age estimates. The latter are implemented, as neither MIST and PARSEC reproduce very young stellar populations sufficiently, whereby BHAC15 and Dartmouth were specifically designed for young clusters. For parameter inference, PARSEC v1.2s isochrones are used via the *ezpadova* package by Fouesneau (2025), while PARSEC v2.0 isochrones by Nguyen et al. (2022) are manually fetched from the CMD 3.8 web form (`https://stev.oapd.inaf.it/cgi-bin/cmd`), as they are not yet supported by official python tools. Even though PARSEC v1.2s is used for parameter inference, the derived parameters are applied to PARSEC v2.0 isochrones, as they include updated parameter treatment, yielding improved results (Nguyen et al., 2022). This section will initially outline the implementation of MIST and PARSEC v2.0 models, combined with an MCMC sampler to automatically derive cluster parameters, after which the implementation of BHAC15 and Dartmouth models will be introduced.

4.4.1 PARSEC and MIST

The MESA Isochrones and Stellar Tracks (MIST) developed by Choi et al. (2016) are employed using the *MIST_Isochrone* class from the *isochrone* package by Morton (2015). It initiates MIST grids based on age, distance, metallicity, and extinction values. The Padova and Trieste Stellar Evolution Code (PARSEC) v1.2s (Bressan et al., 2012; Y. Chen et al., 2014; Tang et al., 2014) is obtained via the *ezpadova.get_isochrone* class. This functions fetches isochrones based on age and metallicity inputs from the official CMD 3.8 input form.

Markov Chain Monte Carlo Sampler

To obtain the best-fitting parameters, this thesis employs a custom Markov Chain Monte Carlo (MCMC) sampler via the *sampler.get_chain* class from the official *emcee* python library (Foreman-Mackey et al., 2013). The likelihood function interpolates the observed Gaia DR3 photometry $(G, G_B - G_R)$ to the synthetic isochrones and is based on a Student's-t-distribution, to account for non-Gaussian residuals and improved robust outlier handling. If necessary, magnitude weights can be applied to compensate mismatches between models and observations in certain CMD regions. The posterior distribution is based on six free parameters, including log(age), metallicity ([Fe/H]), distance, extinction (A_V), a scaling factor for photometric uncertainties, and the degree of freedom (ν) of the t-distribution. This allows the sampler to find the best combination, which reproduces the observed data. Since extinction is treated as a free parameter, the sampler applies extinction corrections to all Gaia bands, when comparing observed magnitudes and colors to the isochrones. The used band-dependent conversions are handled internally by the MIST and PARSEC v1.2S models, accounting for the effects of extinction in the synthetic photometry. In these models, extinction is accounted as an additional term A in the basic magnitude relation (Equation 3.1), but the actual correction is far more complex, since its influence differs across each photometric band and includes bolometric corrections (Jordi et al., 2010; Choi et al., 2016; Yang Chen et al., 2019). The priors for these parameters are set according to literature values or can be validated through extensive sampler tests. The MCMC sampler established two hyperparameters, including the number of walkers (*nwalkers*) and the number of steps per walker (*nsteps*). In this thesis, 32 walkers with several thousand steps were used, balancing computational efficiency while preserving sufficient sample results. Best-fit values are then extracted from the posterior median, with uncertainties derived from the Student's-t-distribution. Diagnostics tools like the Gelman-Rubin statistic (R-hat) and χ^2 are used to check convergence and fit quality.

4.4.2 MADYS

The Manifold Age Determination for Young Stars (MADYS) python tool by V. Squicciarini and M. Bonavita (2022) (Tool: Vito Squicciarini and Mariangela Bonavita (2022)) estimates stellar ages and masses for young stellar objects. In this work, it is used to employ the BHAC15 isochrones by Baraffe et al. (2015) and the Dartmouth isochrones developed by Dotter et al. (2008) and Feiden (2016) for λ Ori. The BHAC15 model assumes a solar mixture of $Y = 0.271$ and $Z = 0.015$, and covering an age range of $0.5 - 10000.0$ [Myrs], while Dartmouth adopts a solar mixture of $Y = 0.2755$ and $Z = 0.0187$, over an age range of $1 - 10000.0$ [Myrs]. Both rely on 2MASS photometry (K, H, and J), which MADYS automatically cross-matches to input Gaia DR3 sources via the *madys.SampleObject* function. Extinction corrections are also applied using the 3D dust map of Lallement et al. (2018). Therefore, MADYS provides everything necessary to overlay synthetic isochrones at different ages and masses onto observed Gaia DR3 photometry. Additionally, the *.get_params* function is used to derive individual age, masses, and extinction values with respective uncertainties, based on a Bayesian posterior sampling.

5. Results

5.1 Comparison: HDBSCAN vs GMM

Figure 5.1 and Figure 5.2 compare the membership classifications obtained from HDBSCAN and GMM, both in astrometric space and photometric space for all three clusters. In Figure 5.1, astrometric space includes equatorial position, proper motion, and parallax for the Pleiades (a - c), M67 (d - f), and for λ Ori (g - i). The blue points represent the initial candidate members (ICMs) identified by HDBSCAN, while the orange points illustrate the final members refined by GMM from the ICMs. This visualization therefore displays the transition from density-based clustering to probabilistic membership refinement, highlighting its notable reduction in membership. The spatial position and proper motion panels for the Pleiades (a, b) and for M67 (d, e) also empathize the presence of a well-defined core structure, while λ Ori (g, h) shows a more loosely defined core structure. For the Pleiades and M67, the parallax panels also include the full Source Data distribution, while for λ Ori the Source Data parallax distribution could not be captured due to too many sources.

Complementary to this, Figure 5.2 displays the same classification in the CMD, illustrating the respective stellar evolutionary sequences for the Pleiades (a), M67 (b), and λ Ori (c). Possibly unresolved binaries or otherwise unreliable astrometric sources, flagged with RUWE > 1.4, were also identified. For the Pleiades and M67, these sources systematically lie slightly above the main sequence or beyond it, while for λ Ori, they appear slightly above the pre-main sequence, even though the separation is less noticeable.

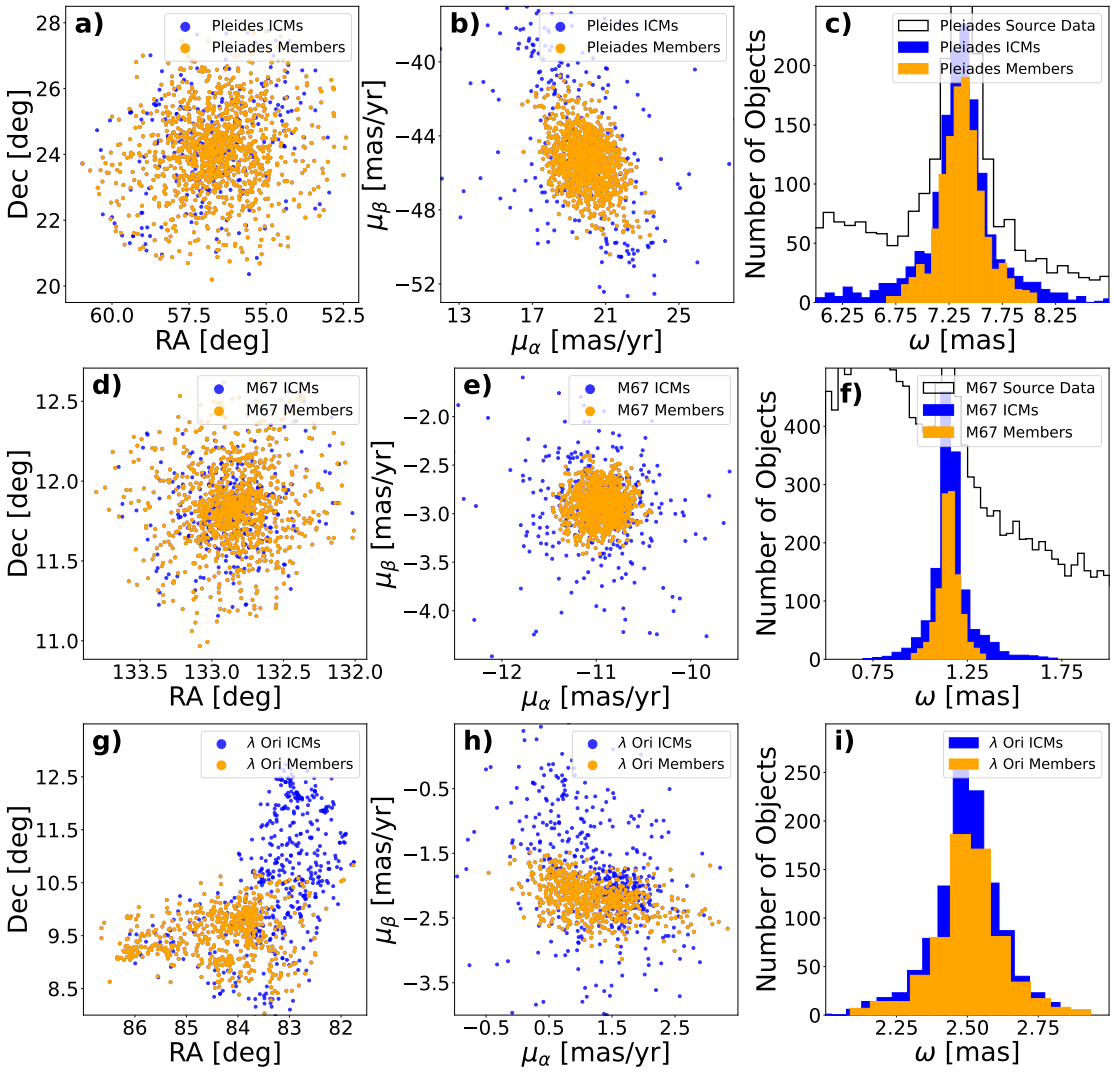

Figure 5.1. Membership classification distribution for the Pleiades (a - c), M67 (d - f), and λ Ori (g - i) shown in spatial coordinates (left), proper motion (middle), and parallax (right). Initial candidate members (ICMs, blue) are identified by HDBSCAN clustering, while high-confidence members assigned by the GMM are shown in orange.

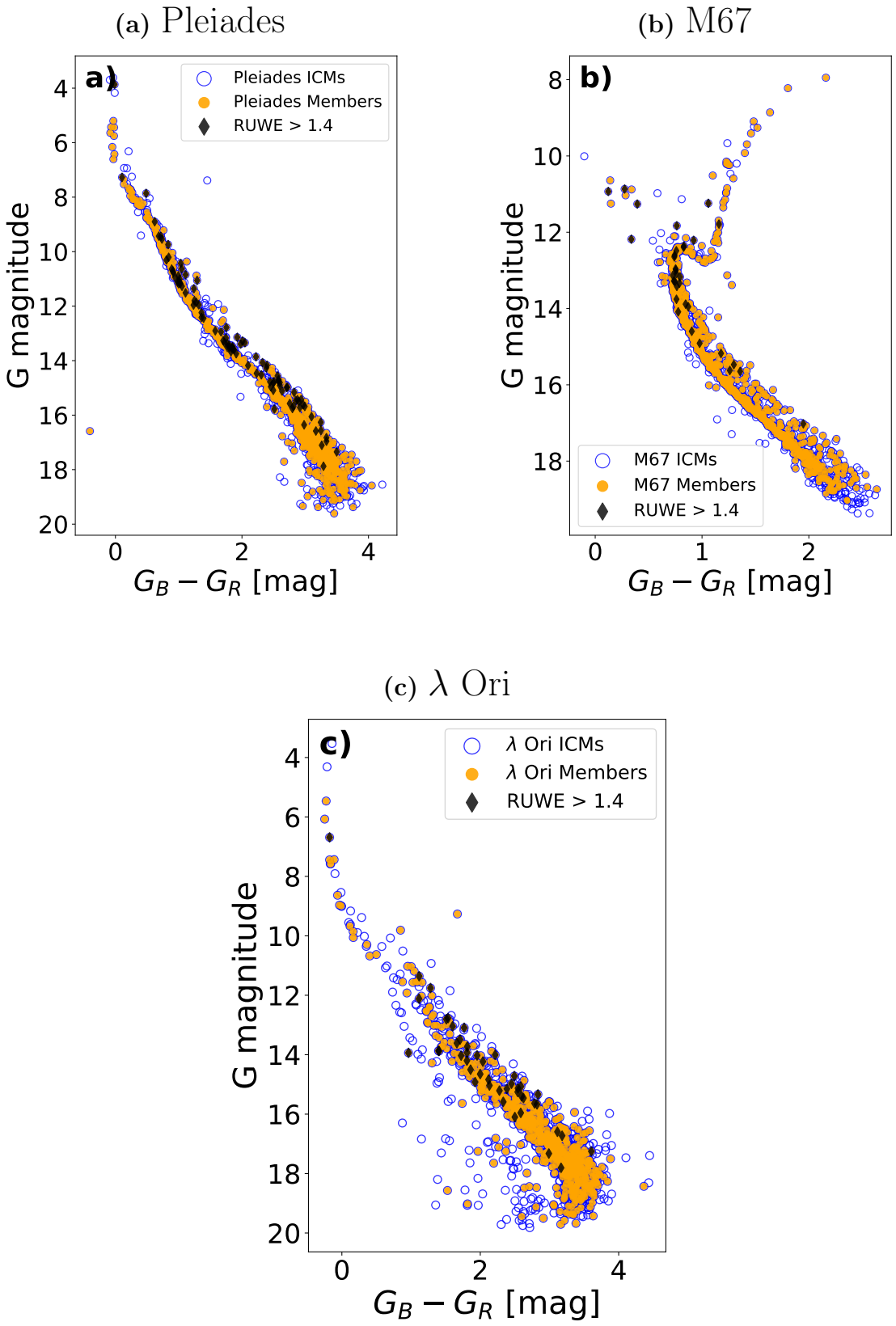

Figure 5.2. Comparison of the derived membership by HDBSCAN (ICMs, blue) and the final members of each respective cluster (orange) obtained by GMM for the Pleiades (a), M67 (b), and λ Ori (c). Sources with RUWE > 1.4 are also highlighted in black.

5.2 Membership comparison with literature catalogs

Table 5.1 summarizes the total number of cluster members reported by each respective study for the Pleiades, M67, and λ Ori. In addition, it lists the cross-matched member counts with this work's memberships and indicates the number of sources with RUWE > 1.4. These sources indicate binary stars or unreliable astrometric measurements and account for only a small fraction of the sample. Still, they are excluded from the subsequent isochrone fitting procedure, since they can bias age and distance estimates by appearing brighter. For all three clusters, the membership identified in this work generally lies between the numbers reported by T. Cantat-Gaudin et al. (2020b) and those by Emily L. Hunt and Sabine Reffert (2024) and Liu et al. (2025), with the latter two showing close agreement. For Emily L. Hunt and Sabine Reffert (2024), the cross-matched members correspond to this work's membership, while for T. Cantat-Gaudin et al. (2020b) the cross-matched members are slightly lower than their reported totals for the Pleiades and M67, and considerably lower for λ Ori.

Table 5.1. Comparison of identified cluster member counts from reference catalogs and this work. For the membership identified in this study, sources with RUWE $>$ 1.4 are also reported, denoted as *Flagged*. Cross-matching was performed for each respective cluster to the reference membership.

	This study		Hunt 2024		Cantat-Gaudin 2020		Liu 2025	
Cluster	Members	Flagged	Total M.	*Matches*	Total M.	*Matches*	Total M.	*Matches*
Pleiades	1188	98	1721	*1188*	952	*889*	1763	*1188*
M67	1092	33	1844	*1089*	598	*550*	—	—
λ Ori	653	45	1247	630	620	395	—	—

Figure 5.3 visualizes these membership trends in proper motion space and within the CMD for the Pleiades (a,b), M67 (c,d), and for λ Ori (e,f). Compared to this work, the catalog of Hunt 2024 identifies significantly more candidate members for all three clusters, including stars with a broader proper motion distribution and fainter sources in the redder regime of the CMD. The same trend is observed for Liu 2025 as well. In contrast, Cantant-Gaudin 2020 identified fewer members showing a more compact core in proper motion space and an exclusion of sources fainter than $G = 18$ (see Sec. 2). Overall, the membership derived in this work represents a compromise between the more inclusive catalog of Emily L. Hunt and Sabine Reffert (2024) (and Liu et al. (2025)) and the more restrictive selection of T. Cantat-Gaudin et al. (2020b).

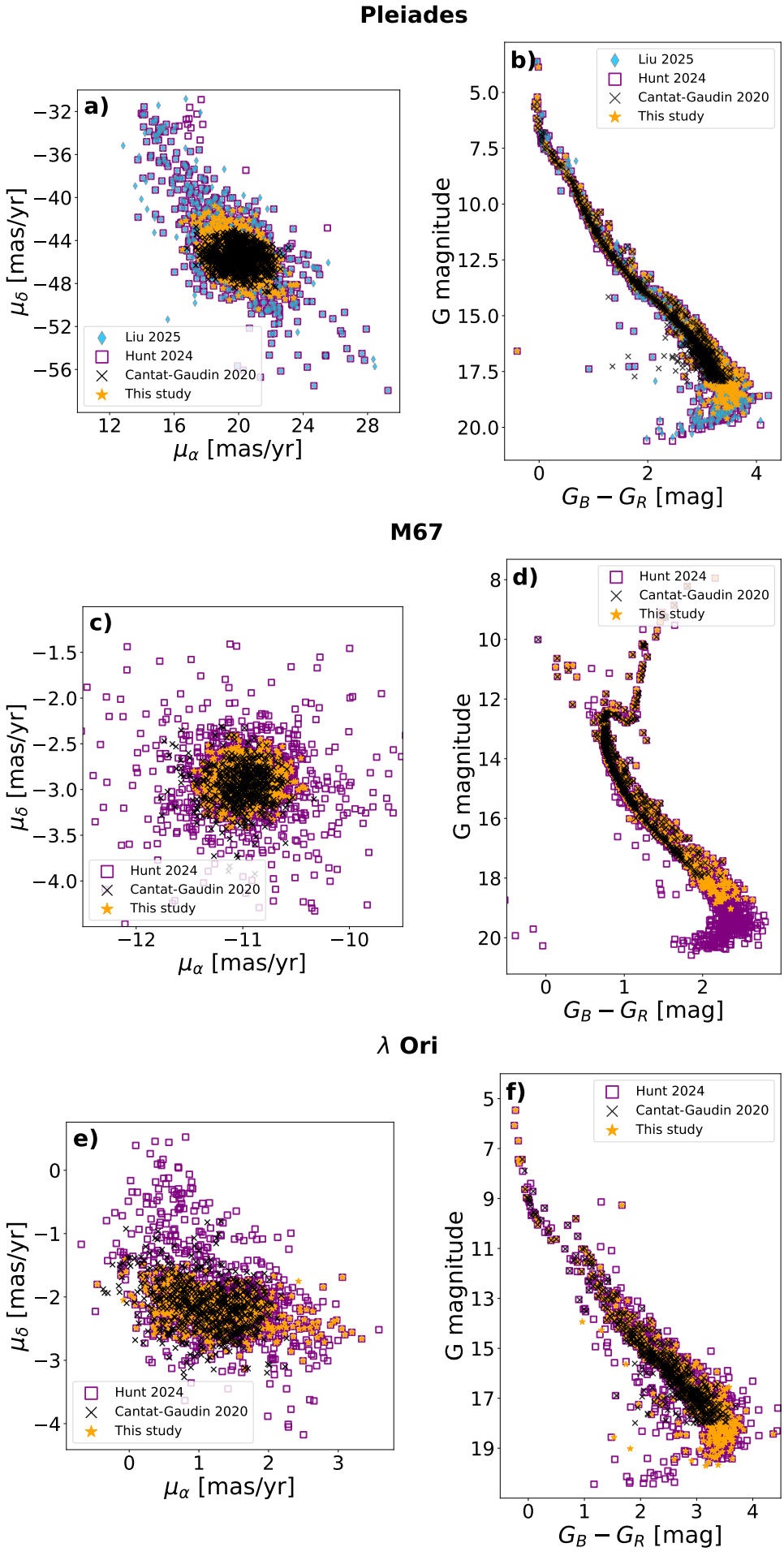

Figure 5.3. Membership distribution for the Pleiades (a,b), M67 (c,d), and λ Ori (e,f) shown in proper motion (left) and CMD (right). Literature membership catalogs for each cluster are derived by Liu et al. (2025) and Emily L. Hunt and Sabine Reffert (2024), and T. Cantat-Gaudin et al. (2020b).

5.3 Mean cluster parameters

The astrometric Gaia DR3 data of the derived memberships for all three clusters were used to calculate the mean equatorial position, proper motion, and parallax, with uncertainties corresponding to a $1\,\sigma$ deviation of these distributions. The results are summarized in Table 5.2.

Table 5.2. Basic mean cluster parameters in equatorial position, proper motion, and parallax derived from high-confidence members identified by GMM from initial cluster members. Reported uncertainties correspond to $1\,\sigma$ deviation.

	Mean Position	Mean Proper Motion [mas/yr]	Mean Parallax [mas]
Pleiades	$\alpha = 03^{\mathrm{h}}46^{\mathrm{m}}26^{\mathrm{s}} \pm 0^{\mathrm{h}}05^{\mathrm{m}}28^{\mathrm{s}}$ $\beta = 24^\circ\,07'\,09'' \pm 1^\circ\,16'\,19''$	$\mu_\alpha = 19.93 \pm 1.11$ $\mu_\beta = -45.41 \pm 1.44$	7.3683 ± 0.2228
M67	$\alpha = 08^{\mathrm{h}}51^{\mathrm{m}}24^{\mathrm{s}} \pm 0^{\mathrm{h}}01^{\mathrm{m}}03^{\mathrm{s}}$ $\beta = 11^\circ\,50'\,04'' \pm 0^\circ\,15'\,15''$	$\mu_\alpha = -10.96 \pm 0.18$ $\mu_\beta = -2.91 \pm 0.17$	1.1518 ± 0.00591
λ Ori	$\alpha = 05^{\mathrm{h}}37^{\mathrm{m}}19^{\mathrm{s}} \pm 0^{\mathrm{h}}03^{\mathrm{m}}41^{\mathrm{s}}$ $\beta = 09^\circ\,32'\,38'' \pm 0^\circ\,27'\,40''$	$\mu_\alpha = 1.20 \pm 0.59$ $\mu_\beta = -2.19 \pm 0.27$	2.5047 ± 0.1247

5.4 Isochrone Fitting

After deriving astrometric parameters of all three clusters using Gaia DR3 data and comparing them to reported membership catalogs, theoretical isochrone fitting is applied to estimate additional astrophysical properties such as stellar age, metallicity ([Fe/H]), extinction in the V-Band (A_V), and distance.

Pleiades and M67

Figure 5.4 shows the best-fitting MIST and PARSEC v2.0 isochrones overlaid on the cluster members of the Pleiades (a) and M67 (b). Both stellar evolutionary models succeed in reproducing the overall evolutionary sequence, including the main-sequence and the turn-off point. In the low-mass, redder regime ($G_B - G_R \gtrsim 1.5$), residuals indicate systematic deviations with theoretical colors leaning towards brighter magnitudes compared to the observations (see Figure 7.3 and Figure 7.4, Appendix C). However, the PARSEC v2.0 model provides a slightly better fit in this regime, while MIST performs slightly better near the turn-off point.

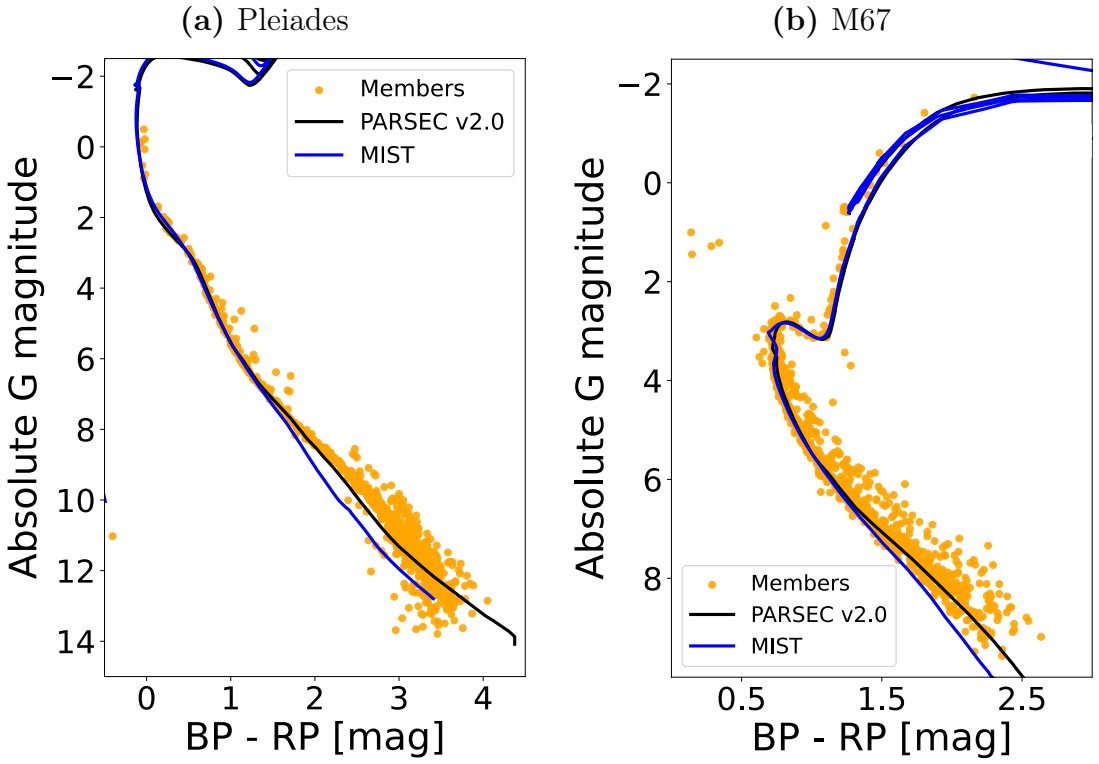

Figure 5.4. Best-fitting theoretical MIST (blue) and PARSEC v2.0 (black) isochrone against cluster members (orange) in a Color-Magnitude Diagram for the Pleiades (a) and M67 (b), obtained by MCMC sampling.

The best-fitting parameters obtained from this isochrone fitting are summarized in Table 5.3 for both clusters, listing the logarithmic age, metallicity ([Fe/H]), extinction (A_V), and distance. The reported uncertainties correspond to approximately 1 σ values derived from the posterior standard deviations under a Student's-t likelihood model. Within the uncertainties, the parameters derived from the MIST and PARSEC v1.2S models are consistent, with only minor differences in the extinction values. The fitted distances also show strong agreement with the mean parallaxes reported in Table 5.2, calculated as $d = \omega^{-1}$.

Table 5.3. Astrophysical parameters for the Pleiades and M67, obtained after isochrone fitting via MIST and PARSEC v1.2S isochrones, combined with a MCMC sampler. Uncertainties reported correspond to approximately $1\,\sigma$, derived from the posterior standard deviations under a Students-t likelihood model.

	Pleiades		M67	
Parameters	MIST	PARSEC v1.2S	MIST	PARSEC v1.2S
log(age)	8.126 ± 0.029	8.111 ± 0.011	9.647 ± 0.070	9.658 ± 0.047
[Fe/H]	0.031 ± 0.020	0.021 ± 0.017	0.038 ± 0.024	0.033 ± 0.025
A_V	0.098 ± 0.032	0.121 ± 0.056	0.090 ± 0.047	0.109 ± 0.073
Distance [pc]	134.7 ± 3.4	135.8 ± 8.2	866.6 ± 6.0	867.1 ± 1.5

Additional corner plots for both the Pleiades and M67 (Figure 7.6 - Figure 7.9, Appendix D) visualize the posterior distributions and covariances of the fitted parameters for both MIST and PARSEC v1.2S MCMC runs. The first diagonal panels show the 1D distributions for each parameter, while the off-diagonal panels present the 2D projections of the combined posterior indication correlations between parameters. The posterior reveals significant deviations and degenerated solutions across all parameter combinations. However, the overall distributions remain well defined and continuous. The dominant and most persistent clump of points reflects the global parameter trends, yielding robust constraints despite local irregularities. Elevated uncertainties for some parameters, especially the metallicity and extinction, are reflected by the impact of these degeneracies on the fitting procedure.

λ Ori

Figure 5.5 shows the best-fitting MIST and PARSEC v2.0 isochrones overlaid on λ Ori cluster members within the CMD. Both stellar evolutionary models effectively reproduce the evolutionary sequence for stars with $G_B - G_R \gtrsim 2$, which represents the majority of identified members. In this regime, PARSEC v2.0 performs slightly better than the MIST model. However, for bluer stars ($G_B - G_R \lesssim 2$), both models increasingly deviate from the observed sequence. This mismatch is quantified by the residuals in Figure 7.5 (Appendix C), where PARSEC v2.0 yields slightly smaller deviations than MIST, although both models struggle significantly to reproduce the observed sequence with high accuracy.

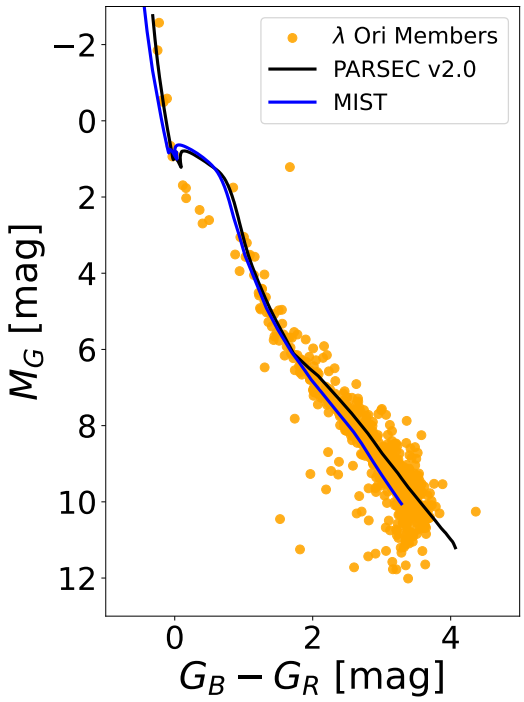

Figure 5.5. Best-fitting theoretical MIST (blue) and PARSEC v2.0 (black) isochrones against cluster members (orange) in a Color-Magnitude Diagram for λ Ori, obtained by MCMC sampling.

Figure 5.6 compares various BHAC15 (a) and magnetic Dartmouth (b) isochrones over a range of ages (0.5 - 10 Myr) overlaid on λ Ori members in the CMD. The BHAC15 model indicates that the majority of cluster members are located between the $1 - 5$ Myr isochrones, whereas the Dartmouth model suggests a broader age spread, with most members located between $1 - 10$ Myr. However, a few outliers are observed beyond these ranges. The mean cluster age is estimated to be approximately ~ 4.7 Myr for BHAC15 and ~ 9 Myr using the Dartmouth model.

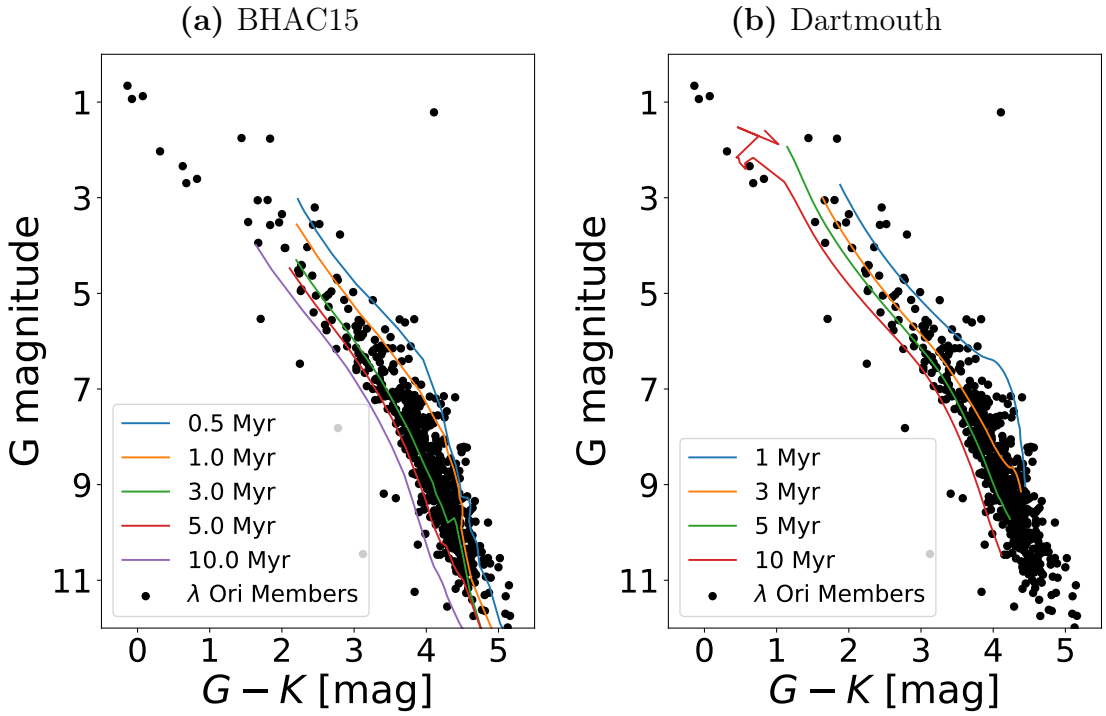

Figure 5.6. Best-fitting BHAC15 (a) and magnetic Dartmouth (b) isochrones with various against λ Ori cluster members within a CMD.

Figure 5.7 shows the λ Ori members projected onto the STILISM 3D extinction map by Lallement et al. (2018). Here, most members exhibit extinction values between $\sim 0.180 - 0.233$ [mag], while some outliers show slightly higher extinction values. The lower bound, however, corresponds to the resolution limit of the STILISM map and reflects the minimum possible extinction.

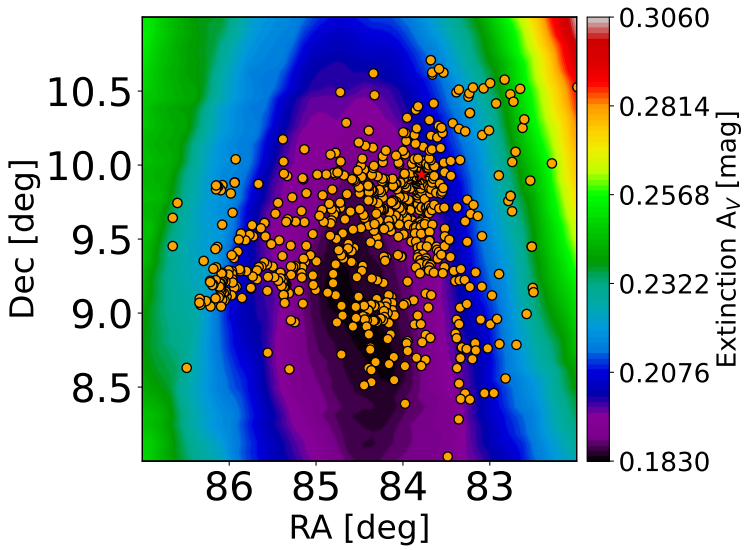

Figure 5.7. λ Ori cluster Members overlayed on top of the STILISM 3D extinction map by Lallement et al. (2018). Alongside the λ Orionis star in red.

The corresponding best-fitting parameters from all four models are summarized in Table 5.4, which lists logarithmic age, distance, extinction, and metallicity. Within the uncertainties, both MIST and PARSEC v2.0 yield consistent values for the age, distance, and metallicity. In contrast, extinction estimates vary significantly, ranging from ~ 0.063 [mag] to ~ 0.26 [mag]. These differences are further visualized in the corner plots (Figure 7.10 and Figure 7.11, Appendix D), showing significant deviations and degeneracies. These are especially significant in metallicity for both models and in log(age) for the PARSEC v1.2S model. As a result, these parameters exhibit elevated uncertainties. The distances derived from isochrone fitting are in strong agreement with the derived mean parallaxes reported in Table 5.2.

Table 5.4. Astrophysical parameters for λ Ori, obtained after theoretical isochrone fitting via MIST and PARSEC v1.2S, combined with a MCMC sampler

Parameters	MIST	PARSEC v1.2S	BHAC15	Dartmouth
log(age)	6.609 ± 0.051	6.588 ± 0.046	$6.647^{+0.141}_{-0.084}$	$6.955^{+0.147}_{-0.111}$
[Fe/H]	0.008 ± 0.028	0.025 ± 0.032	—	—
A_V [mag]	0.083 ± 0.040	0.263 ± 0.066	~ 0.065	~ 0.064
Distance [pc]	402.6 ± 3.8	400.4 ± 8.3	—	—

6. Discussion

Membership comparison

The membership counts derived in this work show good overall agreement with those reported by previous studies (T. Cantat-Gaudin et al., 2020b; Emily L. Hunt and Sabine Reffert, 2024; Liu et al., 2025), although clear differences remain (see Table 5.1). These variations are primarily caused by methodological choices and sample limits. The study of Emily L. Hunt and Sabine Reffert (2024) employed a neutral network based clustering with various diagnostic tools and included fainter magnitudes down to $G \sim 20\,\mathrm{mag}$. This recovered a large sample size including faint, low-mass members and extended halo populations. In contrast, T. Cantat-Gaudin et al. (2020b) employed UPMASK to Gaia DR3 sources down to $G \sim 18\,\mathrm{mag}$ and a more conservative $p > 0.7$ cut, thus reporting smaller and core-dominated members. In this study, astrometric and photometric quality cuts were applied prior to HDBSCAN clustering and GMM refinement, allowing a conservative membership, which retains some halo stars, yielding results that are located between the extended census of Emily L. Hunt and Sabine Reffert (2024) and the compact core of T. Cantat-Gaudin et al. (2020b). This compromise is particularly evident in proper motion space and within the CMD distributions (Figure 5.3), where my membership lies between the extended halo populations of Emily L. Hunt and Sabine Reffert (2024) and Liu et al. (2025) and the compact core of T. Cantat-Gaudin et al. (2020b). It also visualizes the influence of the different G magnitude cuts, explaining the large difference in membership lists. Especially notable is the large discrepancies for λ Ori, where the combined cross-matches between my list and T. Cantat-Gaudin et al. (2020b) is significantly lower than the individual counts would suggest. This mostly reflects the youth and dynamical complexity of λ Ori, making the separation of true members from field contamination more challenging. This complexity is further enhanced by the distinct and expanding proper motion pattern of the region (Figure 5.1, g and h). Several studies include additional sources that are likely true cluster members, but which are not recovered in this work (Kounkel, 2020, Figure 1), (Armstrong and Tan, 2024, Figure 8). This is most likely due to λ Ori's dispersed kinematics, so that some members deviate from the dominant cluster motion. This reflects the young, dynamically evolving nature of λ ori, where expansion and substructure lead to broader proper motion distributions. Conversely, the Pleiades and M67 exhibit smoother cross-matching and more robust proper motion patterns, most likely since they are more evolved, dynamically separated against field stars, and show well defined cores. Overall, the comparison indicates that my methodology is robust and effectively balances completeness with reliability.

Mean cluster parameters

The mean astrometric parameters derived for the Pleiades, M67, and λ Ori (Table 5.2) agree well with those reported in previous studies (Table 7.1, Appendix A). Minor discrepancies in position and parallax are within the reported uncertainties and primarily reflect the differences in methodology choice and source selection such as magnitude limits and various astrometric quality cuts. Therefore, the slightly elevated proper motion values compared to Emily L. Hunt and Sabine Reffert (2024) may be the result of the inclusion of halo populations, since these stars introduce higher-velocity outliers and broaden the overall distribution. However, T. Cantat-Gaudin et al. (2020b) reports values slightly closer, likely due to their more restrictive membership selection. In addition, λ Ori shows larger differences in derived proper motion values, likely caused by λ Ori's young and dynamic nature. In contrast, the Pleiades and M67 show close consistency in position and proper motions, consistent with their more dynamically relaxed evolutionary stages. Overall, the close match in position and proper motion across all three clusters demonstrates that HDBSCAN and GMM reliably identified the dominant kinematic and spatial populations, while minimizing field contamination. This highlights the robustness of the derived membership lists.

Isochrone Fitting

Isochrone fitting using a MCMC approach with a Student's-t distribution likelihood, yielded the best-fit cluster parameters, including logarithmic age, extinction A_V, metallicity ([Fe/H]), and distance for all three clusters.

The Pleiades and M67

For the Pleiades and M67, MIST and PARSEC v2.0 both reproduce the evolutionary sequence in the region of $G_B - G_R < 1.5$, capturing the main sequence and turn off point quite well (Figure 5.4). However, they struggle to reproduce the evolutionary sequence in the faint, low mass regime ($G_B - G_R > 1.5$). G magnitude residuals (Figure 7.3 and Figure 7.4) show that PARSEC v2.0 performs slightly better in the faint, low-mass, and redder regime, while MIST captures the observed photometry marginally better at the turn-off. This reflects their differences in the treatment of low-mass stellar physics (Y. Chen et al., 2014; Nguyen et al., 2022; Choi et al., 2016), since both grids were calibrated primarily for intermediate- and high-mass stars while low-mass stellar physics for PARSEC v2.0 was improved.

The best-fit parameters obtained by this process (Table 5.3) indicate and age of $\sim 133\,\mathrm{Myr}$, a distance of $\sim 135\,\mathrm{pc}$, extinction of $\sim 0.1\,\mathrm{mag}$, and a metallicity slightly above solar for the Pleiades across both models. For M67, the process

yields an age of $\sim 4.4\,\mathrm{Gyr}$, a distance of $\sim 866\,\mathrm{pc}$, extinction of $\sim 0.09\,\mathrm{mag}$ and a near-solar metallicity. These results agree broadly with literature (Table 7.1, Appendix A), although the metallicity values are slightly above solar, while literature reports solar metallicity values. In addition, distances derived for M67 are slightly larger than literature suggests ($\sim 837\,\mathrm{pc}$). These offsets are likely due to the absence to Gaia parallax zero-point corrections in my analysis, which account for $\sim 15 - 30\,\mathrm{pc}$ (Emily L. Hunt and Sabine Reffert, 2024; T. Cantat-Gaudin et al., 2020b; Reyes et al., 2024). However, mean Gaia parallaxes remain consistent across all studies, indicating agreement.

Parameter uncertainties for both clusters remain non-negligible. While ages and distances are well constrained ($< 1\%$) across both clusters and models, extinction values have uncertainties of $\sim 30 - 46\%$ for the Pleiades and $\sim 50 - 67\%$ for M67. Metallicity is also strongly degenerate, with uncertainties of $60 - 80\%$ for the Pleiades and $\sim 60 - 75\%$ for M67. These issues are visible in the corner plots (Figure 7.6 - Figure 7.9), showing strong degeneracies and bad covariances in the posterior distributions, especially between metallicity and extinction. Poor MCMC convergence likely exacerbate these uncertainties, since chain lengths were insufficient to fully explore the parameter space. This resulted in poorly developed modes, artificial halos around the contours, and local maxima. In addition, fixed global extinction laws and even small errors in photometry can make it difficult to determine extinction, whereby the determination of metallicity is limited by the models and the lack of spectroscopy.

λ Ori

For λ Ori, both MIST and PARSEC v2.0 models reproduce the lower pre-main sequence (PMS) ($G_B - G_R > 1.5$) quite well, while failing to capture the sequence at bluer colors $G_B - G_R \lesssim 1$, and $M_G \lesssim 3$. This is expected, as neither model is optimized for PMS stars, thus neglecting magnetic effects and accretion history (Feiden, 2016; Sills et al., 2018). The best-fit parameters from these models yielded distances of $\sim 400\,\mathrm{pc}$, ages of $\sim 3.5 - 4.1\,\mathrm{Myr}$, near-solar metallicity, and extinction estimates ranging from ~ 0.08 mag (MIST) to ~ 0.26 mag (PARSEC). The results for age, distance, and metallicity values agree with those in the literature, while only the extinction derived by PARSEC is consistent with literature values (Table 7.1, Appendix A) and the STILISM extinction map (Figure 5.7). However, to determine age more precisely, I additionally employed PMS-optimized isochrones, including BHAC15 and magnetic Dartmouth (5.6). These reproduce the observed photometry significantly better, yielding mean ages of $\sim 4.4\,\mathrm{Myr}$ (BHAC15) and $\sim 9\,\mathrm{Myr}$ (Dartmouth). The discrepant likely arises from model limitations rather than astrophysical differences, since the Dartmouth PMS tracks do not extend uniformly across the full range of observed members,

which artificially extends older isochrones across the CMD, and by accounting for magnetic activity and convection. However, both models suggest age spreads of $\sim 1 - 10\,\mathrm{Myr}$, expected by ongoing star formation, and firm with literature values (Cao et al., 2022). Uncertainties for λ Ori are especially large when compared to the Pleiades and M67, with metallicity deviations of $\sim 130 - 350\%$ and extinction showing deviations of $\sim 15 - 48\%$, reflecting the aforementioned methodological limitations and the astrophysical complexity regarding PMS stars. These issues are clearly visible in the corner plots (Figure 7.10 and Figure 7.11), where strong degeneracies emerge in the posterior distributions and are supported by the G magnitude residuals (Figure 7.5 and Figure 7.5, Appendix C), showing mismatches to observed photometry, especially dominant for MIST. As discussed for the Pleiades and M67, these are caused by poor MCMC convergence and insufficient chain lengths. Extinction estimates by BHAC15 and Dartmouth show systematic discrepancy, since they were not corrected for the distance modulus. Despite these challenges, the consistency across all models support that λ Ori is a young, dynamically developed cluster with ongoing star formation, whereas the Pleiades and M67 exhibit well-defined main sequences and turn-off points consistent with their more advanced evolutionary stages. Overall, the general agreement with literature values for all clusters emphasizes the sturdiness of the identified member by HDBSCAN and GMM. Thus, the method is validated across three distinct regimes and therefore demonstrates robustness and flexibility.

7. Conclusion

This thesis aimed to identify reliable stellar membership for the Pleiades, M67, and λ Ori stellar clusters using unsupervised clustering methods, particularly HDBSCAN and GMM. By employing these methods to three clusters at distinct evolutionary stages, the study tested whether they can match or even surpass traditional approaches such as parallax cuts. For this purpose, the thesis reviews the methodology applied in previous studies used as reference framework, followed by the employment of HDBSCAN and GMM for each cluster. The main part then compares the resulting memberships to those reported in literature, and derives basic cluster parameters. The effectiveness of the method was finally evaluated using these parameters, as well as additional derived properties obtained through isochrone fitting. The analysis shows that HDBSCAN and GMM identified memberships are consistent with published catalogs, while balancing between completeness and reliability. HDBSCAN generally produced a more extended selection, including halo populations, which were then refined by GMM to yield robust core members. This demonstrates that HDBSCAN is effective in excluding contamination in large datasets, while the probabilistic refinement provided by GMM further enhances membership reliability. The derived mean positions, proper motions and parallaxes from Gaia DR3 agree well with literature values, further validating the methodology. However, for λ Ori, the comparison with previous studies revealed that some additional genuine members were identified, which were not captured by HDBSCAN and GMM in this work. This likely highlights the cluster's expanding kinematic structure and broader proper motion distributions (Kounkel, 2020; Armstrong and Tan, 2024), which future work could adapt or extend these methods to better recover disperse memberships in young, dynamically evolving clusters.

Isochrone fitting using MIST and PARSEC v1.2S model grids provided astrophysical parameters broadly consistent with previous studies. For the Pleiades, the method yields an age of $\sim 133\,\mathrm{Myr}$, a distance of $\sim 135\,\mathrm{pc}$, $A_V \approx 0.1\,\mathrm{mag}$ and slightly sub solar metallicity, within the uncertainties across both models with only slight differences. For M67, the results report an age of $\sim 4.4\,\mathrm{Gyr}$, a distance of $\sim 866\,\mathrm{pc}$, $A_V \approx 0.09\,\mathrm{mag}$, and a near-solar metallicity. However, the distance for M67 slightly differs relative to literature values, likely due to the lack of Gaia zero-point corrections, but remain firm with uncertainties. For λ Ori, MIST and PARSEC v1.2S reproduce the lower pre-main sequence ($G_B - G_R > 1.5$), but fail at bluer colors, as anticipated given the lack of PMS physics. These models suggest

ages of $\sim 3.5 - 4.1$ Myr, distances near ~ 400 pc, and near-solar metallicity, while PARSEC v1.2S suggest an extinction of ~ 0.26 mag consistent with literature and the STILISM dust map, and MIST underestimating the extinction significantly. PMS-optimized models such as BHAC15 and magnetic Dartmouth, however, provide more realistic age estimates, suggesting an age spread of $1 - 10$ Myr with mean ages of 4.4 Myr and ~ 9 Myr, respectively. This is consistent with literature values and with ongoing star formation. The study thus emphasizes both the strengths and limitations of the approach. Membership identification with HDBSCAN and GMM is reliable, flexible, and robust across three clusters with very different properties. However, isochrone fitting remains challenging due to parameter degeneracies caused by poor MCMC convergence and the struggle to reproduce faint, low-mass stars and PMS populations for MIST and PARSEC, which exacerbates uncertainties.

Future work could improve on these challenges by systematically optimizing the choice of hyperparameters for HDBSCAN and GMM (minimum cluster size, Number of Gaussian components, etc.), employing longer and more sturdy MCMC sampling and implementing bolometric, Gaia zero-point, and further systematic corrections. In addition, expanding the dataset to include radial velocities and multi-band photometry at different wavelengths (2MASS, APOGEE, etc.) would allow for a more detailed and robust analysis, while extending model grids would refine parameter determination. Applying the method to a wider range of clusters would enable a more comprehensive validation of its direct applicability.

In summary, this thesis demonstrates that HDBSCAN combined with GMM provides a reliable and effective alternative to classical methods for stellar cluster membership determination, thereby answering the research question. The derived astrophysical parameters are broadly consistent with literature and the approach confirms reliability and robustness across three distinct evolutionary stares, thus offering a promising methodology for future cluster studies.

Acknowledgments

I am very thankful to Dr. Scharringhausen, my first examiner, for their support and feedback throughout this project, and to Prof. Dr. Lämmerzahl for serving as my second examiner. I am also deeply grateful to Research Prof./ Astronomer Dr. Kim from the Steward Observatory, University of Arizona, serving as my primary supervisor and for their guidance and support, without which this work would not have been possible. Furthermore, I acknowledge all the authors and contributors of publicly available software packages and tools, whose work greatly facilitated this project. I also thank all anonymous test readers who carefully reviewed this work for accuracy. This work used data from the European Space Agency (ESA) Gaia mission (Gaia Collaboration et al., 2016), processed by the Gaia Data Processing and Analysis Consortium. Especially the third data release was used in this study (Gaia Collaboration et al., 2023).

Bibliography

Alfonso, Jeison and Alejandro García-Varela (2023). "A Gaia astrometric view of the open clusters Pleiades, Praesepe, and Blanco 1". In: *Astronomy & Astrophysics* 677, A163. ISSN: 1432-0746. DOI: 10.1051/0004-6361/202346569. URL: http://dx.doi.org/10.1051/0004-6361/202346569.

Armstrong, Joseph J. and Jonathan C. Tan (2024). "Expansion kinematics of young clusters: I. Lambda Ori". In: *Astronomy & Astrophysics* 692, A166. DOI: 10.1051/0004-6361/202451538.

Baraffe, I. et al. (Apr. 2015). "New evolutionary models for pre-main sequence and main sequence low-mass stars down to the hydrogen-burning limit". In: *Astronomy & Astrophysics* 577, A42. DOI: 10.1051/0004-6361/201425481.

Bossini, D. et al. (2019). "Age determination for 269 Gaia DR2 open clusters". In: *Astronomy & Astrophysics* 623, A108. DOI: 10.1051/0004-6361/201834693.

Bressan, A. et al. (2012). "PARSEC: stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code". In: *Monthly Notices of the Royal Astronomical Society* 427.1, pp. 127–145. DOI: 10.1111/j.1365-2966.2012.21948.x.

Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander (2013). "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining, PAKDD 2013*. Ed. by Jian Pei et al. Vol. Part II, Lecture Notes in Computer Science 7819. Springer, pp. 160–172. DOI: 10.1007/978-3-642-37456-2_14.

Cantat-Gaudin, T. et al. (2020a). "A detailed view of the open cluster population in the Galactic disc with Gaia DR2". In: *Astronomy & Astrophysics* 633, A99. DOI: 10.1051/0004-6361/201936691.

Cantat-Gaudin, T. et al. (2020b). "Painting a portrait of the Galactic disc with its stellar clusters". In: *Astronomy & Astrophysics* 640, A1. DOI: 10.1051/0004-6361/202038192. URL: %5Curl%7Bhttps://vizier.cds.unistra.fr/viz-bin/VizieR?-source=J/A+A/640/A1%7D.

Cantat-Gaudin, T. et al. (2020). *Gaia DR2 open clusters in the Milky Way (Cantat-Gaudin, 2018)*. Accessed: 2025-07-07. VizieR Online Data Catalog. URL: https://vizier.cds.unistra.fr/viz-bin/VizieR?-source=J/A+A/640/A1.

Cao, Lyra et al. (2022). "Age Spreads and Systematics in λ Orionis with Gaia DR2 and the SPOTS Tracks". In: *The Astrophysical Journal* 924, p. 84. DOI: 10.3847/1538-4357/ac307f.

Chen, Y. et al. (Nov. 2014). "Improving PARSEC models for very low mass stars". In: *Monthly Notices of the Royal Astronomical Society* 444.3, pp. 2525–2543. DOI: 10.1093/mnras/stu1605.

Chen, Yang et al. (Dec. 2019). "YBC: a stellar bolometric corrections database with variable extinction coefficients: Application to PARSEC isochrones". In: *Astronomy & Astrophysics* 632, A105. DOI: 10.1051/0004-6361/201936612. URL: http://dx.doi.org/10.1051/0004-6361/201936612.

Choi, Jieun et al. (2016). "MESA Isochrones and Stellar Tracks (MIST). I. Solar-scaled Models". In: *The Astrophysical Journal* 823.2, p. 102. DOI: 10.3847/0004-637X/823/2/102. URL: https://doi.org/10.3847/0004-637X/823/2/102.

Dotter, A. et al. (Sept. 2008). "The Dartmouth Stellar Evolution Database". In: *The Astrophysical Journal Supplement Series* 178.1, pp. 89–101. DOI: 10.1086/589654.

ESA/Hubble (n.d.). *Open cluster*. https://esahubble.org/wordbank/open-cluster/. ESA/Hubble Word Bank (accessed 25 August 2025).

Ester, Martin et al. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. Portland, OR: AAAI Press, pp. 226–231.

Feiden, Gregory A. (Sept. 2016). "Magnetic inhibition of convection and the fundamental properties of low-mass stars: III. A consistent 10 Myr age for the Upper Scorpius OB association". In: *Astronomy & Astrophysics* 593, A99. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201527613. URL: http://dx.doi.org/10.1051/0004-6361/201527613.

Foreman-Mackey, Daniel et al. (2013). "emcee: The MCMC Hammer". In: *Publications of the Astronomical Society of the Pacific* 125.925, pp. 306–312. DOI: 10.1086/670067. arXiv: 1202.3665 [astro-ph.IM].

Fouesneau, Morgan (2025). *ezpadova*. https://github.com/mfouesneau/ezpadova. Version 2.0.4. Released on April 8, 2025.

Fraley, Chris and Adrian E. Raftery (2002). "Model-based clustering, discriminant analysis, and density estimation". In: *Journal of the American Statistical Association* 97.458, pp. 611–631. DOI: 10.1198/016214502760047131.

Gaia Collaboration et al. (2016). "The Gaia mission". In: *Astronomy & Astrophysics* 595, A1. DOI: 10.1051/0004-6361/201629272.

Gaia Collaboration et al. (2023). "Gaia Data Release 3: Summary of the contents and survey properties". In: *Astronomy & Astrophysics* 674, A1. DOI: 10.1051/0004-6361/202243940.

Ghosh, Esan Mouli et al. (2022). "Membership and age determination of M67 open cluster using GAIA EDR3 data". In: *Journal of Physics: Conference Series* 2214.1, p. 012009. DOI: 10.1088/1742-6596/2214/1/012009.

Gossage, Seth et al. (2018). "Age Determinations of the Hyades, Praesepe, and Pleiades via MESA Models with Rotation". In: *The Astrophysical Journal* 863.1, p. 67. ISSN: 1538-4357. DOI: 10.3847/1538-4357/aad0a0. URL: http://dx.doi.org/10.3847/1538-4357/aad0a0.

Hao, C. J. et al. (2023). "Unveiling the initial conditions of open star cluster formation". In: *Research in Astronomy and Astrophysics* 23.7, p. 075023. DOI: 10.1088/1674-4527/acd58d.

Hunt, E. L. and S. Reffert (2024). *Member stars of open clusters (Hunt+, 2024)*. Accessed: 2025-07-07. VizieR Online Data Catalog.

Hunt, Emily L. and Sabine Reffert (2023). "Improving the open cluster census. II. An all-sky cluster catalogue with Gaia DR3". In: *Astronomy & Astrophysics* 673, A114. DOI: 10.1051/0004-6361/202346285.

– (2024). "Improving the open cluster census: III. Using cluster masses, radii, and dynamics to create a cleaned open cluster catalogue". In: *Astronomy &*

Astrophysics 686, A42. DOI: 10.1051/0004-6361/202348662. URL: https://doi.org/10.1051/0004-6361/202348662.

Jordi, C. et al. (Nov. 2010). "Gaia broad band photometry". In: *Astronomy & Astrophysics* 523, A48. DOI: 10.1051/0004-6361/201015441. arXiv: 1008.0815 [astro-ph.IM]. URL: http://dx.doi.org/10.1051/0004-6361/201015441.

Kounkel, Marina (2020). "Supernovae in Orion: The Missing Link in the Star-forming History of the Region". In: *The Astrophysical Journal* 902, p. 122. DOI: 10.3847/1538-4357/abb6e8.

Kounkel, Marina et al. (2018). "The APOGEE-2 Survey of the Orion Star-forming Complex. II. Six-dimensional Structure". In: *The Astronomical Journal* 156, p. 84. DOI: 10.3847/1538-3881/aad1f1.

Kuhn, Michael A. et al. (2019). "Kinematics in Young Star Clusters and Associations with Gaia DR2". In: *The Astrophysical Journal* 870, p. 32. DOI: 10.3847/1538-4357/aaef8c.

Lada, Charles J. and Elizabeth A. Lada (2003). "Embedded Clusters in Molecular Clouds". In: *Annual Review of Astronomy and Astrophysics* 41, pp. 57–115. DOI: 10.1146/annurev.astro.41.011802.094844.

Lallement, R. et al. (Aug. 2018). "Three-dimensional maps of interstellar dust in the Local Arm: using Gaia, 2MASS, and APOGEE-DR14". In: *Astronomy & Astrophysics* 616, A132. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201832832. URL: http://dx.doi.org/10.1051/0004-6361/201832832.

Lindegren, Lennart, Søren Madsen, and Dainis Dravins (2000). "Astrometric radial velocities. II. Maximum-likelihood estimation of radial velocities in moving clusters". In: *Astronomy & Astrophysics* 356, pp. 1119–1135. DOI: 10.1051/0004-6361:20000136.

Liu, Penghui et al. (2025). "Revisiting Open Clusters within 200 pc in the Solar Neighbourhood with Gaia DR3". In: *The Astronomical Journal*. DOI: 10.3847/1538-3881/adceda. arXiv: 2504.08179 [astro-ph.SR].

Madsen, Søren, Dainis Dravins, and Lennart Lindegren (2002). "Astrometric radial velocities. III. Hipparcos measurements of nearby star clusters and associations".

In: *Astronomy & Astrophysics* 381, pp. 446–463. DOI: 10.1051/0004-6361:20011458.

McInnes, Leland and John Healy (2017). "Accelerated Hierarchical Density Based Clustering". In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, pp. 33–42. DOI: 10.1109/ICDMW.2017.12.

McInnes, Leland, John Healy, and Steve Astels (2017). "hdbscan: Hierarchical density based clustering". In: *Journal of Open Source Software* 2.11, p. 205. DOI: 10.21105/joss.00205.

Morton, Timothy D. (Mar. 2015). *Isochrones: Stellar model grid package*. Astrophysics Source Code Library, record ascl:1503.010. Astrophysics Source Code Library, ascl:1503.010. URL: http://ascl.net/1503.010.

Murdin, Paul and M. V. Penston (1977). "The λ Orionis association". In: *Monthly Notices of the Royal Astronomical Society* 181, pp. 657–665.

Nguyen, C. T. et al. (2022). "PARSEC V2.0: Stellar tracks and isochrones of low- and intermediate-mass stars with rotation". In: *Astronomy & Astrophysics* 665, A126. DOI: 10.1051/0004-6361/202244166.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Rangwal, Geeta et al. (2017). "Interstellar extinction in twenty open star clusters". In: *Publications of the Astronomical Society of Australia* 34, e068. DOI: 10.1017/pasa.2017.64.

Reyes, Claudia et al. (2024). "Isochrone Fitting to the Open Cluster M67 in the Era of Gaia and Improved Model Physics". In: *arXiv preprint* arXiv:2407.03526. arXiv: 2407.03526 [astro-ph.SR]. URL: https://arxiv.org/abs/2407.03526.

Reynolds, Douglas A. (2009). "Gaussian Mixture Models". In: *Encyclopedia of Biometrics*. Ed. by Stan Z. Li and Anil K. Jain. Boston, MA: Springer, pp. 659–663. DOI: 10.1007/978-0-387-73003-5-196.

Santos-Silva, T. and J. Gregorio-Hetem (2012). "Characterisation of young stellar clusters". In: *Astronomy & Astrophysics* 547, A107. DOI: 10.1051/0004-6361/201219695.

Sarajedini, A., A. Dotter, and A. Kirkpatrick (2009). "Deep 2MASS Photometry of M67 and Calibration of the Main Sequence J-K_s Color Difference as an Age Indicator". In: *Astrophysical Journal* 698, pp. 1872–1884. DOI: 10.1088/0004-637X/698/2/1872.

Sills, Alison et al. (2018). *Dynamical evolution of stars and gas of young embedded stellar sub-clusters*. arXiv e-print. arXiv:1803.04301 [astro-ph.GA].

Squicciarini, V. and M. Bonavita (2022). "MADYS: the Manifold Age Determination for Young Stars. I. Isochronal age estimates and model comparison". In: *Astronomy & Astrophysics* 666, A15. DOI: 10.1051/0004-6361/202244193. arXiv: 2206.02446 [astro-ph.SR].

Squicciarini, Vito and Mariangela Bonavita (2022). *MADYS: Manifold Age Determination for Young Stars*. Version 0.3.2. ASCL:2206.018. DOI: 10.5281/zenodo.6807358. URL: https://github.com/vsquicciarini/madys.

Stauffer, John R. et al. (2007). "Near- and Mid-Infrared Photometry of the Pleiades and a New List of Substellar Candidate Members". In: *The Astrophysical Journal Supplement Series* 172.2, pp. 663–681. DOI: 10.1086/518961.

Tang, J. et al. (Dec. 2014). "New PARSEC evolutionary tracks of massive stars up to 350 $M_\odot$ at metallicities $0.0001 \leq Z \leq 0.04$". In: *Monthly Notices of the Royal Astronomical Society* 445.1, pp. 4287–4305. DOI: 10.1093/mnras/stu2029.

Tsantaki, M. et al. (2022). "Survey of Surveys. I. The largest compilation of radial velocities for the Galaxy". In: *Astronomy & Astrophysics* 659, A95. DOI: 10.1051/0004-6361/202141702. arXiv: 2110.09316.

Appendix

Appendix A: Reported reference cluster parameters

Table 7.1. Collection of cluster parameters from literature in comparison to the finding of this study. Parameters include the number of stars, position and proper motion in RA and Dec, log(age), metallicity [Fe/H], extinction, and distance.

	# of Members	RA [deg]	Dec [deg]	μ_α [mas/yr]	μ_β [mas/yr]	log(age) [Myr]	[Fe/H] [dex]	A_v [mag]	ω [pc]
					Pleiades				
(MIST)	1188	56.61 ± 1.37	24.12 ± 1.27	19.93 ± 1.11	−45.41 ± 1.44	8.126 ± 0.029	0.031 ± 0.020	0.098 ± 0.032	134.7 ± 3.4
(PARSEC)						8.111, ± 0.011	0.021 ± 0.017	0.121 ± 0.056	135.8 ± 8.2
(1)	952	56.60	24.12	20.08 ± 1.06	−45.50 ± 1.18	7.89	—	0.18	182
(2)	1721	56.68	24.11	19.96 ± 1.00	−45.46 ± 1.14	$8.08^{+0.21}_{-0.34}$	—	$0.102^{+0.064}_{-0.102}$	134.84 ± 0.02
(2)	1763	56.68	24.11	19.96	-45.46	$8.09^{+0.01}_{-0.01}$	−0.01 ± 0.05	—	136.4 ± 4.6
(4)	1026 - 1041	—	—	$19.95^{+0.04}_{-0.04}$	$-45.41^{+0.05}_{-0.04}$	7.99	∼ 0.02	—	135.74 ± 0.10
(5)	—	—	—	—	—	8.04 − 8.20	8.04 − 8.20	0.1054	∼134
					M67				
(MIST)	1092	132.85 ± 0.26	11.84 ± 0.25	−10.96 ± 0.18	−2.91 ± 0.17	9.647 ± 0.070	0.038 ± 0.024	0.090 ± 0.047	866.6 ± 6.0
(PARSEC)						9.658 ± 0.047	0.033 ± 0.025	0.109 ± 0.073	867.1 ± 1.5
(1)	598	132.85	11.81	−10.99 ± 0.19	−2.96 ± 0.20	9.63	—	0.07	889
(2)	1844	132.85	11.82	−10.97 ± 0.19	−2.91 ± 0.19	$9.23^{+0.19}_{-0.17}$	—	$0.102^{+0.055}_{-0.095}$	$837.3^{+0.06}_{-0.07}$
(6)	488	132.85	11.81	−11.0	−2.9	9.60	0.08 ± 0.03	0.127 ± 0.012	837 ± 19
(7)	1269	132.85 ± 0.39	11.82 ± 0.38	−10.96 ± 0.20	−2.90 ± 0.19	9.630 ± 0.033	∼ 0.11	0.112 ± 0.025	869.79 ± 4.51
(8)	—	—	—	—	—	9.60	−0.009 ± 0.009	0.127 ± 0.012	—
					λ Ori				
(MIST)	653	84.33 ± 0.92	9.54 ± 0.46	1.20 ± 0.59	−2.19 ± 0.27	6.609 ± 0.051	0.008 ± 0.028	0.083 ± 0.040	402.6 ± 3.8
(PARSEC)						6.588 ± 0.046	0.025 ± 0.032	0.263 ± 0.066	400.4 ± 8.3
(BHAC15)						$6.647^{+0.141}_{-0.084}$	—	∼ 0.065	—
(Dartmouth)						$6.955^{+0.147}_{-0.111}$	—	∼ 0.064	—
(1)	620	83.79	9.81	1.19 ± 0.55	−2.12 ± 0.39	7.1	—	0.25	416
(2)	1247	83.81	9.89	1.22 ± 0.53	−2.05 ± 0.46	$6.72^{+0.13}_{-0.19}$	—	$0.416^{+0.14}_{-0.22}$	$393.8^{+0.2}_{-0.2}$
(9)	563	83.79 ± 0.04	9.82 ± 0.03	1.14 ± 0.02	−2.08 ± 0.02	6.61 ± 0.01	—	—	—
		SPOTS f = 0	SPOTS f = 0.34	SPOTS f = 0.5	MIST	BHAC15	DESEP	Feiden DSEP	PARSEC
(10)	357	5.20 - 6.40	6.40 - 6.61	6.52 - 6.72	6.23 - 6.40	5.20 - 6.41	5.20 - 6.40	6.49 - 6.67	6.34 - 6.52

Notes: Values taken from previous studies are as reported in the respective publications. "This Study" refers to parameters derived using Gaia DR3 data (1-5) and cluster parameters (6 - 9) derived via various stellar evolutionary models. The respective values are marked by the model used, eg. (MIST) refers to This Study via the MIST model. Note, that source (10) only reports log(age) derived by various models.

References: (1) T. Cantat-Gaudin et al. (2020b), (2) Emily L. Hunt and Sabine Reffert (2024), (3) Liu et al. (2025), (4) Alfonso and García-Varela (2023), (5) Gossage et al. (2018), (6) Reyes et al. (2024), (7) Ghosh et al. (2022), (8) Sarajedini, Dotter, and Kirkpatrick (2009), (9) Armstrong and Tan (2024), (10) Cao et al. (2022).

Appendix B: Operational principles of HDBSCAN and GMM

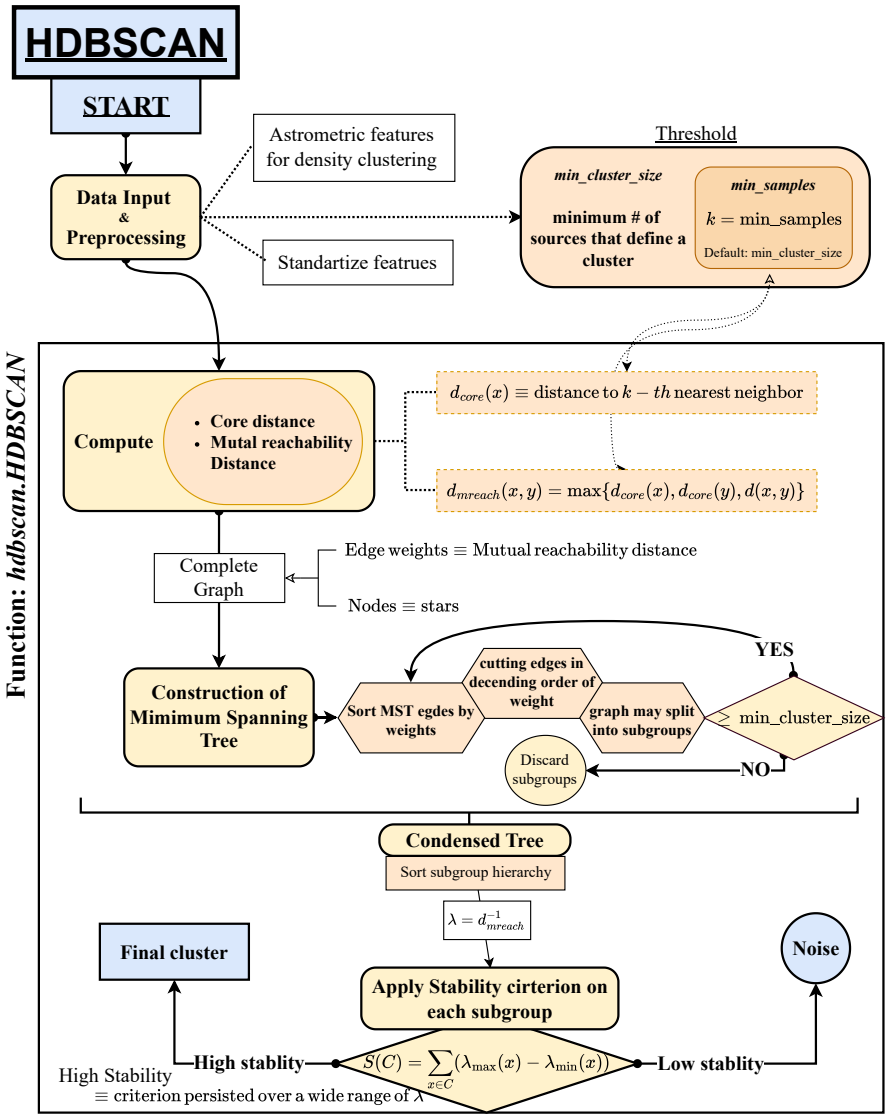

Figure 7.1. Schematic flowchart of the general operational principle of the Hierarchical Density-Based Spatial Clustering Application with Noise.

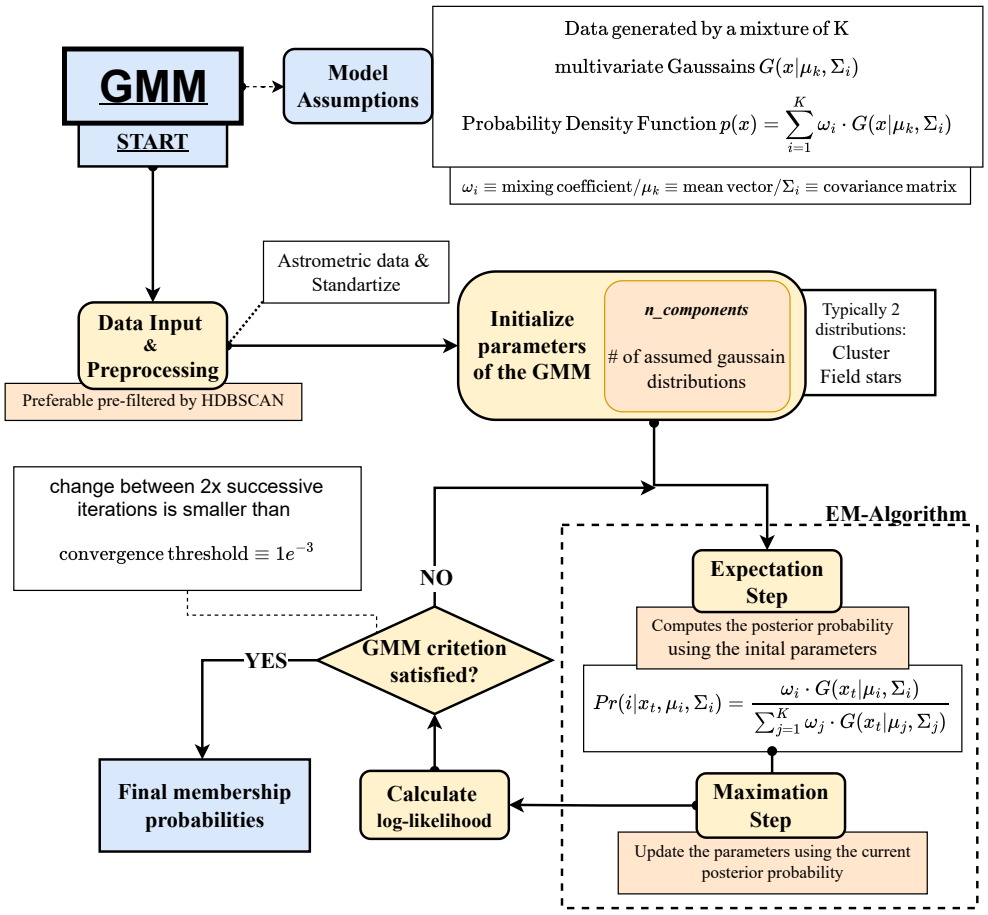

Figure 7.2. Schematic flowchart of general operational principle of the Gaussian Mixture Model.

Appendix C: Absolute G magnitude residuals

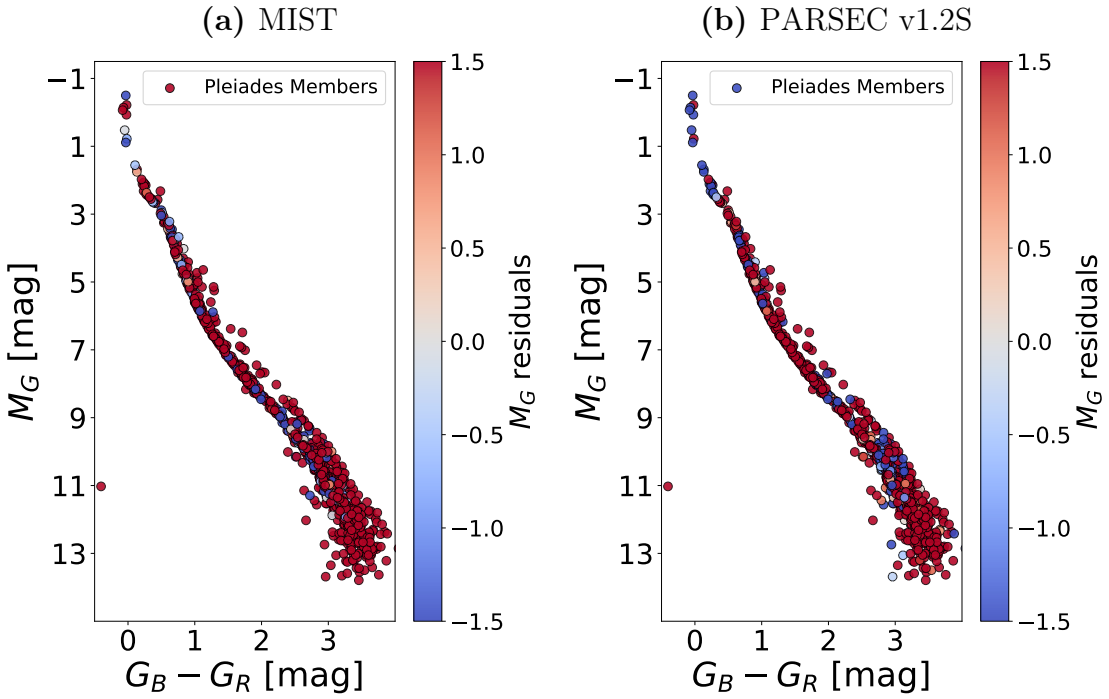

Figure 7.3. Residuals of observed and modeled photometric data $(G, BP - RP)$ color coded onto Pleiades cluster members with a CMD, showing the MIST model (a) and the PARSEC v1.2S model (b). Brighter sources are color coded in red, while blue points correspond to fainter sources.

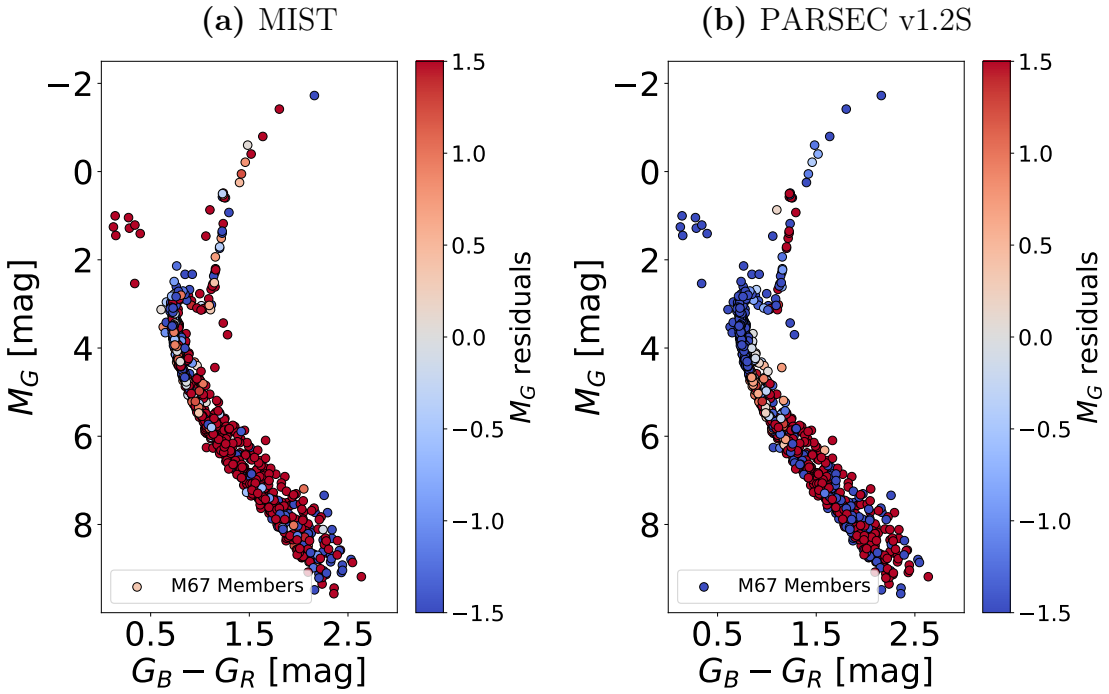

Figure 7.4. Residuals of observed and modeled photometric data $(G, BP - RP)$ color coded onto M67 cluster members with a CMD, showing the MIST model (a) and the PARSEC v1.2S model (b). Brighter sources are color coded in red, while blue points correspond to fainter sources.

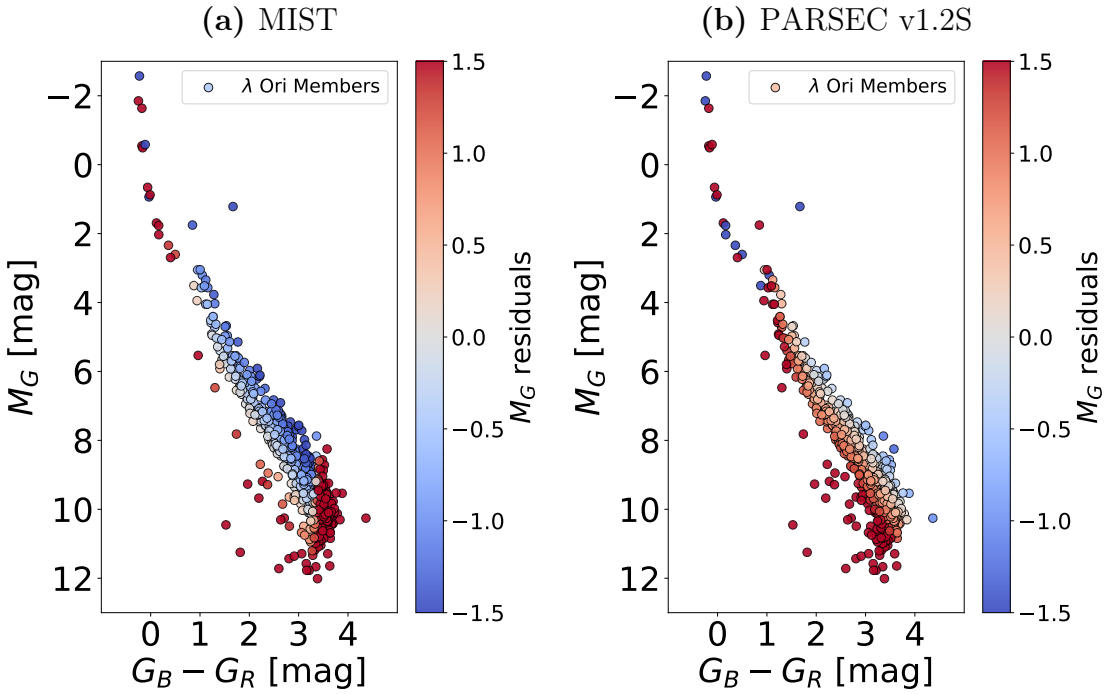

Figure 7.5. Residuals of observed and modeled photometric data $(G, BP - RP)$ color coded onto λ Ori cluster members with a CMD, showing the MIST model (a) and the PARSEC v1.2S model (b). Brighter sources are color coded in red, while blue points correspond to fainter sources.

Appendix D: Corner plots derived from isochrone fitting

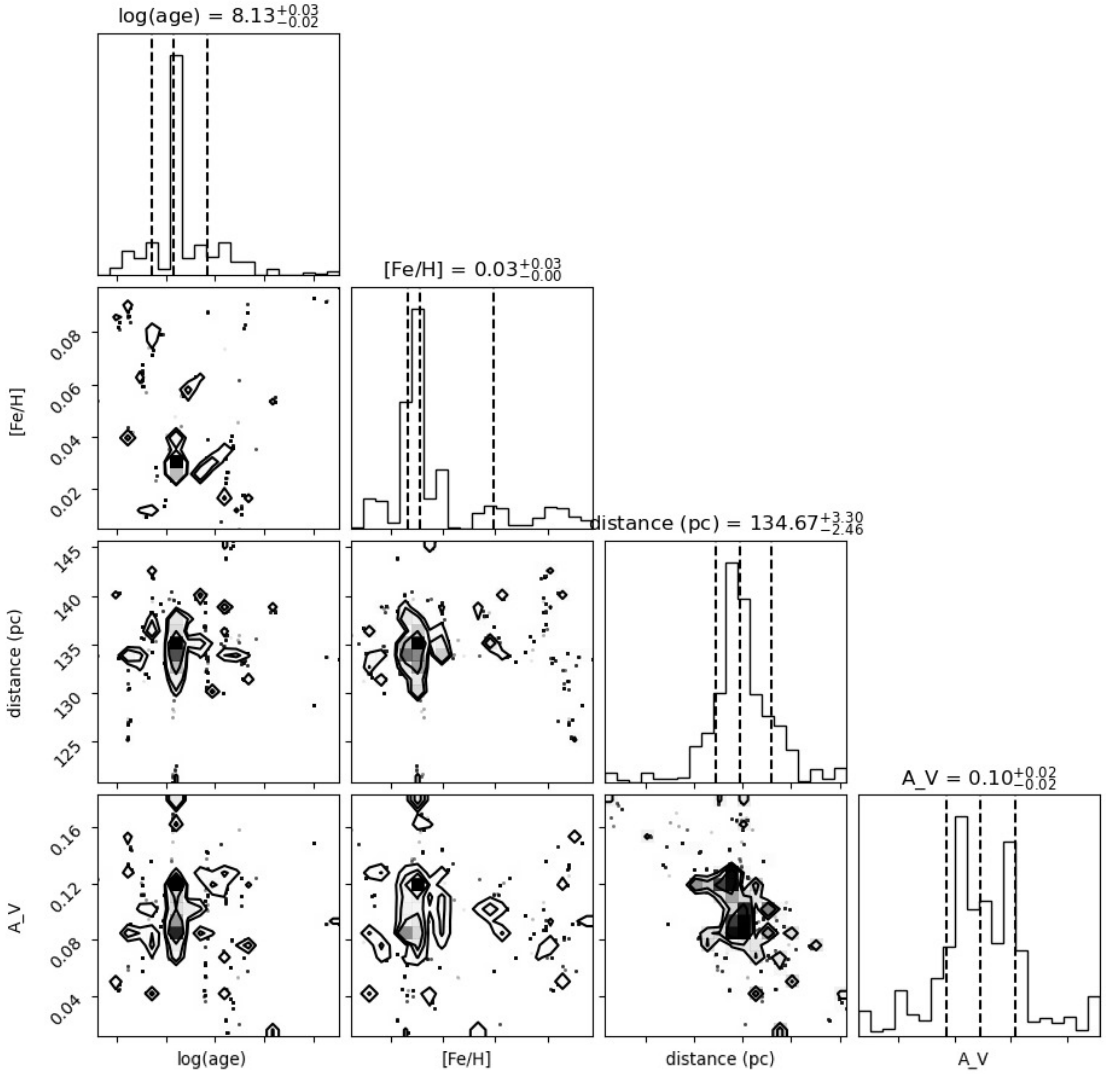

Figure 7.6. Posterior distributions and covariances for the six parameters fitted for Pleiades members, including log(age), [Fe/H], A_V, distance, and the additional diagnostic parameters σ and ν. These were derived from MCMC sampling using a Students-t-distribution likelihood and MIST isochrones. First diagonal panels show the 1D distribution for each parameter, including the median (50%), 84%, and 16% confidence intervals. Off-diagonal panels illustrate 2D projections of the combined posterior, indicating correlations between parameters.

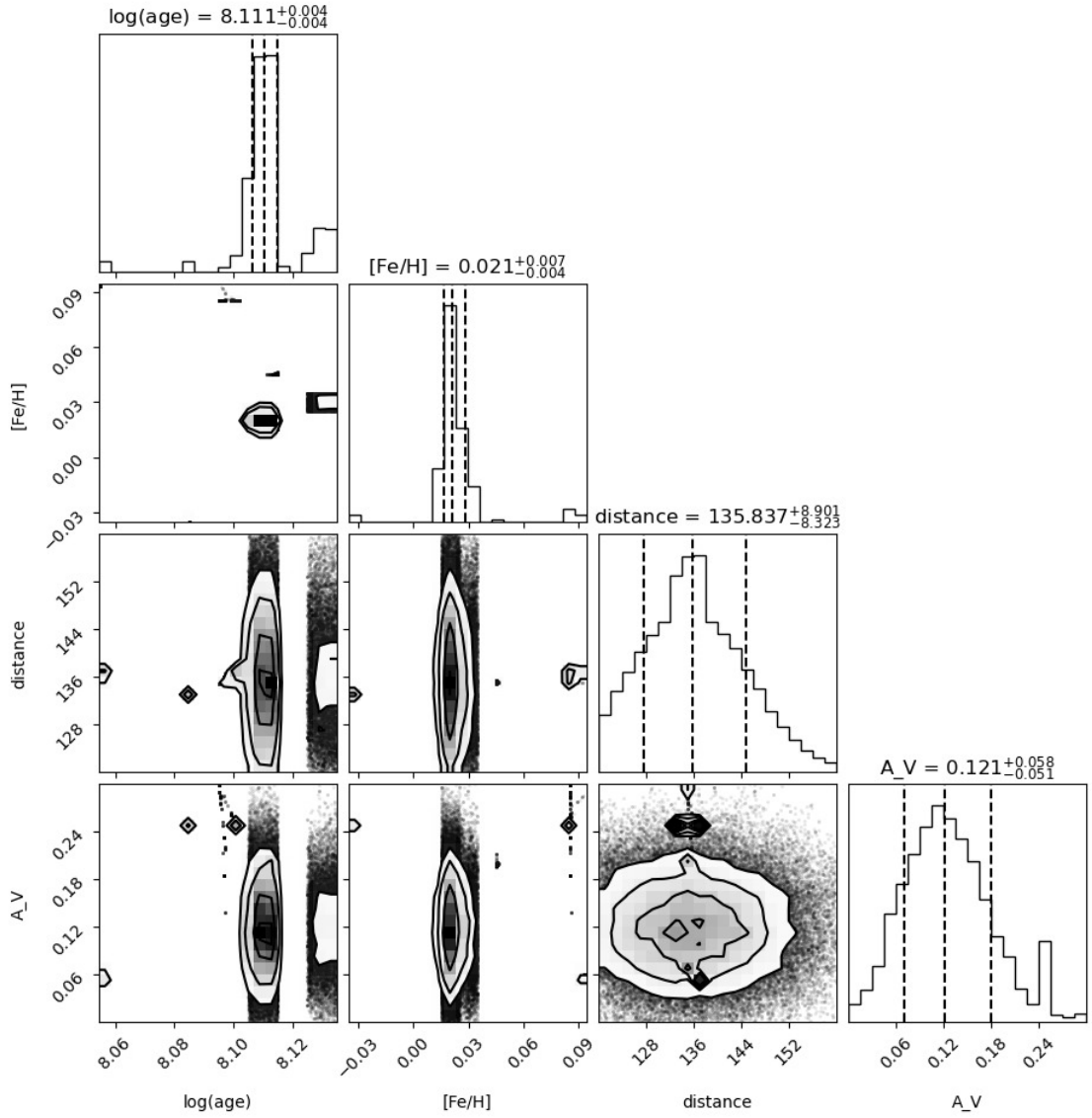

Figure 7.7. Posterior distributions and covariances for the six parameters fitted for Pleiades members, including log(age), [Fe/H], A_V, and distance. These were derived from MCMC sampling using a Students-t-distribution likelihood and PARSEC v1.2S isochrones. First diagonal panels show the 1D distribution for each parameter, including the median (50%), 84%, and 16% confidence intervals. Off-diagonal panels illustrate 2D projections of the combined posterior, indicating correlations between parameters.

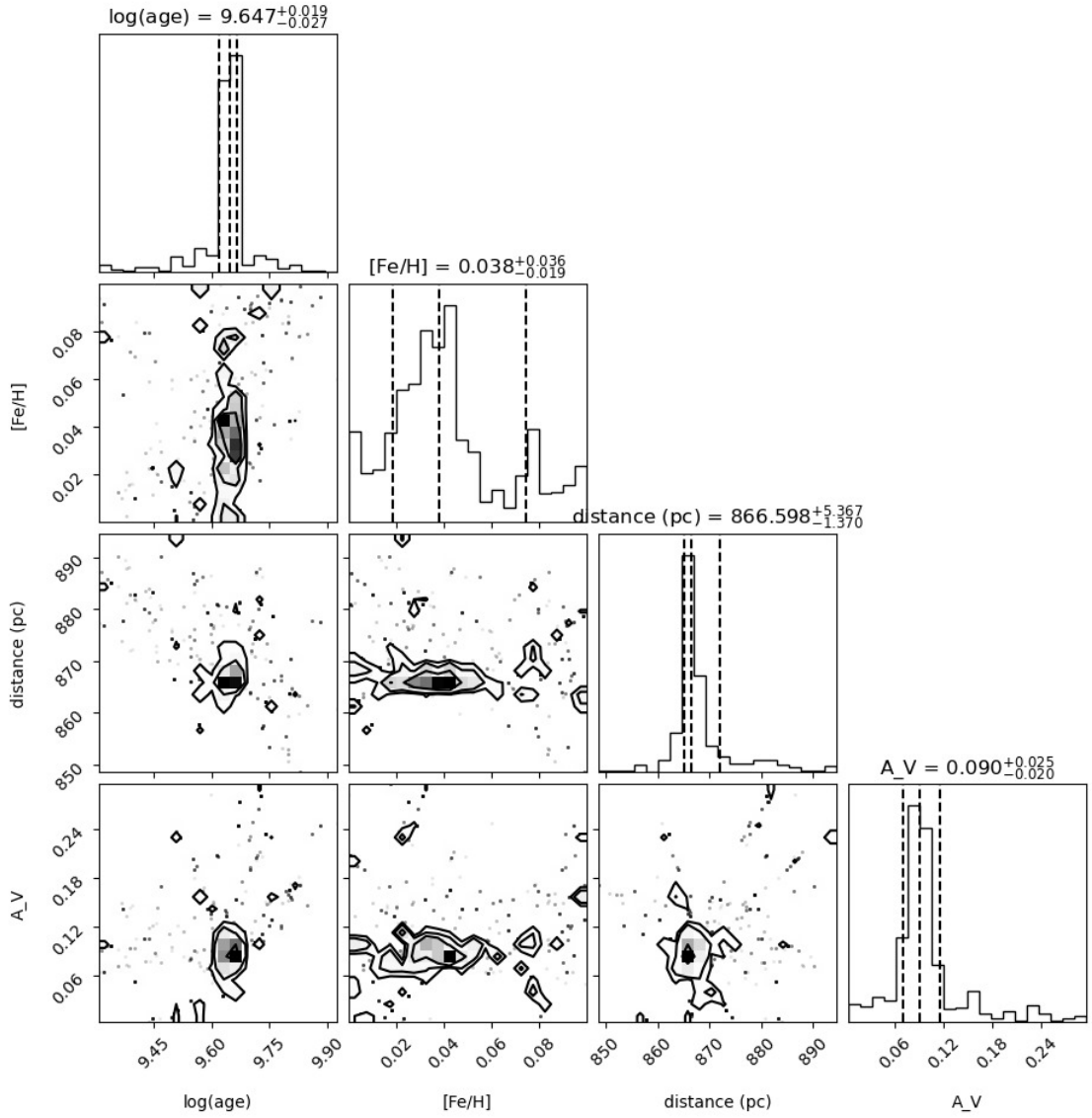

Figure 7.8. Posterior distributions and covariances for the six parameters fitted for M67 members, including log(age), [Fe/H], A_V, and distance. These were derived from MCMC sampling using a Students-t-distribution likelihood and MIST isochrones. First diagonal panels show the 1D distribution for each parameter, including the median (50%), 84%, and 16% confidence intervals. Off-diagonal panels illustrate 2D projections of the combined posterior, indicating correlations between parameters.

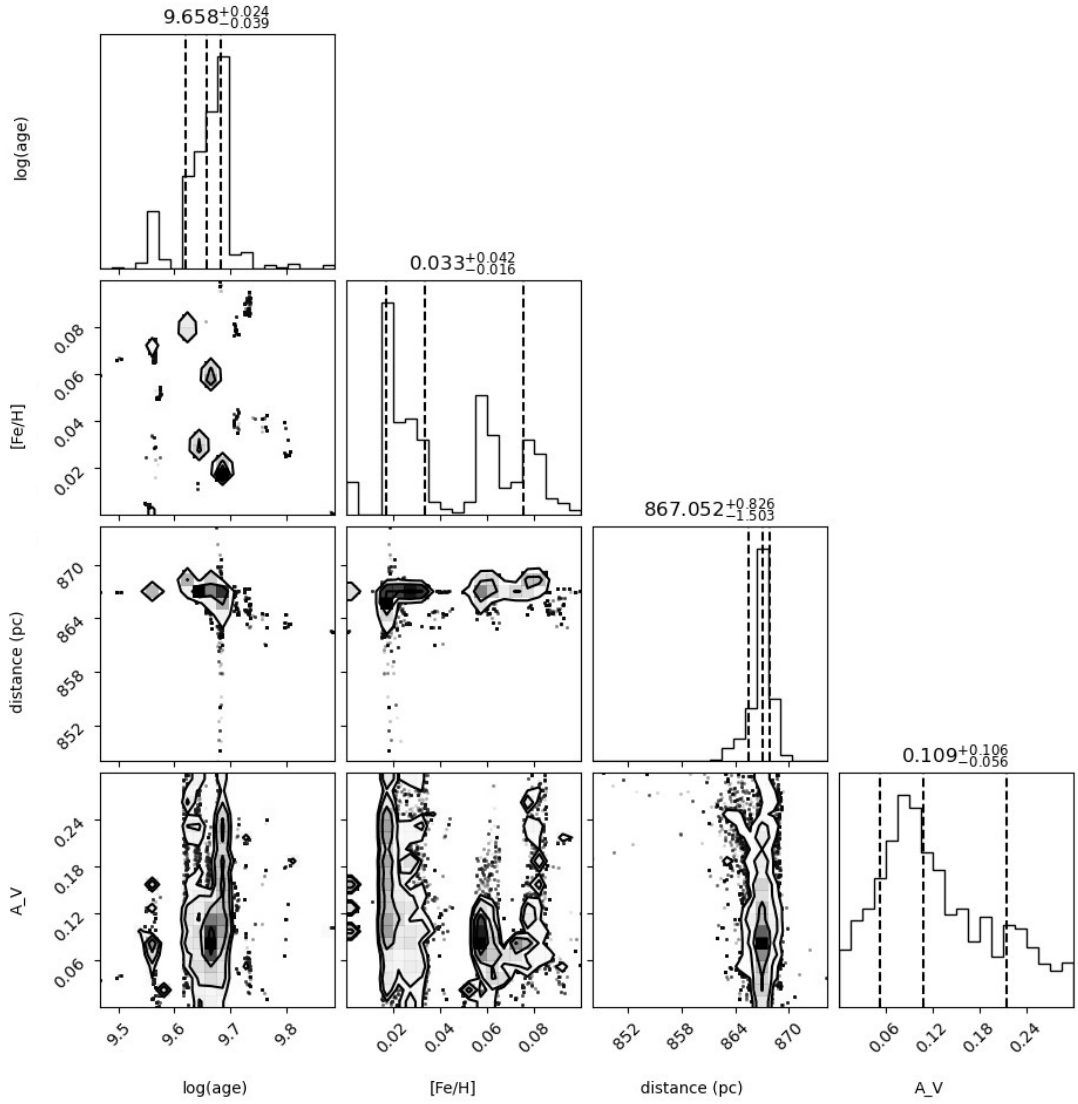

Figure 7.9. Posterior distributions and covariances for the six parameters fitted for M67 members, including log(age), [Fe/H], A_V, and distance. These were derived from MCMC sampling using a Students-t-distribution likelihood and PARSEC v1.2S isochrones. First diagonal panels show the 1D distribution for each parameter, including the median (50%), 84%, and 16% confidence intervals. Off-diagonal panels illustrate 2D projections of the combined posterior, indicating correlations between parameters.

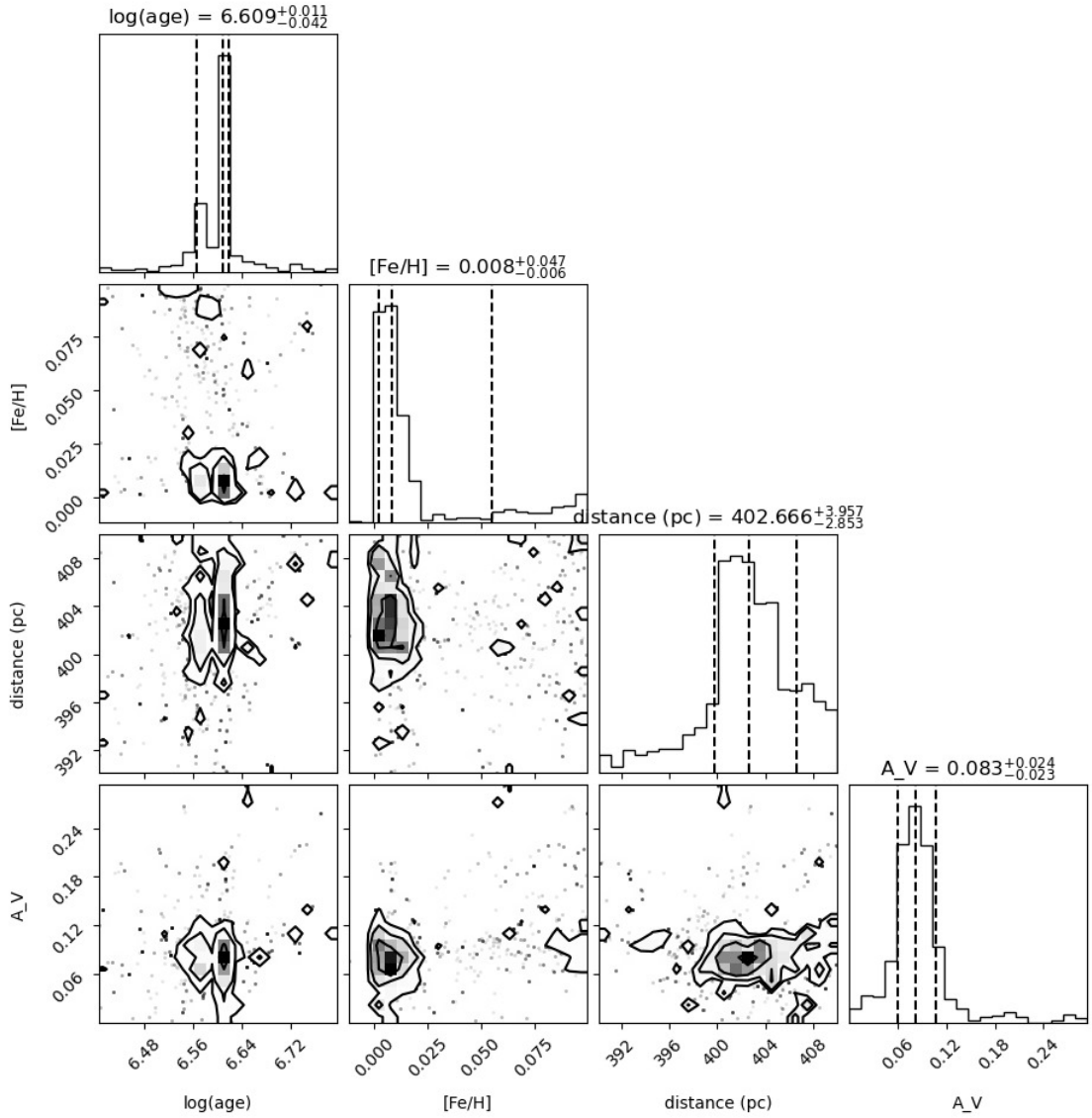

Figure 7.10. Posterior distributions and covariances for the six parameters fitted for λ Ori members, including log(age), [Fe/H], A_V, and distance. These were derived from MCMC sampling using a Students-t-distribution likelihood and MIST isochrones. First diagonal panels show the 1D distribution for each parameter, including the median (50%), 84%, and 16% confidence intervals. Off-diagonal panels illustrate 2D projections of the combined posterior, indicating correlations between parameters.

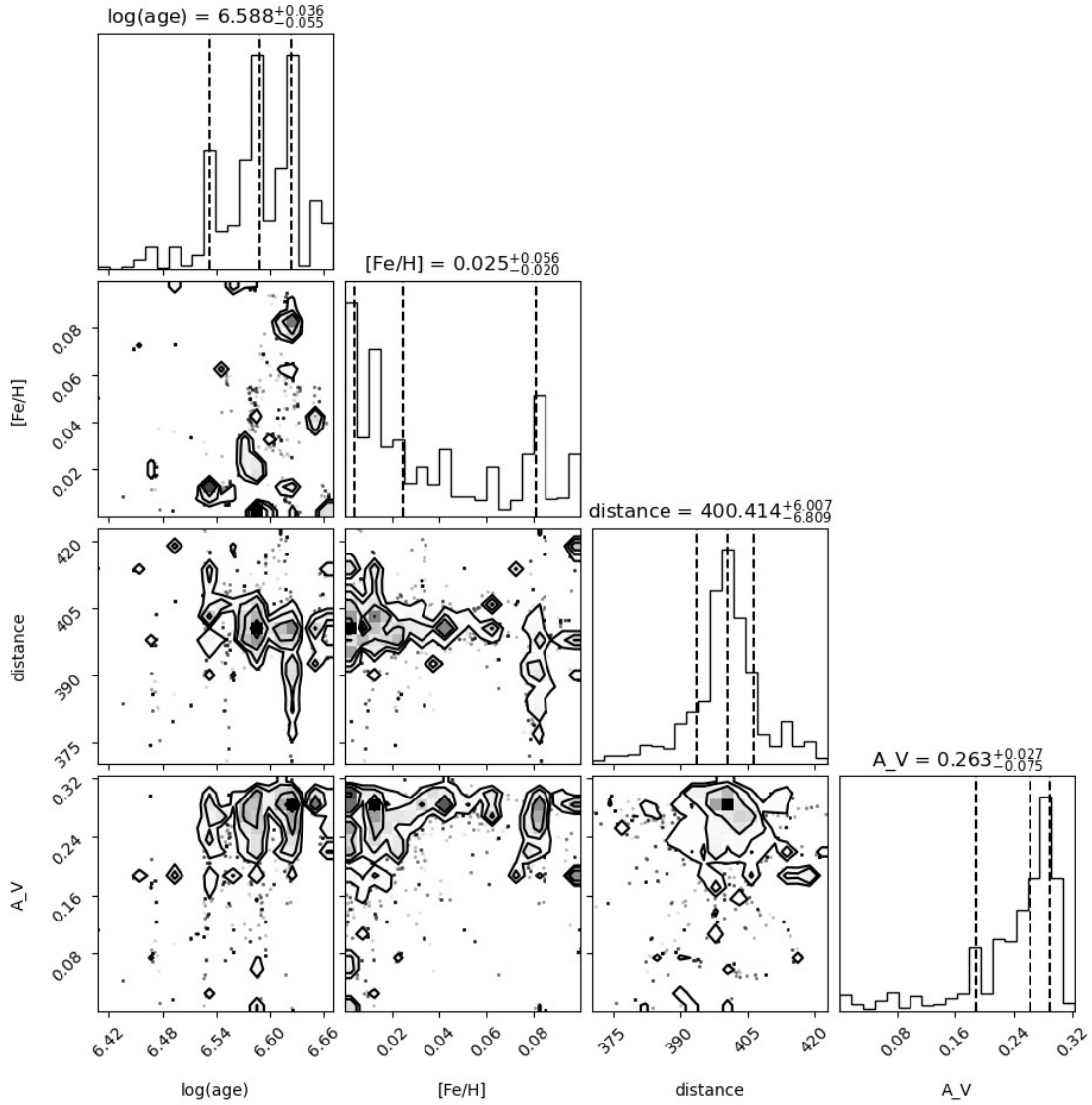

Figure 7.11. Posterior distributions and covariances for the six parameters fitted for λ Ori members, including log(age), [Fe/H], A_V, and distance. These were derived from MCMC sampling using a Students-t-distribution likelihood and PARSEC v1.2S isochrones. First diagonal panels show the 1D distribution for each parameter, including the median (50%), 84%, and 16% confidence intervals. Off-diagonal panels illustrate 2D projections of the combined posterior, indicating correlations between parameters.

Appendix E: Utilized AI-based tools

Table 7.2. Utilized AI-based tools used during the thesis writing process.

AI-based Tool	Purpose	Aspect of Work Affected	Prompt (Entry)	Comment
ChatGPT / Microsoft Copilot	Assistance with writing and improving code	Code	Example: Based on the code provided above, can you add the calculation of residuals for the G-band magnitude within the CMD? or: Can you fix what the error code is indicating?	Mainly error messages from Python were inserted for AI to fix the reported problem.
ChatGPT	Generation of compatible BibTeX entries matching the required style	Bibliography	Hey ChatGPT, I would like you to help me format my bibliography for my bachelor thesis. In general, it should follow this pattern: @article{heiter2014metallicity, author = {Heiter, U. and Soubiran, C. and Netopil, M. and Paunzen, E.}, title = {On the metallicity of open clusters – II.}, journal = {Astronomy & Astrophysics}, year = {2014}, volume = {562}, pages = {A93}, doi = {10.1051/0004-6361/201322559} }	ChatGPT was provided with metadata of the respective works (from publishers) to correctly format the sources. This saved significant time.
ChatGPT	Inquiry about important aspects of scientific writing	Scientific writing	What are the most important things to watch for in scientific writing for my Bachelor's Thesis? Especially in sentence structure, would you keep it rather short and precise?	The answer, combined with reading scientific articles on my topic, helped me to adapt my writing style.
ChatGPT	Assistance with structuring the Conclusion	Conclusion	On what aspects do I need to focus for my Conclusion in my Bachelor Thesis? The general structure is: introduce the research goal, summarize the main results, and finally address limitations and outlook, right?	This was used to check whether my Conclusion had the right structure.
ChatGPT	Assistance with sentence structure, especially repetition, convoluted phrasing, and typos	All chapters, especially Discussion and Conclusion	Would you be able to analyze my text and point out bad sentence structures that I can improve (in the context of scientific writing)? Please pay particular attention to repetition and convoluted sentences.	The AI highlighted paragraphs and sentences with excessive repetition or too many commas, helping me to rephrase and shorten them.

Appendix F: Official Declarations Form

Official Declarations from

Name: Lennart Rathjen Matrikelnr: 6187492

A) Declaration of Authorship

I hereby affirm that I have written the present work independently and have used no sources or aids other than those indicated. All parts of my work that have been taken from other works, either verbatim or in terms of meaning, have been marked as such, indicating the source. The same applies to drawings, sketches, pictorial representations and sources from the Internet, including AI-based applications or tools. The work has not yet been submitted in the same or a similar form as a final examination paper. The electronic version of this work corresponds to the printed version. I am aware that false statements will be treated as deception.

☐ I have used AI-based applications and/or tools and documented them in the appendix Üse of AI-based applications".

B) Declaration regarding the publication of bachelor's or master's thesis

Two years after graduation, the thesis is offered to the archive of the University of Bremen for permanent archiving. The following are archived:

1. Master's theses with a local or regional focus, as well as per subject and academic year 10% of all master's theses
2. Bachelor's theses for the first and last bachelor's degrees per subject and year.

☐ I agree that my thesis may be viewed by third parties in the university archive for academic purposes.

☐ I agree that my thesis may be viewed by third parties for academic purposes in the university archive after 30 years (in accordance with §7 para. 2 BremArchivG).

☐ I do **not** consent to my thesis being made available in the university archive for third parties to view for academic purposes.

Declaration of consent for electronic checking of the work for plagiarism

Submitted papers can be checked for plagiarism using qualified software in accordance with §18 of the General Section of the Bachelor's or Master's Degree Examination Regulations of the University of Bremen. For the purpose of checking for plagiarism, the upload to the server is done using the plagiarism software currently used by the University of Bremen.

- ☐ I agree that the work I have submitted and written will be stored permanently on the external server of the plagiarism software currently used by the University of Bremen, in a library belonging to the institution (accessed only by the University of Bremen), for the above-mentioned purpose.
- ☐ I do **not** consent to the work I submitted and wrote being permanently stored on the external server of the plagiarism software currently used by the University of Bremen, in a library belonging to the institution (accessed only by the University of Bremen), for the above-mentioned purpose.

Consent to the permanent storage of the text is voluntary. Consent can be withdrawn at any time by making a declaration to this effect to the University of Bremen, with effect for the future. Further information on the checking of written work using plagiarism software can be found in the data protection and usage concept. This can be found on the University of Bremen website.

With my signature, I confirm that I have read and understood the above explanations and confirm the accuracy of the information provided.

Bremen, 07.10.2025
Place, Date

Signature