Research in Astronomy and Astrophysics, 24:115021 (16pp), 2024 November https://doi.org/10.1088/1674-4527/ad8055

13 Galactic Star Clusters in Gaia DR3 Identified by An Improved FoF and UPMASK Hybrid Method Using MvC

Huanbin Chi[1,2], Zebang Lai[1,2], Feng Wang[1,2], Zhongmu Li[3], and Ying Mei[1,2]
[1] Center for Astrophysics, Guangzhou University, Guangzhou 510006, China; fengwang@gzhu.edu.cn
[2] Peng Cheng Laboratory, Shenzhen 518000, China
[3] Institute of Astronomy and Information, Dali University, Dali 671000, China
Received 2024 August 7; revised 2024 September 14; accepted 2024 September 23; published 2024 November 26

Abstract

Open clusters (OCs) serve as invaluable tracers for investigating the properties and evolution of stars and galaxies. Despite recent advancements in machine learning clustering algorithms, accurately discerning such clusters remains challenging. We re-visited the 3013 samples generated with a hybrid clustering algorithm of FoF and pyUPMASK. A multi-view clustering (MvC) ensemble method was applied, which analyzes each member star of the OC from three perspectives—proper motion, spatial position, and composite views—before integrating the clustering outcomes to deduce more reliable cluster memberships. Based on the MvC results, we further excluded cluster candidates with fewer than ten member stars and obtained 1256 OC candidates. After isochrone fitting and visual inspection, we identified 506 candidate OCs in the Milky Way. In addition to the 493 previously reported candidates, we finally discovered 13 high-confidence new candidate clusters.

Key words: galaxies: star clusters: general – (Galaxy:) open clusters and associations: general – methods: data analysis

1. Introduction

Open clusters (OCs), or Galactic disk star clusters, are groups of stars formed from the same giant molecular cloud simultaneously. The Gaia (Gaia Collaboration et al. 2018, 2022; Riello et al. 2021) mission has identified many OCs. The discovery process is ongoing and not yet finished. The release of Gaia substantially boosted OC identification and research efforts. The number of OCs reported in the scientific literature is over 7000. A total of 4000 OCs have been released (Castro-Ginard et al. 2018, 2019, 2020; Liu & Pang 2019; Li et al. 2022) based on Gaia Data Release 2 (DR2; Gaia Collaboration et al. 2018) and Early Data Release 3 (EDR3; Lindegren et al. 2021). Chi et al. (2023c) reported 46 OCs in Gaia EDR3 and 83 OCs (Chi et al. 2023a) and 1179 OCs (Chi et al. 2023b) in Gaia Data Release 3 (DR3). During the preparation of this work, Hunt & Reffert (2024) derived completeness-corrected photometric masses for 6956 clusters from their work (Hunt & Reffert 2023) and found that only 5647 (79%) of the clusters from their previous catalog are compatible with bound OCs by calculating cluster masses and Jacobi radii.

Although numerous studies have been conducted to identify OCs, applying machine learning algorithms to obtain the most appropriate parameters for the models remains challenging. We applied the Friends-of-Friends (FoF) algorithm to find the corresponding OC identification in the previous works. As a complicated clustering method, FoF cannot handle noisy data. If FoF parameters (e.g., linking length) are not properly set, cluster members mixed with field star pollution can be easily obtained during cluster identification. After the initial identification of OC candidates, especially using the FoF method, unsupervised photometric membership assignment algorithm (UPMASK, Krone-Martins & Moitinho 2014), Random Forest (Mužić et al. 2022; Chi et al. 2023c), and deep set neural network (van Groeningen et al. 2023) are generally used for further screening of member stars, but these methods still have certain limitations. Some OCs cannot reasonably be described by the classic King model (King 1962) embedded in UPMASK or pyUPMASK (Zhong et al. 2022). Arunima et al. (2023) showed that the much-used method of distance and velocity cutoffs for membership determination often leads to false negatives and positives, and membership determination is still challenging for young star clusters. It is also still difficult to distinguish member stars of star clusters and associations between the foreground and background populations (Gagné et al. 2018).

In our previous work, we utilized the FoF method and pyUPMASK to identify OC candidates (Chi et al. 2023b). However, how to further improve member star identification accuracy has also been a key issue troubling us. In this study, we introduce a multi-view clustering (MvC) ensemble method to further enhance OC membership determination accuracy. The rest of the paper is structured as follows. In Section 2, we describe the methodology developed for determining OC membership. We then introduce the production of OC samples

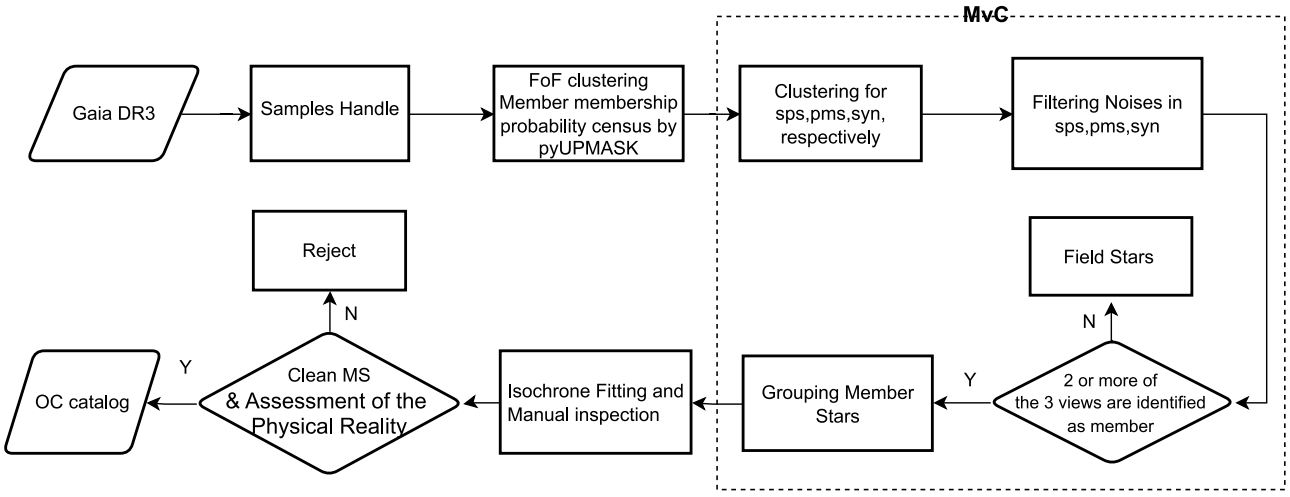


Figure 1. Diagram of MvC method. The sps, pms and syn are three views defined according to Section 2.1, respectively.

in Section 3. Section 4 presents our results and the newly found OCs. The MvC algorithm is discussed in Section 5. The conclusion is provided in Section 6.

2. Multi-view Clustering Ensemble Method

Distance metrics in high-dimensional spaces would lose efficacy (Beyer et al. 1999; Hinrichs et al. 2014), resulting in poor density-based clustering. According to Chang et al. (2014), high-dimensional data have the characteristics of being sparse, a dimensional disaster, and noise, which reduce the possibility of class recognition in all dimensions, making the traditional clustering algorithm unsuitable for high-dimensional data clustering. With the subview approach (Zhao et al. 2017), since each viewpoint characterizes the same subject differently when all viewpoint features are concatenated together, they can be considered as a description of the same subject from a new viewpoint. The learned consistency graph can describe the structure of all perspective features.

Inspired by multi-view learning (Zhao et al. 2017), we introduced the MvC method and applied it to determine OC memberships. The flowchart of applying MvC is illustrated in Figure 1. Note that MvC is only used to process OC candidate data initially identified by FoF and pyUPMASK. This is done by clustering the data from multiple views separately and then using the voting mechanism to obtain a more plausible member star result.

2.1. Definitions of Subviews

In general, two principles should be considered while using MvC. (1) The consistency principle: we must ensure consistency among multiple views. For example, many views should share the same category structure. (2) The complementarity principle: each view of multi-view data may contain information or knowledge that others do not.

Considering that, in a real physical system, an OC should have a stellar overdensity in the proper motion space and a stellar overdensity in the sky position space (Piatti et al. 2022), we divided the traditional high-dimensional single view (SYN) (l, b, μ_α, μ_δ, ϖ) into two low-dimensional subviews, i.e., proper motion subview (PMS) and sky position subview (SPS), for each star.

$$\begin{aligned} \mathrm{SPS} &= \{l, b, \varpi\} \\ \mathrm{PMS} &= \{\mu_\alpha, \mu_\delta, \varpi\} \\ \mathrm{SYN} &= \{l, b, \mu_\alpha, \mu_\delta, \varpi\}. \end{aligned} \tag{1}$$

According to the requirements of MvC, subviews are related and not independent of each other. Members of an OC should have similar distances within the Gaia DR3 parallax uncertainties. Both view spaces are affected by parallax. Therefore, we use parallax as the association between all views.

2.2. Member Star Identification Based on Subviews

We use the clustering algorithm to perform secondary clustering on all members in the PMS and SPS views, separately. For example, for the SPS view, we use the HDBSCAN algorithm to cluster $\{l, b, \varpi\}$, with cluster label of 1 for member stars and cluster label of -1 for field stars. Since the data in the PMS and SPS views are the consequence of FoF clustering, we simply used the HDBSCAN algorithm with the hyperparameter ξ having a value of twice the feature space dimension, which is the same as Ghosh & Sulistiyowati (2022).

The clustering on PMS could ensure that the OC has a stellar overdensity in the proper motion space. This is one of the typical physical characteristics of OC because OC originates

from the same dense molecular cloud undergoing the same starburst. The clustering on SPS ensures that the identified clusters have physical characteristics of overdensity distribution in the sky position space.

Integration strategy has been successfully applied in astronomy (Chi et al. 2022). Inspired by this, we further combined the clustering results of the three views. We use a voting integration strategy to integrate the clustering results of the three views. For each cluster member, we use clustering results of three views to conduct voting statistics, and we reserve the final member star which wins more than or equal to 2 votes. For example, one star with more than two views of SPS, PMS, and SYN as a member will eventually be considered a current member. In contrast, a star with less than one view as a member will be regarded as a field star. It should be noted that we filtered the intersection of the noise points of the three views prior to integration. These points are recognized as background noise data on all three views and should be removed. After eliminating the noise the HDBSCAN clustering algorithm identified, we further eliminated the clustering results of subviews with a small number of member stars. When we used HDBSCAN to cluster in subviews, we discarded star groups with less than 10 member stars after subview clustering according to Hunt & Reffert (2021) who suggested that the minimum possible size of a star cluster is set to 10 for HDBSCAN.

3. Data Preparation

Based on Gaia DR3, Chi et al. (2023b) filtered out faint stars ($G > 18$ mag), limited parallax (ϖ) from 0.14 to 5 kpc, and obtained more than 20 million target sources. Based on these stellar data sources, rough grid clustering was performed using the FoF algorithm. A member probability census was carried out using pyUPMASK. Then, according to the quality of the fit with the isochrone from high to low, the data sets for the OC candidates are divided into three categories (Class A (1194), Class B (5252), and Class C (5925)). The samples in Class C have a loose color–magnitude diagram (CMD) distribution and poor-fitting results ($\sigma_{d^2} > 0.04$ or $\bar{d}^2 > 0.02$). Therefore, by focusing on the data analysis of Class A and B, Chi et al. (2023b) identified a total of 3763 high-confidence OC candidates, including 2584 clusters that have already been published, and 1179 OCs that are new discoveries.

We did not apply FoF or pyUPMASK to search for OC candidates again in the study. We just re-visited the rest of the data (3013 OC candidates in total) in Class A and B which were not identified and reported by Chi et al. (2023b). These 3013 samples were meticulously analyzed for membership using the MvC method.

4. OC Identification and Results

Based on the OC candidates prepared in Section 3, we applied the MvC method to identify member stars of each possible OC. According to Section 2, we first created three views (SPS, PMS, SYN) for each OC candidate. Subsequently, in each of these three views, we performed HDBSCAN to cluster all the members of each OC again to further determine the member stars and field stars of each OC more precisely. To integrate the clustering results of the three views, we used a voting integration strategy. After the processing of MvC, we selected 1256 more plausible candidate OCs from 3013 candidate OCs.

4.1. Isochrone Fitting

Based on these 1256 OC candidates, we performed isochrone-fitting for each new result, following the methods described in Chi et al. (2023c). Due to the small number of stars with G magnitudes less than 17 mag in the ID3041, ID12455, and ID1446 candidate clusters, there would be large uncertainties in the fit results. We discard these candidate clusters and 1069 OC candidates are thus fitted.

4.2. Visual Inspection

We adopted the methods described in Chi et al. (2023a) to perform manual validation. After manual validation, we obtained 506 OC candidates with high confidence. An example is shown in Figure 2. For comparison, we also present the membership probabilities obtained with the same pyUPMASK as in Chi et al. (2023b) in the color bar of each subplot.

Figure 2 shows that a few member stars deviate from the main sequence, which could be caused by blue or yellow stragglers, variable stars, or dust extinction. These are usually special members of some OCs but these members are also identified by our method.

4.3. Cross-matched and New Open Clusters

Considering only the galactic longitude (l) and latitude (b) provided by the star catalog, our approach to cross-matching is to deem an observed star cluster as coincident with a cataloged cluster if the centers of both clusters fall within a 0.°5 radius in both the galactic longitude and latitude coordinates. If the cluster catalog from literature also includes information on individual cluster members, we then proceed to a more detailed analysis of the members of both clusters, as illustrated in Figure 3.

We cross-matched with 26 main star cluster catalogs, i.e., Cantat-Gaudin et al. (2018, 2019, 2020), Castro-Ginard et al. (2018, 2019, 2020, 2022), Bica et al. (2019), Ferreira et al. (2019, 2020, 2021), Liu & Pang (2019), Torrealba et al. (2019), Cantat-Gaudin & Anders (2020), Hao et al. (2020, 2021, 2022a, 2022b), Casado (2021), Dias et al. (2021), He et al. (2021, 2022a, 2022b, 2022c), Hunt & Reffert (2021), Jaehnig et al. (2021), Qin et al. (2021), Vasiliev & Baumgardt (2021), Tarricq et al. (2022), Li et al. (2022), Chi et al. (2023a, 2023b, 2023c), Hunt & Reffert (2023), Li & Mao (2023) and Sim et al. (2019).

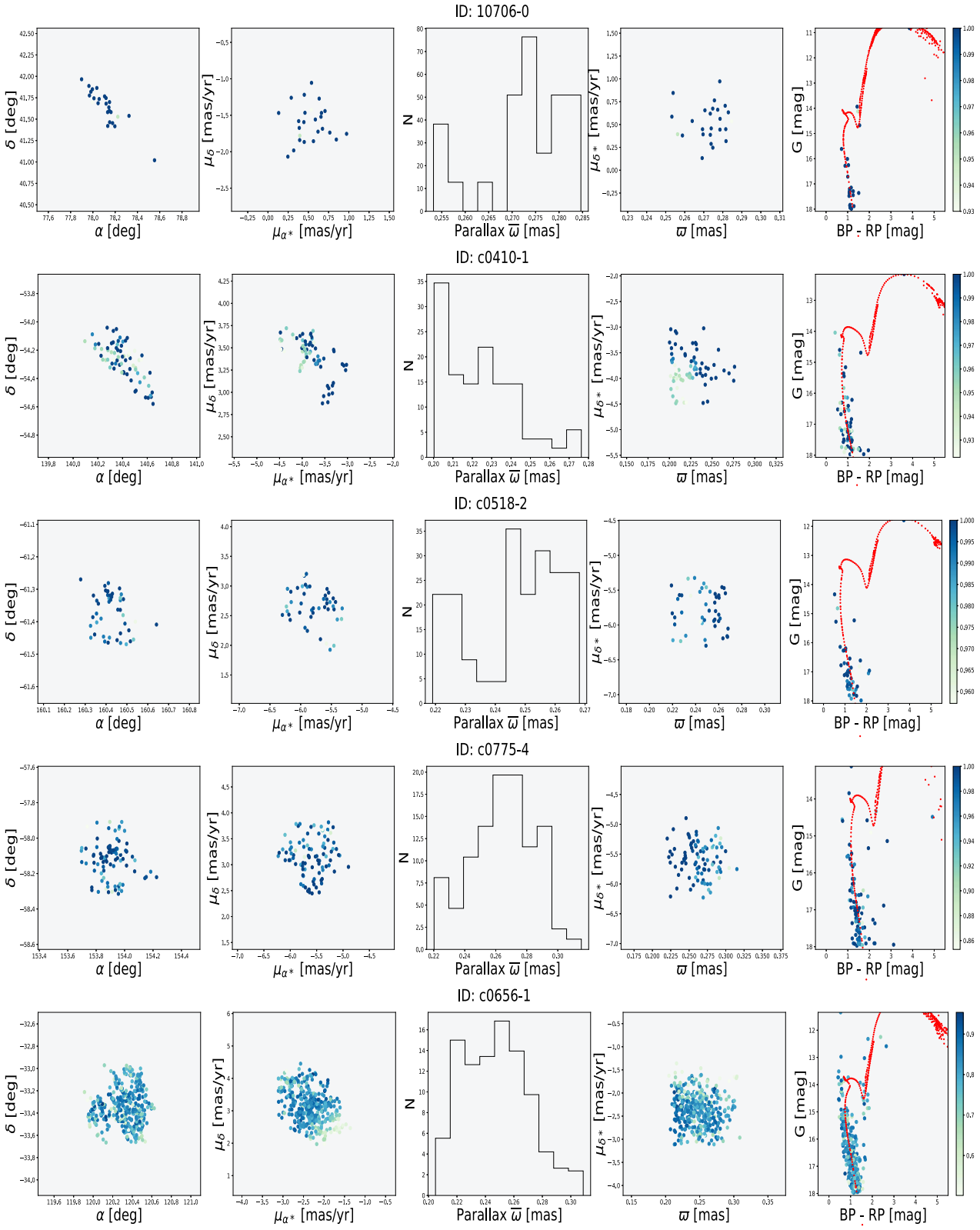


Figure 2. Examples of newly found OCs. From left to right, the subplots show the spatial distribution, proper-motion distribution, parallax statistics, parallax distribution, and CMD with the best-fitting isochrone line, respectively. The color bars represent the cluster probability of the member stars calculated by pyUPMASK. (The complete figure set (13 images) is available in Appendix.)

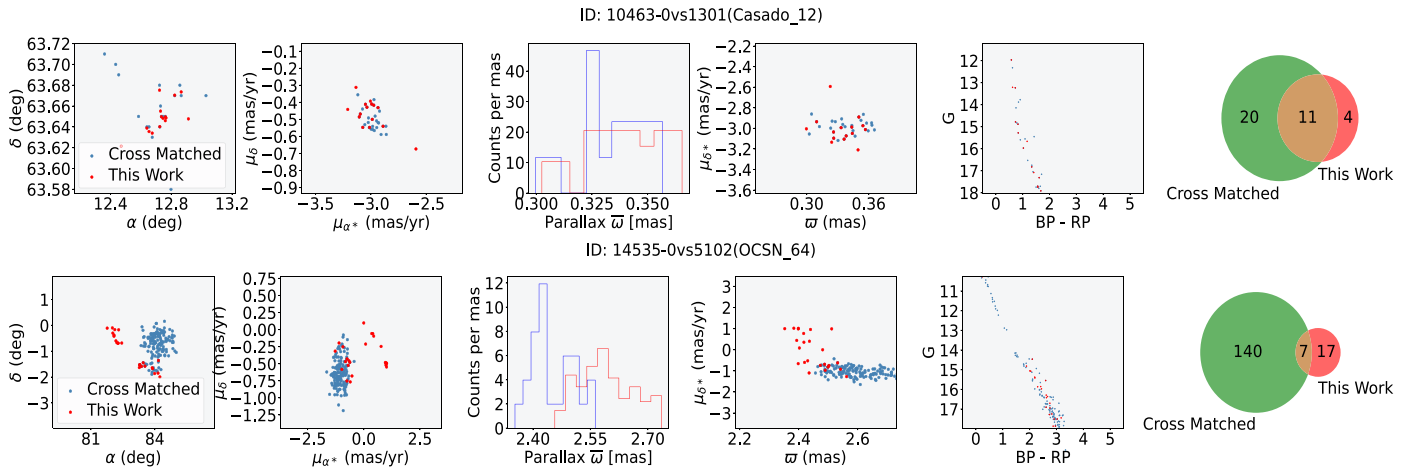


Figure 3. Two examples of member analysis for star clusters matched within a 0.°5 range. Although there are only six common members in the diagram, by combining the distributions of space, proper motion, and parallax, we still believe that Casado 12 and field 10463-0 constitute a star cluster.

Recently, Qin et al. (2023) reported 101 star clusters within 500 pc, and we also performed a cross match. Li & Mao (2024) reported the discovery of 83 new star clusters. Given that the star catalog only provides galactic longitude (l) and latitude (b), our strategy for cross-matching involves considering an observed star cluster (OC) to be positionally coincident with a cataloged one if the centers of both are within a circular region of 1.°5 radius in both the galactic longitude and latitude coordinates. To exclude previously reported star clusters as much as possible, we cross-match the most recently updated UCC catalog (Perren et al. 2023) (containing 16,179 clusters) using the same method. After cross-matching, 490 OC candidates were reported in the previous literature. We are confident that the remaining 13 OC candidates have never been reported. The parameter information of 13OCs is given in Table 1, and the member information of star clusters is given in Table 2.

5. Discussions

No matter how machine learning algorithms might identify OCs, each OC will have to be manually verified. However, machine learning algorithms can significantly decrease the number of potential candidates, reducing manual identification. The hybrid approach of FoF+pyUPMASK+MvC shows excellent application results in this study. First, MvC reduces the number of OC candidates that need to be manually confirmed from 3013 to 1256. Furthermore, the FoF+pyUPMASK+MvC hybrid algorithm addresses the problem that FoF needs to be more accurate in identifying field stars.

5.1. Multi-view Voting Mechanism for Member Star Determination

Chang et al. (2014) pointed out that high-dimensional data have the characteristics of sparsity, dimensional catastrophe, and noise variables, reducing the possibility of identifying classification in all dimensions, making the traditional clustering algorithms less effective when facing high-dimensional data clustering. Thus, we have learned from Xie et al. (2019)'s approach, i.e., constructing multiple views (global view + multiple subviews) for clustering, which can significantly improve the clustering effect. We introduce this approach to the task of identifying star clusters in astronomy by clustering the global view (l, b, μ_α, μ_δ, ϖ) and two subviews (l, b, ϖ) and (μ_α, μ_δ, ϖ) for MvC. The clustering of the two subviews reduces the dimensionality of the clustered features, which overcomes the "curse of dimensionality" (Hinrichs et al. 2014) in favor of improving the effectiveness of downstream density-based clustering algorithms. Second, we refer to Jiang et al. (2018) for clustering integration to get more robust clustering results. Specifically, the integration is performed on the clustering results of three views: (l, b, ϖ), (μ_α, μ_δ, ϖ) and (l, b, μ_α, μ_δ, ϖ). First, we eliminate the members identified as noise points in all three views achieving the elimination of highly reliable field stars. Then, we discard potential field stars by voting to retain members with two or more views and consider member stars as the final cluster members. This mitigates the sensitivity of the downstream clustering algorithm to the parameters to obtain reliable cluster member stars.

Based on the clustering results in the three subviews, voting is performed to give the final results. The application is very effective from a practical point of view. To a certain extent, it compensates for the difficulties of parameter optimization in traditional clustering algorithms and enhances the robustness of the algorithm. The HDBSCAN method used in the subview can also be replaced by other clustering algorithms, such as Gaussian Mixture Models (GMMs) or DBSCAN. HDBSCAN is chosen in this study because it has been successfully applied to cluster mining and identified many OCs (Chi et al. 2023a; Hunt & Reffert 2023).

Table 1
Parameters of Final 13 OCs

ID	R.A. (deg)	ra_std (deg)	Decl. (deg)	dec_std (deg)	plx (mas)	plx_std (mas)	pmra (km s^{-1})	pmra_std (km s^{-1})	pmdec (km s^{-1})	pmdec_std (km s^{-1})	l (deg)	b (deg)	$N_{\rm mem}$	Age $\left(\log\left(\frac{\rm age}{\rm yr}\right)\right)$	Z $\left(\log\left(\frac{Z}{Z_\odot}\right)\right)$
10428-2	359.528	0.059	63.409	0.001	0.306	0.006	−3.194	0.175	−0.881	0.101	117.006	1.152	21	8.70	0.40
10507-1	31.811	0.094	62.879	0.003	0.265	0.018	−1.068	0.040	−0.039	0.042	131.448	1.268	32	8.82	−0.10
10706-0	78.111	0.017	41.646	0.038	0.272	0.009	0.524	0.036	−1.605	0.060	165.685	1.410	25	9.09	0.02
c0410-1	140.403	0.016	−54.272	0.016	0.224	0.019	−3.822	0.122	3.416	0.040	275.308	−3.066	72	8.52	0.12
c0514-0	161.415	0.047	−61.664	0.010	0.240	0.019	−5.623	0.251	2.428	0.086	288.583	−2.348	190	8.82	−0.28
c0518-2	160.425	0.005	−61.373	0.004	0.247	0.015	−5.786	0.075	2.665	0.089	288.030	−2.313	46	8.52	0.20
c0656-1	120.320	0.026	−33.349	0.025	0.248	0.022	−2.342	0.141	3.124	0.292	250.073	−1.582	378	8.88	0.07
c0775-4	153.907	0.010	−58.124	0.011	0.264	0.021	−5.591	0.083	3.158	0.128	283.537	−1.244	91	8.70	0.30
c1409-1	163.218	0.016	−60.510	0.001	0.236	0.009	−5.880	0.085	2.236	0.054	288.839	−0.924	33	8.85	0.44
c2915-0	28.744	0.206	63.062	0.003	0.307	0.018	−1.126	0.177	−0.130	0.040	130.056	1.073	44	8.28	0.16
c3059-1	27.416	1.145	65.054	0.008	0.242	0.019	−1.043	0.123	0.140	0.082	129.024	2.877	190	8.88	−0.40
c5076-0	157.134	0.036	−59.851	0.001	0.236	0.015	−5.812	0.168	2.976	0.096	285.870	−1.801	86	8.70	0.23
c6080-1	0.772	0.044	61.215	0.001	0.281	0.022	−2.784	0.117	−1.245	0.066	117.143	−1.112	20	8.82	−0.88

Note. $N_{\rm mem}$ is the number of cluster members.

(This table is available in its entirety in machine-readable form in the online article.)

Table 2
Clustered Sources

	source_id	R.A. (deg)	Decl. (deg)	pmra (mas yr^{-1})	pmdec (mas yr^{-1})	parallax (mas)	mag_g (mag)	mag_bp (mag)	mag_rp (mag)	rv (km s^{-1})	ra_err (km s^{-1})	dec_err (deg)	pmra_err (mas yr^{-1})	pmdec_err (mas yr^{-1})	parallax_err (mas)	rv_err (km s^{-1})	probs_final	Cluster_ID
0	2013108453632539520	359.8399	63.3928	−4.021	−1.077	0.3	15.92	16.271	15.286		0.03	0.03	0.039	0.039	0.036		0.72	10428-2
1	2013108659790975360	359.9929	63.4063	−3.745	−0.941	0.296	15.963	16.294	15.456		0.025	0.024	0.034	0.03	0.03		0.51	10428-2
2	2016107371592207744	359.3409	63.3839	−2.894	−0.275	0.308	17.741	18.524	16.907		0.067	0.07	0.088	0.094	0.085		0.99	10428-2
3	2016107921348269440	359.1857	63.3899	−2.432	−1.222	0.305	15.128	15.464	14.598		0.016	0.019	0.023	0.027	0.022		1	10428-2
4	2016108093146707840	359.3094	63.3829	−3.299	−0.548	0.314	17.644	18.32	16.84		0.065	0.069	0.086	0.091	0.085		0.98	10428-2
5	2016108196225893760	359.3602	63.4094	−2.843	−0.653	0.305	15.397	16.391	14.413	−60.11	0.02	0.021	0.026	0.029	0.026	5.762	0.99	10428-2
6	2016108436744022016	359.334	63.4291	−3.051	−0.963	0.309	16.443	16.946	15.76		0.032	0.034	0.043	0.044	0.042		0.97	10428-2
7	2016110635767366272	359.6142	63.4243	−3.65	−0.479	0.318	14.622	15.542	13.676	−68.424	0.014	0.014	0.017	0.019	0.017	3.202	0.92	10428-2
8	2016111013724563328	359.761	63.4045	−3.155	−1.129	0.316	16.136	16.636	15.469		0.026	0.028	0.036	0.035	0.033		1	10428-2
9	2016111013724567808	359.737	63.3969	−3.808	−1.16	0.313	16.508	16.979	15.868		0.031	0.033	0.042	0.043	0.04		0.99	10428-2

(This table is available in its entirety in machine-readable form in the online article.)

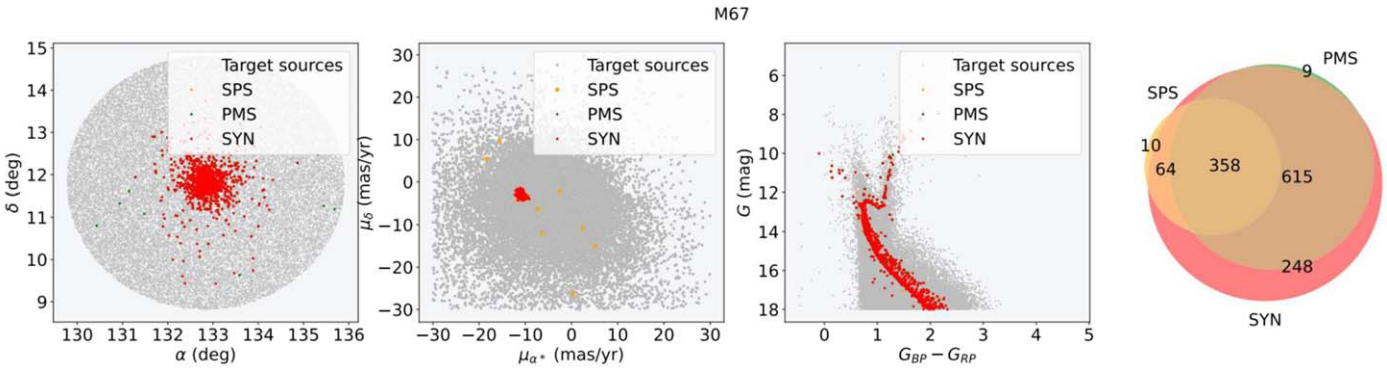


Figure 4. Clustering with SPS (orange dots), PMS (green dots) and SYN (red dots) based on the pre-processed data obtained by the cone query method in Gaia DR3 (target sources).

To further illustrate the effect of clustering in multiple views, we selected a well-studied OC (i.e., M67 (NGC 2682)) whose members have been meticulously studied by many researchers (Castro-Ginard et al. 2020; Agarwal et al. 2021; Jadhav et al. 2021; Ghosh & Sulistiyowati 2022). We downloaded a conical source with a radius of 50 pc around the OC center from Gaia DR3. We pre-processed it according to Section 2. Then, SPS, PMS, and SYN were clustered separately using HDBSCAN with the hyperparameter $m_{\rm clSize}$ set to 4. We obtained 432 (SPS), 982 (PMS), and 1285 (SYN) member stars, respectively (see Figure 4). Compared with the results of Agarwal et al. (2021), the accuracy of the member stars of SPS, PMS, and SYN was 38.1%, 86.7%, and 83.5%, respectively. After voting, the total accuracy of the member stars reached 91.69%.

5.2. The Improvement of the FoF+pyUPMASK+MvC

The application of the FoF and pyUPMASK algorithms has provided a list of cluster candidates with different probabilities. Based on probabilities, Chi et al. (2023b) successfully identified 1760 OCs from high probability candidates. However, are there OCs among the low-probability candidates, and if so, what causes FoF+pyUPMASK to give low probability?

In Figure 5, we analyzed this issue using the NGC (from CG20) star cluster. In the FoF+pyUPMASK clustering method, we kept stars with membership probabilities greater than 0.5 (gray points) after the pyUPMASK membership probability survey. For clusters like NGC 581, FoF+pyUPMASK is ineffective at eliminating mixed-field stars (see left panel), making the cluster's center appear significantly biased. The reason should be the hyperparameter $b_{\rm FoF}$ (Liu & Pang 2019; Chi et al. 2023b) used in FoF+pyUPMASK, which is an empirical value set for simulated cosmology but is ineffective in identifying some clusters. After further use of MvC, we can obtain a clear and clean main sequence, which is relatively consistent with CG20 and fits well with the theoretical isochrones (red dots).

Based on the same approach, we further analyzed other OC candidates. Figure 6 shows three randomly selected samples. The gray dots are the member stars found by FoF. The blue color is the result after further removal of field stars using MvC based on FoF results. The final fitted results are shown in red. Obviously, the MvC algorithm can effectively remove the field stars, thus effectively improving the quality of the sample fitting. The ID of "c1084-32" in Figure 6(a) corresponds to NGC 581 of CG20, and as well as "c2594-16" corresponds to NGC 146 in CG20. This shows that the unidentified OCs in Chi et al. (2023b) are correctly identified after MvC's processing, proving that the judgment of the field stars using MvC is correct. To a certain extent, it solves the problem of low fitting quality of some samples in Chi et al. (2023b).

5.3. Correctness Analysis of the MvC Method

To validate the result of the MvC method, we tested MvC using published OC samples. Agarwal et al. (2021) presented four types of OCs, i.e., NGC 2539, IC 4651, NGC 2141, and Berkeley 18, which have difficulties in member star determinations using the GMM method because those four clusters have a low concentration of cluster members and/or the parameter peaks coincide (Deb et al. 2022). In addition, we also chose NGC 2682 (M67) and Blanco 1 as test OCs, because M67 is well-studied with DR2 and (E)DR3, such as Cantat-Gaudin et al. (2020), Agarwal et al. (2021), Jadhav et al. (2021), Ghosh & Sulistiyowati (2022) and Blanco 1 is a near-OC that has been well-studied with Gaia DR3 (Zhang et al. 2020; Alfonso & García-Varela 2023) recently. For a fair comparison, we reserved the cluster members we identified only for those with $G < 18$ and with the same Gaia dataset (DR3) in works Tarricq et al. (2022, hereafter T22) and Hunt & Reffert (2023, hereafter H23).

Based on these four OCs data as the validation set, we performed the HDBSCAN/subviews method to identify member stars to verify the usability and accuracy of the MvC

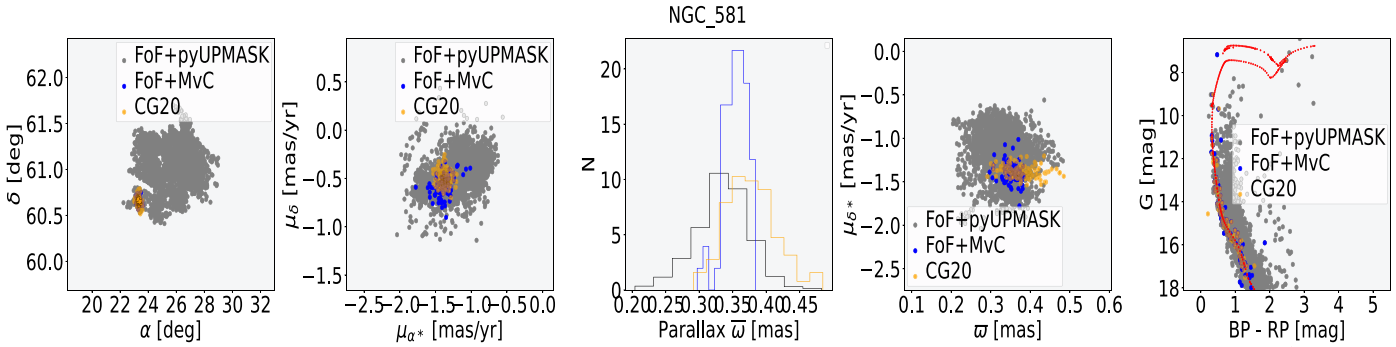
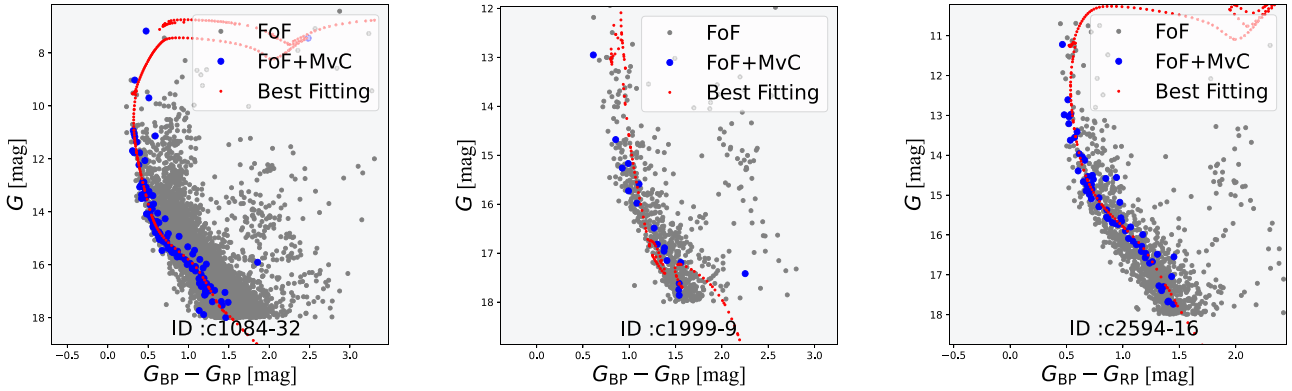


Figure 5. Schematic diagram of improving FoF+PyUPMASK+MvC.



(a) (b) (c)

Figure 6. The diagram of fitting result with/without MvC.

method. We first conducted a cone query around the center of each OC within 2° to obtain the source data sets. We then applied the MvC method to the sources mentioned. The results are shown in Figure 7. To fairly compare the scenes found by different methods, we constrain the distribution range of the member stars in member star matching. The radii are all controlled within the MvC spatial range.

Figure 7 demonstrates that the MvC method can effectively identify most of the member stars (core members) of the four clusters, which are more concentrated in spatial distribution and proper motion distribution. The CMD also indicates that these core members are more tightly distributed in the main sequence.

The third row of Figure 7 affirms that the MvC can capture more diffuse potential members at the same parallax compared with T22 and H23.

The MvC ensemble clustering method can identify most member stars, which is considered an effective method for star cluster member identification. MvC ensemble clustering is able to identify reliable member stars, but for some clusters, it might ignore members of substructures in the core region of the cluster, especially those far from the cluster center.

5.4. Assessment of the Physical Reality of 13 OCs

To ascertain the physical authenticity of the 13 clusters in question, we conducted a comprehensive evaluation based on five critical dimensions as outlined by Piatti et al. (2022). This multifaceted approach includes: (1) Spatial Distribution: Analyzing the spatial arrangement of the clusters to discern whether they exhibit characteristic distribution patterns indicative of genuine OCs. (2) Radial density profile (RDP) Fitting: Assessing the RDP of the clusters to evaluate their density distribution, which confirms their status as authentic OCs. (3) CMD Morphology: Examining the CMD morphology to understand the stellar composition within the clusters, thereby determining their legitimacy as true OCs. (4) Age–Mass Relationship: Investigating the correlation between the age and mass of the stellar entities within the clusters to verify their physical reality as OCs. (5) Mass–Radius Relationship: Exploring the relationship between the mass and radius of the clusters to further substantiate their classification as genuine OCs. This rigorous, multi-dimensional analysis facilitates a more precise determination of cluster authenticity as OCs.

The RDP serves as an essential instrument in the examination of spatial distribution within star clusters. Our

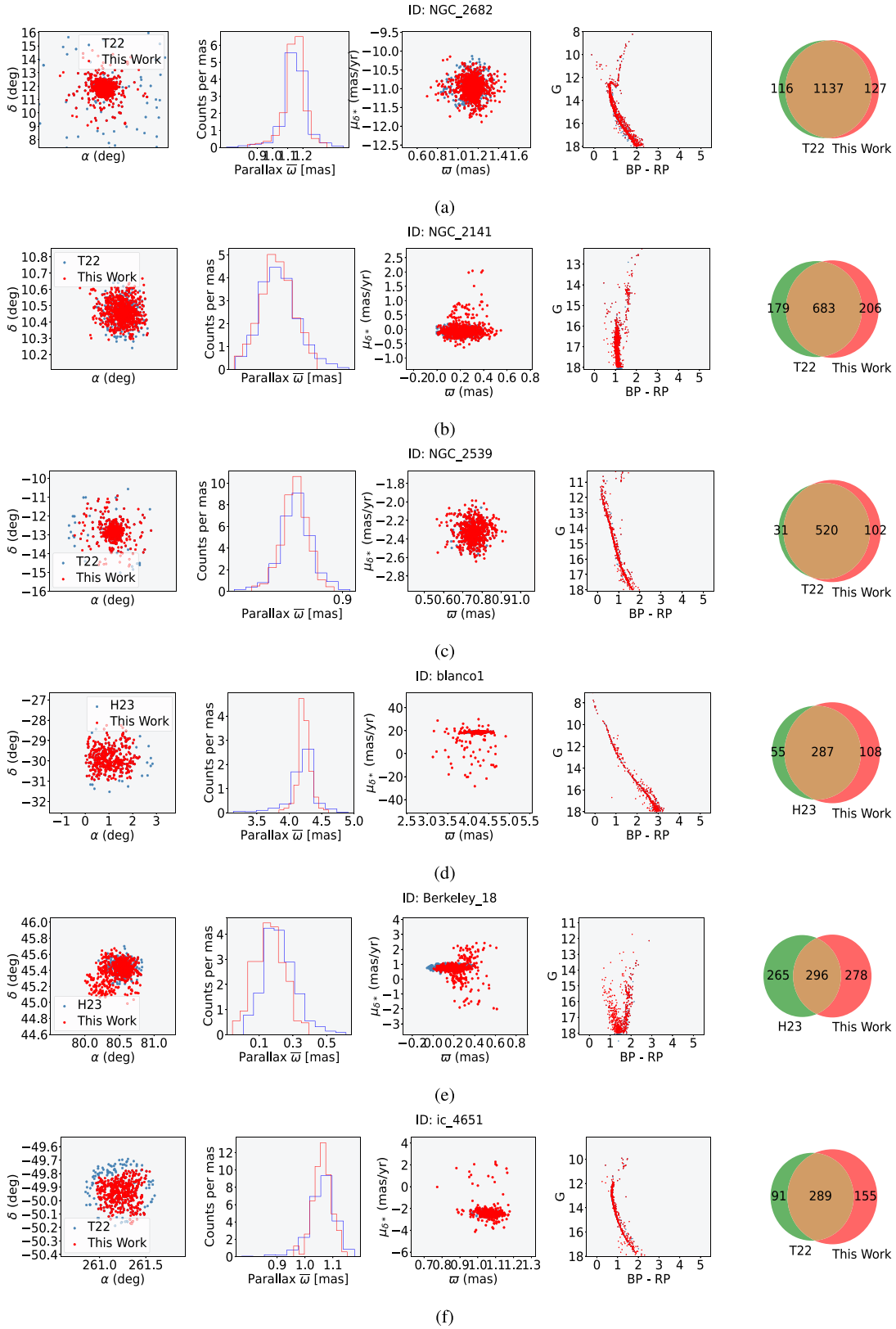


Figure 7. From left to right, the subplots show spatial distribution, proper-motion distribution, and CMD with best-fitting isochrone line and member star matching diagram, respectively.

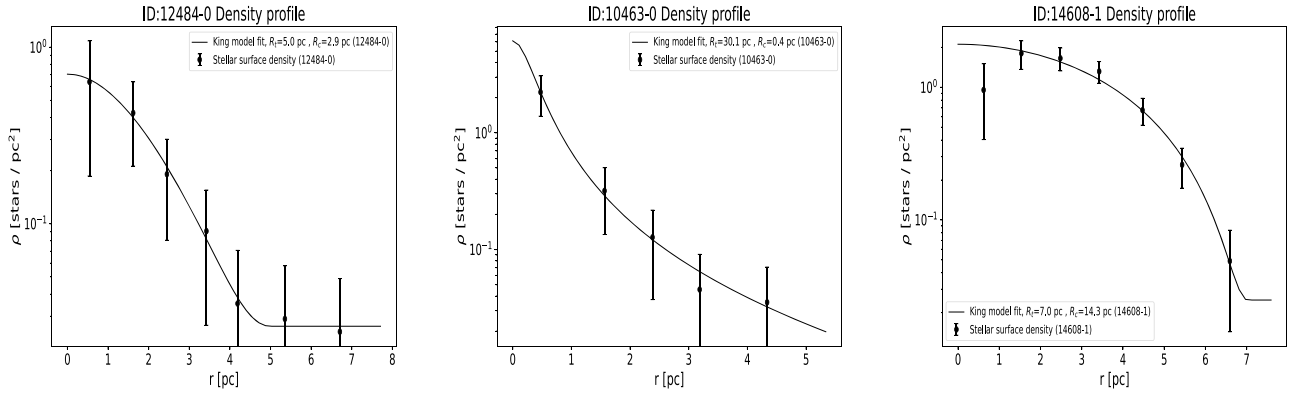


Figure 8. RDP of the cluster's constituent stars is marked by black dots. The black line superimposed on the graph signifies the fitting outcome based on the renowned King model (King 1962).

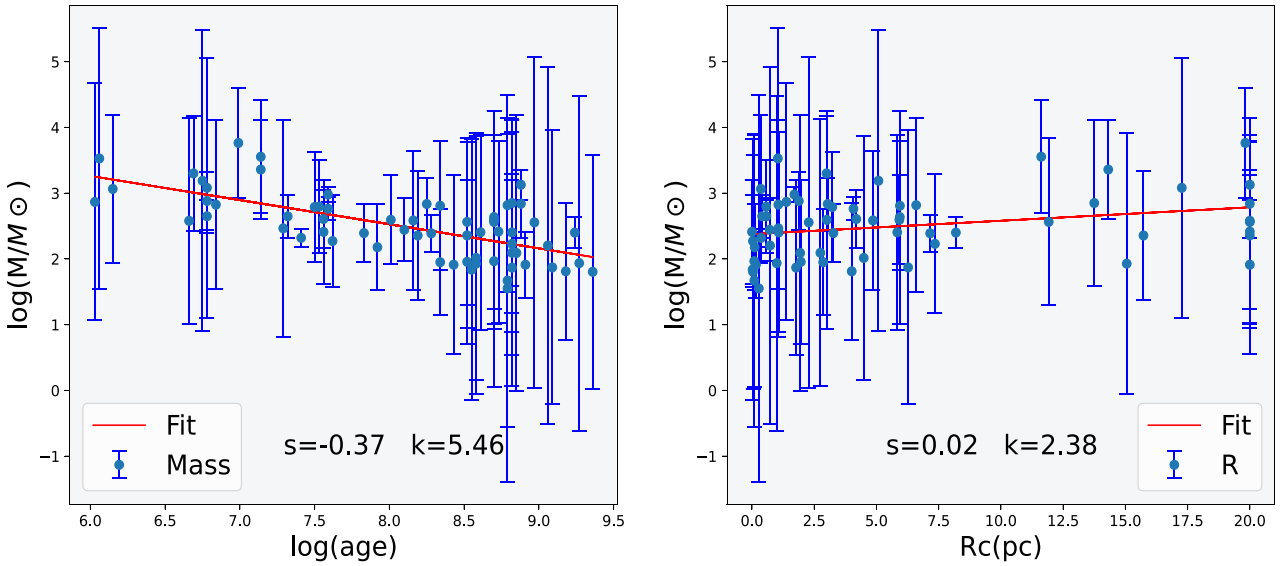


Figure 9. The left panel illustrates the correlation between age and mass, while the right panel depicts the relationship between mass and radius. The error bars are derived from the standard deviation of the mass measurements of the star clusters. "*s*" is the slope fitting value, and "*k*" is the intercept by fitting. OC masses and errors are calculated by the method of Almeida et al. (2023). The diagram includes 68 star cluster samples (consisting of 13 new clusters) randomly selected from the 506 high-confidence star clusters in this study.

methodology begins by establishing the star cluster center as a reference point, or the origin. Subsequently, the cluster is meticulously segmented into a series of concentric "*i*-rings." The procedure then involves calculating the stellar surface density encompassed within the *i*th ring. This is achieved through the formula

$$\rho_i = N_i/\pi(r_{i+1}^2 - r_i^2). \quad (2)$$

Here, N_i represents the count of stars situated within the *i*th ring, which is bounded by the inner radius r_i and the outer radius r_{i+1}. This calculation allows for a detailed analysis of how star density varies across different regions of the cluster, providing valuable insights into its structural composition and potential dynamics. In this study, we have utilized the King (1962) model to accurately determine the stellar surface density values. As shown in Figure 8, the three candidate OCs can be well-fitted by the King function.

In previous research, Joshi et al. (2016) identified an age–mass relationship through an in-depth analysis of nearly 1300 star clusters listed in the Milky Way Star Clusters (MWSC) catalog, which are located within a distance of 1.8 kpc from our solar system. This relationship is mathematically articulated as

$$\log(M/M_\odot) = -0.36(\pm0.05)\log T + 5.3(\pm0.4), \quad (3)$$

where $M_\odot$ signifies the mass of the Sun, T represents the cluster's age, and M serves as cluster mass. The left panel of Figure 9 presents a comparative examination of our empirical findings juxtaposed with the theoretical model proposed by

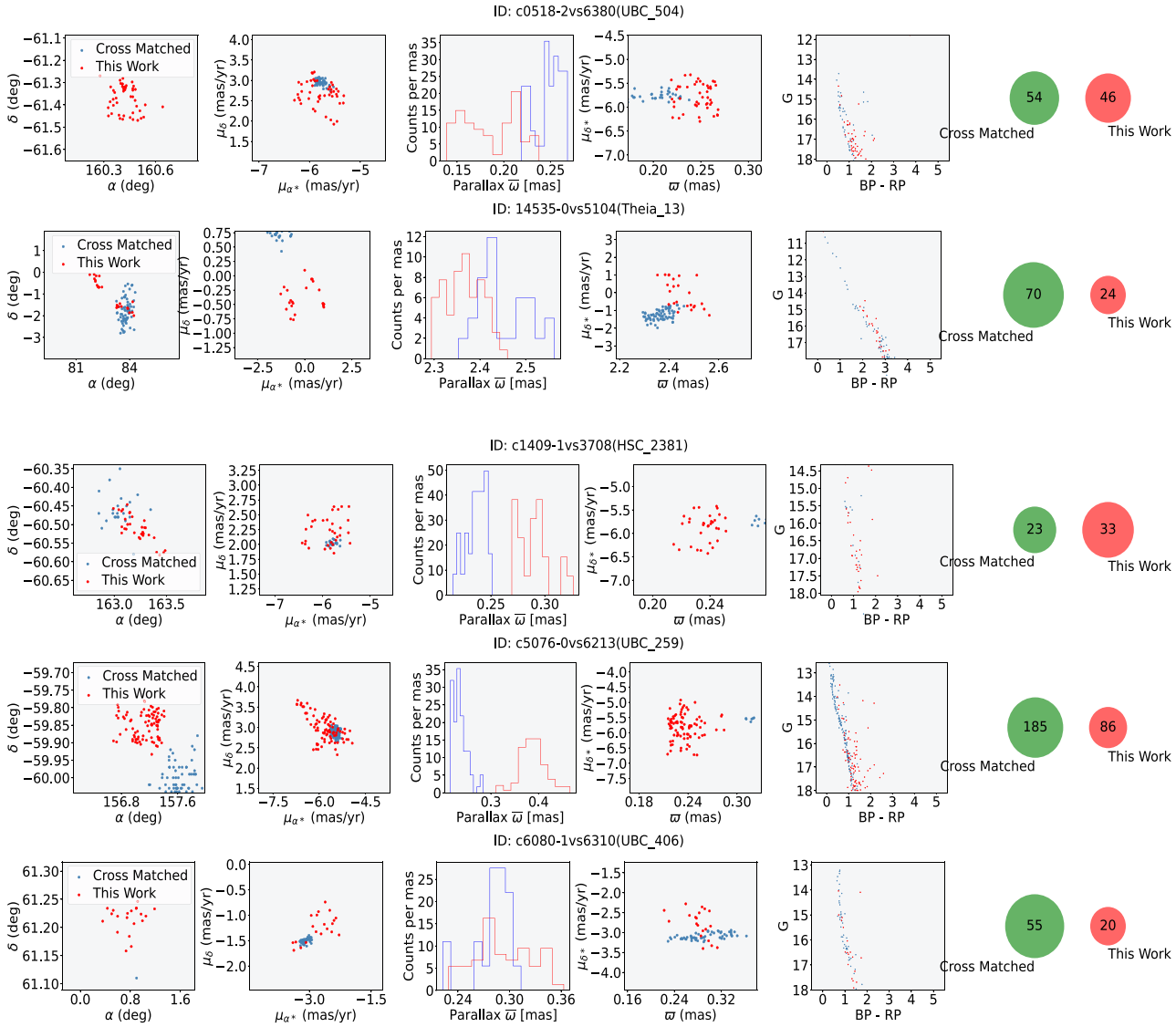


Figure 10. Five BOCs. The blue dots represent the members identified by H23, and the red dots represent the cluster members identified in this work.

Joshi et al. (2016). It is observable that there is a striking congruence between our results and the theoretical function's pattern, thereby validating the reliability of our analysis. Furthermore, examining the mass–radius relationship is essential for gaining insights into the temporal progression and development of the star cluster. Joshi et al. (2016) delineated that the mass–radius relationship in star clusters is characterized by two distinct distribution functions. The first function is linear and is expressed as

$$R = 2.08(\pm 0.10)\log(M/M_{\odot}) - 0.64(\pm 0.27). \quad (4)$$

Here, R represents the cluster's radius, M is its mass, and $M_{\odot}$ is the solar mass, which is used as a unit of measurement. The second function, applicable to clusters situated within the solar orbit, adheres to a power-law relationship, which is described as

$$R \propto M^{1/3}. \quad (5)$$

The right panel of Figure 9 offers a visual representation of the mass–radius distribution for 68 candidate OCs. It is evident that the majority of these clusters are well-fitted by the power-law function, highlighting the prevalence of this relationship in the observed data. Finally, a comprehensive decision table for cluster authenticity is given in Table A1. The comprehensive evaluation shows that the 13 clusters reported have high confidence.

5.5. Limitations of the MvC

Accurately identifying cluster members is challenging work. Uncertainties about membership and stellar properties are the

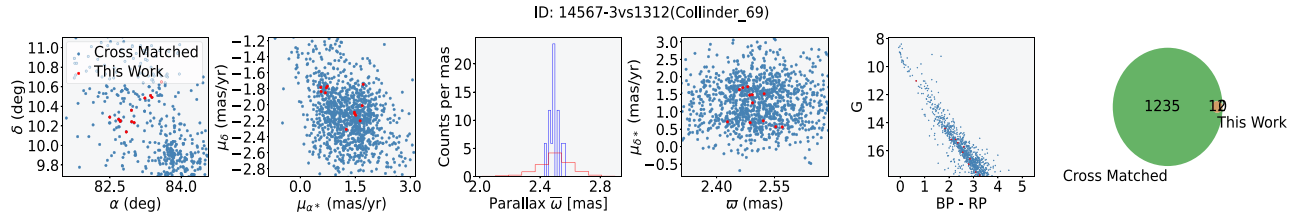


Figure 11. Collinder 69 cross-matched with 14567-3. The blue dots represent the members identified by H23 for Collinder 69, and the red dots represent the cluster members identified in this work.

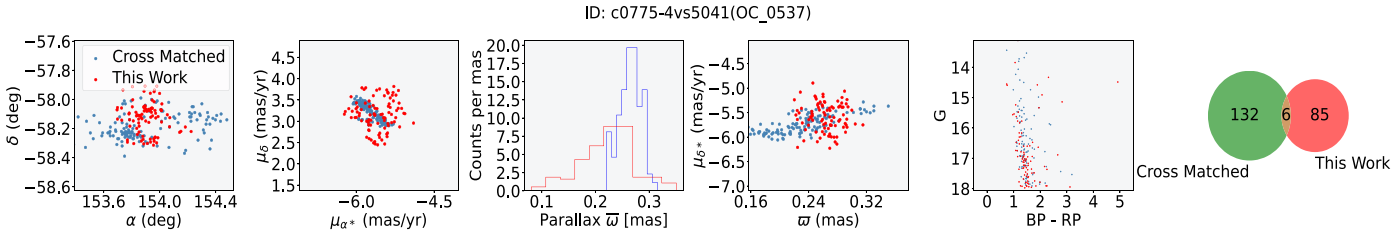


Figure 12. OC 0537 cross-matched with c0775-4. (The blue dots represent the members identified by H23 for OC 0537, and the red dots represent the cluster members identified in this work.)

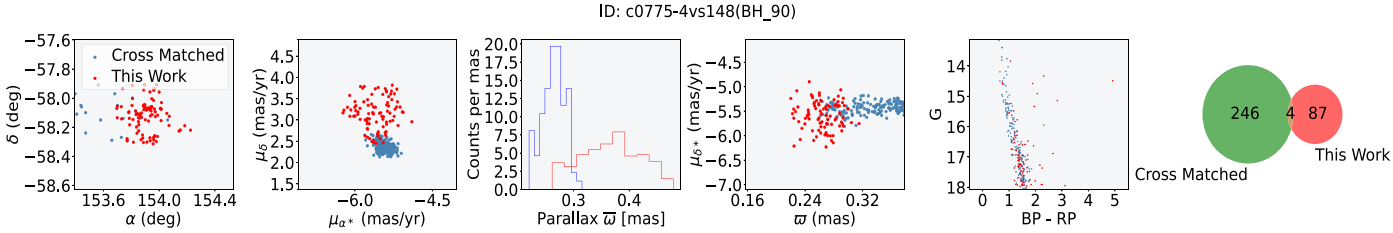


Figure 13. BH 90 cross-matched with c0775-4. (The blue dots represent the members identified by H23 for BH 90, and the red dots represent the cluster members identified in this work.)

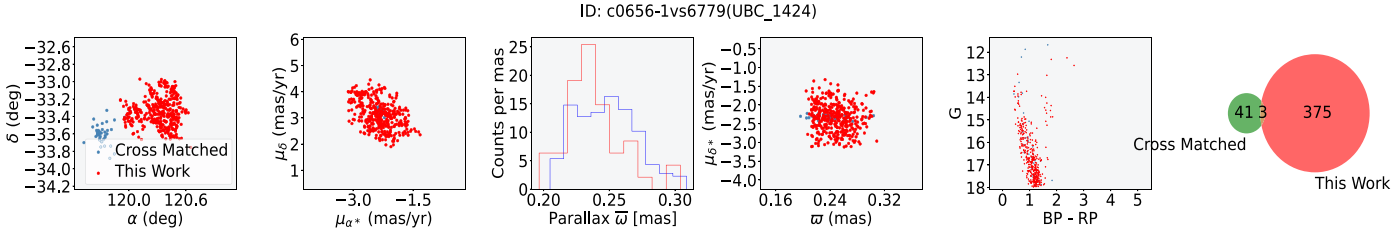


Figure 14. UBC 1424 cross-matched with c0775-4. (The blue dots represent the members identified by H23 for UBC 1424, and the red dots represent the cluster members identified in this work.)

primary issues in identifying cluster members. As described by Penev & Schussler (2022), modeling these uncertainties is difficult. Although MvC has made progress in accurately identifying cluster members, limitations are also obvious. In general, the number of OC member stars based on the FoF +pyUPMASK+MvC algorithm is smaller than the number obtained by directly using DBSCAN or HDBSCAN. This is caused by multiple views, which reduce the number of member stars while improving member star reliability.

We also note that in some cases, MvC identifies members more concentrated in the middle of the OC. Those high-reliability members are usually in the OC's high-density region (the cluster's core body). For this reason, we realize that MvC results have a considerable degree of confidence. However, low-density periphery members might be missed. Members located in the low density of the extended structure of OC, such as "halos," "strings," "coronae," "outskirts," and "tidal tails," need a specific study of a single OC combined with cluster

morphology and dynamics. After using FoF+pyUPMASK +MvC to identify OCs, it is worth considering further employing other methods, such as GMM, to further search for member stars. This will make OC data more accurate.

5.6. Results Analysis

In the cross-matching comparison check of the 13 clusters reported in this study, we identified five binary open cluster candidates (BOCs). Figure 10 presents five pairs of BOCs. Because the focus of this paper is not on the study of binary OCs, these five pairs of binary OCs are worthy of further dynamical study using the method in Li & Zhu (2024) and *N*-body simulations in the future.

The central $0\overset{\circ}{.}5$ matching strategy is too small to be applicable for clusters with a large number of members, and suggesting that it is necessary to expand it. In Figure 11, when conducting a cross-matching of cluster members, although the central coordinates of the cluster 14567-3 and Collinder 69 differ by more than $0\overset{\circ}{.}5$, the members of cluster 14567-3 are clearly part of the known OC Collinder 69. This indicates that the previous work's approach of using a $0\overset{\circ}{.}5$ match based on the cluster center is not applicable for matching corresponding large-scale clusters.

From Figure 3, some previously reported cluster members are evidently incomplete. Figure 12 indicates that the members of the cluster c0775-4 found by our method overlap with the members of the known cluster OC 0537. Considering other views, it is highly likely that they belong to the same cluster. If this is the case, our method has re-identified 62% of the new members. In this work, we are more inclined to consider them as a pair of binary clusters with the same physical evolution, because the number of shared members is only six, which constitutes a small proportion. We believe that the intersection of the six members is likely due to the algorithmic selection effect. Similar to this, there is also OC 0537 cross-matched with c0775-4 (Figure 13), and BH 90 cross-matched with c0775-4 (Figure 14).

6. Conclusions

We proposed an effective clustering method for identifying reliable cluster member stars. The identification results show that the MvC algorithm is capable of effectively identifying reliable member stars. This is the first attempt at using MvC to develop an ensemble clustering technique for hunting star clusters. As a result of re-identification, 13 reliable OCs are reported in this work through isochrone-fitting and visual inspection.

The newly found objects enrich our understanding of the Galactic OC population and indicate that the present OC sample is far from complete. It is anticipated that many new OCs can still be detected through careful observational data analysis.

Acknowledgments

Thanks for the valuable suggestions from the anonymous reviewers. This work is supported by the National Key Research And Development Program of China (No. 2022YFF0711500), the National Natural Science Foundation of China (NSFC, Grant No. 12373097), the Basic and Applied Basic Research Foundation Project of Guangdong Province (No. 2024A1515011503), and the Guangzhou Science and Technology Funds (2023A03J0016). This work is also supported by the Major Key Project of PCL.

Appendix
Additional Figures and Tables

Figure A1 shows a complete list of figures for 13 new OCs. Blue points represent cluster members. The red dotted curves are the best-fitting isochrones. The color bars represent the cluster probability of the member stars calculated by pyUPMASK. Table A1 presents the assessment results of the physical reality of 13 OCs.

Table A1
Assessment of the Physical Reality of the 13 OCs

Cluster ID	Spatial Distribution	RDP	CMD	Age–Mass Relation	Mass–Radius Relation	Adapted
10428-2	Y	Y	Y	Y	Y	Y
10507-1	Y	Y	Y	Y	Y	Y
10706-0	Y	Y	Y	Y	Y	Y
c0410-1	Y	Y	Y	Y	Y	Y
c0514-0	Y	Y	Y	Y	Y	Y
c0518-2	Y	Y	Y	Y	Y	Y
c0656-1	Y	Y	Y	Y	Y	Y
c0775-4	Y	Y	Y	Y	Y	Y
c1409-1	Y	Y	Y	Y	Y	Y
c2915-0	Y	Y	Y	Y	Y	Y
c3059-1	Y	Y	Y	Y	Y	Y
c5076-0	Y	Y	Y	Y	Y	Y
c6080-1	Y	Y	Y	Y	Y	Y

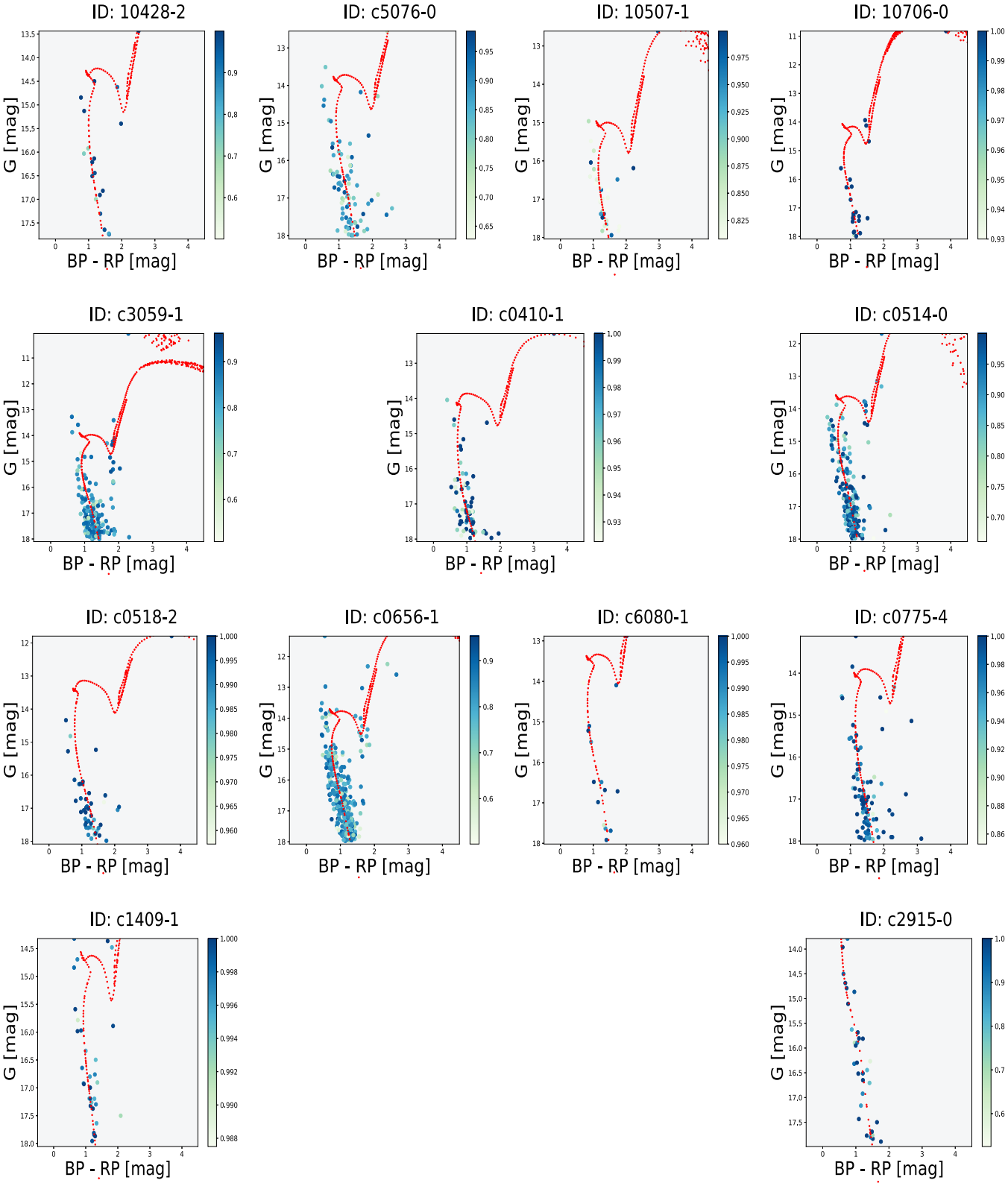


Figure A1. Full list of 13 OCs. Blue points represent cluster members. The red dotted curves are the best-fitting isochrones. The color bars represent the cluster probability of the member stars calculated by pyUPMASK.

ORCID iDs

Huanbin Chi https://orcid.org/0000-0001-7343-7332

References

Agarwal, M., Rao, K. K., Vaidya, K., & Bhattacharya, S. 2021, MNRAS, 502, 2582
Alfonso, J., & García-Varela, A. 2023, A&A, 677, 11
Almeida, A., Monteiro, H., & Dias, W. S. 2023, MNRAS, 525, 2315
Arunima, A., Pfalzner, S., & Govind, A. 2023, A&A, 670, 14
Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. 1999, in Database Theory—ICDT'99: 7th Int. Conf. (Berlin Heidelberg: Springer), 217
Bica, E., Pavani, D. B., Bonatto, C. J., & Lima, E. F. 2019, AJ, 157, 12
Cantat-Gaudin, T., & Anders, F. 2020, A&A, 633, A99
Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. 2020, A&A, 640, A1
Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, A&A, 618, A93
Cantat-Gaudin, T., Krone-Martins, A., Sedaghat, N., et al. 2019, A&A, 624, A126
Casado, J. 2021, RAA, 21, 117
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59
Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., & Balaguer-Núñez, L. 2019, A&A, 627, A35
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, A&A, 635, A45
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2022, A&A, 661, A118
Chang, X., Wang, Y., Li, R., & Xu, Z. 2014, Statistica Sinica, 28, 3
Chi, H., Li, Z., & Zhao, W. 2022, Advances in Intelligent Automation and Soft Computing (Berlin: Springer), 495
Chi, H., Wang, F., & Li, Z. 2023a, RAA, 23, 065008
Chi, H., Wang, F., Wang, W., Deng, H., & Li, Z. 2023b, ApJS, 266, 36
Chi, H., Wei, S., Wang, F., & Li, Z. 2023c, ApJS, 265, 20
Deb, S., Baruah, A., & Kumar, S. 2022, MNRAS, 515, 4685
Dias, W. S., Monteiro, H., Moitinho, A., et al. 2021, MNRAS, 504, 356
Ferreira, F. A., Corradi, W. J. B., Maia, F. F. S., Angelo, M. S., & Santos, J. F. C. J. 2020, MNRAS, 496, 2021
Ferreira, F. A., Corradi, W. J. B., Maia, F. F. S., Angelo, M. S., & Santos, J. F. C. J. 2021, MNRAS, 502, L90
Ferreira, F. A., Santos, J. F. C., Corradi, W. J. B., Maia, F. F. S., & Angelo, M. S. 2019, MNRAS, 483, 5508
Gagné, J., Mamajek, E. E., Malo, L., et al. 2018, ApJ, 856, 23
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, 22
Gaia Collaboration, Drimmel, R., Romero-Gomez, M., et al. 2023, A&A, 674, 35
Ghosh, E. M., Sulistiyowati, Tucio, P., & Fajrin, M. 2022, JPhCS, 2214, 012009
Hao, C., Xu, Y., Wu, Z., He, Z., & Bian, S. 2020, PASP, 132, 034502
Hao, C. J., Xu, Y., Hou, L. G., et al. 2021, A&A, 652, A102
Hao, C. J., Xu, Y., Wu, Z. Y., et al. 2022a, A&A, 660, A4
Hao, C. J., Xu, Y., Wu, Z. Y., et al. 2022b, A&A, 668, 13
He, Z., Li, C., Zhong, J., et al. 2022a, ApJS, 260, 8
He, Z., Liu, X., Luo, Y., Wang, K., & Jiang, Q. 2022b, ApJS, 264, 12
He, Z., Wang, K., Luo, Y., et al. 2022c, ApJS, 262, 7
He, Z.-H., Xu, Y., Hao, C.-J., Wu, Z.-Y., & Li, J.-J. 2021, RAA, 21, 093
Hinrichs, A., Novak, E., Ullrich, M., & Woźniakowski, H. 2014, JCom, 30, 117
Hunt, E. L., & Reffert, S. 2021, A&A, 646, A104
Hunt, E. L., & Reffert, S. 2023, A&A, 673, 31
Hunt, E. L., & Reffert, S. 2024, A&A, 686, A42
Jadhav, V. V., Pennock, C. M., Subramaniam, A., Sagar, R., & Nayak, P. K. 2021, MNRAS, 503, 236
Jaehnig, K., Bird, J., & Holley-Bockelmann, K. 2021, ApJ, 923, 129
Jiang, Z., Yuan, H., & Min, W. U. 2018, Computer Engineering and Applications, 54, 150
Joshi, Y. C., Dambis, A. K., Pandey, A. K., & Joshi, S. 2016, A&A, 593, A116
King, I. 1962, AJ, 67, 471
Krone-Martins, A., & Moitinho, A. 2014, A&A, 561, A57
Li, Z., Deng, Y., Chi, H., et al. 2022, ApJS, 259, 19
Li, Z., & Mao, C. 2023, ApJS, 265, 3
Li, Z., & Mao, C. 2024, RAA, 24, 16
Li, Z., & Zhu, Z. 2024, arXiv:2405.02530
Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021, A&A, 649, A2
Liu, L., & Pang, X. 2019, ApJS, 245, 32
Mužić, K., Almendros-Abad, V., Bouy, H., et al. 2022, A&A, 668, A19
Penev, K. M., & Schussler, J. A. 2022, MNRAS, 516, 6145
Perren, G. I., Pera, M. S., Navone, H. D., & Vázquez, R. A. 2023, MNRAS, 526, 4107
Piatti, A. E., Illesca, D. M. F., Massara, A. A., et al. 2022, MNRAS, 518, 6216
Qin, S., Li, J., Chen, L., & Zhong, J. 2021, RAA, 21, 045
Qin, S., Zhong, J., Tang, T., & Chen, L. 2023, ApJS, 265, 12
Riello, M., De Angeli, F., Evans, D. W., et al. 2021, A&A, 649, A3
Sim, G., Lee, S. H., Ann, H. B., & Kim, S. 2019, JKAS, 52, 145
Tarricq, Y., Soubiran, C., Casamiquela, L., et al. 2022, A&A, 659, A59
Torrealba, G., Belokurov, V., & Koposov, S. E. 2019, MNRAS, 484, 2181
van Groeningen, M. G. J., Castro-Ginard, A., Brown, A. G. A., Casamiquela, L., & Jordi, C. 2023, A&A, 675, 10
Vasiliev, E., & Baumgardt, H. 2021, MNRAS, 505, 5978
Xie, D., Gao, Q., Wang, Q., & Xiao, S. 2019, IEEEA, 7, 31197
Zhang, Y., Tang, S.-Y., Chen, W. P., Pang, X., & Liu, J. Z. 2020, ApJ, 889, 99
Zhao, J., Xie, X., Xu, X., & Sun, S. 2017, Information Fusion, 38, 43
Zhong, J., Chen, L., Jiang, Y., Qin, S., & Hou, J. 2022, AJ, 164, 54