A&A 668, A16 (2022)
https://doi.org/10.1051/0004-6361/202243316
© ESO 2022

Astronomy & Astrophysics

The *Gaia*-ESO Survey: Chemical tagging in the thin disk

Open clusters blindly recovered in the elemental abundance space⋆

L. Spina[1], L. Magrini[2], G. G. Sacco[2], G. Casali[3,4], A. Vallenari[1], G. Tautvaišienė[5], F. Jiménez-Esteban[6], G. Gilmore[7], S. Randich[2], S. Feltzing[8], R. D. Jeffries[9], T. Bensby[8], A. Bragaglia[4], R. Smiljanic[10], G. Carraro[11], L. Morbidelli[2], and S. Zaggia[1]

[1] INAF – Padova Observatory, Vicolo dell'Osservatorio 5, 35122 Padova, Italy
e-mail: spina.astro@gmail.com
[2] INAF – Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Firenze, Italy
[3] Dipartimento di Fisica e Astronomia, Università degli Studi di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy
[4] INAF – Osservatorio di Astrofisica e Scienza dello Spazio, Via P. Gobetti 93/3, 40129 Bologna, Italy
[5] Institute of Theoretical Physics and Astronomy, Vilnius University, Sauletekio Av. 3, 10257 Vilnius, Lithuania
[6] Departamento de Astrofísica, Centro de Astrobiología (CSIC-INTA), 28692 Villanueva de la Cañada, Madrid, Spain
[7] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[8] Lund Observatory, Department of Astronomy and Theoretical Physics, Box 43, 221 00 Lund, Sweden
[9] Astrophysics Group, Keele University, Keele, Staffordshire ST5 5BG, UK
[10] Nicolaus Copernicus Astronomical Center, Polish Academy of Sciences, ul. Bartycka 18, 00-716 Warsaw, Poland
[11] Dipartimento di Fisica e Astronomia, Università di Padova, Vicolo dell'Osservatorio 3, 35122 Padova, Italy

Received 11 February 2022 / Accepted 24 March 2022

ABSTRACT

Context. The chemical makeup of a star provides the fossil information of the environment where it formed. Under this premise, it should be possible to use chemical abundances to tag stars that formed within the same stellar association. This idea – known as chemical tagging – has not produced the expected results, especially within the thin disk where open stellar clusters have chemical patterns that are difficult to disentangle.

Aims. The ultimate goal of this study is to probe the feasibility of chemical tagging within the thin disk population using high-quality data from a controlled sample of stars. We also aim at improving the existing techniques of chemical tagging and giving some kind of guidance on different strategies of clustering analysis in the elemental abundance space.

Methods. Here we develop the first blind search of open clusters' members through clustering analysis in the elemental abundance space using the OPTICS algorithm applied to data from the *Gaia*-ESO survey. First, we evaluate different strategies of analysis (e.g., choice of the algorithm, data preprocessing techniques, metric, space of data clustering), determining which ones are more performing. Second, we apply these methods to a data set including both field stars and open clusters attempting a blind recover of as many open clusters as possible.

Results. We show how specific strategies of data analysis can improve the final results. Specifically, we demonstrate that open clusters can be more efficaciously recovered with the Manhattan metric and on a space whose dimensions are carefully selected. Using these (and other) prescriptions we are able to recover open clusters hidden in our data set and find new members of these stellar associations (i.e., escapers, binaries).

Conclusions. Our results indicate that there are chances of recovering open clusters' members via clustering analysis in the elemental abundance space, albeit in a data set that has a very high fraction of cluster members compared to an average field star sample. Presumably, the performances of chemical tagging will further increase with higher quality data and more sophisticated clustering algorithms, which will likely became available in the near future.

Key words. astrochemistry – methods: statistical – stars: abundances – Galaxy: abundances – Galaxy: disk

1. Introduction

The idea of chemical tagging originated about 20 years ago from the work of Freeman & Bland-Hawthorn (2002), in which it was suggested that destroyed clusters could be found by means of their unique chemical signatures. Thus, the global star formation history of the Galactic disc could be reconstructed, also detecting possible debris of satellite systems. The first principle of the theory is based on the observational evidence of the chemical homogeneity of open clusters and moving groups (De Silva et al. 2006; Sestito et al. 2007; Bovy 2016; Poovelil et al. 2020). The second axiom is about the uniqueness of the chemical pattern of each cluster, and therefore the possibility, using a sufficient number of elements, to find stars born in the same cluster, although it no longer exists (Bland-Hawthorn et al. 2010). Thus, the feasibility of chemical tagging relies on the presence and magnitude of cluster-to-cluster differences (inter-cluster, see Reddy et al. 2020) and on the intra-cluster homogeneity. The chemical tagging based only on the use of chemical information is often called

⋆ Based on observations collected with the FLAMES instrument at VLT/UT2 telescope (Paranal Observatory, ESO, Chile), for the *Gaia*-ESO Large Public Spectroscopic Survey (188.B-3002, 193.B-0936, 197.B-1074).

Article published by EDP Sciences

strong chemical tagging, and aims at recovering each unique site of star formation. A weak version of the chemical tagging aims at recovering broader stellar populations, like the thin and the thick discs, the halo, and their substructures (Martell & Grebel 2010; Hawkins et al. 2015; Wojno et al. 2016; Schiavon et al. 2017; Recio-Blanco et al. 2017).

Progress in recent years has shown that reality is more complex than anticipated. On the one hand, the advent of large spectroscopic surveys, such as *Gaia*-ESO (Gilmore et al. 2012; Randich et al. 2013), APOGEE (Majewski et al. 2017) and GALAH (De Silva et al. 2015), which have devoted significant fractions of their time to observe star clusters, has shown that there can be measurable differences in the abundances of members of clusters in different phases. Those differences can be due to physical reasons, as, e.g., diffusion (see, for instance Bertelli Motta et al. 2018; Souto et al. 2018; Liu et al. 2019), internal dredge-up (e.g., Charbonnel 1995; Charbonnel & Zahn 2007; Lagarde et al. 2012), planet engulfment events (e.g., Maia et al. 2019; Spina et al. 2015, 2018, 2021), or due to analysis techniques (see, e.g., Casali et al. 2020a).

On the other hand, the question of the uniqueness of the chemical pattern of each cluster is still a matter of debate (see, e.g., Blanco-Cuaresma et al. 2015; Price-Jones & Bovy 2018). The most plausible way to test it is definitely by trying to refind members of known clusters, mixed between themselves and with stars belonging to the field. In addition, the uniqueness is closely linked to the elements chosen to characterise the chemical pattern. Indeed, not all elements work equally well: in particular those produced at short scale times by massive stars, such as the α elements, and that have enriched the disk mostly in remote times, vary in similar way in all open clusters, making them of little use as diversity tracers (see. e.g., Ness et al. 2022). On the other side, elements released only recently in the interstellar medium (ISM) by low-mass stars and with a strong metallicity-dependence of their yields (see, e.g., Casali et al. 2020b; Magrini et al. 2021), such as slow neutron capture elements, contribute differently to the chemical composition of the clusters of various ages and located in diverse parts of the disc, and might be better tracers.

There have been several attempts in recent years to recover known stellar associations through clustering analysis in the elemental abundance space or to find the birth sites of stars born in disrupted open clusters. These works have shown disadvantages and advantages of the various types of approach (see, e.g., Mitschang et al. 2014; Ting et al. 2015; Ness et al. 2018). However, some works showed encouraging results. Hogg et al. (2016) were able to identify (by the k-means algorithm) known globular clusters (such as M 13 and M 5), stellar streams, and other undetermined overdensities of stars in the 15-dimensional chemical-abundance space delivered by The Cannon (Ness et al. 2016). Globular clusters were also recovered by Chen et al. (2018) with the Density-Based Spatial Clustering of Applications with Noise (DBSCAN; Ester et al. 1996) algorithm. However, while Hogg et al. (2016) only used the chemical abundance information, Chen et al. (2018) carried out the clustering analysis in a chemical-velocity space. Furthermore, both these works are not "blind", meaning that known members of stellar clusters were used to tune the parameters of the clustering algorithms. Finally, Hogg et al. (2016) highlighted that it is easier for clustering analysis to find structures at low metallicity (e.g., in the Galactic halo) than at solar metallicity (e.g., the Galactic disk). In fact, abundance-space features are more prominent at the edges of the distribution than in the center (see e.g., Ting et al. 2016).

Nevertheless, strong chemical tagging has been tested in the Galactic disk as well. Blanco-Cuaresma & Fraix-Burnet (2018) performed a phylogenetic analysis to recover members of known star clusters, reconstructing most of the original open clusters using differential abundances with respect to M 67, and discussing the best set of elements to recover clusters. Price-Jones & Bovy (2019) identified groups of stars in a synthetic set of birth clusters using DBSCAN and their percentage of clusters with more than 10 members recovered is 40% using only abundances. In a following paper, Price-Jones et al. (2020) repeated the experiment on the APOGEE data set identifying in the abundance space 21 candidate stars clusters, but they were not able to recover any known open cluster. Garcia-Dias et al. (2019) carried out a test on eight clustering algorithms and with four dimensionality reduction strategies applied on a data set composed by members of both open and globular clusters. This work shows how different data analysis strategies must be carefully evaluated before attempting chemical tagging. Finally, Casamiquela et al. (2021) used abundances of member stars in open clusters from high-resolution spectroscopy to recover them with HDBSCAN: they could recover about 30% of the analysed clusters in groups containing at least 40% of the actual cluster members. Notably, these attempts of chemical tagging in the thin disk are non-blind searches of stellar associations within a data set exclusively composed by known members of open clusters. The unique exception in the literature is the study from Price-Jones et al. (2020), which used a sample of 182 538 stars belonging to all the Galactic components. Their inability of recovering any of the known open clusters present in their data set clearly demonstrates the difficulty of the task. This is especially true when noisy data significantly blur the clusters we wish to recover. In fact, Price-Jones et al. (2020) noted how real members of known open clusters have a larger dispersion in the abundance space than those of the 21 groups they identified.

Outline of the paper. Inspired by the study of Price-Jones et al. (2020), in the present work we devise an experiment that simulates a realistic attempt of strong chemical tagging in the thin disk, attempting to improve the existing methodology and giving some kind of guidance. Our full data set is composed by both field stars and members of 40 open clusters. This data set – described in Sect. 2 – is smaller than the APOGEE sample used by Price-Jones et al. (2020), but it constitutes a very controlled set of abundances derived through higher-resolution spectra.

In the first part of the paper (Sect. 3), we use a subset of our data composed exclusively by members of 13 open clusters. The overarching aims of this Section are (i) considering all the most common methods used so far to tackle the problem of chemical tagging, (ii) identifying their pros and cons, (ii) ultimately selecting the one we consider to be the most appropriate for our problem. Firstly, we discuss the rationale behind the choice of the OPTICS algorithm and behind important preprocessing techniques of our data set. Another important point we address is that of the choice of metric, which is often overlooked, assuming the Euclidean one (Hogg et al. 2016; Blanco-Cuaresma & Fraix-Burnet 2018; Price-Jones et al. 2020; Casamiquela et al. 2021), but which can give important improvements in the identification of stars belonging to common formation regions. The choice of the Manahattan metric, which ensures the best contrast in distance between the different points, is perfectly suited to highlight the contrast between clusters and field stars. Then, we focus on defining the abundance space, by tackling the issue of number of elements vs quality

of measurement, and especially the choice of elements based on their uniqueness in terms of nucleosynthesis channels (see, e.g., Adibekyan et al. 2015; Blanco-Cuaresma & Fraix-Burnet 2018). Initial work on chemical tagging considered that between 10 and 15 elements were needed to guarantee uniqueness across populations (Bland-Hawthorn & Freeman 2004; De Silva et al. 2007). We aim at demonstrating with our experiment that we can have some advantage in reducing the overall dimensionality to only the strongest variant dimensions. As already pointed out by Mitschang et al. (2014), one might think that increasing dimensionality would bring improvements. However, in practice, adding elements that are difficult to measure (with few or weak lines, uncertain atomic data, etc.) or elements which duplicate the information since they have a common origin, can only lead to increased uncertainties and scatter.

In the second part of the paper (Sect. 4) we apply the most performing strategies of data analysis to the full data set. First we identify the best set of algorithm's hyper-parameters that maximise the number of stellar associations that are recovered in the abundance space among the 13 open clusters analysed in Sect. 3. The resulting model is then applied to the rest of the data set with the aim of recovering the other known open clusters and other eventual groups of stars of potential interest. With this approach, the algorithm is going to identify in the abundance space a number of clusters, some of which are real stellar associations, while others are just spurious overdensities. Thus, we further investigated only the groups with the highest density in the space of orbital actions. Interestingly, a few of these latter coincide with open clusters that are blindly recovered through our analysis. We also recover new stellar members of open clusters that were lost from the membership analysis. Finally, we identify groups of stars clustering both in the abundance space and in orbital actions, which should be further investigated with follow-up observations.

Finally, in Sect. 5 we provide a summary of our main results and we discuss future perspectives of strong chemical tagging.

In order to avoid confusion with the terminology, in the following sections we will refer to "open cluster" or "stellar association" as an ensemble of stars sharing the same origin and formation site, regardless of whether they are still physically bound or dispersed. Instead, with "cluster" or "group" we will refer to a group of stars identified by the clustering algorithm in the elemental abundance space, regardless of whether they are all part of a real "stellar association" or not.

2. Data set

In this paper we make use of the sixth internal data release (iDr6) of the *Gaia*-ESO survey (Gilmore et al. 2012; Randich et al. 2013). More specifically we use the catalog of the targets observed in the high resolution mode with the spectrograph UVES (resolving power, $R = 47\,000$, and spectral range 480.0−680.0 nm). The catalog includes parameters and abundances for 6917 stars. From this initial sample we only consider the stars with a radial velocity determination, and analysed by the working group (WG) in charge for the analysis of stars with spectral type F-G-K, WG 11 (see Smiljanic et al. 2014). The selected targets are identified with `REC_WG` equal to "WG11". Furthermore, in the data set that we are going to analyse in this paper we only include evolved stars as their abundances are not affected by either atomic diffusion (Souto et al. 2018; Bertelli Motta et al. 2018) or by planet engulfment events (Spina et al. 2021), therefore we select only the stars with surface gravity $\log g \leq 3.5$ dex and effective temperature $4500 < T_{\rm eff} \leq 5500$ K. Also, given that for our analysis we require a controlled sample of very high-quality abundance determinations, we only consider the stars with microturbulence $\xi < 2.0\,{\rm km\,s^{-1}}$, with the number of nodes that provided estimations of iron abundances `N_FEH` > 1, and signal to noise ratio $S/N \geq 50\,{\rm pixel^{-1}}$. In fact, chemical abundances of these stars are more reliable than those derived for the rest of the sample. Since our study is focused on the thin disk population, we consider only the stars with a distance above the Galactic midplane $|z| < 1.5$ kpc (derived using *Gaia* EDR3 distances). In addition to that, we only include stars having abundance determinations of all the following elements: Fe I, Na I, Mg I, Al I, Ca I, Sc II, Ti I, V I, Cr I, Mn I, Co I, Ni I, Cu I, Zn I, Ba II, Y II, Nd II, and Eu II. Finally, we discarded the stars whose parameters had convergence problems in at least one of the analysis nodes. We also discarded the entries for which one node provided results which differs from the all Nodes mean by more than 500 K in $T_{\rm eff}$, or 0.5 dex in $\log g$, or 0.5 dex in [Fe/H]. The list of these low quality results is provided in an internal report of WG11, and are flagged in the final table. Thus, our final sample is composed of 937 stars that fulfill all these criteria.

Our data set still contains a number of stars in globular clusters. Since our experiment is about chemical tagging in the thin disk, we want to remove such objects. Therefore, assuming the membership probabilities $P_{\rm GCs}$ from Vasiliev & Baumgardt (2021) we reject all the stars with $P_{\rm GCs} > 0.9$. This reduces the data set to 800 stars which are either in open clusters or in the field.

With this study we want to evaluate our capability of recovering stars in open clusters hidden in our data set. Therefore, before starting with the clustering analysis, it is of fundamental importance to establish the membership of stars to these associations. In a recent study, Jackson et al. (2022) made use of the *Gaia* astrometric solutions and *Gaia*-ESO radial velocities to determine 3D membership probabilities $P_{\rm 3D}$ of several stars in our data set. When available, here we use these membership probabilities to infer whether or not each single star is member of an open cluster. When these probabilities are not available, which is the case for clusters with few observed stars, we use membership probabilities computed as in Magrini et al. (2021) that consider as members stars within 2-σ from the average radial velocity, proper motion and parallax of the cluster. The first from *Gaia*-ESO iDR6 and the others from *Gaia* EDR3. Given these probabilities, we consider members of open clusters all the stars with $P \geq 0.9$. From these cluster members we discard the stars in clusters younger than 500 Myr. This is necessary because chemical abundances in younger clusters suffer of low accuracy as their metal content appears significantly lower than what it actually is depending on the level of stellar chromospheric activity (Yana Galarza et al. 2019; Baratella et al. 2020, 2021; Spina et al. 2020; Zhang et al. 2021). Therefore, after this further selection we count 306 stars in open clusters. There are 40 open clusters in our data set. Among these, 13 include at least 10 stars: these cluster members will be used to choose the best strategies of data preprocessing and clustering analysis (see Sect. 3) and to select the best set of hyperparameters for the algorithm (see Sect. 4.1). Thus, hereafter we will refer to this sample that counts 190 stars as the optimization sample. In addition to that, there are 23 open clusters with a number of members ranging between 2 and 9. These are the clusters that are hidden in our data set and that we want to "recover" through our analysis. Hereafter we will refer to this sample of 112 stars as the hidden clusters sample. Finally, there are four open clusters with only one member. For practical reasons these four stars will

be considered as field stars along with all the other stars without an estimation of membership probability to any cluster or with $P \leq 0.1$. The resulting sample of field stars counts 456 objects. All the stars with membership probability $P \in (0.1, 0.9)$ are excluded from our data set. The removal of stars with an indefinite membership is necessary to maximise the accuracy of the metrics used to evaluate our ability to recover open clusters through chemical tagging.

3. Methods

The search for clusters in a multidimensional space relies on some fundamental decisions that have to be taken. These decisions, which can significantly improve or deteriorate the efficiency of chemical tagging (see Garcia-Dias et al. 2019), are about the choice of the clustering algorithm, the criteria to set its hyperparameters, the techniques of data preprocessing, and the choice of the space of parameters that needs to be analysed. Unfortunately there is no a general method of data clustering nor a rule of thumb that ensure good results for a reasonably broad range of cases or situations. Instead, the best course of actions is strongly influenced by the data set one has at hand and by the type of cluster one aims to recover. For that reason it is necessary to think very carefully to each step of the analysis and evaluate whether or not it can boost the performances of the clustering algorithm.

In this section we want to explore different approaches for clustering in the abundance space. Our aim is to identify the empirical and conceptual criteria that should drive our choices, and to outline – when possible – general strategies that can improve the efficiency of chemical tagging.

3.1. The algorithm

Cluster is a tricky concept in data science which varies depending on the type of problem that is tackled. For instance, a data set may be formed only by clusters or instead it may contain clusters plus scattered data. Clusters may have the same density profiles or be significantly different from each other. In some cases we may also be interested in detecting sub-clusters within the same agglomerate of data points. Given all this diversity, various techniques of cluster searches have been developed and several clustering algorithms are now available to us. Clusters found by one algorithm will definitely be different from clusters found by another algorithm, however there is no best algorithm for clustering analysis because their outcomes can be evaluated with very different criteria and under various points of view. The best way to get oriented in the vast panorama of machine learning algorithms is always understanding the type of problem that we need to solve and the data set that we have at hand.

First, for practical reasons we restrict the choice of the algorithm to the several clustering techniques that are already implemented in PYTHON. Second, our specific problem – chemical tagging – implies the search of clusters in a data set containing points that do not belong to any agglomerate. Thus, we must exclude the algorithms that are unable to process scattered or noisy instances, such as affinity propagation, agglomerative clustering, k-means, mini-batch k-means, spectral clustering, Gaussian mixing models, WARD, and the balanced iterative reducing and clustering using hierarchies (BIRCH)[1]. Third, as we will demonstrate in Sect. 3.7, open clusters can have different densities in the abundance space when they are seen by large spectroscopic surveys. Therefore, we must also exclude DBSCAN as it assumes that all clusters have the same density. This leaves our choice to two algorithms: OPTICS and HDBSCAN.

The philosophy behind OPTICS and HDBSCAN is very similar. Both these algorithms work by detecting areas where points are concentrated more than other areas that are empty or sparse. These algorithms start the analysis with the first point of the data set and keep expanding to the neighbours points. While considering a specific point in a data set, the algorithm assigns to it a value called reachability distance, which can be understood as a measure of the average density of the neighbourhood around the point under consideration. The complete description of OPTICS and HDBSCAN and how they work is beyond the scope of this paper, but it can be found in the seminal papers by Ankerst et al. (1999), Campello et al. (2013). What is important to highlight here is that both OPTICS and HDBSCAN do not produce a clustering of a data set explicitly, but instead they create an augmented ordering of the data points representing its density-based clustering structure. In practice, all the points of the data set are (linearly) ordered such that spatially closest points become neighbors in the new linear ordering. In addition, the reachability distance assigned to each point gives a measure of the density of its neighbourhood. Therefore, since a cluster can be defined as a region of high density separated by regions of low density, two points that are adjacent in the new linear ordering and that have very different reachability distances can be seen as the "beginning" or the "end" of a cluster. That explains why the new indices of the reordered data set and the reachability distance are the fundamental information used by both OPTICS and HDBSCAN to extract clusters and sub-clusters. Once this information is reached, the two algorithms differ in the technique used to identify these agglomerates.

For our analysis we decide to use the OPTICS algorithm. This choice is driven by a few reasons. First, the OPTICS code that is implemented in PYTHON as part of the `scikit-learn` package allows the user to visualize in a very simple and intuitive way the information that is at the basis of the cluster extraction. This information is fully represented in the reachability plot that shows the reachability distance of every point in the database – with the exception of the starting point for which the reachability distance is not defined – as a function of the new index assigned by the algorithm (see an example in Fig. 1). The clusters are identified as the valleys or dips in the reachability plot, while the saddles are cluster boundaries or noise. Thus, the reachability plot represents a stand-alone tool to get key insights into the distribution of a data set, very useful to evaluate how different data preprocessing steps and different strategies for clustering analysis can improve (or not) the efficiency of the clustering algorithm. This is a very precious piece of information because, as we will see in the following sections, the many insights that we will learn from the reachability plots are directly applicable also to HDBSCAN and other algorithms as well. Furthermore, the OPTICS algorithm is simpler than HDBSCAN. As we will show in Sect. 3.3, OPTICS mostly relies on one important hyperparameter that is unknown a priori. Instead HDBSCAN has more parameters that have to be tuned. Finally, HDBSCAN has been already used in other studies (e.g., Garcia-Dias et al. 2019; Casamiquela et al. 2021). Instead, there are no experiments of chemical tagging with OPTICS. Thus, we want to take advantage of the present study as an opportunity to fill this gap.

[1] For more details on these different methods of clustering analysis, see Rokach & Maimon (2005).

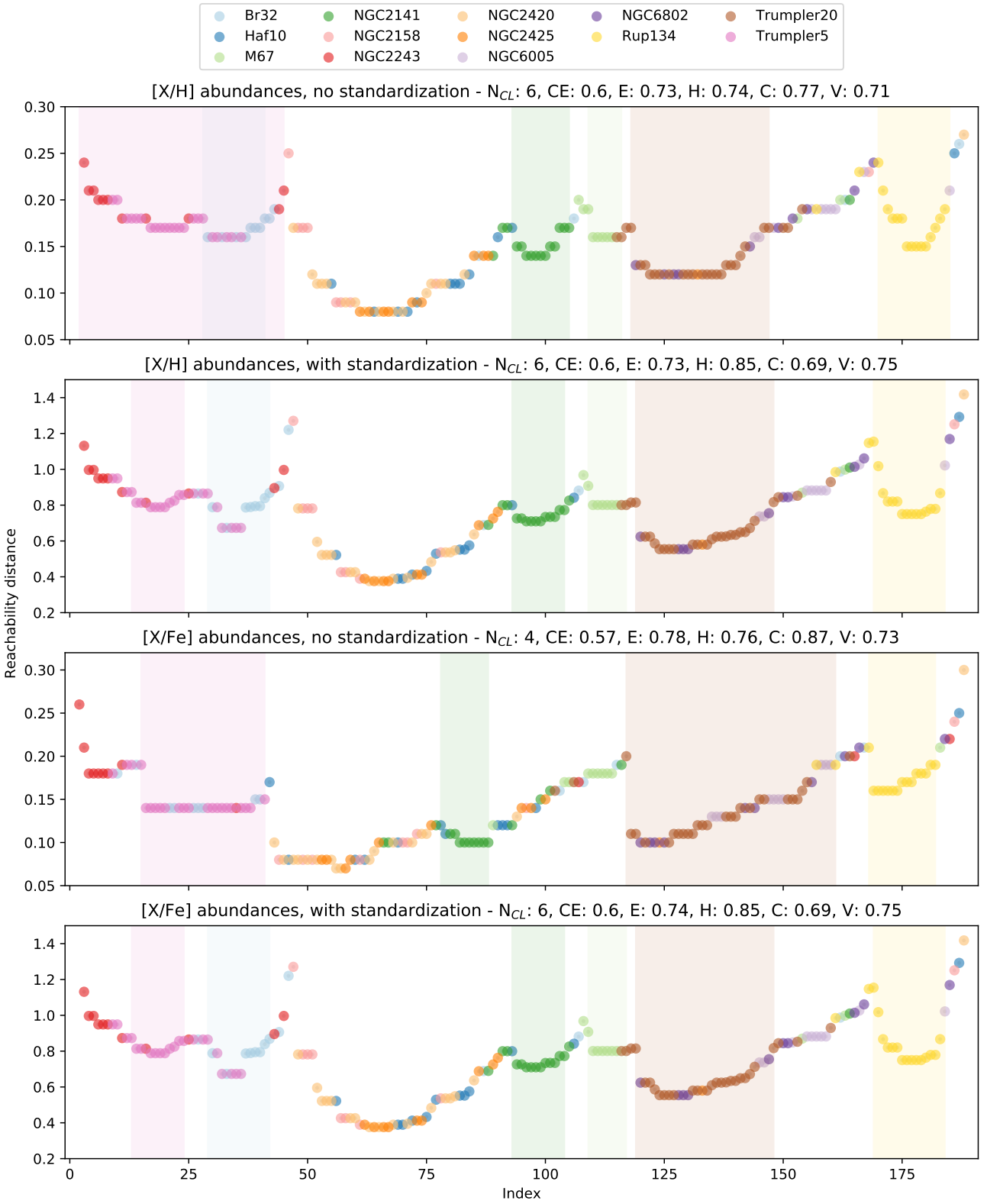

Fig. 1. The panels show the reachability plots resulting from clustering with different strategies of standardization. The points are color coded depending on the cluster they are member of. Vertical colored strips indicate the groups associated to the recovered open clusters. The relevant coefficients and scores are listed on top of each panel. *Top panel*: reachability plot obtained from clustering analysis in the [Fe/H]–[Si/H]–[Ni/H]–[Co/H]–[Ba/H] space without standardization. *Second panel*: reachability plot obtained from clustering analysis in the [Fe/H]–[Si/H]–[Ni/H]–[Co/H]–[Ba/H] space with standardization. *Third panel*: reachability plot obtained from clustering analysis in the [Fe/H]–[Si/Fe]–[Ni/Fe]–[Co/Fe]–[Ba/Fe] space without standardization. *Bottom panel*: reachability plot obtained from clustering analysis in the [Fe/H]–[Si/Fe]–[Ni/Fe]–[Co/Fe]–[Ba/Fe] space with standardization.

3.2. The hyperparameters

The OPTICS algorithm relies on the following hyperparameters:

– `min_samples`. This parameter is strictly related to the reachability distances. In particular, it defines how large a neighbourhood of a specific point is. A small `min_samples` value will result in small neighbourhoods, thus the resulting reachability distances will be more sensitive to noise. Instead, a large `min_samples` value generates large neighbourhoods, with a reachability distance that is more stable against noise, but is also less sensitive to data sparsity. Therefore a "good" value of the `min_samples` is the one that can capture the sparsity of our data set while filtering the noise out. This is the only hyperparameter that we do not set to a fixed value, as it is unknown a priori. Every time that we use OPTICS, we will search for the "best" `min_samples` across the integer interval between 2 and 30. The "best" `min_samples` value is determined accordingly to the evaluation metrics defined in Sect. 3.3.

– `max_eps`. The maximum distance between two samples for one to be considered as in the neighborhood of the other. We assign ∞ to this parameter, so that OPTICS can identify clusters across all scales.

– `metric`. Metric to use for distance computation. In our analysis we use the Manhattan metric. The reason of this choice is explained in Sect. 3.6.

– `xi`. Determines the minimum steepness on the reachability plot that constitutes a cluster boundary. In other words, the `xi` parameter can be intuitively understood as a contrast parameter that establishes the increase in density expected for a cluster relatively to the noise. This parameter directly controls the number and the types of clusters we will obtain. A higher `xi` value can be used to find only the most significant clusters and it could be used to separate – for instance – the thin and thick disk populations in the chemical space. Instead, a lower `xi` value is better suited find less significant clusters, such as open clusters. Given that open clusters are the smallest building blocks of our Galactic disk and that they should be perfectly homogeneous except for the noise intrinsically present in the abundance determinations, we set `xi` to the smallest value that it can take. Namely, we empirically find that any $\texttt{xi} \leq 0.001$ does not affect the results of the clustering analysis applied to our data set. Instead, any $\texttt{xi} > 0.001$ can reduce the number of clusters found by the algorithm. Therefore we set $\texttt{xi} = 0.001$.

– `min_cluster_size`. Describes the number of points required to form an OPTICS cluster. Thus, the value of this parameter depends on the type of clusters the user is looking for. When evaluating the different strategies of clustering analysis, we will aim at extracting at least half of the members of open clusters composed by 10 members or more. Therefore, in this case we will set $\texttt{min_cluster_size} = 5$. Instead, in Sect. 4, when we attempt the proper experiment of chemical tagging, the number of members of each stellar association is unknown, therefore we will set $\texttt{min_cluster_size} = 2$, which is the lowest value it can take.

3.3. The evaluation metrics

The consequence of setting $\texttt{max_eps} = \infty$ and $\texttt{xi} = 0.001$ is that OPTICS finds a hierarchical structure of clusters encompassing from extremely large clusters that include almost all the points in our data set down to very small clusters composed only by few objects. The $\texttt{min_cluster_size} = 5$ filters out the clusters with less than 5 points. We also remove the clusters formed by >100 points. We define as N_g the number of the remaining groups found by OPTICS. Below we also define the evaluation metrics that we use to search the "best" `min_samples` value and to assess different strategies of data analysis.

– `Entropy`. Let $\{idx\}_i$ be the indices assigned by OPTICS to the n_i stellar members of the ith open cluster OC_i (e.g., among the 13 open clusters of the optimization sample). Let $[1 \ldots n_i]$ be the integer interval between 1 and n_i. The entropy score of the ith open cluster is defined as

$$E_i = \frac{\sigma(\{idx\}_i) - \sigma([1 \ldots n_i])}{\sigma(\{idx\}_i) + \sigma([1 \ldots n_i])}, \tag{1}$$

where σ is the standard deviation. The entropy score ranges within [0,1]. A low E_i value indicates that OPTICS has effectively reorganized the data set in a way that members of OC_i are neighbors in the reachability plot. Instead, a high E_i value implies that OPTICS has not efficiently recognised the similarity of the members of OC_i. The median of the E_i derived for all the open clusters defines the global entropy score E.

– `Homogeneity` and `completeness`. Given a group g found by OPTICS, the homogeneity score is defined as

$$H_i = \frac{\text{Number of } \mathrm{OC}_i \text{ members included in g}}{\text{Number of components of g}}. \tag{2}$$

Instead the completeness score is defined as

$$C_i = \frac{\text{Number of } \mathrm{OC}_i \text{ members included in g}}{\text{Number of members of } \mathrm{OC}_i}. \tag{3}$$

These scores are derived only for the open clusters with H_i and $C_i \geq 0.5$. These are the open clusters that – we say – have been recovered by OPTICS. The median of these two scores of the open clusters for which they are defined respectively give the global homogeneity score H and the global completeness score C.

– `V measure`. It is the harmonic mean of two the desirable aspects of clustering: homogeneity and completeness. It is defined as it follows:

$$V_i = \frac{2 \times H_i \times C_i}{H_i + C_i}. \tag{4}$$

The V measure is defined only for the open clusters with H_i and $C_i \geq 0.5$. The median of V_i gives the global V measure V.

– `Number of recovered OCs`. This is the number N_{rec} of open clusters that have been recovered by OPTICS. An open cluster OC_i is "recovered" when both H_i and C_i are ≥ 0.5.

– `Clustering Efficiency`. This score is defined as the ratio between the number of clusters that have been recovered and the number of groups found by OPTICS:

$$\mathrm{CE} = \frac{N_{\mathrm{rec}}}{N_{\mathrm{g}}}. \tag{5}$$

Clustering algorithms are extremely sensitive to the choice of their hyperparameters. As we have discussed in Sect. 3.2, we fix them all with the only exception of `min_samples`. This is a very delicate parameter that has a great impact on the final results. When `min_samples` is too high, the algorithm does not detect any variation in density across the data set. As a consequence, the distribution of points in the reachability plot is flat, and the algorithm does not find any group. On the other hand, when `min_samples` is too low, the reachability plot would show a significant level of granularity. Therefore, the code would probably find the real clusters along with a myriad of other groups in the noise. Thus, tuning the `min_samples` parameter to very low values would give the idea that the algorithm is able to recover a

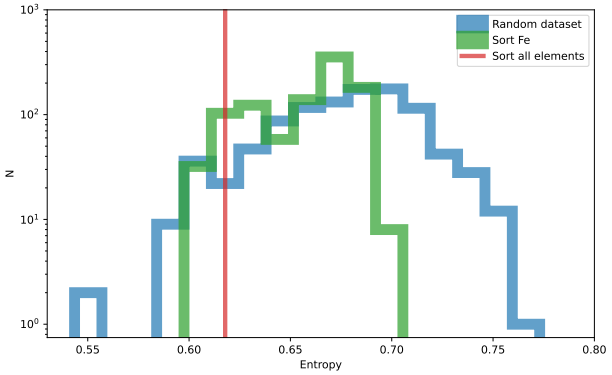

Fig. 2. The entropy values resulting from the 1000 trials of clustering analysis carried out by OPTICS on a data set that is randomly ordered (blue histogram), and on a data set that was sorted as a function of [Fe/H] (green histogram). Finally the red vertical line shows the entropy that is obtained when the data set is sorted as a function of the [X/H] abundances, starting with [Fe/H] followed by all the other elements.

large fraction of clusters, however that may happen because the algorithms is generating several possible combinations of groups (e.g., Casamiquela et al. 2021) and not because it is able to efficiently recognize patterns in the data set. It is not desirable a model that detects several clusters, many of which are noise and just a few are real. That model is unable to properly generalise when applied to a real data set. Therefore, in order to avoid overfitting, when evaluating the response of OPTICS to the variation of `min_samples` we only consider the cases where CE is above a certain threshold. Among these cases alone we select the `min_samples` value that maximises $N_{\rm rec}$. When the clustering analysis is applied to the optimization sample, we impose CE > 0.5. In other words, we consider only the cases when more than half of the groups identified by OPTICS correspond to real open clusters. Instead, when analysing the extended data set which also includes field stars and the hidden clusters, it is reasonable to assume a lower threshold because the number of stellar associations that one can possibly recover is larger than the number open clusters that we use for the optimization (for more details see Sect. 4).

3.4. Reproducibility

The distance- and density-based algorithms – such as k-means, DBSCAN, HDBSCAN and OPTICS – are not entirely deterministic. The sequence of points in the linearly re-ordered data set, thus also the stars belonging to each group, slightly depend on how the data set is initially ordered. Therefore, to obtain results that are fully reproducible and objective it is strictly necessary to define a criterion that unequivocally pre-orders the data set before it is fed to the clustering algorithm. All that also implies that there must be a way to pre-order the data set that can make the algorithm more efficient than just analysing a random sequence of instances.

In Fig. 2 we show the distribution of the entropy coefficients obtained by repeating 1000 times the clustering analysis on a randomly ordered data set (in blue), and on a data set sorted as a function of [Fe/H] (in green). At each iteration, the clustering analysis is carried out following the recipe detailed in Sect. 3.3 and using all the 19 elemental abundances as the input features. When the data set is randomly ordered the resulting entropy spans a large range of values between 0.55 and 0.80. However, when we sort input data set as a function of Fe abundance, the entropy distribution is mostly restricted between 0.6 and 0.7. Finally, when we sort the data set as a function of [Fe/H] and of all the other abundances [X/H], we always obtain the same entropy of 0.62 (see red line). This last entropy is one of the lowest value that we have obtained across all the previous iterations. Therefore, pre-ordering the data set as a function of [X/H], starting with [Fe/H] and followed by all the other elements, is a good strategy to improve the performances of the algorithm. In light of these results, we will apply this data preprocessing step in the following analysis.

3.5. Standardization

Standardization is a scaling technique in which a variable's distribution is transformed in order to have its mean equal to zero and a unit standard deviation. Data standardization makes data dimensionless. By standardizing one attempts to give all variables an equal weight, in the hope of achieving objectivity. In fact, if one feature has a range of values much larger than the others, clustering would be completely dominated by that single feature. Therefore, standardizing data is recommended because otherwise the range of values in each feature will act as a weight when determining how to cluster data, which is undesired in most cases. However, in practice standardization is not always strictly necessary.

In Fig. 1 we compare the effects of standardization in the analysis of our data set. Points are color coded depending on the open cluster they are member of. In the first panel, we show the reachability plot derived by OPTICS which has searched for clusters in the abundance space [Fe/H]–[Si/H]–[Ni/H]–[Co/H]–[Ba/H]. No standardization is applied to the data set before the analysis. Instead, the second panel shows the reachability plot obtained through the same analysis, but on the standardised data set. On top of each panel we list the metrics and coefficients, while the colored bands highlight the clusters that have been recovered by OPTICS.

Solely based on the number of recovered open clusters and the other metrics and coefficients, standardization does not produce a significant improvement in the performances of the clustering algorithm. This is because all the features in the abundance space that we have used span similar ranges of values and have similar distributions. Therefore the OPTICS will give them similar weights even without standardization. However, from a more careful analysis of the reachability plots it is possible to notice that standardization allows us to detect smaller density variations within each cluster. This is particularly evident for Trumpler 20, whose members mostly share the same reachability distance when no standardization is applied. Instead, when data are standardized the densest region of the cluster, corresponding to the points with a lower reachability distance, is more easily identifiable.

In order to make the benefits of standardization appear more evident we use the following features: [Fe/H]–[Si/Fe]–[Ni/Fe]–[Co/Fe]–[Ba/Fe]. In this case [Fe/H] spans a much larger range of values than all the other features. In fact, when we apply OPTICS without any standardization, we obtain much worse results (see the third panel): the algorithms is able to recover only four open clusters instead of the six of the previous analysis. In this case standardization can significantly improve the efficiency of the algorithms (see the fourth panel).

In conclusion, although there are cases when standardization would not produce significant improvements in terms of the

metrics that we are using to evaluate our results, this preprocessing technique will always positively contribute in finding overdensities in the abundance space. Therefore, it should always be applied even when features have very similar distributions. These conclusions apply for OPTICS, and many other clustering algorithms (e.g., k-means, DBSCAN, HDBSCAN, etc.).

Hereafter, we always standardize the data set before the clustering analysis.

3.6. The metric

The metric is another critical ingredient in clustering analysis. The metric determines how distances between points are estimated. Therefore the contrast in distance between points in a dataset heavily depends on which metric is chosen. All previous experiments on chemical tagging through cluster analysis in the abundance space have used the Euclidean metric (Hogg et al. 2016; Blanco-Cuaresma & Fraix-Burnet 2018; Price-Jones et al. 2020; Garcia-Dias et al. 2019; Casamiquela et al. 2021), which is certainly the most familiar one and for that reason it could appear as the most natural choice. However, there are cases where the Manhattan metric may be preferable to Euclidean distance. In fact, the Euclidean metric is much more sensitive against to noise than the Manhattan metric. That is readily evident when considering how the two distances between points $\boldsymbol{p}$ and $\boldsymbol{q}$ is calculated. The Euclidean distance is defined as

$$d(\boldsymbol{p}, \boldsymbol{q})_{\mathrm{E}} = \sqrt{\sum_i (p_i - q_i)^2}, \tag{6}$$

while the Manhattan distance is

$$d(\boldsymbol{p}, \boldsymbol{q})_{\mathrm{M}} = \sum_i |p_i - q_i|. \tag{7}$$

Because of the square exponent, the Euclidean metric gives more emphasis to the outliers regardless of whether they are just noisy data or not. Instead, the Manhattan is more stable against noise.

Certainly there are applications where a high sensitivity to noise is desirable, for instance when the features are measured with a precision that is much higher than the typical separation of clusters. Unfortunately, this is not our case. Chemical abundances from spectroscopic surveys are obtained through automatic pipelines built to analyse, in a reasonable amount of time, thousands of spectra with very different signal-to-noise ratios and from stars having a broad range of atmospheric parameters, rotational velocities, and compositions. Thus, Galactic surveys typically prioritise large statistics over precision and accuracy. Therefore, given the unavoidable presence of noise, the Euclidean metric blurs clusters in the abundance space more than the Manhattan metric. On top of that, when we are working in a space that has more than a few dimensions – which is exactly our case – the Manhattan metric is the one that ensures the best contrast in distance between the different points (Aggarwal et al. 2001), hence also the best density contrast between clusters and empty regions.

In Fig. 3, we compare the results obtained though the two different metrics. The abundance space considered here is defined by the [X/H] abundances of all the 19 elements (Fe, Na, Mg, Al, Si, Ca, Sc, Ti, V, Cr, Mn, Co, Ni, Cu, Zn, Ba, Y, Nd, and Eu). We notice some key differences between the outcomes of the two analysis. First, the Manhattan distance gives a smaller entropy score E, meaning that with this metric the data set is more efficiently ordered. This is expected as Manhattan is less affected by noisy data than the Euclidean metric. Second, the reachability distance in the Manhattan metric spans a wider range of values than in the Euclidean metric. As a result of this higher contrast in the distances between points, dips and valleys in the reachability plot obtained with the Manhattan metric will be deeper than those derived with the Euclidean distance. When dips and valleys are deeper, it is easier to identify clusters in the data set.

In conclusion, the Manhattan metric is a better option than the Euclidean for our specific application. This conclusion holds even for other distance- or density-based clustering algorithms, such as k-means, DBSCAN, HDBSCAN, etc. Therefore, in the rest of this paper we will always use Manhattan distance.

3.7. The elemental abundance space

The definition of the elemental abundance space is one of the most important decisions that has to be taken before applying the clustering analysis. The option that seems the most obvious and natural is to use all the elemental abundances that one has in hand to define the space where to look for clusters. That strategy is the one typically used in the past (Hogg et al. 2016; Chen et al. 2018; Price-Jones et al. 2020; Casamiquela et al. 2021).

The first panel in Fig. 4 shows the results obtained with all the 19 elemental abundances that have been measured in our sample. As we notice, the results are not entirely satisfactory as only three open clusters out of 13 have been recovered. Instead, when we decrease the dimensionality of the abundance space to the abundances [X/H] of only five elements (i.e., Fe, Si, Ni, Co, and Ba), the algorithm is able to recover up to six open clusters (see second panel). This significant difference in the performance is due to some common difficulties arising when attempting clustering analysis in a high-dimensional space. These issues are known as "the curse of dimensionality" (Bellman 1957).

First, when we increase the number of dimensions, our data becomes more sparse: every new dimension increases the volume of the abundance space, giving our data a higher differentiation chance. As a result, it is also more difficult to find groups of stars sharing a common chemistry.

Second, as we increase the number of dimensions, the average distance between two points of our data set increases as well. Therefore, the relative distances between points get blurred. These two effects are visible by comparing the first two panels in Fig. 4. For instance, while the distribution of members of the cluster Ruprecht 134 is flat in the first panel, it clearly shows a dip in the second panel. Meaning that in a lower-dimensional space the algorithm is more efficient in recognizing overdensities.

Third, as we increase the number of dimensions, we likely introduce features that do not add much value to our model. When this is the case, the model will learn from noisy or irrelevant features. This can lead to a reduction in performance of the clustering algorithm, especially when abundances are derived with a limited precision such as in the case of those produced by large spectroscopic surveys.

On top of the "the curse of dimensionality", there are two other aspects that they should always be considered, although they do not affect algorithms' performance. Namely, working in a higher-dimensional space can significantly increase the time of the analysis and also lead to more complex models that are harder to interpret.

In conclusion, the abundance space defined by all the 19 elements contains a redundant number of dimensions. The simplest

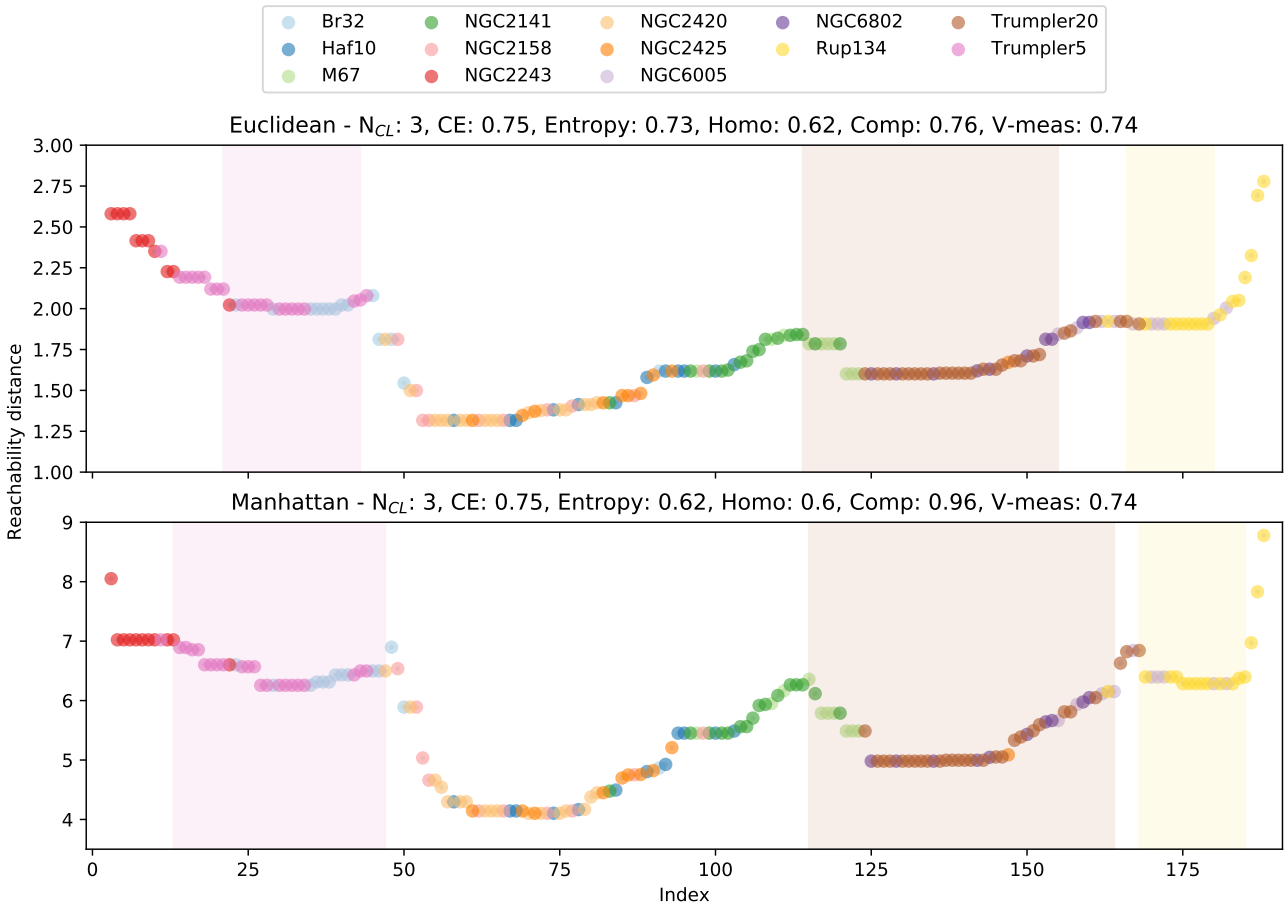

Fig. 3. The panels show the reachability plots resulting from clustering with different metrics. The points are color coded depending on the cluster they are member of. The relevant coefficients and scores are listed on top of each panel. *Top panel*: reachability plot obtained with the Euclidean metric. *Bottom panel*: reachability plot obtained with the Manhattan metric.

way to define a space with a smaller number of dimensions is to identify which elements that are the most meaningful among those that we have detected. To do so, we calculate for each element the Calinski-Harabasz score (Caliński & Harabasz 1974), which is defined as the ratio of the sum of between-cluster dispersion and of within-cluster dispersion for all clusters. These scores are calculated on all stars belonging to the optimization sample. The values are listed in the second column of Table 1. The higher is the Calinski-Harabasz score and the more the [X/H] abundance can differentiate between members of different clusters. Note that this score is also influenced by the typical uncertainty of abundance determinations for each element. In fact, under the assumption that stellar association are chemically homogeneous, a small within-cluster dispersion implies a high precision in the abundance determination. In other words, a chemical element with very uncertain abundance determinations is expected to have a low Calinski-Harabasz even though it is highly distinctive of different stellar associations. Thus, the elemental abundances with the highest Calinski-Harabasz score are expected to be the most distinctive of stellar associations and the most precisely determined.

However, considering the elements with the highest Calinski-Harabasz score is not enough. In fact, two elements with an equally high score might have been produced from the same nucleosynthetic channel (e.g., see Mg and Si), therefore the inclusion of both of them would not add much information to the system (Ness et al. 2022). Since each element has its own story to tell, in addition of using the Calinski-Harabasz score, we also want elements mostly produced though different nucleosynthetic processes (Ting & Weinberg 2022). The main nucleosynthetic origin of each element is inferred from Kobayashi et al. (2020) and they are listed in the third column of Table 1: Core Collapse supernovae (CC), Type Ia supernovae (Type Ia), HyperNovae (HNe), and Asymptotic Giant Branch stars (AGB). Given the CH score and the nucleosynthetic channel, we can select the five most meaningful elements among the 19 available: Si, Fe, Co, Ni, and Ba. By reducing the dimensions of the abundance space, from 19 to five, the performances of the OPTICS algorithm significantly increase, as it is evident from Fig. 4-mid panel. Note that these five elements are not necessarily the best ones to recover the largest number of clusters from our data set, however, they certainly produce better results than just simply taking all the elements. Although our choice of the most meaningful elements is driven by some objective criteria, one could also adopt a more heuristic approach increasing or decreasing the number of features, testing different combinations of elements and empirically verifying how the result change. What we want to highlight here is just that, for our specific problem, we observe a significant improvement in performance when we select fewer more meaningful features.

There are other techniques specifically designed to automatically reduce the dimensionality of a data set. The most popular one certainly is the Principal Component Analysis (PCA; Pearson 1901). PCA projects the original space of $n_{\rm orig}$ parameters onto

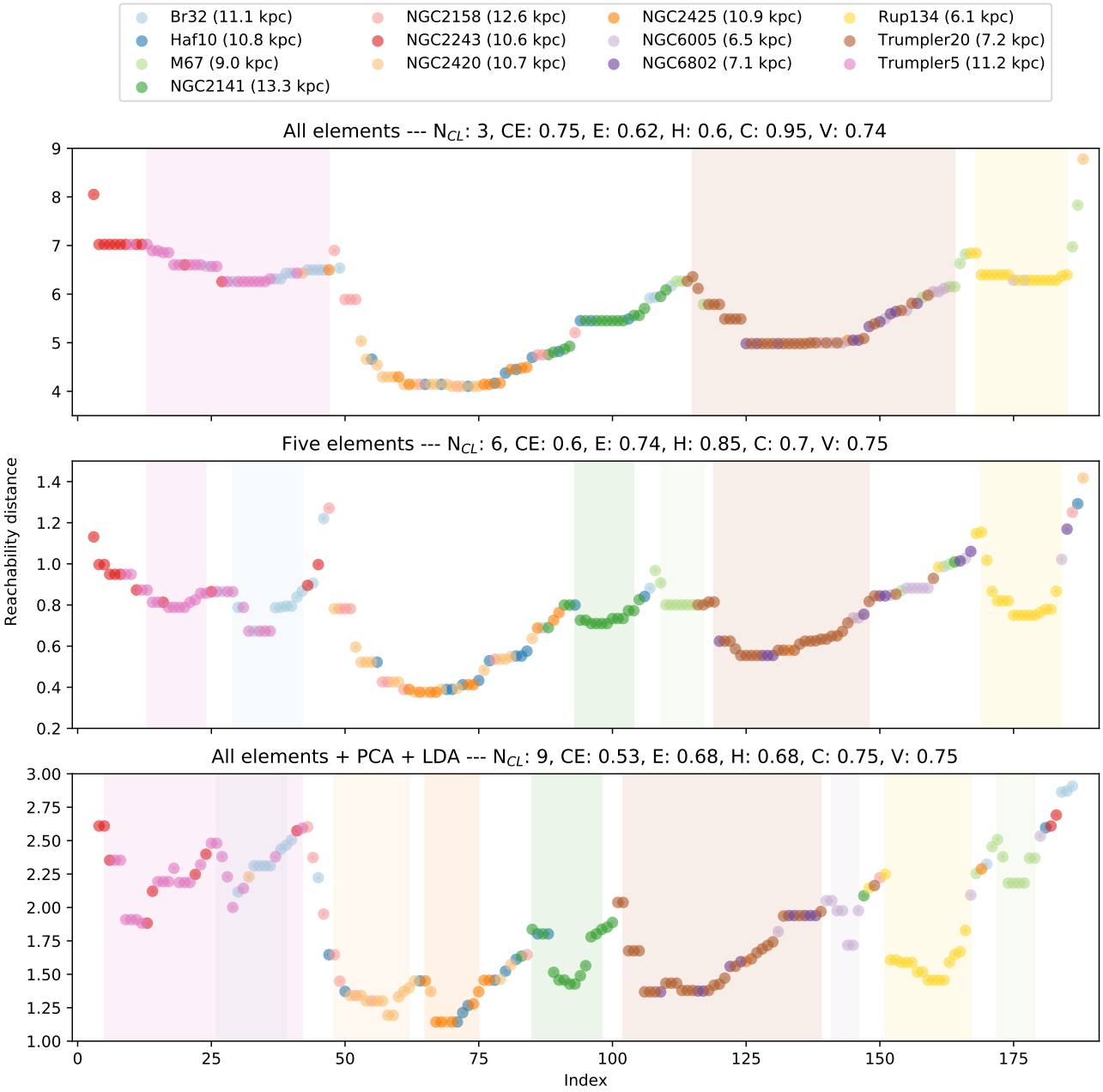

Fig. 4. The panels show the reachability plots resulting from clustering on different abundance spaces. The points are color coded depending on the cluster they are member of. The relevant coefficients and scores are listed on top of each panel. *Top panel*: reachability plot obtained from the 19-dimensional space defined by the chemical abundances of Fe, Na, Mg, Al, Si, Ca, Sc, Ti, V, Cr, Mn, Co, Ni, Cu, Zns, Ba, Y, Nd, and Eu. *Mid panel*: reachability plot obtained from the 5-dimensional space defined by the chemical abundances of Fe, Si, Ni, Co, and Ba. *Bottom panel*: reachability plot obtained from the 5-dimensional space found by PCA and LDA.

a new set of $n_{\rm PCA} \leq n_{\rm orig}$ features which captures the maximum variation in the data set. In other words, while the original data set may contain variables that are not particularly meaningful because they are highly correlated, the PCA finds a new space of parameters where these correlations are minimized. However, capturing the direction of maximum variation in the data set does not necessarily lead to a larger separation of groups in the new space of parameters and, therefore, to better results.

There are other algorithms for dimensionality reduction that are better suited than PCA for clustering analysis, such as the Linear Discriminant Analysis (LDA; Fisher 1938). By analysing the distribution of known groups (or classes) of points in the original space of parameters, LDA finds a new set of $n_{\rm LDA} \leq n_{\rm orig}$ features that maximizes the ratio between class variation and within-class variation and that can best characterize all the different classes. As a result, the different classes will be more separated in this new space of parameters. That is exactly what can boost the performances of a clustering algorithm.

The main difference between PCA and LDA is illustrated in the example of Fig. 5. The top panel shows the distribution of points belonging to three different classes in a 2D space. We also overplot the projection axis found by the PCA (in red) and LDA (in blue). While PCA projects data on the axis with the highest variation, LDA projects on the axis that ensures a better

Table 1. Criteria used to select the most meaningful elements.

Element	CH score	Nucleosynthetic origin
Na I	205	CC
Mg I	232	CC
Al I	156	CC
Si I	379	CC
Ca I	101	CC and Type Ia
Sc II	136	CC
Ti I	69	CC
V I	142	CC and Type Ia
Cr I	170	CC and Type Ia
Mn I	189	Type Ia
Fe I	390	CC and Type Ia
Co I	173	HNe
Ni I	303	Type Ia
Cu I	27	HNe
Zn I	141	HNe
Y II	59	AGB
Ba II	81	AGB
Nd II	29	CC
Eu II	20	CC

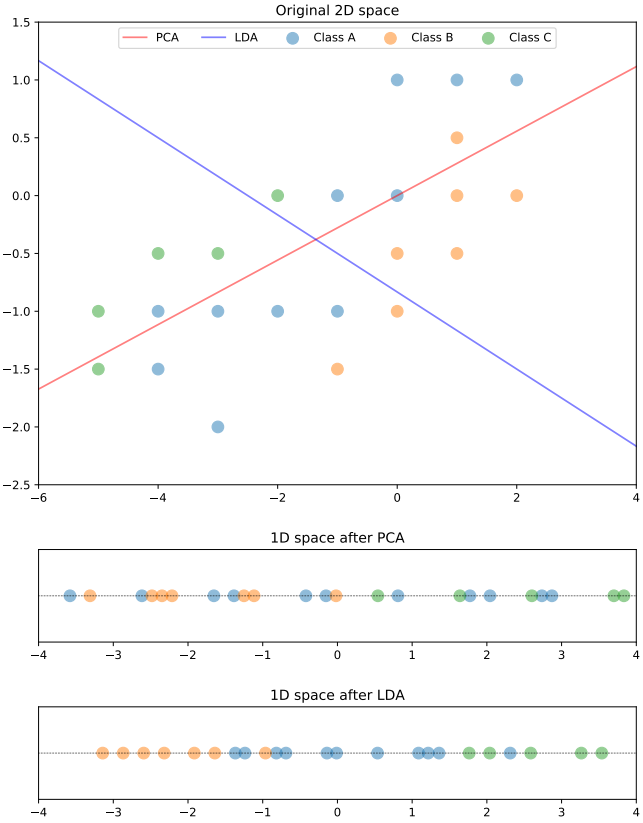

Fig. 5. *Top panel*: distribution of points belonging to three different classes in a 2D space. The lines represent the projection axis found by PCA (in red) and LDA (in blue). *Middle panel*: distribution of points in the 1D space found by the PCA. *Bottom panel*: distribution of points in the 1D space found by the LDA.

separation of the three classes. The results of the two different approaches for dimensionality reduction are shown in the panels below: while there is a significant overlap between the three classes in the 1D space produced by PCA (mid panel), LDA has found a new space that is much better suited for classification problems and clustering analysis (bottom panel). However, we also stress that there are two possible difficulties in using LDA instead of PCA. First, while PCA is an unsupervised technique, LDA needs to be trained on a set of points whose classes are previously known. In principle this is not always possible. However, in practice, to train LDA here we can use the known members of the open clusters that have been excluded from optimization data set. The second difficulty is due to the fact that LDA very easily tends to overfit the data especially when applied on a large number of highly correlated variables. One way to tackle this issue is to use PCA to reduce dimensionality first and then apply LDA only to the principal components. Here we test this latter strategy and we demonstrate that it can further improve the performances of the clustering analysis.

First, we apply the PCA in order to reduce the dimensions of the abundance space from $n_{\rm orig} = 19$ to $n_{\rm PCA} = 8$. Second, we apply the LDA to further reduce the dimensions to $n_{\rm LDA} = 5$. The LDA is trained on the 18 open clusters of the hidden data set and that are older than 0.8 Gyr. Note that these open clusters have ages up to 7 Gyr and Galactocentric distances roughly between 6 and 15 kpc. For each of these open clusters we calculate the mean abundances and standard deviations relative to all the 19 elements. Given this information, we randomly drawn 1000 points from the multivariate Gaussian distribution defined for each open cluster in the 19-dimensional abundance space. These 18 000 points are those used to train the LDA. However, before that, we have applied the same standardization and PCA models used on the optimization data set. This is necessary in order to ensure that the training set is preprocessed in exactly the same way as the data set what we want to analyse. Once this latter has been transformed by the LDA model, we apply the standard clustering analysis whose results are shown in the bottom panel of Fig. 4. This technique of reducing the number of dimensions of the abundance space through PCA + LDA outperforms all our previous tests. Now we are able to recover 9 open clusters out of 13. This result is extremely satisfactory, as it also outperforms similar attempts of recovering open clusters in the abundance space (e.g., Casamiquela et al. 2021).

4. Blind chemical tagging in the thin disk

The analysis of the optimization sample carried out in the previous section has allowed us to determine the best strategies and preprocessing steps that can improve chances of chemical tagging. Instead, in this Section we perform a completely new study using the full data set described in Sect. 2. The overarching aim of this additional analysis is to conduct an experiment that simulates – to the extent possible – a realistic attempt of strong chemical tagging in the thin disk. There is no need to say that all the lessons learnt in the previous Section are applied and exploited here with the objective of achieving higher performances from the clustering analysis. In practice, (i) we make use of the OPTICS algorithm, (ii) we pre-order and standardize the data set, (iii) we use the Manhattan metric, and (iv) we reduce the abundance space dimensions with PCA+LDA.

A typical data set that one could use for chemical tagging is composed by a number of known members of open clusters, plus a number of stars with no association to any open cluster. The first serve as a control sample, necessary for tuning the algorithm's hyperparameters and testing the efficacy of the analysis.

Instead, the stars with unknown association include field stars and members of the hidden associations that we wish to recover. Our data set, as it is described in Sect. 2, includes 190 stars from the optimization sample which are members of the 13 open clusters. These stars will be our control sample. Furthermore, our data set also includes 112 stars from the hidden cluster sample which are members of 23 open clusters. These will be the associations that we will try to identify in the elemental abundance space. These cluster members will be "diluted" in a sample of 456 field stars, also included in our data set.

Therefore, the goal of this new analysis is to recover as many stellar associations as possible among those of the hidden cluster sample. This will be a blind search, as the information on the stellar membership to the open clusters that we wish to recover is never used during the analysis. The only information we are going to use is the one related to the membership of the stars from the optimization sample, which will be useful to set the algorithm's hyperparameters.

4.1. Searching the hyperparameters

There are three hyperparameters that need to be tuned: `min_samples`, $n_{\rm PCA}$, and $n_{\rm LDA}$. They are searched within the following intervals: `min_samples` $\in [2 \ldots 30]$, $n_{\rm PCA} \in [5 \ldots 15]$, and $n_{\rm LDA} \in [5 \ldots n_{\rm PCA}]$. All the other hyperparameters are set to the values listed in Sect. 3.2. Before applying the clustering analysis we follow the preprocessing steps described above. Namely, we sort and standardize the data set, we train the LDA on the optimization sample and then we apply the dimensionality reduction algorithms PCA + LDA similarly to the procedure described in Sect. 3.7. One significant difference from our previous analysis is that LDA is now trained on the optimization sample. The OPTICS algorithm is then applied to the whole data set of 767 stars including the field stars, the hidden cluster sample, and the optimization sample. Given the fact that this data set contains four times the number of stars used for the analysis in Sect. 3, here we relax the CE threshold considering all the solutions with CE ≥ 0.05. Therefore, we tune the hyperparameters in order to maximize the number of recovered open clusters among those of the optimization sample in a data set which includes also the field stars and the hidden cluster sample. The hyperparameters satisfying this criterion are the following: `min_samples` $= 3$, $n_{\rm PCA} = 11$, and $n_{\rm LDA} = 8$.

Specifically, using these hyperparameters we are able to recover four open clusters at $H \geq 0.5$ and $C \geq 0.5$ from the optimization sample. These are NGC 2420, NGC 2141, NGC 2425, and Trumpler 5. The resulting reachability-plot is shown in Fig. 6 from which we can observe that other open clusters from optimizing sample – such as Berkeley 32, Ruprecht 134, M 67, and Trumpler 20 – are correctly ordered by OPTICS although they are not fully recovered at the $H \geq 0.5$ and $C \geq 0.5$ threshold. In fact, the inclusion of field stars and members of other open clusters into this data set, which is much larger than that analysed in Sect. 3, is a factor of "confusion" for the clustering algorithm, which is thus able to recover a smaller number of open clusters. On the other hand, we also notice that there are dips and valleys filled by stars that are either field stars or members of the hidden cluster sample. This indicates that the algorithm is able to detect other overdensities among the rest of the data set.

4.2. Clustering analysis

Once we have found the best set of hyperparameters, we proceed with the clustering analysis on the data set exclusively composed by the 456 field stars plus the 112 stars from the hidden clusters sample. In fact, our goal is to recover as many open clusters as possible from those of the hidden clusters sample. Therefore, there is no need of further considering the optimization sample which is then removed from the analysis.

The preprocessing steps are the same of Sect. 4.1, including the LDA which was trained on the 13 open clusters from the optimizing sample. Also, the hyperparameres are those found in Sect. 4.1: `min_samples` $= 3$, $n_{\rm PCA} = 11$, and $n_{\rm LDA} = 8$. The reachability-plot resulting from this analysis is shown as a whole in the top panel of Fig. 7, while the four panels below show smaller windows of 150 stars each. In these latter, the actual members of the open clusters we wish to recover are clearly indicated as colored circles. This information is not used in the analysis, but it helps the reader to distinguish the stars we are targeting from the field stars represented as black points. In fact, in Fig. 7 we notice that stars have been ordered in a way that members of the same open cluster are – in many cases – aligned to adjacent positions. This is mostly evident for open clusters such as Berkeley 31 (index ~ 45), NGC 3960 (index ~ 110), Trumpler 23 (index ~ 200), NGC 4337 (index ~ 210), and Berkeley 21 (index ~ 515). The reachability plot also shows several dips and valleys, indicating the presence of overdensities in the abundance space. As expected, some of these overdensities are mainly formed by members of the same open cluster, such as in the cases of the stellar associations mentioned above. However, we also notice that other dips in the reachability-plot are mostly – or even entirely – composed by field stars. Therefore, our data set may contain other interesting groups of stars characterised by a similar chemistry. On the other hand, it is also possible that these latter groups are just overdensities detected in the noise.

4.3. Analysis of the overdensity groups found by OPTICS

Though our analysis we identify 122 groups of stars clustered in the abundance space. Some of them may be possibly related to real stellar associations. However, there must also be several other groups that are actually spurious clusters due to noise. Furthermore, given that OPTICS identifies clusters in a hierarchical structure, some of the groups produced by the algorithm must be too high in this hierarchy and be composed by two or more subclusters. For instance, stellar associations sharing a similar origin within the disk are likely to be very close to each other in the reachability plot and eventually form a unique big valley. Eventually these big valleys are then further divided in smaller dips, which represent the actual open clusters. For instance, although NGC 2420 and NGC 2425 are two individual open clusters, they are enclosed within the same valley in the reachability plots of Fig. 5. That happens because they share a similar composition, probably because they formed at similar $R_{\rm GC}$ and have similar ages (Cantat-Gaudin et al. 2020). However, the aim of our analysis is to identify the smallest groups of stars sharing the same identical origin. We are not aiming at more generalized stellar populations, such as the thin and thick disk, or the inner and outer disk. Therefore, we need some criteria to separate the groups of potential interest from those that are just spurious or not relevant.

In order to do so, we first carry out a selection based on the number of stars included in each groups found by OPTICS. Thus, among the initial 122 groups, we consider only those composed by a number of members smaller than 20. This allows us to

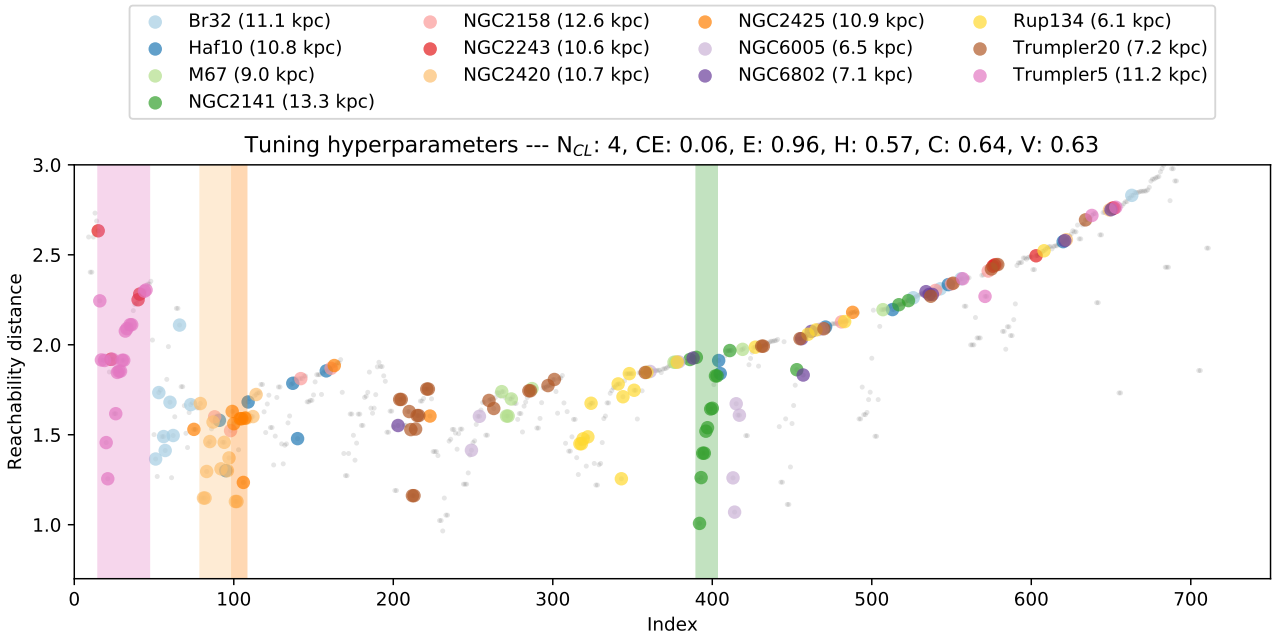

Fig. 6. The figure shows the reachability-plot resulting from the clustering analysis of the data set comprising the 190 stars of the optimizing sample, 112 stars of the hidden cluster sample, and the 465 stars of the field stars sample. The open cluster members of the optimizing sample are shown with coloured points. All the other stars are represented as gray points. The solution represented here corresponds to the one that maximises the number of recovered clusters from the optimizing sample ($N_{\rm cl}$). The overdensities found by OPTICS corresponding to the recovered open clusters are highlighted with vertical coloured bands.

reject all the groups that are too high in the hierarchy. In fact, given the extent and the nature of our data set, it is extremely unlikely that it contains by chance more than 20 members of a single stellar association. This criterion reduces our sample to 77 groups. A further selection is made by studying how the members within each of the remaining groups are distributed in the space of orbital actions $L_{\rm Z} - J_{\rm R} - J_{\rm Z}$. Namely, the groups that are denser than a certain threshold will be considered as possible stellar association, while all the others will be discarded.

A potential issue in studying the dispersion of different groups in the space of actions is due to the fact that the distributions of stars in $J_{\rm R}$ and $J_{\rm Z}$ are strongly peaked at low values (see top panels in Fig. 8). That poses a problem when using simple measures of spread, such as the standard deviation or the median absolute deviation (MAD), which take a symmetric view of the dispersion giving equal importance to negative and positive deviations from the centre of the distribution. Therefore, in order to alleviate this problem we apply a Yeo-Johnson power transformation (Yeo & Johnson 2000) to the $J_{\rm R}$ and $J_{\rm Z}$ variables using the `PowerTransformer` function from the `Scikit-learn` PYTHON library. As we can observe in Fig. 8-mid panels, the transformed $J_{\rm R}$ and $J_{\rm Z}$ distributions are now more similar to Normal distributions, although they are not entirely Gaussian. Finally, we standardize the three variables using `RobustScaler` from `scikit-learn`. The resulting distributions are shown in Fig. 8-bottom panels. These new variables define the space of actions $L^*_{\rm Z} - J^*_{\rm R} - J^*_{\rm Z}$ that we use to discriminate whether a group found by OPTICS contains a potential stellar association or not.

For each i-group we calculate the MADs in the three variables $\{{\rm MAD}(L^*_{{\rm Z},i}), {\rm MAD}(L^*_{{\rm Z},i}), {\rm MAD}(L^*_{{\rm Z},i})\}$. The same quantities are derived for each of the 13 open clusters from the optimization sample. Then, we define a metric D_i that we use to discriminate between groups of interest and those that will be rejected, based on the typical internal dispersion that we observe within the open clusters of the optimization sample. Namely, D_i is defined as it follows:

$$D_i = \sqrt{\left(\frac{{\rm MAD}(L^*_{{\rm Z},i})}{T_{L_{\rm Z}}}\right)^2 + \left(\frac{{\rm MAD}(J^*_{{\rm R},i})}{T_{J_{{\rm R},i}}}\right)^2 + \left(\frac{{\rm MAD}(J^*_{\rm Z})}{T_{J_{\rm Z}}}\right)^2}, \quad (8)$$

where $T_{\rm X}$ are threshold values for each action coordinate. They are defined as the average of the MADs calculated for the 13 open clusters in the optimization sample plus three times the standard deviation:

$$T_{\rm X} = \langle\{{\rm MAD}(X_{{\rm OC}_j\,{\rm opt}})\}\rangle + 3 \times \sigma(\{{\rm MAD}(X_{{\rm OC}_j\,{\rm opt}})\}), \quad (9)$$

where $X_{{\rm OC}_j\,{\rm opt}}$ can be one of the coordinates $L^*_{\rm Z}$, $J^*_{\rm R}$, $J^*_{\rm Z}$ observed for the members of the jth open clusters of the optimization sample. In other words, we use the $\{D_i\}$ metric to rank the dispersion observed in the $L^*_{\rm Z} - J^*_{\rm R} - J^*_{\rm Z}$ space for each of the 77 groups based on what is observed in the optimization sample. All groups with $D_i > 1$ are rejected, while the 14 groups with $D_i \leq 1$ are considered as groups of interest or potential stellar associations. The location of these 14 groups is indicated in the reachability plots of Fig. 7 with vertical shaded bands.

The number of sigmas used to define $T_{\rm X}$ is arbitrary. A very small number would allow one to recover in the abundance space only the stellar associations that are still physically bound. A very large number would allow one to discover stellar associations that are now fully disrupted, but it would also produce several spurious groups. Here we opt for 3-sigmas, which is an intermediate value.

4.4. The 14 groups of interest

Here we provide a detailed analysis of the stellar content of the 14 groups identified through our method. Our main aim is testing whether or not we are able to recover known clusters hidden

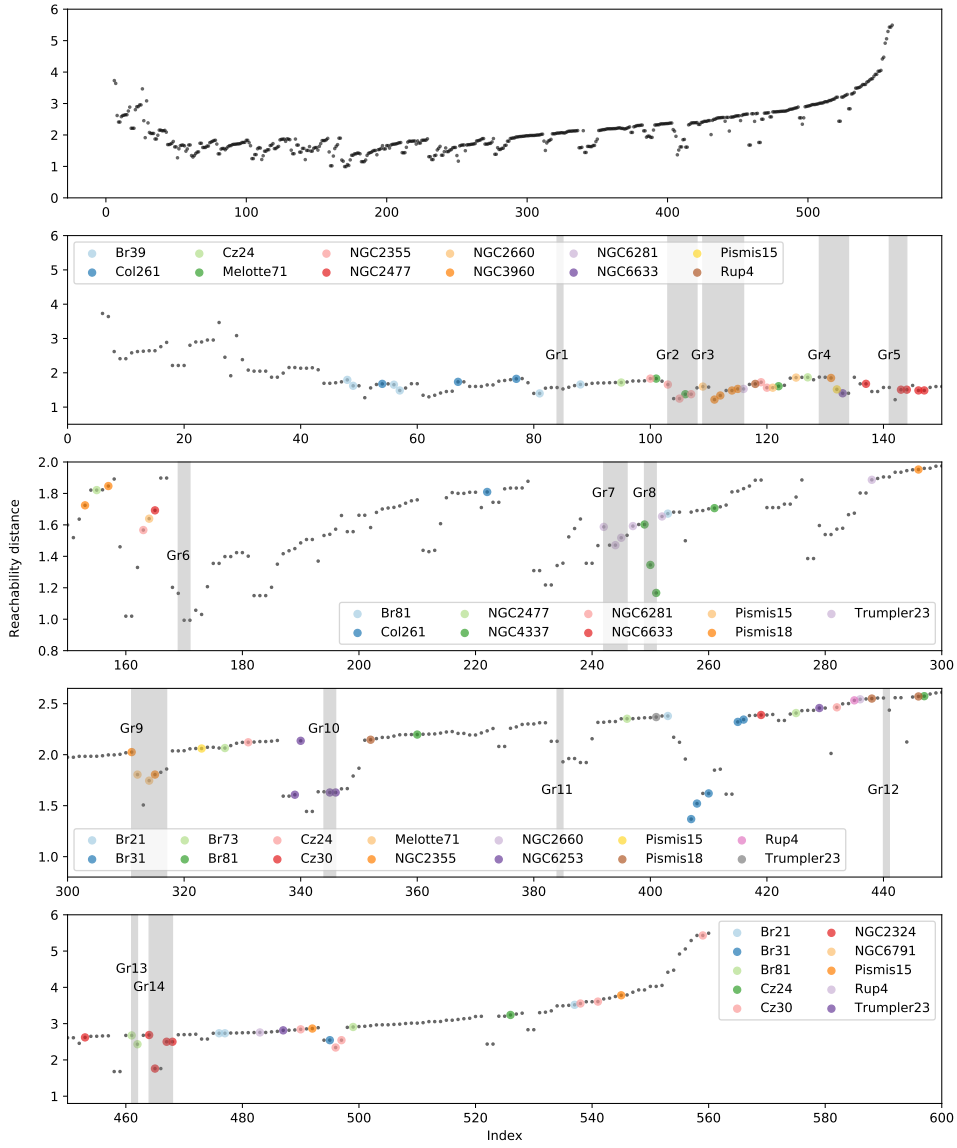

Fig. 7. The panels show the reachability plots resulting from the blind clustering analysis of the data set composed by field stars and hidden open clusters. *Top panel*: entire reachability plot across the full data set. The *following panels* show smaller portions of the reachability plot of 150 instances each. Stellar members of the 23 open clusters of the hidden sample are shown as coloured points. The gray points correspond to the field stars. The shaded vertical bands highlight the location of the 14 groups of interest identified as in Sect. 4.3.

among the field stars. We are also interested in identifying eventual escapers or binary members of open clusters. These stars may have been classified as non-members because of kinematics that are discrepant than that of the open cluster. Finally, our data set may also contain disrupted stellar associations that one could potentially recover in the abundance space.

In Figs. 9–11, we show the location of the 14 groups in three planes: J_R versus L_Z (left-side panels), J_Z versus L_Z (central panels), and z versus R_{GC} (right-side panels). The first two columns of panels tell us the kinematic similarity of the stars, while the last column indicate the spatial relation between them.

The [C/N] abundance ratio is an excellent proxy of stellar age that can be used for giant stars (Salaris et al. 2015). In fact, carbon and nitrogen produced into the stellar interior are brought up to in the stellar surface during the first dredge-up in quantities that depend on the stellar age. In the [C/N] VS age relation calibrated by Casali et al. (2019), the abundance ratios can vary from −1.0 for a 100 Myr old star to −0.2 for a 10 Gyr old star. Therefore, given that carbon and nitrogen have not been used in the clustering analysis, we use this chemical clock when available to provide fundamental and independent information on the real nature of the stellar groups we have identified. The [C/N] ratios are computed from *Gaia*-ESO DR6 abundances as in Casali et al. (2019). We also use their [C/N] VS age relation to estimate the stellar ages. Note that this relation can be used only for stars that have already passed through the first dredge-up.

Group 1. It contains two stars, one belonging to one of the CoRoT field and one to the K2. Although they share the same chemical composition, the two stars are not spatially close, being about 2 kpc apart. No information about their ages, nor [C/N] is available.

Group 2. It contains four members, one in Melotte 71 and three in NGC 2355. The two clusters are similar for R_{GC}, close to 10 kpc, and age. The group contains also two field stars. Interestingly, the star with CNAME = 07162986+1344207 is in the field of NGC 2355. Although it has radial velocity (70.9 ± 0.4 km s^{-1}) and proper motions ($\mu_\alpha = -0.665 \pm 0.014$ mas yr^{-1}, $\mu_\delta = -0.324 \pm 0.011$ mas yr^{-1}) that are different from those of NGC 2355 (36 ± 1 km s^{-1}, $-3.80 \pm 0.0.14$ mas yr^{-1}, and $-1.09 \pm 0.0.13$ mas yr^{-1}; Cantat-Gaudin et al. 2020), the parallax is consistent within the uncertainties (0.486 ± 0.012 mas). In addition, the star has a [C/N] = -0.54 ± 0.10 dex that is

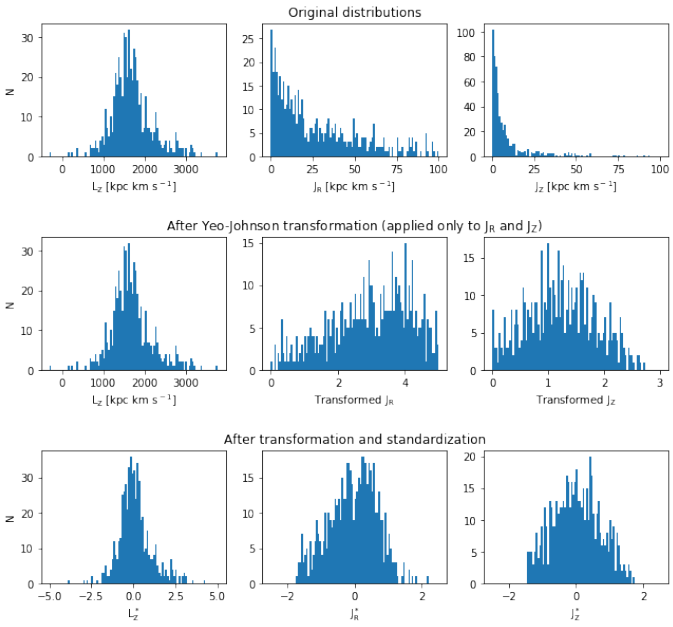

Fig. 8. *Top panels*: distribution of stars (field stars + optimizing sample + hidden cluster sample) as a function of the orbital actions L_Z, J_R, and J_Z. *Middle panels*: the orbital action distributions after that the Yeo-Johnson power transformation is applied to the J_R and J_Z orbital actions. *Bottom panels*: the orbital actions distributions after the power transformation and standardization.

consistent with that of the open cluster (i.e., -0.51 ± 0.04 dex). All these features combined with their chemical similarity suggest that 07162986+1344207 is a true member of NGC 2355 which is escaping from the open clusters. The star was disregarded as part of the open cluster due to the discrepant kinematics, however we assign it to NGC 2355 through strong chemical tagging.

Group 3. It is composed by a mixture of members of NGC 2660, NGC 3960, NGC 6281 and two field stars. The NGC 3960 cluster is recovered within this group with completeness $C = 0.80$, homogeneity $H = 0.50$ and V-value $V = 0.62$. The three clusters whose members have been identified are located at R_{GC} from 7.7 to 8.9 kpc and with log(age) from 8.7 to 9 (Cantat-Gaudin et al. 2020).

Group 4. Members of three clusters with similar ages (NGC 3960, NGC 6633, Pismis15), but a slightly different location in the disc ($R_{GC} = 7.6, 8.6, 8.0$ kpc). In addition, the group contains three field stars. There is not an evident kinematic connection between all these objects.

Group 5. We have correctly re-identified two members of NGC 2477, together with other two field stars. These latter do not show any evident connection to NGC 2477.

Group 6. It contains three stars located in the inner disc, towards the bulge. They have similar R_{GC} and z. Two of them (CNAMEs 18101550−3147076 and 18161182−3339161) are also very similar in their actions and they have similar abundance ratios between carbon and nitrogen, $[C/N] = -0.50 \pm 0.04$ and $[C/N] = -0.61 \pm 0.03$. They can be candidate member of a possible common site of star formation. The third star has a very different $[C/N] = -0.27 \pm 0.06$.

Group 7. We have re-identified three members of Trumpler 23. The group also contains two field stars located in the inner disc, towards the bulge. These latter share the same R_{GC} of Trumpler 23, but different z. Between them the two field stars have similar height in the plane and kinematic properties (possible common site of formation).

Group 8. We have re-identified three members (out of 4) of NGC 4337. The cluster is recovered with $C = 0.75$, $H = 1.0$, and $V = 0.86$.

Group 9. The group contains members of Melotte 71 (2/5) and of NGC 2353 (2/8). It also contains two stars in the field of Melotte 71 (possible escapers for evaporation or binary stars, with difference in radial velocity within 20 km s^{-1} from the mean cluster velocity) and an interesting star in the inner disc, with the same composition, but without any kinematic connection with the clusters' members. The radial velocities of the two stars in the field of Melotte 71 measured in *Gaia*-ESO IDR6 and in *Gaia* EDR3, thus at two different epochs, indicate variations, favouring the hypothesis of binarity. Among the two candidate members of Melotte 71 – CNAMES 07372506−1202241 and 07373561−1202542 – only the first has a determination of $[C/N] = -0.73 \pm 0.09$, which is consistent with that of the open cluster, $[C/N] = -0.63 \pm 0.07$.

Group 10. The group contains two members (2/5) of NGC 6253. The cluster is recovered within this group at $C = 0.40$, $H = 0.67$, and $V = 0.50$. The group also includes one field star that is spatially and kinematically unrelated to NGC 6253.

Group 11. It is composed by two stars in one of the CoRoT fields observed by *Gaia*-ESO, which share similar kinematic properties, location and chemistry. They may have originated from the same star-formation site.

Group 12. The group contains two field stars with similar R_{GC}, but different z.

Group 13. The group contains two members (2/8) of Berkeley 81. The cluster is recovered within this group at $C = 0.33$, $H = 1.0$, and $V = 0.5$.

Group 14. We have re-identified four members of NGC 2324. The cluster is recovered within this group at $C = 0.80$, $H = 0.80$, and $V = 0.80$. The group also includes one star (CNAME = 07040822+0105185) located in the field of NGC 2324. It has a radial velocity 34.9 ± 0.4 km s^{-1}, and proper motion ($\mu_\alpha = -0.27 \pm 0.02$ mas yr^{-1}, $\mu_\delta = -0.06 \pm 0.02$ mas yr^{-1}) that are consistent with those of the open cluster (-0.31 ± 0.11 mas yr^{-1}, and -0.09 ± 0.09 mas yr^{-1}). Only parallax is marginally inconsistent: 0.0284 ± 0.018 mas for the star and 0.19 ± 0.06 mas for the cluster. Unfortunately the [C/N] ratio is unavailable. However, the high similarity in kinematics strongly suggests that 07040822+0105185 is a member of NGC 2324 that we recovered through strong chemical tagging.

Finally, some general remarks on the nature of the groups identified: (i) the chemical homogeneity of each group is in the range 0.01−0.1 dex. If there are members of known star clusters, they are usually more homogeneous than the other stars; (ii) in some cases, groups contain members of stars in known clusters and field stars, which are chemically similar, but there are no evident connections in terms of their location in the disc and kinematic properties; (iii) there are some groups (or sub-groups within groups) composed by GE_BL stars (meaning stars located

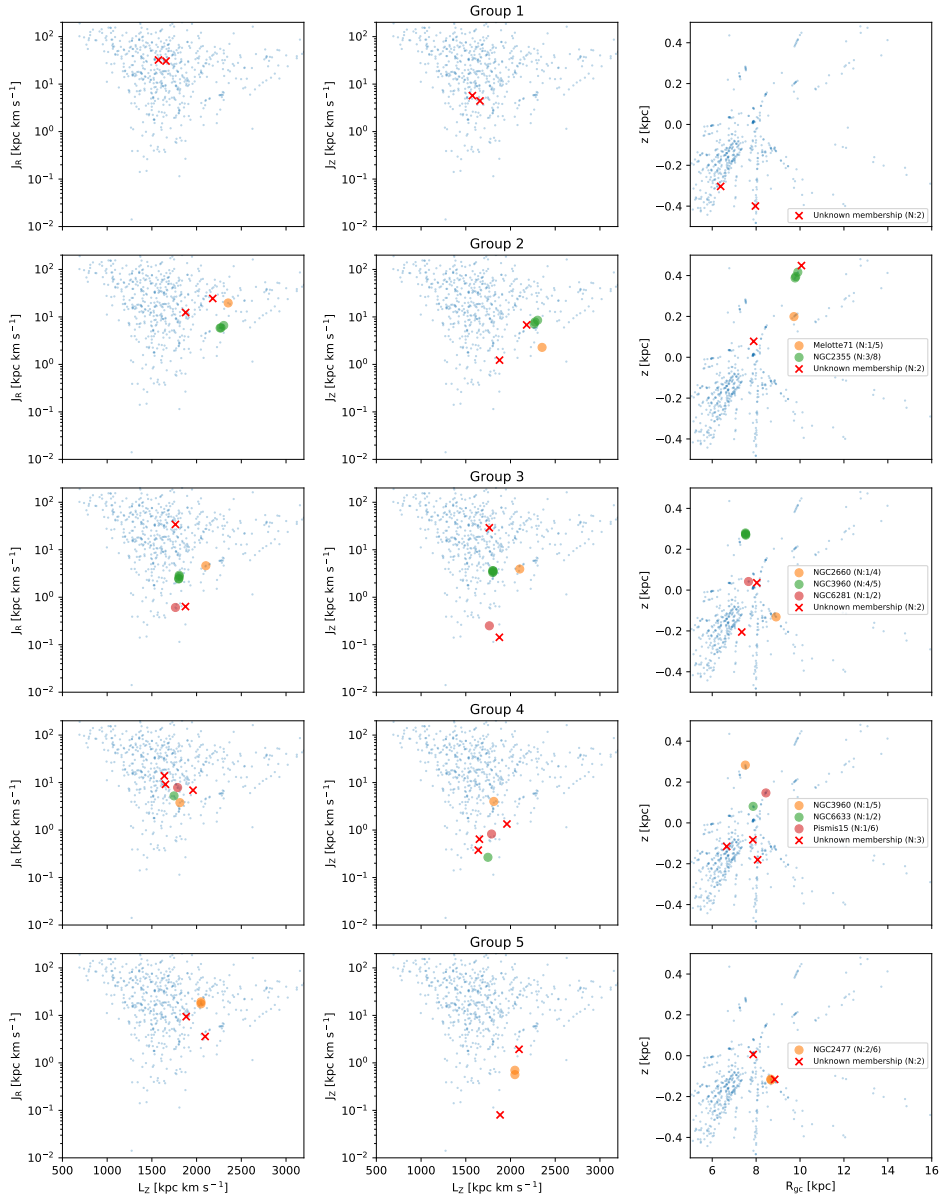

Fig. 9. Location of the groups 1−5 in three planes: J_R versus L_Z (*left-side panels*), J_Z versus L_Z (*central panels*), and z versus R_{GC} (*right-side panels*).

towards the bulge, in the inner disc). They share chemistry and often also R_{GC}, being possible candidate members of disrupted clusters. Although this result is important, we must also remember that we have a bias towards the inner disc when selecting field giants (see group 6, and sub-group 7) because of the *Gaia*-ESO selection function (Stonkutė et al. 2016); (iv) groups with members of more than one cluster usually contain members of clusters located at the same R_{GC}. This may be due to a bias introduced the metric D used to select the groups of interest. However, it may also indicate the major role played by the position in the disc for chemical evolution; (v) we find stars in the field of clusters which are not classified as members, but whose chemical composition is similar to the other members (see groups 2, 9, and 14). They can be binary stars (SB1) or cluster escapers (Moyano Loyola & Hurley 2013), (vi) we find a possible signature of a disrupted cluster in one of the CoRoT fields (group 11).

5. Summary and conclusions

The goal of this study is performing an experiment on strong chemical tagging, testing whether or not we are capable of blindly recovering members of open clusters through clustering analysis in the abundance space. For this experiment we use a controlled sample of giant stars observed with UVES by the *Gaia*-ESO survey. The sample contains both members of open clusters and field stars.

5.1. Practical tips for clustering in the elemental abundance space

The success of clustering analysis relies on some particular decisions that the scientist has to take about the pre-processing techniques and the hyper-parameters' values. Unfortunately, there is no rule of thumb, nor a general strategy that works well for all types of problems. Instead, the correct strategy strongly depends on the type of problem one has to tackle and the type of data set one has at hand. Therefore, as a first step of our analysis, we carry out a preliminary study specifically designed to explore different strategies of analysis. The aim is to identify the best procedure for recovering members of open clusters within our data set. For this study we use a sample of 13 different open clusters, containing at least 10 members each. Given its

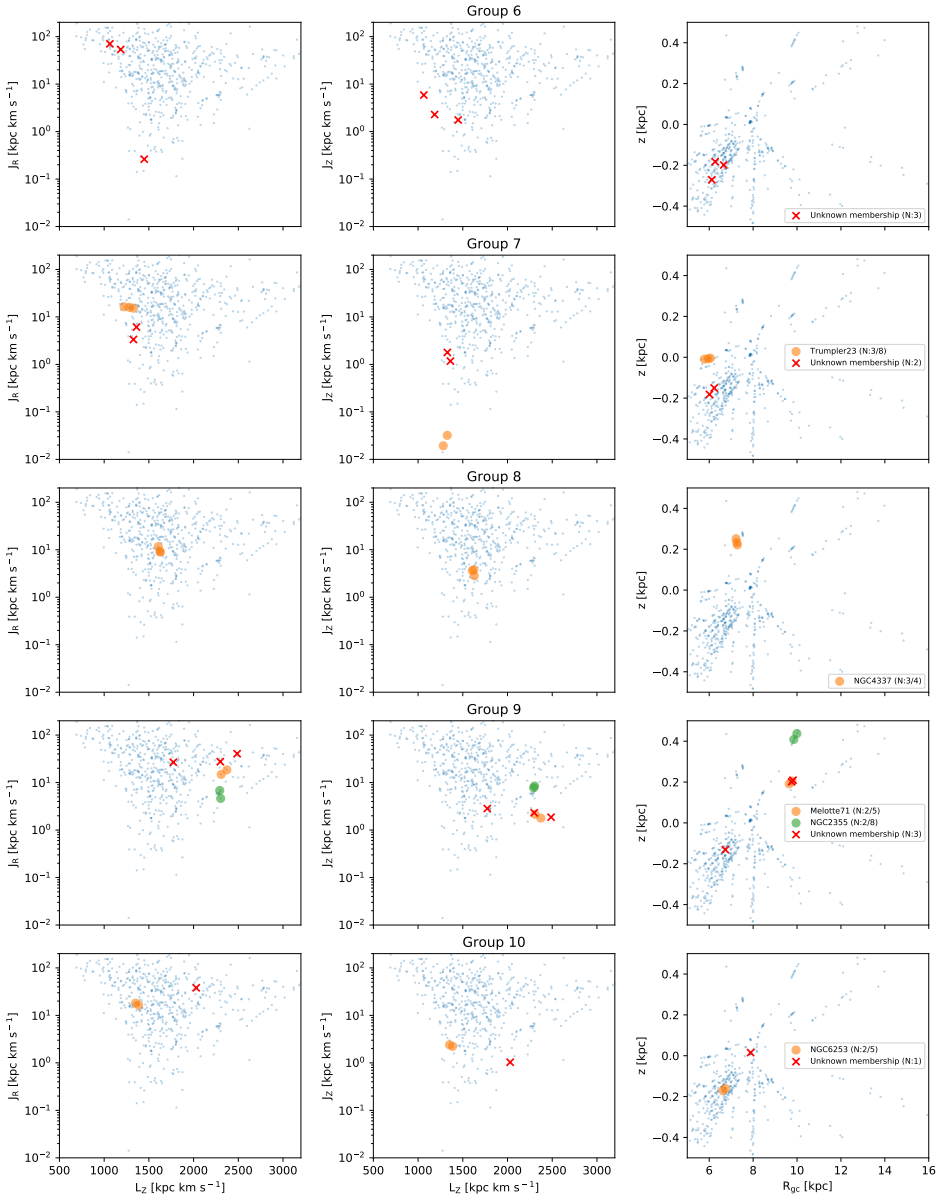

Fig. 10. Location of the groups 6−10 in three planes: J_R versus L_Z (*left-side panels*), J_Z versus L_Z (*central panels*), and z versus R_{GC} (*right-side panels*).

function in our work, the sample is called optimization sample and it comprises 190 stars in total. From the preliminary study we learn a few lessons that are valid for any chemical tagging problem:

- Open clusters have different densities in abundance space. The algorithms already implemented in `Python` that can process a data set containing both clusters and noisy data and that can capture groups of different densities are OPTICS and HDBSCAN. Specifically, we decide to use OPTICS because of its simplicity and versatility (it has a smaller number of hyper-parameters to be tuned than HDBSCAN) and also because there are no chemical tagging experiments with OPTICS in the literature yet.
- The distance- and density-based algorithms – such as k-means, DBSCAN, HDBSCAN and OPTICS – are not entirely deterministic. Their outcomes slightly depend on how the data set is initially ordered. Therefore, with the aim of boosting the performances of the clustering algorithms and of ensuring the reproducibility of our results, we pre-order our dataset sorting as a function of chemical abundances, stating with [Fe/H].
- Standardization is always recommended as it gives all dimensions an equal weight.
- The Manhattan metric should be preferred over the Euclidean metric. This is especially important if we are working with a noisy data set and on a high-dimensional space.
- Using all the chemical abundances that one has at hand is a strategy that does not pay. Instead, the clustering analysis should be carried out using only the most relevant features. The choice of the best elements for chemical tagging depends on the chemical information they carry, but also on the typical precision of their abundance measurements. Therefore, there we expect that the best set of elements could vary from survey to survey. The most relevant features can be chosen by the scientist using criteria such as (i) the nucleosynthetic history of each specific element, and (ii) the Calinski-Harabasz score. There are other methods for an automatic dimensionality reduction, such as the PCA, the LDA, and many others.
- PCA alone is not the most convenient strategy to reduce the dimensionality for chemical tagging problems. This is

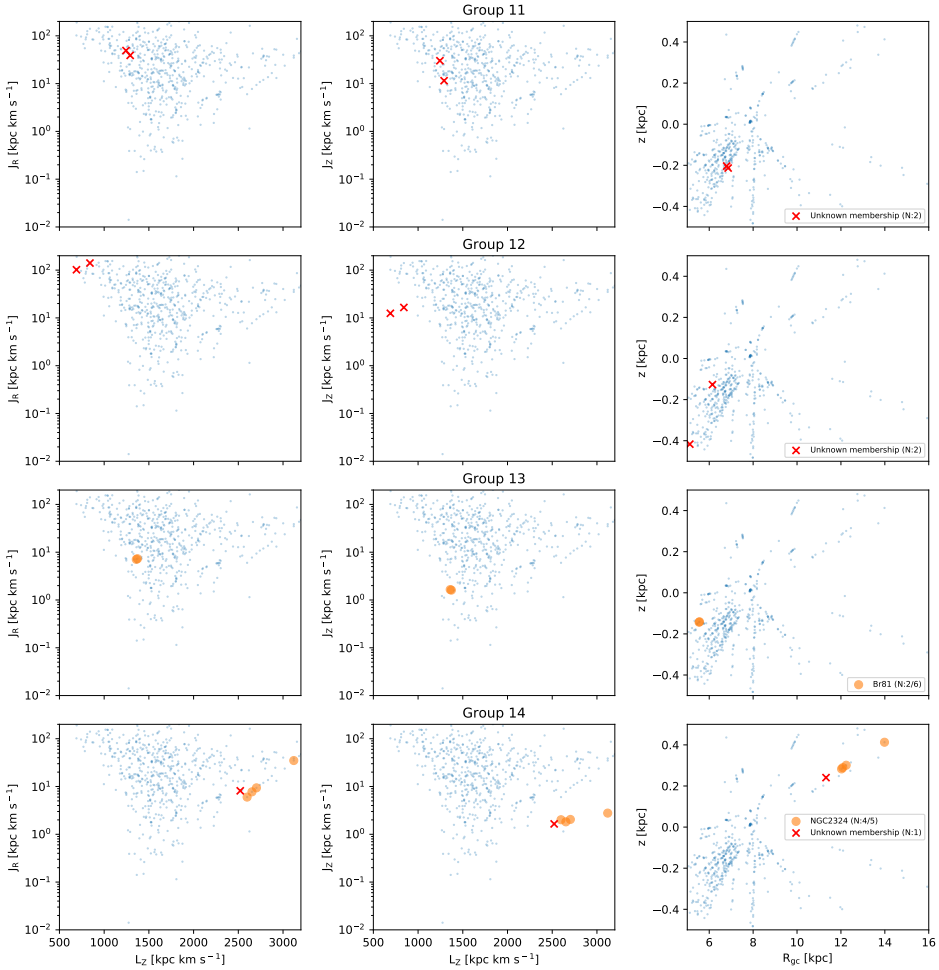

Fig. 11. Location of the groups 11−14 in three planes: J_R versus L_Z (*left-side panels*), J_Z versus L_Z (*central panels*), and z versus R_{GC} (*right-side panels*).

demonstrated in Fig. 5. Instead, better results are expected by applying in sequence PCA and LDA, where LDA can be trained on a set of known cluster members.

By putting these lessons into practice we are able to recover 9/13 open clusters into groups at the 50% of completeness and homogeneity. This result outclass previous studies attempting non-blind chemical tagging on samples solely composed by open clusters' members (e.g., Blanco-Cuaresma & Fraix-Burnet 2018; Garcia-Dias et al. 2019; Casamiquela et al. 2021).

5.2. Blind chemical tagging

One of the main goals of this study is testing whether clustering analysis in the abundance space can recover the building blocks of the thin disk. To do this, we carry out an experiment of blind chemical tagging aiming at the identification of open clusters hidden in our data set.

Thus, following the strategy outlined above, we analyse the data set composed by 456 field stars and 112 members of 23 open clusters. These latter are the hidden clusters that we aim to recover in the abundance space. The analysis is completely "blind", meaning that we never use information of the stellar membership to reconstruct the hidden clusters. In fact, the algorithm's hyperparameters are optimised exclusively on the optimisation sample. Nevertheless, the actual membership of each single star from the hidden clusters sample is known and is used to evaluate our results. The clustering analysis identifies 122 overdensities in the abundance space. We consider only the groups with a number of stars ranging between 2 and 20 and with a dispersion in the space of orbital actions below a certain threshold. We think that these 14 groups are those with the highest chances of coinciding with some of the hidden clusters.

Interestingly three of the 14 groups coincide with open clusters recovered at the $C \geq 0.5$ and $H \geq 0.5$ threshold. These are the groups #3, #8, and #14. The open clusters are NGC 3960, NGC 4337, and NGC 2324, respectively. There are other two groups coinciding with open clusters recovered at the threshold of $V \geq 0.5$. These are the group #10 for NGC 6253 and the group #10 for Berkeley 81. In addition, we identify four stars that are possibly escapers or binary members of NGC 2355 (group #2), Melotte 71 (group #9), and NGC 2324 (group #14). These stars were initially classified as field stars, due to the marginal discrepancy between their kinematics and that of their open cluster.

In conclusion, based on the knowledge that is currently available to us, 7/14 groups carry real information around stars formed within the same stellar association. The other seven groups candidate stellar associations which should be further investigated with follow-up observations.

5.3. The OPTICS execution time

The algorithm execution time is a very important aspect of data clustering. This is especially true when the data set is composed by a large number of instances and features. Our data set is extremely small and the clustering analysis described in Sect. 4.2

(i.e., 568 instances and 8 features) has taken 0.4 s with a 3.1 GHz dual-core processor.

Given that the time complexity[2] of the OPTICS algorithms scales with the square of the number of instances, a data set of 10 000 instances would take around 20 s. Such a data set is more similar to those that could be employed by future spectroscopic surveys. Tuning the hyper-parameters under these conditions – e.g., running the algorithms ~100 times – would be still feasible with our computing resources. Instead, a data set composed by 100 000 instances would easily take around 30 min to run and would require a much faster machine.

However, the OPTICS algorithm includes the `max_eps` hyper-parameter which describes the maximum distance between two points for one to be considered as in the neighborhood of the other and that can be tuned in order to reduce the execution time. The `max_eps` hyper-parameter is set to infinity by default. However, reducing its value reduces the volume considered around each point to find its neighbours, thus it can also speeds up the execution time.

5.4. Concluding remarks

Strong chemical tagging – if fully feasible – would allow us to disentangle the spatial distribution of stars from their radial migration within the disk and would allow us to reconstruct the building blocks of the thin disk. The results outlined above are very promising for the prospect of using chemical abundances to tag stars to their birth clusters. On the other hand, we are still far from being fully satisfied, as the recovery fraction of open clusters in the blind analysis tends not to be particularly high. It is still unclear whether this recovery fraction is intrinsically low due to the small chemical diversity of stellar associations within the thin disk, or instead it is due to other limiting factors that we can overcome with a more performing algorithm and more precise data.

One should also notice that the data set used in our experiment is strongly unbalanced compared to what one would expect under real circumstances. In fact, about 1/4 of the stars in our data set are members of the hidden clusters, which has certainly made easier the recovery of open clusters. Nevertheless, open clusters have been recovered, as well as new members that were lost from the membership analysis. This certainly leaves a door open to feasibility of strong chemical within the thin disk.

In this regard, one should also consider which is the prospect of stellar astronomy and data science for the next decade. The forthcoming large spectroscopic surveys and facilities (e.g., 4MOST, WEAVE, MOONS; Cirasuolo et al. 2014; Dalton et al. 2016; de Jong et al. 2019) are expected to provide huge data sets, but – given their limited spectral resolution – they will probably provide abundance determinations that are similar or even more noisy than those used here. Likely, this is a big limitation for future attempts of chemical tagging. However, it is conceivable that after 2030 we will see the first large high-resolution spectroscopic survey of our Galaxy, which will increase our chances of finally using this technique.

Finally, one should not forget that data science is a discipline that is currently progressing at high speed. This momentum is due to the constant acquisition of data from states, multinationals, industries, etc. It is a practice that will further increase during the next years. Thus, the current techniques of clustering analysis will be certainly improved and other algorithms will be developed. For instance, new techniques of clustering analysis are being developed to deal with noise, outliers, and data sets with missing values (Song et al. 2021; Yan et al. 2021). These are the same limitations that cripple our efforts of chemical tagging. Therefore, we must not lose hope of effective and complete chemical tagging in our Galaxy: the refinement of techniques combined with the big databases that will be available in the next future, will allow us to study the star formation history of the Galactic disk in great detail.

[2] Time complexity is the amount of time taken by an algorithm to run, as a function of the length of the input.

Acknowledgements. Based on data products from observations made with ESO Telescopes at the La Silla Paranal Observatory under programme ID 188.B-3002. These data products have been processed by the Cambridge Astronomy Survey Unit (CASU) at the Institute of Astronomy, University of Cambridge, and by the FLAMES/UVES reduction team at INAF/Osservatorio Astrofisico di Arcetri. These data have been obtained from the *Gaia*-ESO Survey Data Archive, prepared and hosted by the Wide Field Astronomy Unit, Institute for Astronomy, University of Edinburgh, which is funded by the UK Science and Technology Facilities Council. This work was partly supported by the European Union FP7 programme through ERC grant number 320360 and by the Leverhulme Trust through grant RPG-2012-541. We acknowledge the support from INAF and Ministero dell' Istruzione, dell' Università' e della Ricerca (MIUR) in the form of the grant "Premiale VLT 2012" and "Premiale 2016 MITiC". The results presented here benefit from discussions held during the *Gaia*-ESO workshops and conferences supported by the ESF (European Science Foundation) through the GREAT Research Network Programme. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. L.S. is supported by the Italian Space Agency (ASI) through contract 2018-24-HH.0 to the National Institute for Astrophysics (INAF). L.M. thank the COST Action CA18104: MW-*Gaia*. G.C. acknowledges support from the European Research Council Consolidator Grant funding scheme (project ASTEROCHRONOMETRY, G.A. n. 772293, `http://www.asterochronometry.eu`). T.B. was funded by the project grant No. 2018-04857 from the Swedish Research Council.

References

Adibekyan, V., Figueira, P., Santos, N. C., et al. 2015, A&A, 583, A94
Aggarwal, C. C., Hinneburg, A., & Keim, D. A. 2001, Proceedings of the 8th International Conference on Database Theory, ICDT '01 (Berlin, Heidelberg: Springer-Verlag), 420
Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. 1999, Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), Philadelphia, PA, 49
Baratella, M., D'Orazi, V., Carraro, G., et al. 2020, A&A, 634, A34
Baratella, M., D'Orazi, V., Sheminova, V., et al. 2021, A&A, 653, A67
Bellman, R. 1957, Dynamic Programming (New York: Dover Publications)
Bertelli Motta, C., Pasquali, A., Richer, J., et al. 2018, MNRAS, 478, 425
Blanco-Cuaresma, S., & Fraix-Burnet, D. 2018, A&A, 618, A65
Blanco-Cuaresma, S., Soubiran, C., Heiter, U., et al. 2015, A&A, 577, A47
Bland-Hawthorn, J., & Freeman, K. C. 2004, PASA, 21, 110
Bland-Hawthorn, J., Krumholz, M. R., & Freeman, K. 2010, ApJ, 713, 166
Bovy, J. 2016, ApJ, 817, 49
Caliński, T., & Harabasz, J. 1974, Commun. Stat.-Simul. Comput., 3, 1
Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, in Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, Proceedings, Part II, eds. J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Springer), Lect. Notes Comput. Sci., 7819, 160
Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. 2020, A&A, 640, A1
Casali, G., Magrini, L., Tognelli, E., et al. 2019, A&A, 629, A62
Casali, G., Magrini, L., Frasca, A., et al. 2020a, A&A, 643, A12
Casali, G., Spina, L., Magrini, L., et al. 2020b, A&A, 639, A127
Casamiquela, L., Castro-Ginard, A., Anders, F., & Soubiran, C. 2021, A&A, 654, A151
Charbonnel, C. 1995, ApJ, 453, L41
Charbonnel, C., & Zahn, J. P. 2007, A&A, 467, L15
Chen, B., D'Onghia, E., Pardy, S. A., et al. 2018, ApJ, 860, 70
Cirasuolo, M., Afonso, J., Carollo, M., et al. 2014, in Ground-based and Airborne Instrumentation for Astronomy V, eds. S. K. Ramsay, I. S. McLean, & H. Takami, SPIE Conf. Ser., 9147, 91470N

Dalton, G., Trager, S., Abrams, D. C., et al. 2016, in Ground-based and Airborne Instrumentation for Astronomy VI, eds. C. J. Evans, L. Simard, & H. Takami, SPIE Conf. Ser., 9908, 99081G
de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, The Messenger, 175, 3
De Silva, G. M., Sneden, C., Paulson, D. B., et al. 2006, AJ, 131, 455
De Silva, G. M., Freeman, K. C., Asplund, M., et al. 2007, AJ, 133, 1161
De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, MNRAS, 449, 2604
Ester, M., Kriegel, H. P., Sander, J., & Xu, X. 1996, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press), 226
Fisher, R. A. 1938, Ann. Eugen., 8, 376
Freeman, K., & Bland-Hawthorn, J. 2002, ARA&A, 40, 487
Garcia-Dias, R., Allende Prieto, C., Sánchez Almeida, J., & Alonso Palicio, P. 2019, A&A, 629, A34
Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25
Hawkins, K., Jofré, P., Masseron, T., & Gilmore, G. 2015, MNRAS, 453, 758
Hogg, D. W., Casey, A. R., Ness, M., et al. 2016, ApJ, 833, 262
Jackson, R. J., Jeffries, R. D., Wright, N. J., et al. 2022, MNRAS, 509, 1664
Kobayashi, C., Karakas, A. I., & Lugaro, M. 2020, ApJ, 900, 179
Lagarde, N., Decressin, T., Charbonnel, C., et al. 2012, A&A, 543, A108
Liu, F., Asplund, M., Yong, D., et al. 2019, A&A, 627, A117
Magrini, L., Vescovi, D., Casali, G., et al. 2021, A&A, 646, L2
Maia, M. T., Meléndez, J., Lorenzo-Oliveira, D., Spina, L., & Jofré, P. 2019, A&A, 628, A126
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Martell, S. L., & Grebel, E. K. 2010, A&A, 519, A14
Mitschang, A. W., De Silva, G., Zucker, D. B., et al. 2014, MNRAS, 438, 2753
Moyano Loyola, G. R. I., & Hurley, J. R. 2013, MNRAS, 434, 2509
Ness, M., Zasowski, G., Johnson, J. A., et al. 2016, ApJ, 819, 2
Ness, M., Rix, H. W., Hogg, D. W., et al. 2018, ApJ, 853, 198
Ness, M. K., Wheeler, A. J., McKinnon, K., et al. 2022, ApJ, 926, 144
Pearson, K. 1901, London Edinb. Dublin Philos. Mag. J. Sci., 2, 559
Poovelil, V. J., Zasowski, G., Hasselquist, S., et al. 2020, ApJ, 903, 55
Price-Jones, N., & Bovy, J. 2018, MNRAS, 475, 1410
Price-Jones, N., & Bovy, J. 2019, MNRAS, 487, 871
Price-Jones, N., Bovy, J., Webb, J. J., et al. 2020, MNRAS, 496, 5101
Randich, S., Gilmore, G., & Gaia-ESO Consortium 2013, The Messenger, 154, 47
Recio-Blanco, A., Rojas-Arriagada, A., de Laverny, P., et al. 2017, A&A, 602, L14
Reddy, A. B. S., Giridhar, S., & Lambert, D. L. 2020, JApA, 41, 38
Rokach, L., & Maimon, O. 2005, in The Data Mining and Knowledge Discovery Handbook, eds. O. Maimon, & L. Rokach (Springer), 321
Salaris, M., Pietrinferni, A., Piersimoni, A. M., & Cassisi, S. 2015, A&A, 583, A87
Schiavon, R. P., Zamora, O., Carrera, R., et al. 2017, MNRAS, 465, 501
Sestito, P., Randich, S., & Bragaglia, A. 2007, A&A, 465, 185
Smiljanic, R., Korn, A. J., Bergemann, M., et al. 2014, A&A, 570, A122
Song, S., Gao, F., Huang, R., & Wang, Y. 2021, Proceedings of the 2021 International Conference on Management of Data, SIGMOD/PODS '21 (New York: Association for Computing Machinery), 1692
Souto, D., Cunha, K., Smith, V. V., et al. 2018, ApJ, 857, 14
Spina, L., Palla, F., Randich, S., et al. 2015, A&A, 582, L6
Spina, L., Meléndez, J., Karakas, A. I., et al. 2018, MNRAS, 474, 2580
Spina, L., Nordlander, T., Casey, A. R., et al. 2020, ApJ, 895, 52
Spina, L., Sharma, P., Meléndez, J., et al. 2021, Nat. Astron., 5, 1163
Stonkutė, E., Koposov, S. E., Howes, L. M., et al. 2016, MNRAS, 460, 1131
Ting, Y.-S., & Weinberg, D. H. 2022, ApJ, 927, 209
Ting, Y.-S., Conroy, C., & Goodman, A. 2015, ApJ, 807, 104
Ting, Y.-S., Conroy, C., & Rix, H.-W. 2016, ApJ, 816, 10
Vasiliev, E., & Baumgardt, H. 2021, MNRAS, 505, 5978
Wojno, J., Kordopatis, G., Steinmetz, M., et al. 2016, MNRAS, 461, 4246
Yan, A., Wang, W., Ren, Y., & Geng, H. 2021, Front. Neurorobot., 15, 64
Yana Galarza, J., Meléndez, J., Lorenzo-Oliveira, D., et al. 2019, MNRAS, 490, L86
Yeo, I. K., & Johnson, R. 2000, Biometrika, 87, 954
Zhang, R., Lucatello, S., Bragaglia, A., et al. 2021, A&A, 654, A77