Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **473,** 4612–4633 (2018)
Advance Access publication 2017 October 17

doi:10.1093/mnras/stx2637

The GALAH survey: chemical tagging of star clusters and new members in the Pleiades

Janez Kos,[1★] Joss Bland-Hawthorn,[1,2] Ken Freeman,[3] Sven Buder,[4] Gregor Traven,[5] Gayandhi M. De Silva,[1,6] Sanjib Sharma,[1] Martin Asplund,[3] Ly Duong,[3] Jane Lin,[3] Karin Lind,[4,7] Sarah Martell,[8] Jeffrey D. Simpson,[9] Dennis Stello,[1,8,10] Daniel B. Zucker,[9] Tomaž Zwitter,[5] Borja Anguiano,[11] Gary Da Costa,[3] Valentina D'Orazi,[12] Jonathan Horner,[13] Prajwal R. Kafle,[14] Geraint Lewis,[1] Ulisse Munari,[15] David M. Nataf,[16] Melissa Ness,[4] Warren Reid,[9,17] Katie Schlesinger,[3] Yuan-Sen Ting[3] and Rosemary Wyse[16]

Affiliations are listed at the end of the paper

Accepted 2017 October 6. Received 2017 October 5; in original form 2017 July 18

ABSTRACT

The technique of chemical tagging uses the elemental abundances of stellar atmospheres to 'reconstruct' chemically homogeneous star clusters that have long since dispersed. The GALAH spectroscopic survey – which aims to observe one million stars using the Anglo-Australian Telescope – allows us to measure up to 30 elements or dimensions in the stellar chemical abundance space, many of which are not independent. How to find clustering reliably in a noisy high-dimensional space is a difficult problem that remains largely unsolved. Here, we explore *t-distributed stochastic neighbour embedding* (t-SNE) – which identifies an optimal mapping of a high-dimensional space into fewer dimensions – whilst conserving the original clustering information. Typically, the projection is made to a 2D space to aid recognition of clusters by eye. We show that this method is a reliable tool for chemical tagging because it can: (i) resolve clustering in chemical space alone, (ii) recover known open and globular clusters with high efficiency and low contamination, and (iii) relate field stars to known clusters. t-SNE also provides a useful visualization of a high-dimensional space. We demonstrate the method on a data set of 13 abundances measured in the spectra of 187 000 stars by the GALAH survey. We recover seven of the nine observed clusters (six globular and three open clusters) in chemical space with minimal contamination from field stars and low numbers of outliers. With chemical tagging, we also identify two Pleiades supercluster members (which we confirm kinematically), one as far as 6° – one tidal radius away from the cluster centre.

Key words: methods: data analysis – stars: abundances – open clusters and associations: general – open clusters and associations: individual: Pleiades.

1 INTRODUCTION

The use of chemical tagging in Galactic archaeology was first proposed by Freeman & Bland-Hawthorn (2002), who suggested that the abundances of elements in stars could be used as unique signatures over their lifetime to 'reconstruct' stellar groups that have long since dissolved. Theoretical arguments indicate that chemical homogeneity (with the exception of light elements) is guaranteed in open clusters up to 10^5 $M_\odot$ and in globular clusters up to a limit of 10^7 $M_\odot$ (Bland-Hawthorn, Krumholz & Freeman 2010a). In practice, the degree of homogeneity may depend on the initial abundance spread in the collapsing cloud from which the cluster forms (Feng & Krumholz 2014). But, to date, essentially all open clusters appear to be chemically homogeneous to better than 0.1 dex (De Silva et al. 2006; Sestito, Randich & Bragaglia 2007; Bovy 2016). Both young and ancient (up to ∼9 Gyr) open clusters appear to be chemically homogeneous (De Silva et al. 2006, 2007), indicating that pollution from the interstellar medium does not wipe out this information. For chemical tagging to be feasible for field stars, a large amount of high-quality data has to be collected, i.e. of the order of 10^6 observed stars and ∼30 measured elements (Bland-Hawthorn

★ E-mail: janez.kos@sydney.edu.au

& Freeman 2004; Ting, Conroy & Goodman 2015). Indeed, these are the design goals of the GALAH[1] survey on the HERMES instrument at the Anglo-Australian Telescope (Barden et al. 2010; De Silva et al. 2015; Martell et al. 2017). This requirement can be much lower for 'soft' chemical tagging if there is additional information (e.g. kinematics, location) to associate the stars.

To chemically tag stars, one has to search for clustering in chemical space ($\mathcal{C}$-space), i.e. an N-dimensional space determined by the measured number of elemental abundances. Strictly speaking, these dimensions are unlikely to be independent, for example, iron-peak elements are strongly coupled. Different elements also experience a different cosmic spread (e.g. Bensby, Feltzing & Oey 2014), so they cannot all be treated equally. Bensby et al. (2014) only give cosmic spreads for the F and G dwarfs in the solar neighbourhood, and the cosmic spreads for the population observed by GALAH (all stellar types with ∼90 per cent of the stars within 3 kpc) is largely unknown. Sample used in our study covers much larger region, as well as more stellar types. Cosmic spreads that would be useful for our work are therefore largely unknown. In the GALAH survey, there are up to 30 elements for which abundances can be determined in each star, but in our study we will concentrate on a smaller number ($N = 13$) of elements with well determined abundances.

How are we to find substructures in a high-dimensional space? The human brain is excellent at detecting clustering in three or fewer dimensions, but falls short for problems in more dimensions. Most work to date has focused on finding clusters in the original N-dimensional space. For example, Hogg et al. (2016) searched for chemical groups in the APOGEE data by the k-means algorithm and showed that some clusters correspond to groups in phase space. Blanco-Cuaresma et al. (2015) utilized PCA to distinguish between known clusters. Bland-Hawthorn et al. (2010b) used a density-based hierarchical clustering algorithm and introduced the S-statistic to show that clustering exists in a simulated dwarf galaxy. Mitschang et al. (2014) and Quillen et al. (2015) used a probabilistic approach to resolve chemical groups in a blind chemical tagging study, but were unable to determine whether the groups found were, in fact, co-natal (born together), or simply had nearly identical abundances. When key chemical signatures can be confined to two or three dimensions, this classification becomes straightforward. Martell et al. (2016) were able to identify halo stars that originated in globular clusters, and de Silva et al. (2011) placed Hyades supercluster members in one chemical group. De Silva et al. (2013) used chemical tagging to relate the Argus association to IC 2391. A more advanced algorithm, that also provides visualization, was used by Jofré et al. (2017), who applied a method of evolutionary trees to stellar abundances and produced a phylogenetic tree for 21 solar twins and the Sun.

One of the most successful methods in recent years exploits the huge computational power now available in desktop computers. The so-called t-distributed stochastic neighbour embedding (t-SNE) algorithm is a remarkable technique for reducing the dimensions of a problem (van der Maaten & Hinton 2008). That technique embeds each high-dimensional data point into a two dimensional 'visualization' space where 'similar' points are kept together and 'dissimilar' points are moved apart. Once the problem is reduced into two dimensions, the clustering can be identified by eye. We find that this method is highly effective in identifying known and unknown cluster members. The method has some limitations: (i) it is a black box that is difficult to tune or control; and (ii) the abundance measurement errors are not used in the present application. But the method is extraordinarily powerful as demonstrated in recent papers, for example, to efficiently identify peculiar stars and stellar populations in large surveys with a high level of completeness (Lochner et al. 2016; Matijevič et al. 2017; Traven et al. 2017; Valentini et al. 2017).

We describe our data in Section 2 and the method in Section 3. In Section 4, we explore the efficiency of our method on nine known clusters, and in Section 5, we present a more detailed analysis of the Pleiades cluster. In Section 6, we discuss the implications of this work and future development of the field.

2 THE DATA

The data set analysed here has been drawn from three programmes: the GALAH pilot programme, the K2-HERMES survey (Sharma 2017, in preparation) and the main GALAH survey (De Silva et al. 2015; Martell et al. 2017). These three programmes have different selection functions, but share the same observing procedures, reduction pipeline and analysis pipeline (Kos et al. 2017). Stars from all programmes are analysed together so the stellar parameters and the abundances are comparable.

All stars used in this paper have abundances measured by The Cannon (Ness et al. 2015), applying a data-driven approach to estimate stellar parameters and abundances using linear algebra to combine the spectral flux for each pixel. The quadratic spectral model of The Cannon was trained on a representative set of spectra, consisting of benchmark stars (Heiter et al. 2015; Jofré et al. 2015), K2 stars with known seismic gravities (Stello et al. 2016, 2017) as well as stars with high-quality spectra covering the parameter space. For the training set, stellar parameters and abundances were estimated using the spectrum synthesis code SME (Valenti & Piskunov 1996; Piskunov & Valenti 2017).

The complete data set consists of 187 640 stars, mostly dwarfs, observed between 2013 November and 2016 January. 15 601 stars have unreliable stellar parameters (e.g. because of a peculiar spectrum, strong cosmic rays, etc.) and were excluded from the study. As this is the first time The Cannon has been used with GALAH data and represents the first internal release of abundances, the uncertainties of the measured abundances have not been validated yet. A map of the observed fields on the celestial sphere is given in Martell et al. (2017).

Our data set consists of abundances of 13 elements (Na, Mg, Al, Si, K, Ca, Sc, Ti, Cr, Fe, Ni, Cu, Ba) representing groups of light, light odd-Z, alpha, iron peak and s-process elements. All abundances were measured by The Cannon. The number of lines used to measure the abundances of each element is given in Table 1. Stellar parameters ($T_{\rm eff}$, $\log g$, [Fe/H], and radial velocity) were measured by fitting synthetic models of stellar atmospheres to one-dimensional GALAH spectra (Kos et al. 2017). Abundances for all 13 elements were measured for each of 187 640 stars.

Proper motions from UCAC4 are also available for all stars, and parallaxes from *Gaia* TGAS for some. For most stars, we calculated the photometric distances following Zwitter et al. (2010) using APASS and 2MASS photometry (Martell et al. 2017) and our stellar parameters.

3 T-DISTRIBUTED STOCHASTIC NEIGHBOUR EMBEDDING (T-SNE)

t-SNE is an algorithm from a family of manifold learning algorithms. It has been extensively used in data science and made a break

[1] GALAH survey webpage is https://galah-survey.org

Table 1. Number of lines used to measure the abundances of each element.

Element	Number of lines
Na	3
Mg	2
Al	4
Si	5
K	2
Ca	5
Sc	10
Ti	20
Cr	9
Fe	52
Ni	7
Cu	2
Ba	2

into astronomy as a classification algorithm (Lochner et al. 2016; Matijevič et al. 2017; Traven et al. 2017; Valentini et al. 2017), along with other manifold learning algorithms (e.g. Vanderplas & Connolly 2009; Daniel et al. 2011; Bu, Chen & Pan 2014). We extend its use as a pure manifold learning algorithm to find structure in a 13-dimensional $\mathcal{C}$-space.

t-SNE's input is a set of N high-dimensional objects $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. In our case, each $\boldsymbol{x}_i$ will be a collection of 13 abundances for a single star:

$$\boldsymbol{x}_i = \left(\left[\frac{\mathrm{Na}}{\mathrm{Fe}}\right]_i, \left[\frac{\mathrm{Mg}}{\mathrm{Fe}}\right]_i, \ldots, \left[\frac{\mathrm{Ba}}{\mathrm{Fe}}\right]_i, \left[\frac{\mathrm{Fe}}{\mathrm{H}}\right]_i \right). \quad (1)$$

Following van der Maaten & Hinton (2008), we first calculate similarities $\boldsymbol{p}_{ij}$ of the input set:

$$\boldsymbol{p}_{ij} = \frac{\boldsymbol{p}_{i|j} + \boldsymbol{p}_{j|i}}{2N}, \quad \boldsymbol{p}_{j|i} = \frac{\exp\left(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-||\boldsymbol{x}_j - \boldsymbol{x}_k||^2/2\sigma_i^2\right)}, \quad (2)$$

where σ_i is a parameter calculated by t-SNE depending on the perplexity (a parameter that controls how one point relates to others – see Wattenberg, Viégas & Johnson (2016) for a demonstration of how perplexity works) and local density of the data set. Distances $||\boldsymbol{x}_i - \boldsymbol{x}_j||$ and $||\boldsymbol{x}_j - \boldsymbol{x}_k||$ in the original implementation are Euclidean. We modified the code to use Manhattan distances, as they are less sensitive to sporadic outliers. Manhattan distance between two points $\boldsymbol{p}$ and $\boldsymbol{q}$ in n dimensions is defined as

$$d(\boldsymbol{p}, \boldsymbol{q}) = \| \boldsymbol{p} - \boldsymbol{q} \|_1 = \sum_{i=1}^{n} |p_i - q_i|. \quad (3)$$

We aim to produce a lower dimensional map with objects $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ with similarities:

$$\boldsymbol{q}_{ij} = \frac{(1 + ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2)^{-1}}{\sum_{k\neq i}(1 + ||\boldsymbol{y}_k - \boldsymbol{y}_i||^2)^{-1}}. \quad (4)$$

To find the optimal mapping where $\boldsymbol{q}_{ij}$ reflects $\boldsymbol{p}_{ij}$ as well as possible, we minimize the Kullback–Leibler divergence

$$\mathrm{KL}(P||Q) = \sum_{i\neq j} \boldsymbol{p}_{ij} \log \frac{\boldsymbol{p}_{ij}}{\boldsymbol{q}_{ij}}. \quad (5)$$

Kullback–Leibler divergence measures the amount of information lost if distribution Q is used instead of P. Kullback–Leibler divergence is a non-convex function that is minimized by gradient descent initialized randomly. Different runs of t-SNE, even when the same parameters are used, can therefore result in a different mapping.

Table 2. Clusters with observed members in the GALAH and K2 surveys. Six clusters with no given literature for membership have members identified by us (See Appendix C).

Cluster	Number of stars	Type	Notes
47 Tuc	90	GC	Membership from Tucholke (1992).
M30	4	GC	
M67	113	OC	Membership from Geller, Latham & Mathieu (2015). 404 spectra of 113 unique stars.
NGC 288	14	GC	
NGC 362	27	GC	
NGC 1851	7	GC	
NGC 2516	3	OC	Membership from Jeffries, Thurston & Hambly (2001).
Pleiades	27	OC	
ω Cen	230	GC	246 spectra of 230 unique stars.

GC = globular cluster
OC = open cluster

The algorithm is usually run several times and the mapping with the lowest Kullback–Leibler divergence is used.

The scale of the map produced by t-SNE is irrelevant. Only the relative relations between objects and groups on the map hold any information. We therefore refrain from plotting the coordinate system on the maps.

The above algorithm has a time dependence of $\mathcal{O}(N^2)$, because we have to calculate similarities for every pair of objects. This is impractical for most applications, so we employ the Barnes–Hut algorithm to calculate sparse similarities in $\mathcal{O}(N \log(N))$ time (van der Maaten 2014). With such optimization, we can analyse our whole data set on an average desktop computer in less than one hour.

t-SNE is presented in a more intuitive way in Appendix A, together with a comparison with some other algorithms.

4 RECOVERING KNOWN CLUSTERS IN CHEMICAL SPACE

The GALAH survey targeted only a small number of individual clusters (47 Tuc, NGC 288, NGC 1851, M 30, ω Cen, NGC 362 and M 67) as a part of the pilot survey. An additional two clusters (NGC 2516 and the Pleiades) were observed because their stars happen to fall into the magnitude range of the main survey or K2-HERMES survey. A list of the observed clusters is given in Table 2. In the pilot survey only probable members were observed. To confirm the membership of such stars, as well as members of serendipitously observed clusters, we made additional cuts in radial velocity, position, and in some instances, their proper motion. Details are given in Appendix C. The exceptions are three clusters for which we found members in the literature. There are also three more clusters (Hyades, NGC 2243, and NGC 6362) in the observed fields, but we could not find any of their members among survey stars with valid parameters and abundances.

Fig. 1 shows the abundances for all 13 elements in the observed clusters. Note that the scatter for some elements is consistently high, regardless of the cluster. Around half of the scatter is statistical noise (see Table 3). We demonstrate this by measuring the uncertainties of the abundances from repeated observations of field stars and M67. Only measurements made from spectra with signal-to-noise ratio (S/N) ≥45 per pixel were used in order to distinguish between repeats done with the purpose of quality estimation and repeats

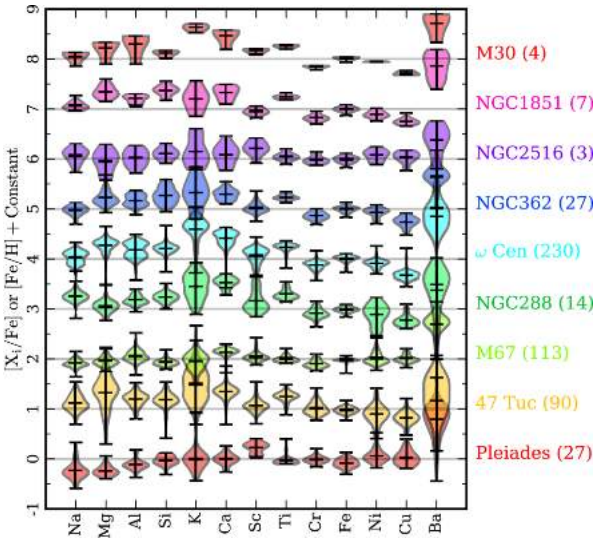

Figure 1. Abundances of 13 elements in 9 studied clusters. A violin plot represents the distribution of measured abundances for all stars that we identified as cluster members. The number of all stars in each cluster is given next to the cluster name. Note how some elements have consistently more scattered distributions.

Table 3. Uncertainties measured from 1579 repeated observations in the whole GALAH sample and 377 repeated observations of M67 stars compared to scatter observed in Fig. 1, and weights assigned to each element. Uncertainties and scatters are expressed as standard deviations.

Element	Uncertainty from all repeats dex	Uncertainty from M67 repeats dex	Scatter in all clusters dex	Weight
Na	0.063	0.063	0.122	1.0
Mg	0.078	0.079	0.162	0.5
Al	0.066	0.063	0.129	1.0
Si	0.053	0.053	0.106	1.0
K	0.099	0.129	0.228	0.25
Ca	0.065	0.056	0.131	0.5
Sc	0.050	0.054	0.115	1.0
Ti	0.044	0.048	0.071	2.0
Cr	0.047	0.049	0.081	2.0
Fe	0.024	0.021	0.060	2.0
Ni	0.056	0.057	0.112	1.0
Cu	0.049	0.036	0.095	2.0
Ba	0.114	0.135	0.230	0.25

done to boost the S/N of some lower quality data. The rest of the scatter is systematic, arising from the abundance determination pipeline being sensitive to temperature variations or dwarf–giant distinction.

The scatter in the Ba and K abundances is highest. Elements like Fe, Ti and Cr have lower uncertainties. It is therefore not fair to treat elements with different uncertainties as equally important dimensions in the $\mathcal{C}$-space. Before we use the abundances in t-SNE, we standardize them so that the distribution of abundances of every element has a zero median and a standard deviation of unity. Standardization is done once for the complete data set (187 640 stars). Then we change the standard deviation of the standardized set based on the weights that are proportional to the scatter we observe in clusters. We are confident that there are no misidentified members contributing to the scatter (see Appendix C). Elements with more scatter will have a narrower distribution, so the distances in those dimensions will always be damped and will not carry as much importance as those for less scattered elements. Because it is hard to quantitatively determine the weight for each element, we will distribute the elements into four groups. Ba and K have by far the highest scatter, so they will be given a weight of 0.25. Mg and Ca also have high scatter, so they will have weights equal to 0.5. Fe, Ti, Cr and Cu have the smallest scatter and will have a weight of 2.0, and the rest of the elements will have a weight of 1.0. Weights, uncertainties measured from the repeated observations, scatter in clusters and related weights are collected in Table 3. Weights are a way to implement uncertainties into the t-SNE, as in our case the uncertainties of individual measurements have not been estimated. Without these weights, there would be fewer groups in the t-SNE map and the stars from known clusters would end up scattered over a larger area.

We use the weighted abundances to produce a t-SNE projection for a region around Pleiades (Fig. 2) and other clusters (Appendix B). In cases where we have more than one measurement for a star, as for most M67 stars and some ω Cen stars, we first calculated the average abundances for each star and used those in the t-SNE. Stars with repeated observations are therefore only plotted once in the t-SNE maps. No other information is used in the projection, even though other stellar parameters (i.e. $T_{\rm eff}$, $\log g$, v_r, ...) are displayed in the colour-coded t-SNE maps. One can pick out many groups in the t-SNE map, some more pronounced than the others. Groups associated with each cluster are marked and we leave a detailed analysis of other pronounced groups for a later study.

Fig. 3 shows only the members in the t-SNE maps for the eight remaining clusters. Out of all nine clusters, we claim that t-SNE gives good results for all but two of them. In NGC 2516 the membership of three stars is not completely certain, so we cannot base our conclusions on this cluster. Note that we only cover an edge of the cluster in one of the observed fields, so a low number of members are expected. We did not find any large groups in $\mathcal{C}$-space for 47 Tuc, so the chemical tagging of this cluster was unsuccessful. It is not clear why only 47 Tuc was so resistant to chemical tagging, since all globular clusters have inhomogeneities in their light-element abundances (e.g. Thygesen et al. 2014), and we were able to successfully tag stars from the other globular clusters. Perhaps the abundances in other globular clusters are distinct enough from field stars that they still form an isolated group.

In most of the tagged clusters, we find a small number of outliers: stars that are cluster members, but do not lie in the same chemical group as the rest of the stars. This is due to our measured abundances being significantly different, so t-SNE did not associate them with the main part of the cluster. Some outliers are expected, as our reduction and analysis pipelines do not produce perfect results. We also expected to see traces of fibre numbers in the t-SNE map (there are two sets of 392 fibres in the fibre positioner, so different stars could be observed with a given fibre with a problematic PSF, which might manifest itself as a systematic error in the measured abundance). With the exception of a few ill-performing fibres, we see no relation between measured abundances and fibre used.

One can also notice that every cluster's chemical group is populated with some stars that are not members. This contamination is expected, as 13 abundances are not enough to completely isolate the cluster (Mitschang, De Silva & Zucker 2012; Ting et al. 2012; Ting, Conroy & Goodman 2015). We explore these stars further in the case of Pleiades in the next section. We chose the Pleiades for this experiment because it is the only young cluster with distinct kinematics that we can use to verify potential new member candidates.

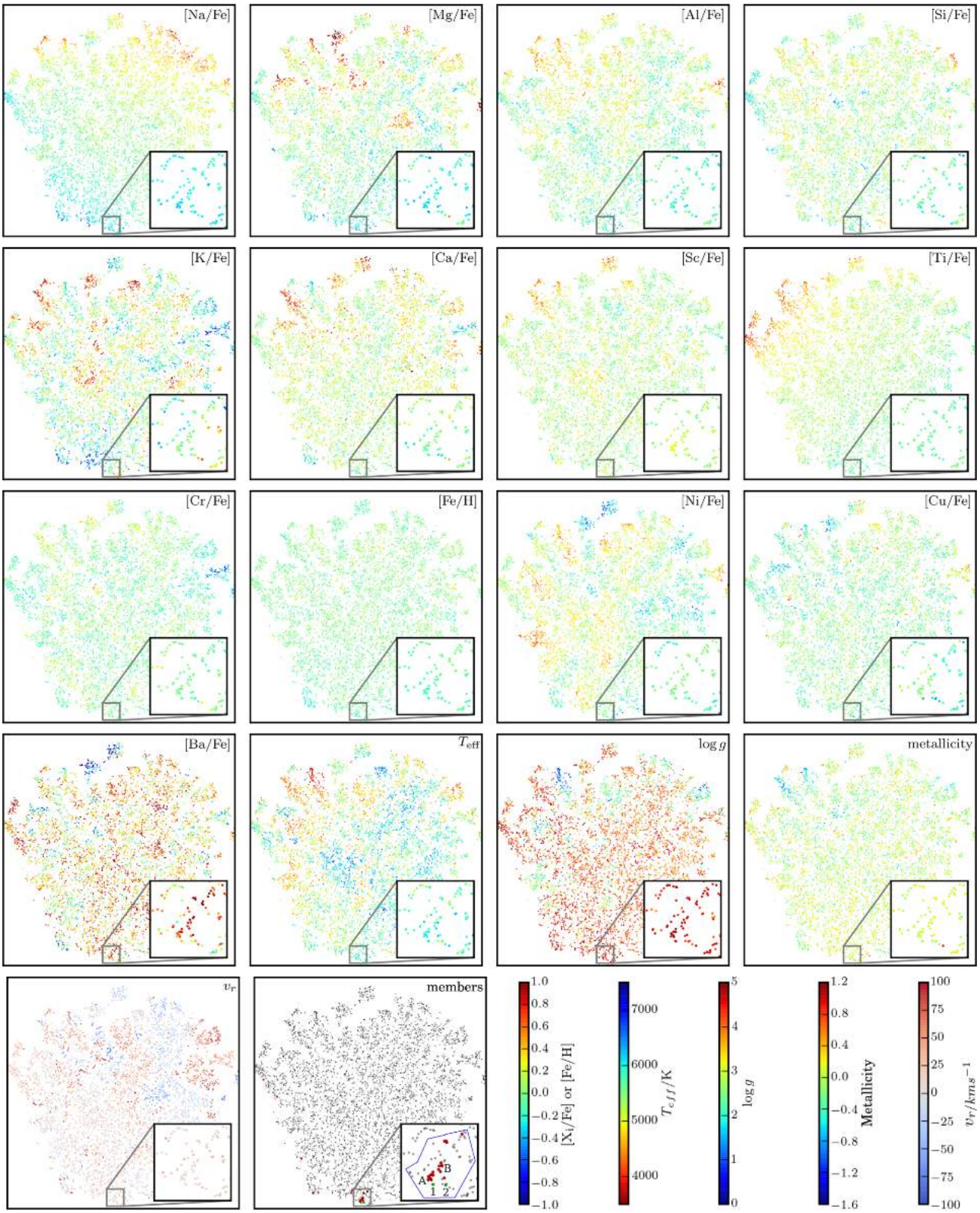

Figure 2. t-SNE projection of 9408 stars in a 40° radius around the Pleiades. Each panel shows the same projection with different colour-codes for different quantities (given in the top right-hand corner of each panel). Abundances of 13 elements, as well as $T_{\rm eff}$, log g, metallicity and radial velocity are colour coded. The panel labeled 'members' shows the stars that belong to the cluster in red and field stars in grey. Two stars marked in green and numbered 1 and 2 are newly discovered Pleiades members discussed in Section 5. They lie slightly away from the rest of the Pleiades because of their slightly different abundances, more clearly illustrated in Fig. 4. 17 out of 27 Pleiades stars lie in the two tight groups in the bottom of the map marked A and B. The blue polygon marks the Pleiades' chemical group (see Section 5 and Fig. 6 for details). Colour version of this figure is available online.

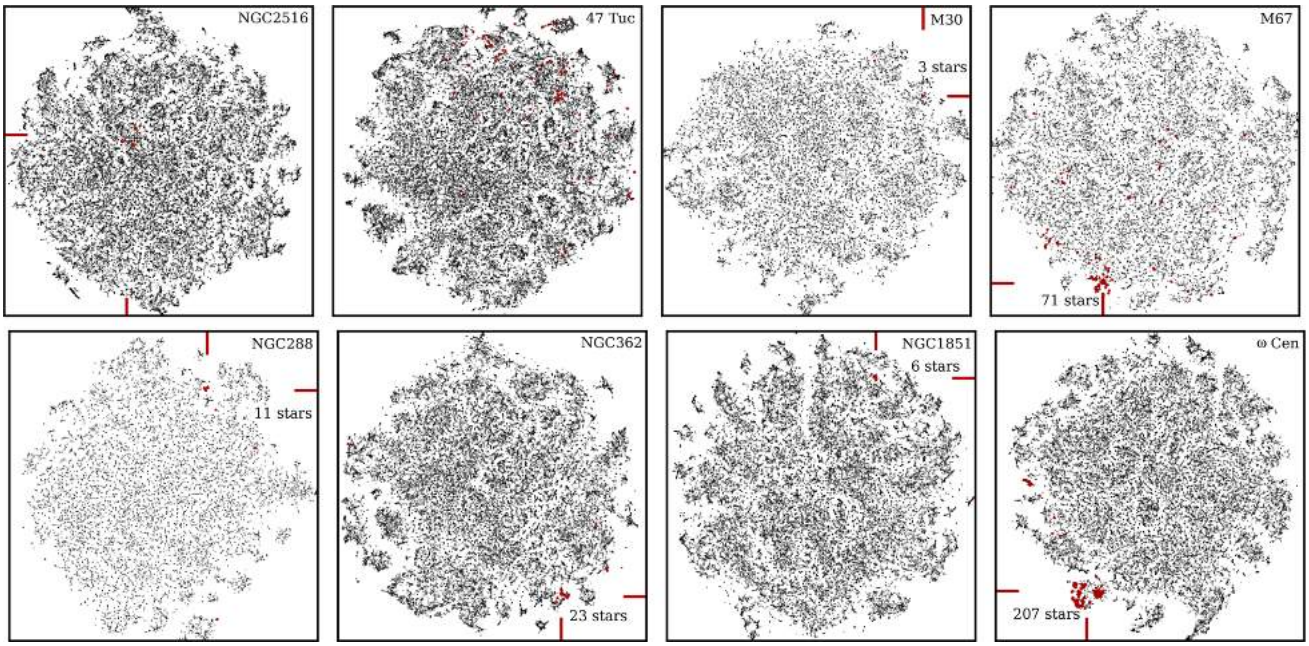

Figure 3. t-SNE projections for regions around the eight remaining clusters. Known members are marked in red and the number in some panels tells the number of the stars in the main group, as the points often overlap. Top left-hand panel, to bottom right-hand panel the following clusters are shown: NGC 2516 (41 106 stars in a 30° radius), 47 Tuc (44 037 stars in a 35° radius), M30 (20 254 stars in a 35° radius), M67 (25 648 stars in a 45° radius), NGC 288 (11 535 stars in a 45° radius), NGC 362 (41 578 stars in a 35° radius), NGC 1851 (33 882 stars in a 35° radius) and ω Cen (33 281 stars in a 30° radius). Red marks on the edge of each panel point towards the position of the main group of members. Colour version of this figure is available online.

5 CHEMICAL POPULATIONS AND NEW MEMBERS OF THE PLEIADES

The Pleiades is a young (Brandt & Huang 2015) cluster for which we expect to find some members well away from the centre of the cluster, yet close enough that we can focus only on a small region around the cluster (Kroupa & Boily 2002). The tidal radius of the Pleiades is ∼6° (Adams et al. 2001), and we do not expect to find any members at distances much larger than this, considering a low number of observed stars in the broader Pleiades region. Despite this, we focus our effort into an area of radius 40° around the Pleiades to demonstrate the method on a larger number of stars. The Pleiades are one of the most northerly objects that GALAH has explored, in one of the K2 fields, so the 40° radius region includes mostly K2-HERMES survey fields, a few pilot survey fields and some regular survey fields at $\delta < +10°$ that have been observed, but most of the 40° region has no observations at all. The Pleiades members were identified by us using cuts in the position, radial velocity and proper motions (see Appendix C). In this way we identified 27 members.

After making the t-SNE map, we can see in Fig. 2 that most Pleiades stars fall into one clump with few contaminating field stars. Closer inspection shows, that the clump consists of two parts (groups marked A and B on Fig. 3) with slightly different abundances of [Sc/Fe], [Ba/Fe] and [Fe/H] (Fig. 4). Stars from both groups are well mixed within 0.5 dex in $\log g$ and within 1000 K in $T_{\rm eff}$, where stars from group A are on average hotter than stars from group B with a large overlap. We tried to verify the two chemical groups by analysing the Pleiades spectra with SME (Valenti & Piskunov 1996; Piskunov & Valenti 2017). We analysed the members in the Pleiades chemical groups, as well as the two new member candidates. With SME, we managed to measure more elements (18), but not for every star. Fig. 5 shows the SME results. We cannot confirm the existence of two chemical groups we see in The Cannon abundances, so they

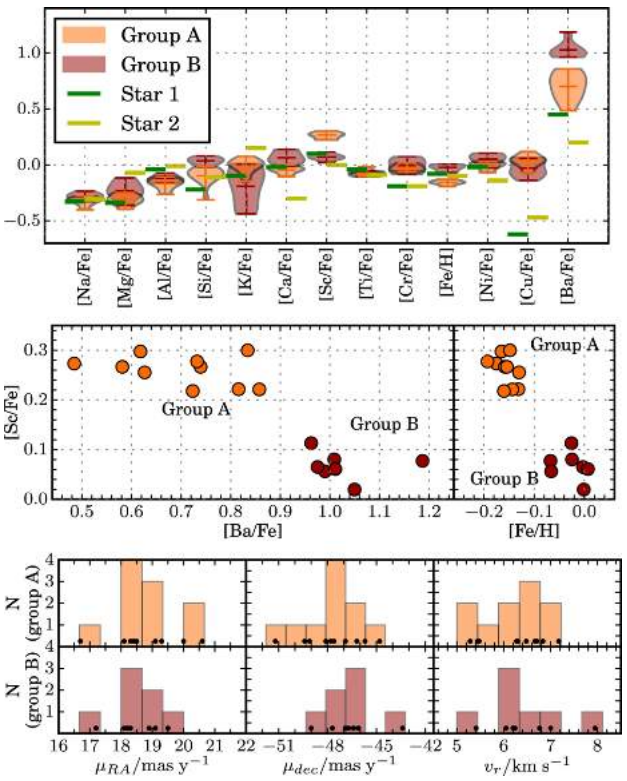

Figure 4. Top panel: the abundances for all 13 elements in groups marked A and B in Fig. 2 and abundances for two new members. Middle panel: Groups A and B are separated in [Sc/Fe], [Ba/Fe] and [Fe/H] abundances. Abundances for each star are plotted. Bottom panel: kinematics for each group.

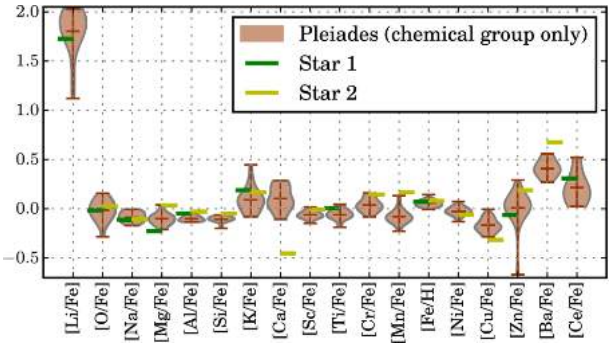

Figure 5. SME abundances for the Pleiades stars from the Pleiades chemical group. Note that there is no bimodality in [Sc/Fe], [Ba/Fe], and [Fe/H] like we see with The Cannon abundances.

are most probably an artefact induced by The Cannon or selection of the training set. The scatter in [Sc/Fe], [Ba/Fe] and [Fe/H] when calculated by SME is similar to what we get with The Cannon, though. Any decisive conclusions will require more data and more careful analysis of abundances. There are similar observations in the literature (Gebran & Monier 2008) matching our large scatter in [Sc/Fe] and [Ba/Fe] abundances, but not confirming two separate groups, which could be due to the low number of observed stars in those studies. Binary clusters do exist (Slesnick, Hillenbrand & Massey 2002) and we can speculate that the Pleiades might be a binary or merged cluster, if two separate chemical groups existed. The idea of two populations in Pleiades has even been proposed before (Stello & Nissen 2001). In any case, we show that features like split chemical groups can be picked out by t-SNE whilst still conserving the hierarchy and putting both groups close together.

Different abundances measured by The Cannon and SME are a good indicator that both methods are susceptible to systematic errors which we were not able to analyse with the current data set. Bots methods give reliable measurements for our study, as the chemical tagging is based on relative abundances only.

We define the Pleiades chemical group by combining all small groups with at least one known Pleiades member that are close to the main group. This decision is arbitrary but conservative. The chemical group is marked with a blue polygon in Fig. 2.

We find a small number of contaminating stars in the Pleiades chemical vicinity. Some contamination is expected, therefore we cannot claim that all the stars in the chemical group are Pleiades members. For clusters with adequate kinematic information, however, 13 abundances are enough, as we can use independent dimensions: radial velocity, amplitude of the proper motion and direction of the proper motion. We also have photometric information that we combine into a single parameter: the distance (Zwitter et al. 2010). These four additional dimensions are enough to select only those stars with kinematics and distances that match the Pleiades'. This leaves us with two stars that we claim are candidate Pleiades members. The process of reducing ∼30 contaminating stars into two candidate members is illustrated in Fig. 6. This can also be confirmed with the SME abundances (see Fig. 5).

There are actually two more stars in the whole 40° radius region that match the Pleiades' kinematics. They are both >20° away from the cluster and do not fall near the Pleiades' chemical group in the t-SNE map. This means that after reducing the number of stars from ∼9400 and ∼2 coincidental stars to ∼30 stars by chemically tagging the cluster, we expect to find $2\frac{30}{9400} \simeq 0.0064$ stars that by chance have the same kinematics as the Pleiades and that fall into the cluster's chemical group. We found two, which are therefore Pleiades members with a high degree of certainty. It must be noted, that one of the newly discovered members (star number 2) is a known supercluster candidate (Mermilliod, Bratschi & Mayor 1997) that escaped our cluster membership determination for being too far from the cluster centre. Star number 1, however, has no relation to the Pleiades in the literature.

6 DISCUSSION

We show that t-SNE is an appropriate algorithm to search for clustering in $\mathcal{C}$-space by demonstrating its performance on all clusters observed in GALAH and K2-HERMES surveys. With reasonable exceptions the method performs well, which we further demonstrate by discovering previously unknown members of Pleiades.

Perhaps an even more important conclusion is that extremely precise abundances are not always needed to successfully chemically tag cluster members or maybe even field stars. It turns out that the ability to find structures in $\mathcal{C}$-space is more valuable than extreme accuracy of the data. Precise abundances help by reducing the number of outliers and increasing the prominence of the chemical groups, but we show that the groups exist and can be isolated even when chemical homogeneity is only of the order of 0.1 dex. This is best shown by being able to match the two new Pleiades members to the cluster, even when the abundances do not always agree with the abundances of Pleiades. Even with large discrepancy in [Cu/Fe], the two stars still lie close to the Pleiades manifold and are therefore correctly identified as members.

We do see some pollution from field stars in the mapped groups associated with the clusters. This is largely due to only 13 elements being used for chemical tagging, two of which (K and Ba) were also given low weights. There are lines of up to 30 elements in the observed wavelength ranges, so in the future we plan to use more elements and reduce the level of pollution. It must be noted, that for some clusters and dissolved groups chemical tagging might remain unfeasible, if their abundances are not distinct enough from the observed population of stars. With the current data, a large fraction of the stars are untaggable. In t-SNE maps they are collected in the middle, with no convincing structure visible that would separate them from each other. Again, this is something that more observed elements can solve. Despite the mentioned limitations, we conclude that clusters tagged by t-SNE experience low pollution and fairly high efficiency, especially when compared to competing methods.

A further demonstration of the power and robustness of t-SNE is the hierarchy seen in some clusters where more than one population is found. Different populations form different groups, but they all compose one larger group that includes the majority of the cluster members. In a two-dimensional map, one can easily see and correctly interpret these structures. This is very hard to do in a higher-dimensional space without a good visualization of all relevant dimensions. Built-in hierarchy also saves us from tagging dwarfs and giants separately, as some studies in the literature do. It can be seen from the $\log g$ panel in Fig. 2, and even better from the figures in Appendix B, that giants are mostly separated from predominantly dwarf-populated t-SNE maps. t-SNE knows nothing about $\log g$ and the result is purely a consequence of abundances being dependent on gravity. This can be either an abundance pipeline issue or a result of different stellar populations observed.

Adopting the Pleiades parallax of $\varpi = 7.48 \pm 0.03$ mas (Gaia Collaboration et al. 2017), median proper motion of our known Pleiades members ($\mu = 48.96\,\mathrm{mas\,y^{-1}}$), and proper motion of the newly discovered members, we can calculate, that these stars were scattered out of the cluster $7.9 \pm^{\infty}_{4.6}$ Myr (star 1) and

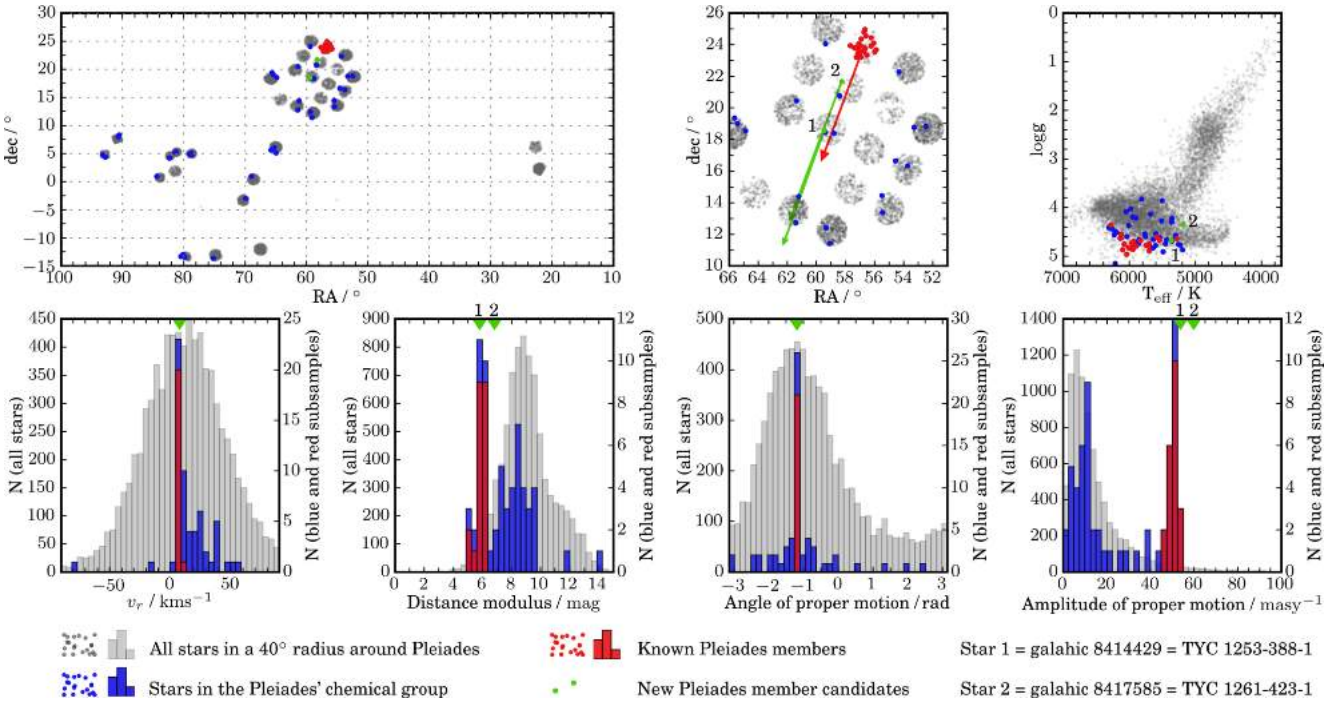

Figure 6. Position of the analysed stars on the sky (top left-hand panel and top middle panel), on the HR diagram (top right-hand panel) and on the radial velocity, distance and proper motion histograms (bottom row). In grey are plotted all the analysed stars in the 40° radius around Pleiades. In blue are the stars that belong to the Pleiades chemical group (inside the blue polygon in Fig. 2). In red are known Pleiades members also marked with red symbols in Fig. 2. Green are two new Pleiades stars that we discovered by chemical tagging. Colour version of this figure is available online.

0.68 ± 0.05 Myr (star 2) ago, at a velocity significantly larger than the escape velocity of the cluster. Considering where in the HR diagram the two stars lie, they are indeed good candidates to be ejected from the cluster due to their low mass.

It was unexpected that our method would work well for globular clusters. Globular clusters have large scatter in light elements and are often chemically inhomogeneous. Results like ones for 47 Tuc were therefore expected. Other globular clusters (with the exception of NGC 2516, where the results are inconclusive) performed well, especially ω Cen. These globular clusters are interesting targets for further studies with chemical tagging as they are obviously easiest and most reliable to tag. Stars from these globular clusters have abundances distinct enough from fields stars that they were successfully tagged. It is possible that t-SNE is robust enough that with more measured element in the future the tagging will work for 47 Tuc as well.

This is an exploratory study in the early years of the GALAH survey to demonstrate that it is feasible to extract homogeneous clusters from a huge stellar survey. There are still numerous improvements to be made to the stellar abundance determinations, including 3D non-LTE atmospheric corrections (Lind et al. 2017), better absorption line measurements using a photonic comb (Bland-Hawthorn et al. 2017) and new data driven techniques to ensure abundance uniformity across the survey (Ness et al. 2015). Thus the efficacy of chemical tagging will only improve in the years to come.

One can see that the maps in Fig. 3 show many more structures than we have analysed in this paper. We explored other chemical groups and observed some regularities and patterns when kinematics and positions on the sky were inspected. There are, however, some contaminating stars in these groups as well and decisive conclusions are hard to make. We leave the topic of pure blind chemical tagging of field stars for future work. Blind chemical tagging will also be much easier on the set of 30 abundances and >300 000 stars soon to be produced by the GALAH collaboration. More observed elements mean much less contamination of chemical groups, so we might soon be able to find long-lost relationships between field stars for the first time with good reliability. We also expect to find many more distinct clusters than we can see in the presented t-SNE maps (Bland-Hawthorn & Sharma 2016).

ACKNOWLEDGEMENTS

JK is supported by a Discovery Project grant from the Australian Research Council (DP150104667) awarded to J. Bland-Hawthorn and T. Bedding. TZ acknowledge the financial support from the Slovenian Research Agency (research core funding No. P1-0188). SM acknowledges support from the Australian Research Council through grant DE140100598. DMN is supported by the Allan C. and Dorothy H. Davis Fellowship. DBZ acknowledges the financial support of the Australian Research Council through grant FT110100793

REFERENCES

Adams J. D., Stauffer J. R., Monet D. G., Skrutskie M. F., Beichman C. A., 2001, AJ, 121, 2053
Barden S. C. et al., 2010, Proceedings of the SPIE, Volume 7735, HERMES: Revisions in the Design for a High-Resolution Multi-Element Spectrograph for the AAT. SPIE, Bellingham, p. 773509,
Bensby T., Feltzing S., Oey M. S., 2014, A&A, 562, A71
Blanco-Cuaresma S. et al., 2015, A&A, 577, A47
Bland-Hawthorn J., Freeman K. C., 2004, PASA, 21, 110
Bland-Hawthorn J., Sharma S., 2016, Astron. Nachr., 337, 894
Bland-Hawthorn J., Krumholz M. R., Freeman K., 2010a, ApJ, 713, 166

Bland-Hawthorn J., Karlsson T., Sharma S., Krumholz M., Silk J., 2010b, ApJ, 721, 582
Bland-Hawthorn J., Kos J., Betters C., De Silva G., O'Byrne J., Patterson R., Leon-Saval S., 2017, Opt. Express, 25, 15614
Bovy J., 2016, ApJ, 817, 49
Brandt T. D., Huang C. X., 2015, ApJ, 807, 58
Bu Y., Chen F., Pan J., 2014, New A, 28, 35
Daniel S. F., Connolly A., Schneider J., Vanderplas J., Xiong L., 2011, AJ, 142, 203
De Silva G. M., Sneden C., Paulson D. B., Asplund M., Bland-Hawthorn J., Bessell M. S., Freeman K. C., 2006, AJ, 131, 455
De Silva G. M., Freeman K. C., Asplund M., Bland-Hawthorn J., Bessell M. S., Collet R., 2007, AJ, 133, 1161
de Silva G. M., Freeman K. C., Bland-Hawthorn J., Asplund M., Williams M., Holmberg J., 2011, MNRAS, 415, 563
De Silva G. M., D'Orazi V., Melo C., Torres C. A. O., Gieles M., Quast G. R., Sterzik M., 2013, MNRAS, 431, 1005
De Silva G. M. et al., 2015, MNRAS, 449, 2604
Feng Y., Krumholz M. R., 2014, Nature, 513, 523
Freeman K., Bland-Hawthorn J., 2002, ARA&A, 40, 487
Gaia Collaboration et al., 2017, A&A, 601, A19
Gebran M., Monier R., 2008, A&A, 483, 567
Geller A. M., Latham D. W., Mathieu R. D., 2015, AJ, 150, 97
Heiter U., Jofré P., Gustafsson B., Korn A. J., Soubiran C., Thévenin F., 2015, A&A, 582, A49
Hogg D. W. et al., 2016, ApJ, 833, 262
Jeffries R. D., Thurston M. R., Hambly N. C., 2001, A&A, 375, 863
Jofré P. et al., 2015, A&A, 582, A81
Jofré P., Das P., Bertranpetit J., Foley R., 2017, MNRAS, 467, 1140
Kharchenko N. V., Piskunov A. E., Schilbach E., Röser S., Scholz R.-D., 2013, A&A, 558, A53
Kos J. et al., 2017, MNRAS, 464, 1259
Kroupa P., Boily C. M., 2002, MNRAS, 336, 1188
Lind K. et al., 2017, MNRAS, 468, 4311
Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, ApJS, 225, 31
Martell S. L. et al., 2016, ApJ, 825, 146
Martell S. L. et al., 2017, MNRAS, 465, 3203
Matijevič G. et al., 2017, A&A, 603, A19
Mermilliod J.-C., Bratschi P., Mayor M., 1997, A&A, 320, 74
Mitschang A. W., De Silva G. M., Zucker D. B., 2012, MNRAS, 422, 3527
Mitschang A. W., De Silva G., Zucker D. B., Anguiano B., Bensby T., Feltzing S., 2014, MNRAS, 438, 2753
Ness M., Hogg D. W., Rix H.-W., Ho A. Y. Q., Zasowski G., 2015, ApJ, 808, 16
Piskunov N., Valenti J. A., 2017, A&A, 597, A16
Quillen A. C., Anguiano B., De Silva G., Freeman K., Zucker D. B., Minchev I., Bland-Hawthorn J., 2015, MNRAS, 450, 2354
Sestito P., Randich S., Bragaglia A., 2007, A&A, 465, 185
Slesnick C. L., Hillenbrand L. A., Massey P., 2002, ApJ, 576, 880
Stello D., Nissen P. E., 2001, A&A, 374, 105
Stello D. et al., 2016, ApJ, 832, 133
Stello D. et al., 2017, ApJ, 835, 83
Thygesen A. O. et al., 2014, A&A, 572, A108
Ting Y.-S., Freeman K. C., Kobayashi C., De Silva G. M., Bland-Hawthorn J., 2012, MNRAS, 421, 1231
Ting Y.-S., Conroy C., Goodman A., 2015, ApJ, 807, 104
Traven G. et al., 2017, ApJS, 228, 24
Tucholke H.-J., 1992, A&AS, 93, 293
Valenti J. A., Piskunov N., 1996, A&AS, 118, 595
Valentini M. et al., 2017, A&A, 600, A66
van der Maaten L., 2014, J. Mach. Learn. Res., 15, 3221
van der Maaten L., Hinton G., 2008, J. Mach. Learn. Res., 9, 85
Vanderplas J., Connolly A., 2009, AJ, 138, 1365
Wattenberg M., Viégas F., Johnson I., 2016, Distill
Zwitter T. et al., 2010, A&A, 522, A54

APPENDIX A: INTRODUCTION TO T-SNE

Dimensionality reduction methods aim to reduce the number of dimensions while preserving the structure of the data that we are interested in. Here, we want to identify groups of data points in a 13-dimensional space. This can be done in the original 13 dimensions, but the visualization would remain a problem. Also, we know that chemical groups are not very distinct or isolated from each other (there are field stars with very similar abundances), so fine-tuning an algorithm in 13 dimensions is nearly impossible.

Linear and non-linear dimensionality reduction. Linear algorithms, like PCA, are not very suitable for chemical tagging, especially if we first intend to project the data into two dimensions. Even though the abundances of different elements are correlated, the relations are not linear, so the projection into only two dimensions is unable to conserve the structure from the high-dimensional space. We illustrate the problem in Fig. A1. A double helix constructed in three dimensions is projected into two dimensions with t-SNE. One can see that both strands of the double helix become separated in the t-SNE projection. A linear method would not be able to

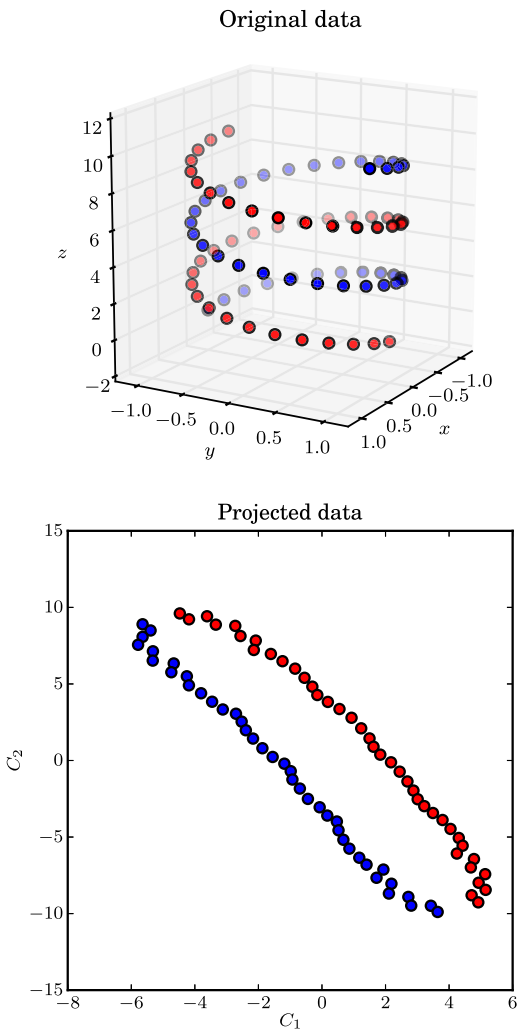

Figure A1. t-SNE projection of a double helix into two dimensions. All information about the shape of the structure is lost, but two strands become separated. Colours are used for the illustration purposes only and are not part of the data set, meaning that t-SNE only knows the coordinates of the points, regardless the colour.

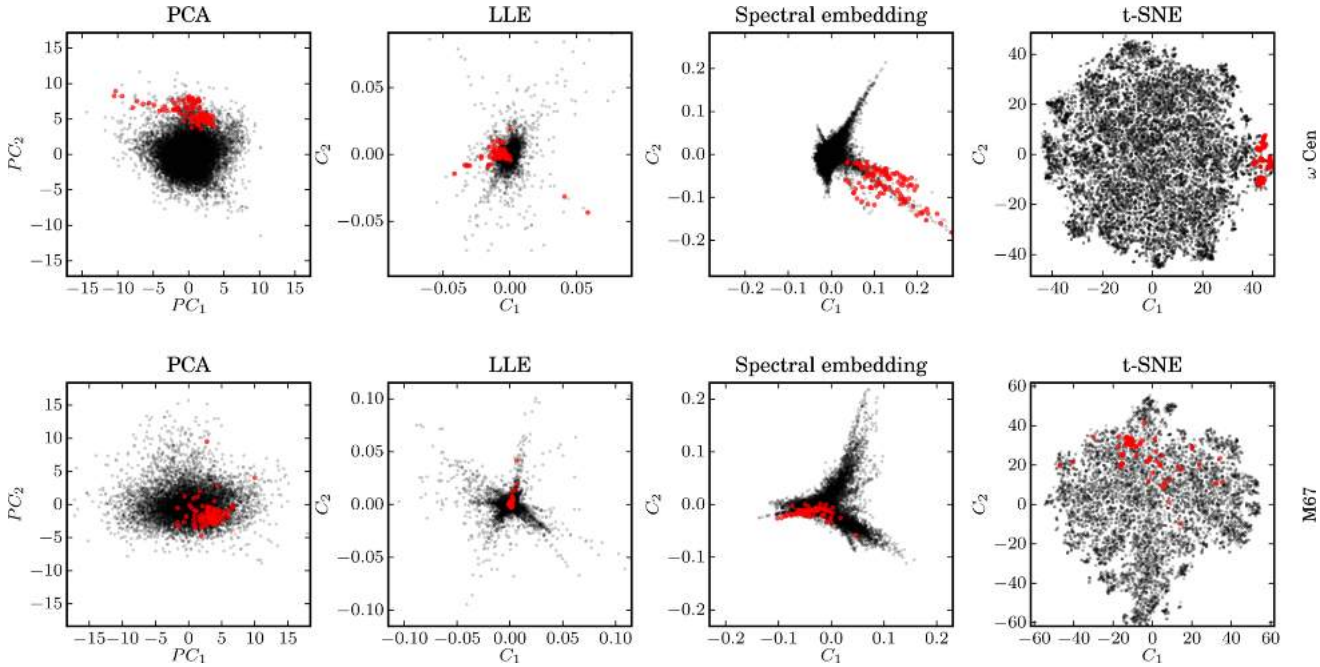

Figure A2. Comparison of the performance of PCA, LLE, Spectral embedding and t-SNE. Methods follow from the most to the least linear (left- to right-hand panels). In the top row, we compare the four methods on the case of ω Cen (an easy case) and in the bottom row for M67 (a harder case). Known cluster members are marked in red. Notice how efficiently t-SNE covers the plane and how many more distinct groups one can see. t-SNE projections in this figure can be compared with projections for the same two clusters in Fig. 3. Projections are not the same, even though the same parameters were used. Chemical groups, however, are almost identical.

produce that. Any linear projection will result in either a ring or two interlocking 'waves' of points. But most non-linear dimensionality reduction algorithms will be able to deal with this example.

Local and global algorithms. Most differences between non-linear algorithms are in the mapping of local and global details. Imagine the previous example, but with added outliers somewhere far away from the double helix. Global algorithms will try to preserve the structure at all scales. Points that are close together will remain close together in the two-dimensional projection and distant outliers will be placed far away. The double helix structure, however, might not be resolved as the distances between the points in the double helix are negligible compared to the distances to the outliers. With local algorithms, the position of each point in the two-dimensional space is determined only by its nearest neighbours. Local algorithms can 'see' the two strands of the double helix, but will fail to map the outliers as their nearest neighbours are points on the double helix. Positions of outliers on the two-dimensional map will therefore be meaningless.

t-SNE is able to adapt itself to local density. It can map data sets with a high variation in density, so it is able to resolve small, local details, as well as the global picture.

Two examples in Fig. A2 show a projection of elemental abundances for 13 elements for stars around ω Cen and M67 made with four different algorithms. ω Cen stars have peculiar enough abundances that they stand out from rest of the stars. M67 stars have abundances that are much closer to field stars, so it is one of the hardest clusters to chemically tag in our sample. One can see in Fig. A2 that only t-SNE is able to create a map where cluster stars lie in the region not densely populated by field stars. Locally linear embedding (LLE) and spectral embedding are able to identify some groups of stars that are mapped into rays extending from the central region. They cannot identify structures inside the central region, while t-SNE distributes stars pretty evenly, identifying many groups where LLE and spectral embedding fail. PCA and spectral embedding are able to tag ω Cen stars, but the pollution from the field stars is much higher than in t-SNE maps.

t-SNE algorithm. The main difference between t-SNE and other methods discussed here is that t-SNE does not use a fixed number of nearest neighbours to determine the position of a point in the two-dimensional map. Instead, a neighbour is only used with a probability that depends on the distance between the data points (equation 2) under a Gaussian distribution with dispersion σ_i. This way, even points far away can be considered to calculate the position of a point in two dimensions (they are used as 'nearest neighbours'). If we imagine a two-dimensional projection, not necessarily an optimal one, a similar probability can be calculated in two dimensions (equation 4), this time requiring that the 'nearest neighbours' resemble a Student's t distribution. A Student's t distribution used instead of the Gaussian makes t-SNE sensitive to fine and global structures. An optimal projection is where the Gaussian and Student's t distributions are as close to each other as possible. This is found by minimizing the sum of Kullback–Leibler divergences for each data point, which is computationally intensive.

Role of perplexity. σ_i in equation (2) is the second moment of the Gaussian distribution (dispersion). It should be smaller for data points in denser regions of the high-dimensional space and larger in sparse regions, assuming regions of different densities are mapped simultaneously. t-SNE finds σ_i for every data point, such that the perplexity

$$\mathrm{Perp}_i = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}} \tag{A1}$$

equals to the perplexity specified by the user. The specified perplexity is typically in the range of 5–50.

Perplexity therefore controls whether t-SNE is more sensitive to large or small structures. Its role is similar to the number of nearest neighbours used by most other dimensionality reduction

methods. See Wattenberg et al. (2016) for an excellent interactive demonstration of the role of perplexity, as well as other caveats of the t-SNE method.

Visualization. t-SNE does not retain distances but probabilities. It is also highly non-linear, so the values for the coordinates of data points in the projected map are meaningless. So are the units. They can be thought of as coordinates of an image. In Figs A1 and A2, we show the two axes with corresponding numerical values to spare the reader any confusion. In the main text, we omit them altogether, so the projected map is actually treated as an image.

Kullback–Leibler divergence is minimized by a gradient descent method initialized by random sampling, so t-SNE will produce a slightly different map on every run, even with the same data and same perplexity. In the case of chemical tagging, usually the only visible difference is a random rotation of the map. The difference can be seen, if Figs 3 and A2 are compared. In general, one can run t-SNE many times and use the projection with the lowest Kullback–Leibler divergence.

At the beginning of this paper, we introduced t-SNE to avoid finding groups in a 13-dimensional space. The problem of how to find groups exists in the two-dimensional map as well. Any group finding algorithm can be used and, thanks to t-SNE, can be inspected visually. Such approach is used in Traven et al. (2017), for example. In this paper, however, we find it unnecessary, as we are only interested in one particular group of stars (one with cluster members) in each two-dimensional map. A more systematic analysis is therefore not needed.

APPENDIX B: T-SNE PROJECTIONS OF THE REMAINING CLUSTERS

t-SNE maps for the eight clusters, not presented in the main text, are collected here.

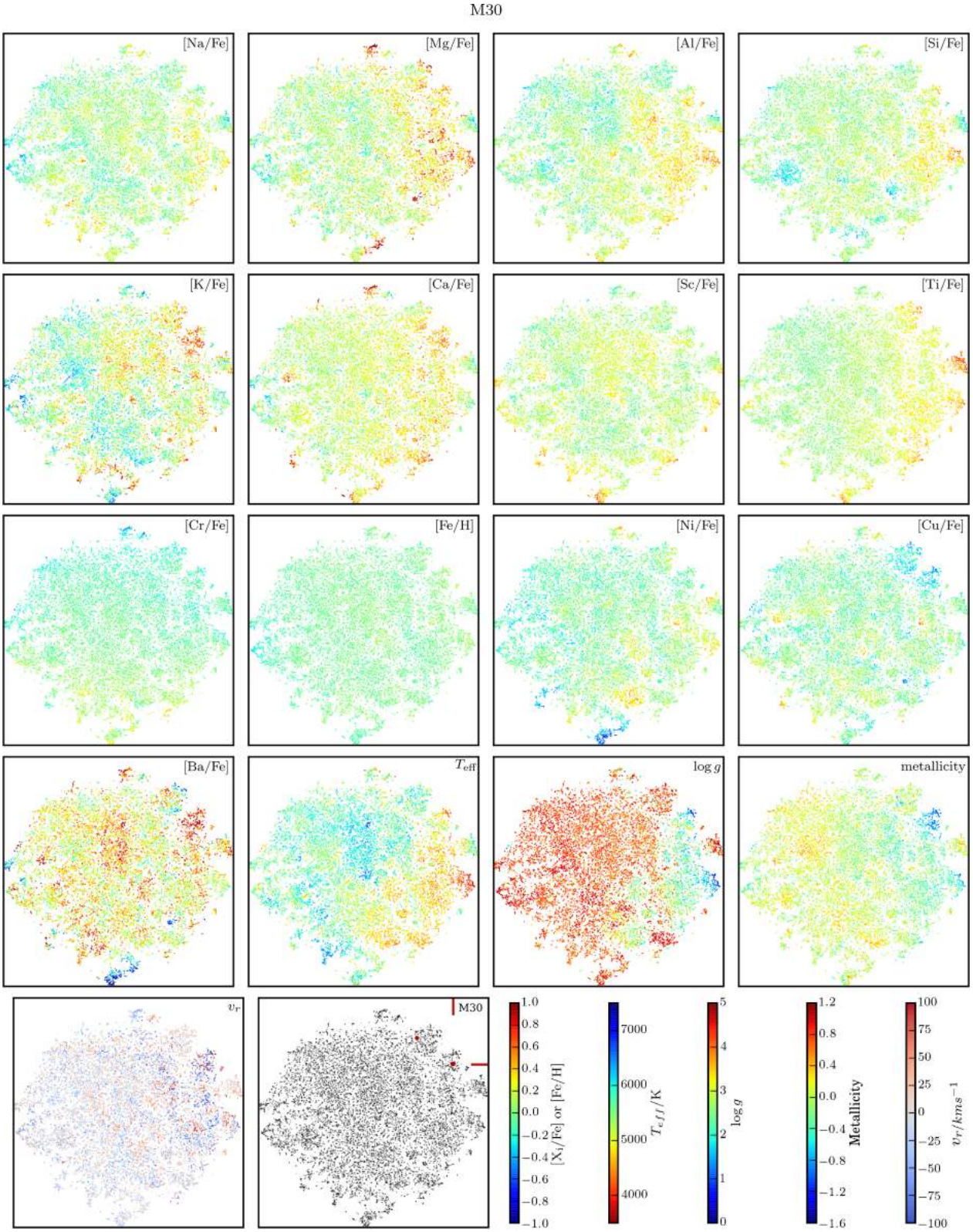

Figure B1. t-SNE projection of 20 254 stars in a 35° radius around M30. Abundances of 13 elements used to create the projection are colour-coded. $T_{\rm eff}$, log g, metallicity and radial velocity colour-codes are also plotted. The last panel shows the stars that belong to the cluster in red and field stars in grey. Three out of four M30 stars lie in a tight group in the top right-hand part of the map (marked with dashes at the edge of the plot).

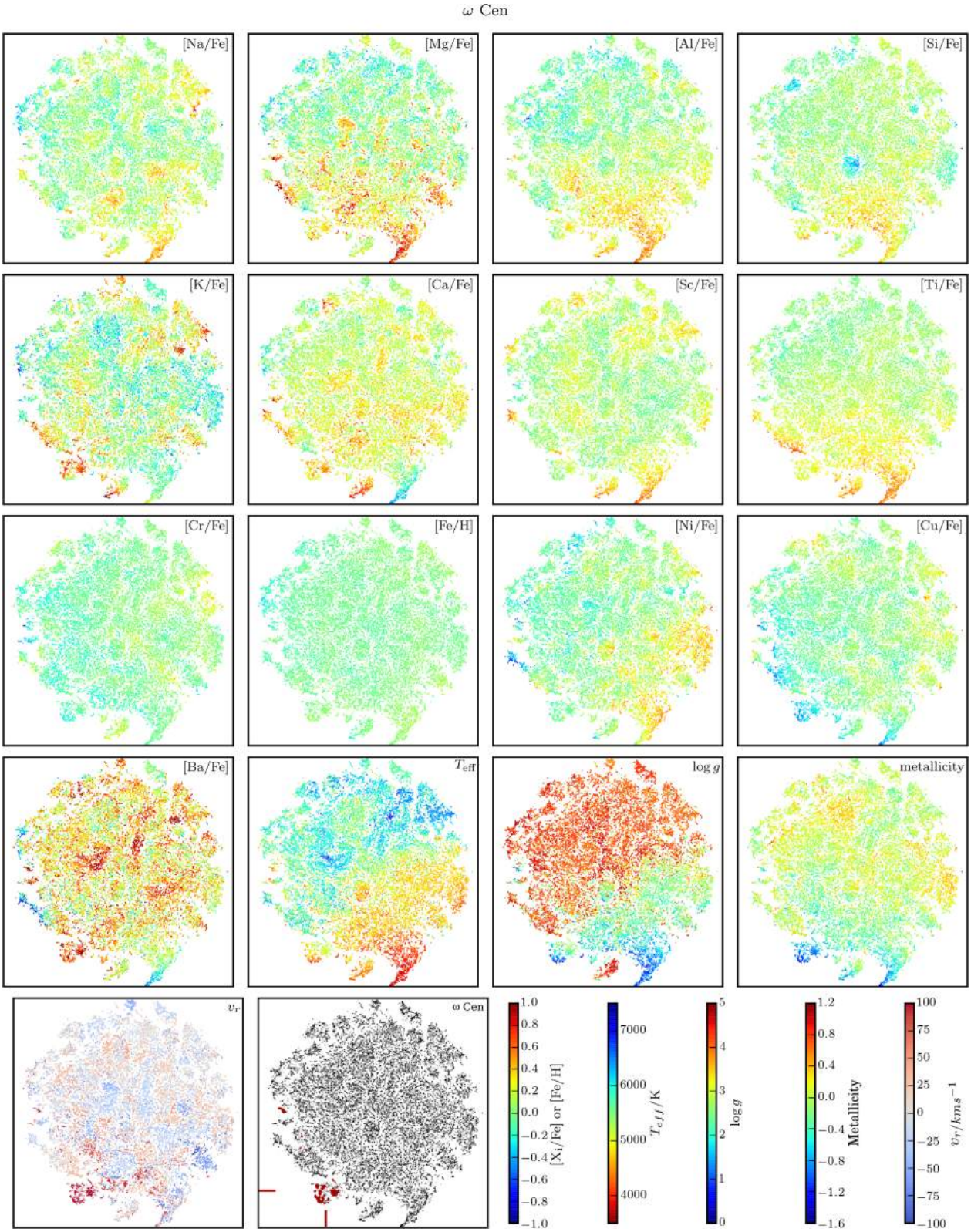

Figure B2. t-SNE projection of 33 281 stars in a 30° radius around ω Cen. Abundances of 13 elements used to create the projection are colour-coded. $T_{\rm eff}$, log g, metallicity and radial velocity colour-codes are also plotted. The last panel shows the stars that belong to the cluster in red and field stars in grey. 101 out of 230 ω Cen stars lie in a tight group in the bottom part of the map (marked with dashes at the edge of the plot) and additional 106 in a more sparse group next to it.

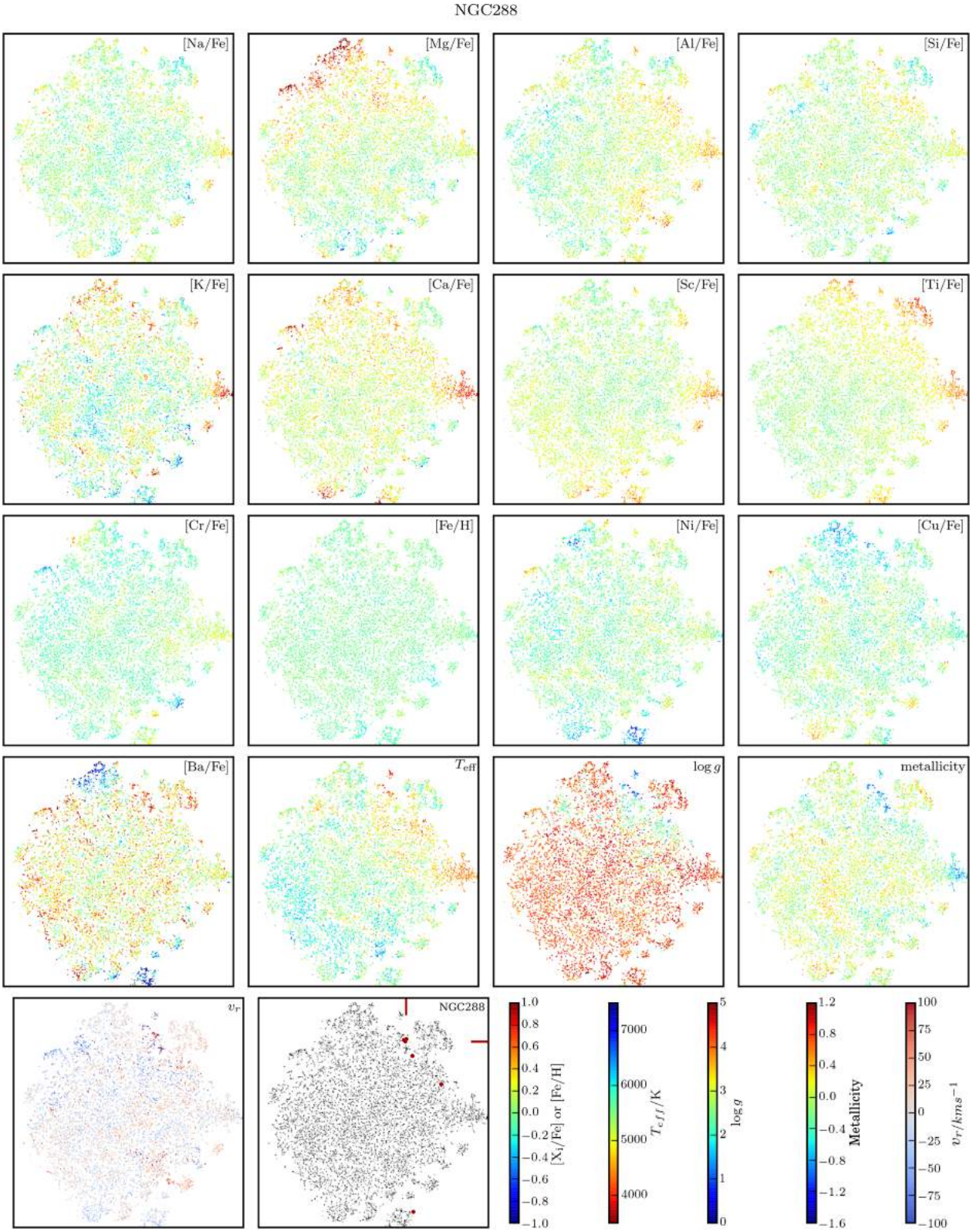

Figure B3. t-SNE projection of 11 535 stars in a 45° radius around NGC 288. Abundances of 13 elements used to create the projection are colour-coded. $T_{\rm eff}$, log g, metallicity and radial velocity colour-codes are also plotted. The last panel shows the stars that belong to the cluster in red and field stars in grey. 10 out of 14 NGC 288 stars lie in a tight group in the top part of the map (marked with dashes at the edge of the plot).

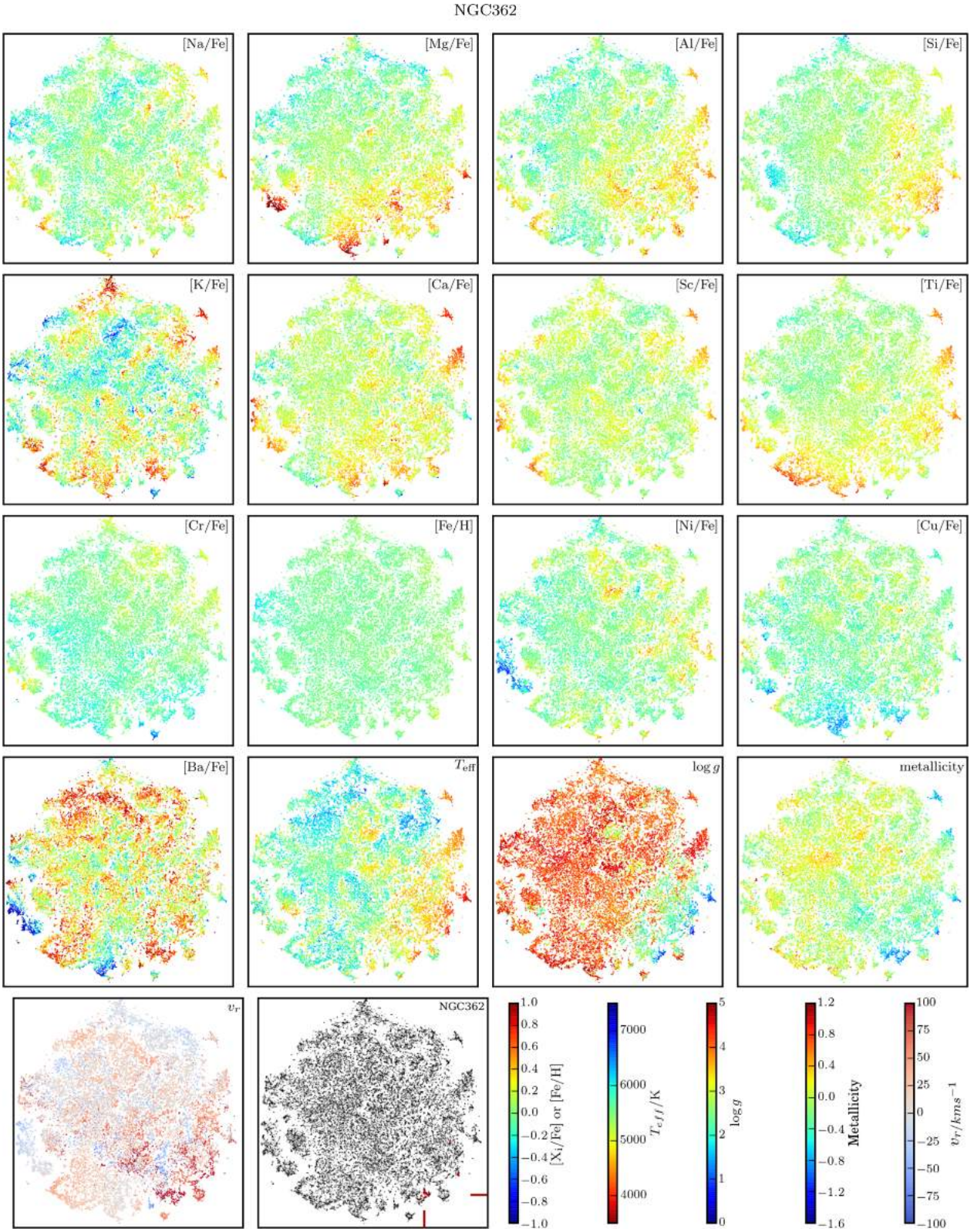

Figure B4. t-SNE projection of 41 578 stars in a 35° radius around NGC 362. Abundances of 13 elements used to create the projection are colour-coded. $T_{\rm eff}$, log g, metallicity and radial velocity colour-codes are also plotted. The last panel shows the stars that belong to the cluster in red and field stars in grey. 23 out of 27 NGC 362 stars lie in a group in the bottom part of the map (marked with dashes at the edge of the plot).

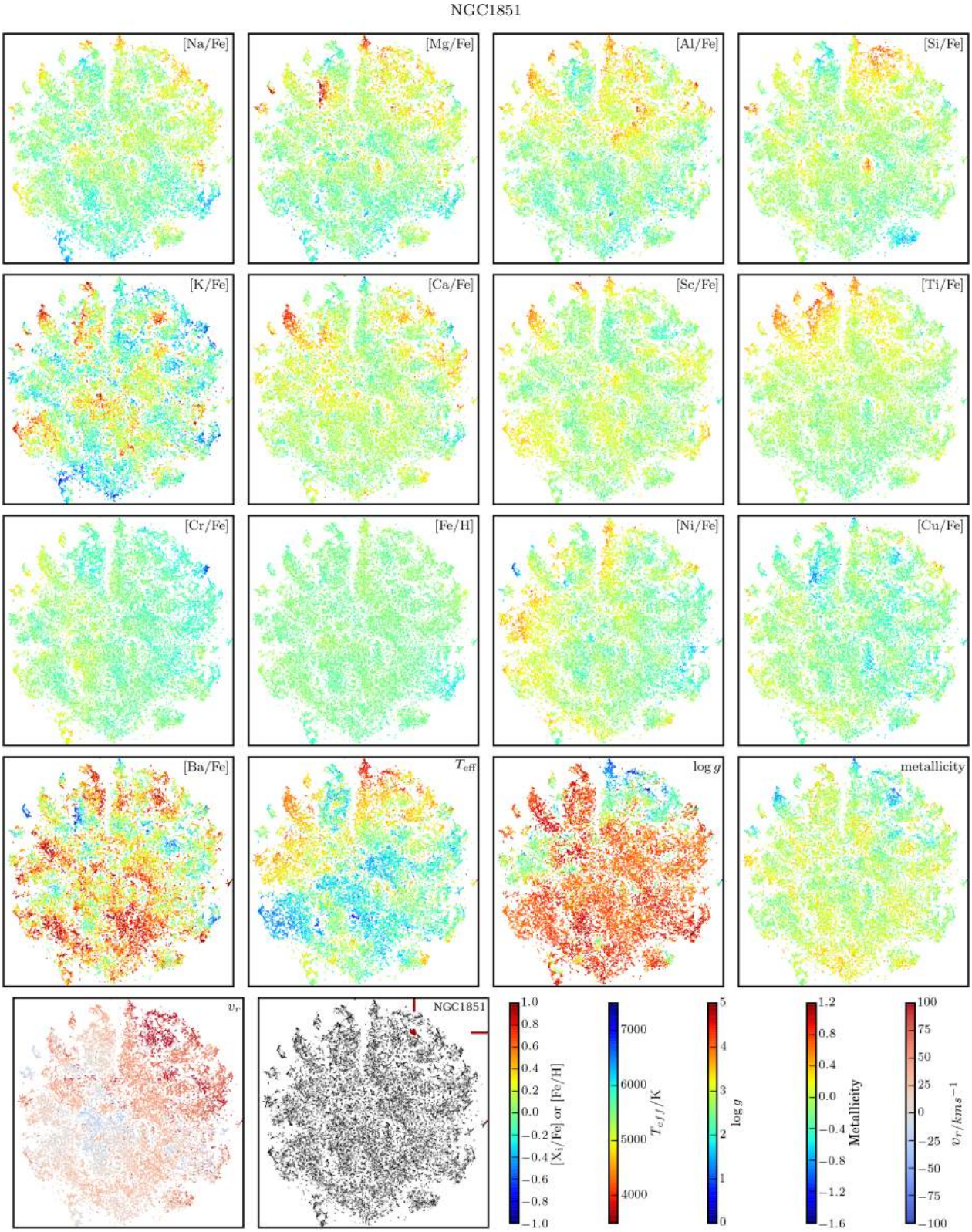

Figure B5. t-SNE projection of 33 882 stars in a 35° radius around NGC1851. Abundances of 13 elements used to create the projection are colour-coded. $T_{\rm eff}$, $\log g$, metallicity and radial velocity colour-codes are also plotted. The last panel shows the stars that belong to the cluster in red and field stars in grey. Six out of seven NGC 1851 stars lie in a tight group in the top right-hand part of the map (marked with dashes at the edge of the plot).

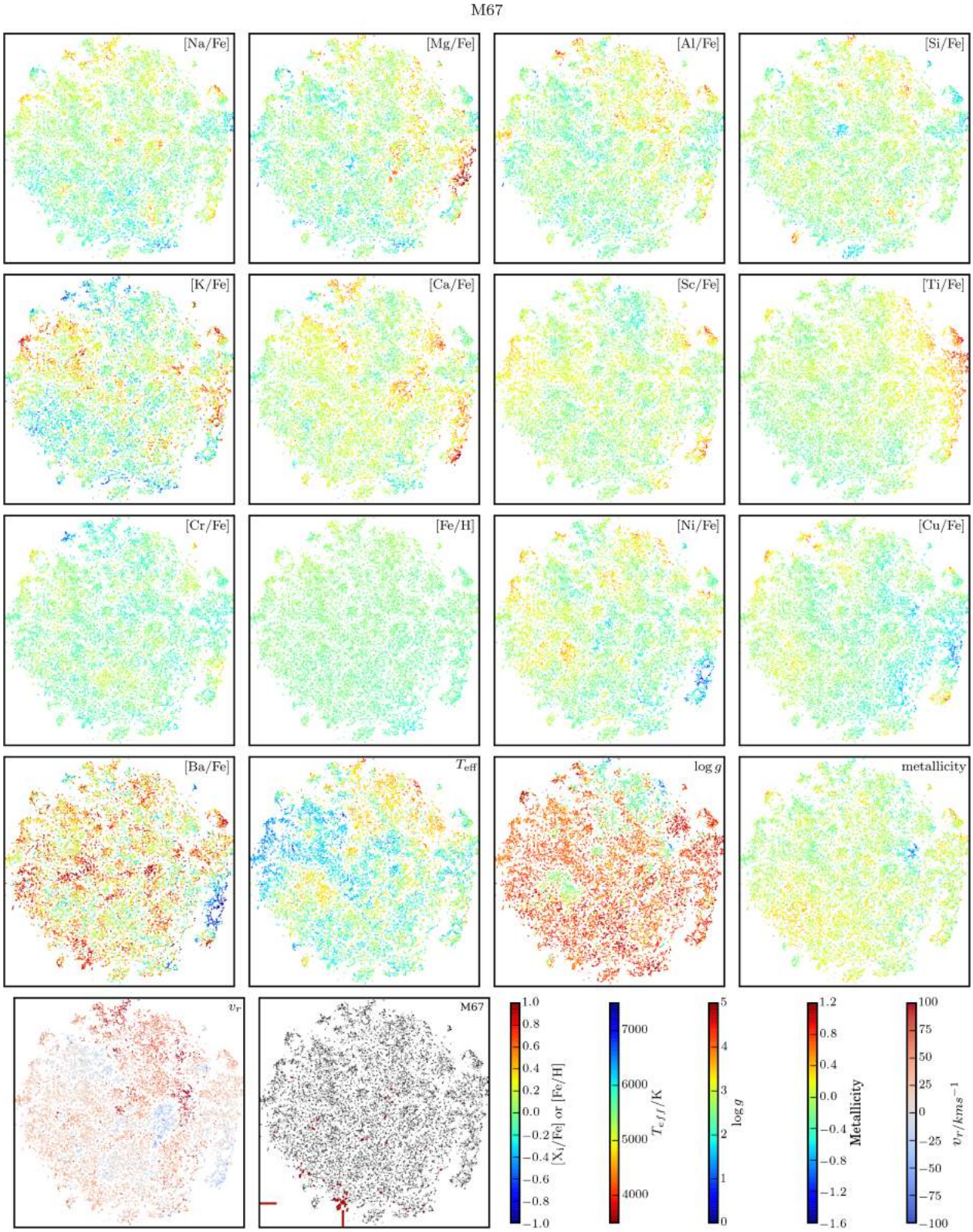

Figure B6. t-SNE projection of 25 648 stars in a 45° radius around M67. Abundances of 13 elements used to create the projection are colour-coded. $T_{\rm eff}$, log g, metallicity and radial velocity colour-codes are also plotted. The last panel shows the stars that belong to the cluster in red and field stars in grey. 71 of the 113 M67 stars lie in the biggest group in the bottom part of the map (marked with dashes at the edge of the plot).

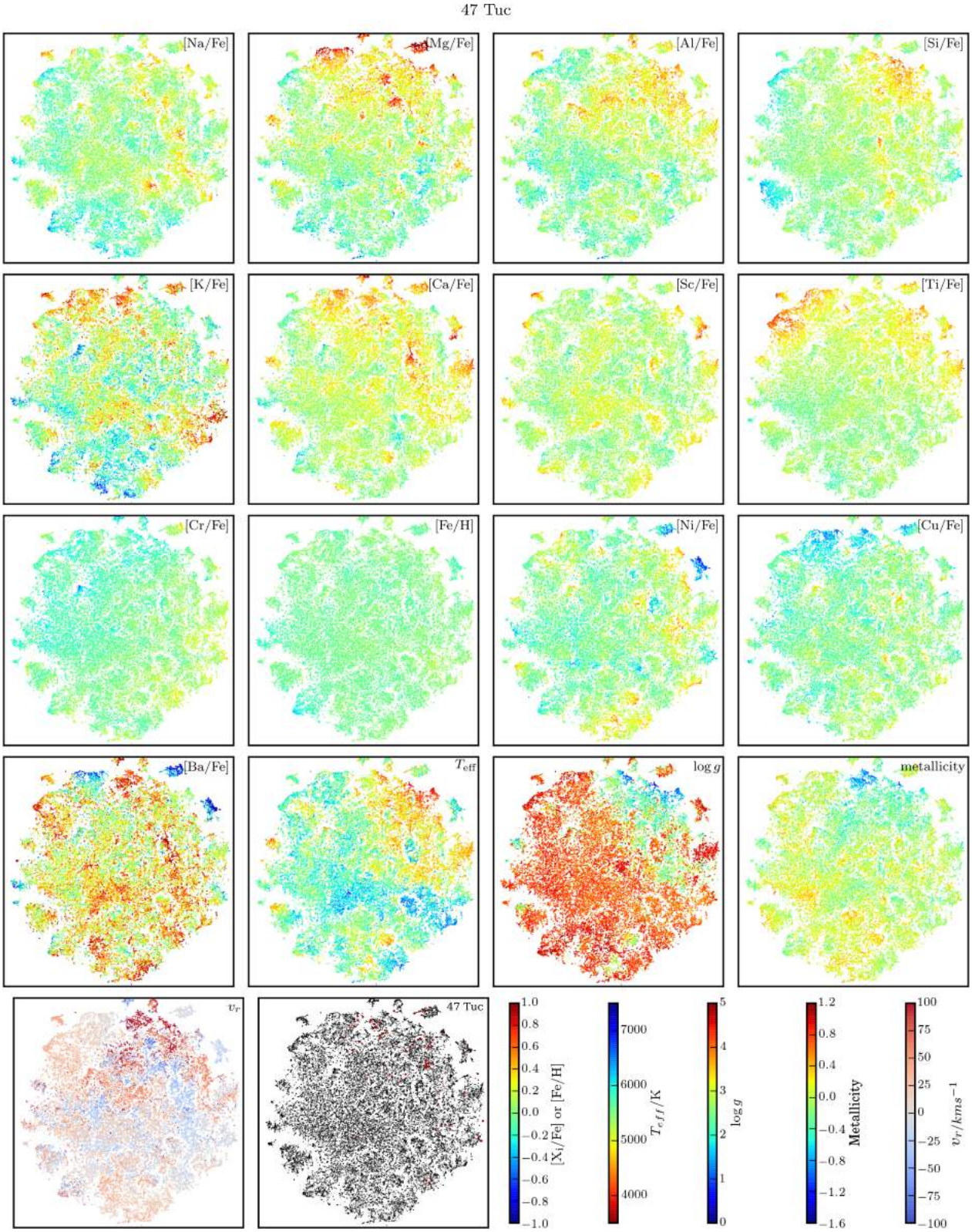

Figure B7. t-SNE projection of 44 037 stars in a 35° radius around 47 Tuc. Abundances of 13 elements used to create the projection are colour-coded. $T_{\rm eff}$, log g, metallicity and radial velocity colour-codes are also plotted. The last panel shows the stars that belong to the cluster in red and field stars in grey. Most 47 Tuc stars do not lie in a single group. The bigest group contains only 21 stars out of 90 47 Tuc stars.

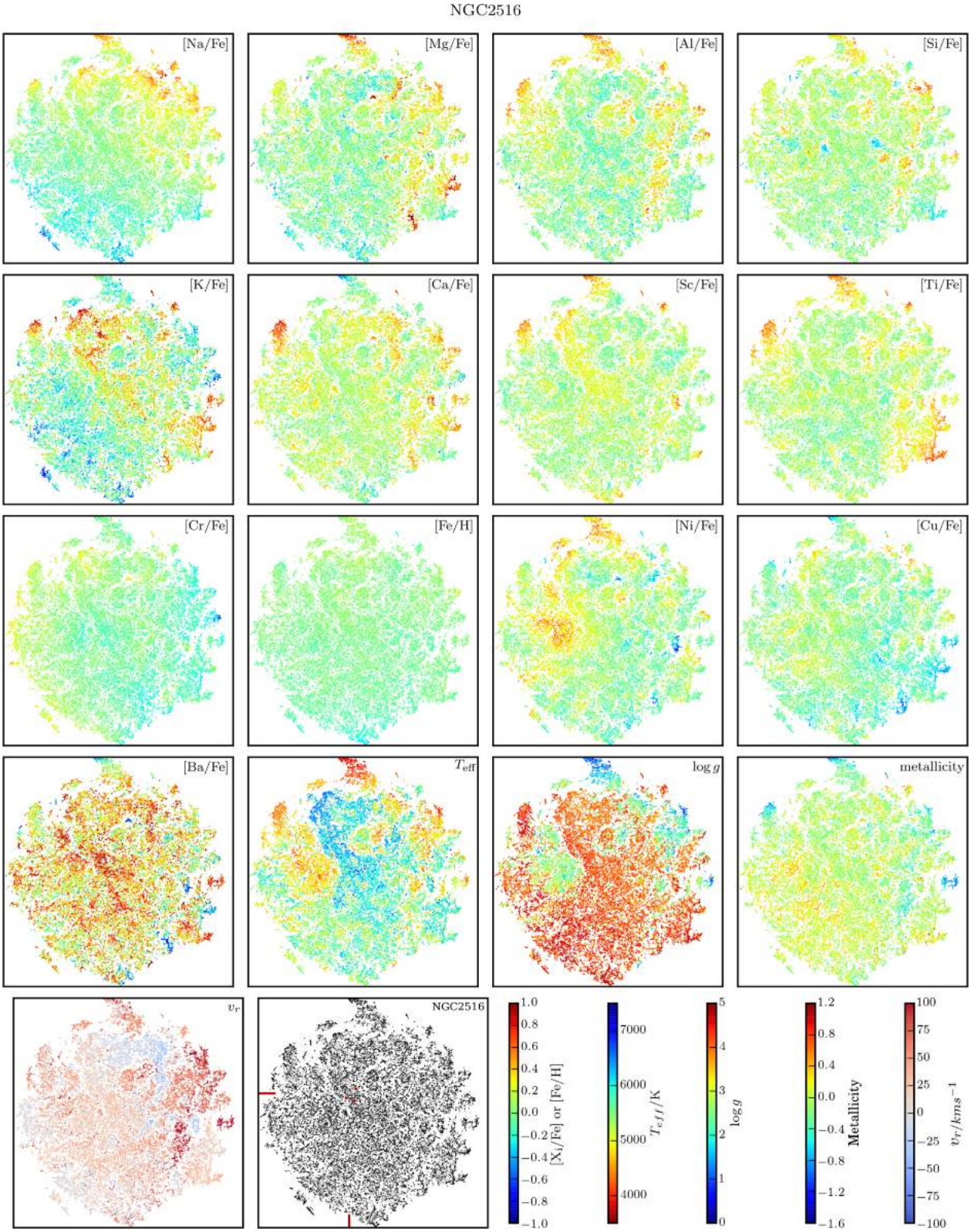

Figure B8. t-SNE projection of 41 106 stars in a 30° radius around NGC 2516. Abundances of 13 elements used to create the projection are colour-coded. $T_{\rm eff}$, log g, metallicity and radial velocity colour-codes are also plotted. The last panel shows the stars that belong to the cluster in red and field stars in grey. We only matched three stars to NGC 2516 of which none has a high membership probability in Jeffries et al. (2001). All stars lie in a middle region of the map where stars that are hardest to classify lie (marked with dashes at the edge of the plot).

APPENDIX C: CLUSTER MEMBERSHIP

For clusters targeted in the pilot survey the observed stars were pre-selected based on their proper motions (47 Tuc, NGC 288, NGC 362, M 67) position on the HR diagram (47 Tuc, NGC 288, ω Cen, NGC 362, M 67) and previous spectroscopic observations (NGC 288, NGC 1851, M 30, ω Cen). See Martell et al. (2017) for details.

Despite the pre-selection of observed stars, we did a further analisys of possible members. Fig. C1 shows position, radial velocities and proper motions used to determine the memberships. Our conditions are simple cuts in position and radial velocity for globular clusters and additional cuts in amplitude and angle of proper motion for Pleiades. Radii r_1 and r_2 (Kharchenko et al. 2013) are used for the radius of the core and radius of the cluster, respectively. For globular clusters, we consider all stars within $1.5r_2$, as there are members expected to be observed outside r_2. Any field stars are then discarded by making a cut in radial velocity. Because all our globular cluster have radial velocities that are distinct from the radial velocity of nearby field stars, we do not expect any misidentified members. This is not true for Pleiades, so we use r_1 as the position criterium and make additional cuts in proper motion. We might miss some members this way, but should keep the selection clear of any field stars. The criteria are conservative, as any misidentified members have more impact on the success of chemical tagging than possible missed members. The following list gives the membership criteria for each cluster:

(i) ω Cen: All stars within $1.5r_2$ and $200 < v_r < 260\,\mathrm{km\,s^{-1}}$.

(ii) Pleiades: All stars within r_1 and $5.0 < v_r < 8.0\,\mathrm{kms^{-1}}$ and $45 < \mu < 55\,\mathrm{mas\,yr^{-1}}$ and $2.70 < \phi_\mu < 2.80$. ϕ_μ is the angle of proper motion expressed in radians in the celestial coordinate system.

(iii) NGC 1851: All stars within $1.5r_2$ and $300 < v_r < 340\,\mathrm{kms^{-1}}$.

(iv) NGC 362: All stars within $1.5r_2$ and $210 < v_r < 230\,\mathrm{kms^{-1}}$.

(v) NGC 288: All stars within $1.5r_2$ and $-55 < v_r < -35\,\mathrm{kms^{-1}}$.

(vi) M30: All stars within $1.5r_2$ and $-175 < v_r < -185\,\mathrm{kms^{-1}}$.

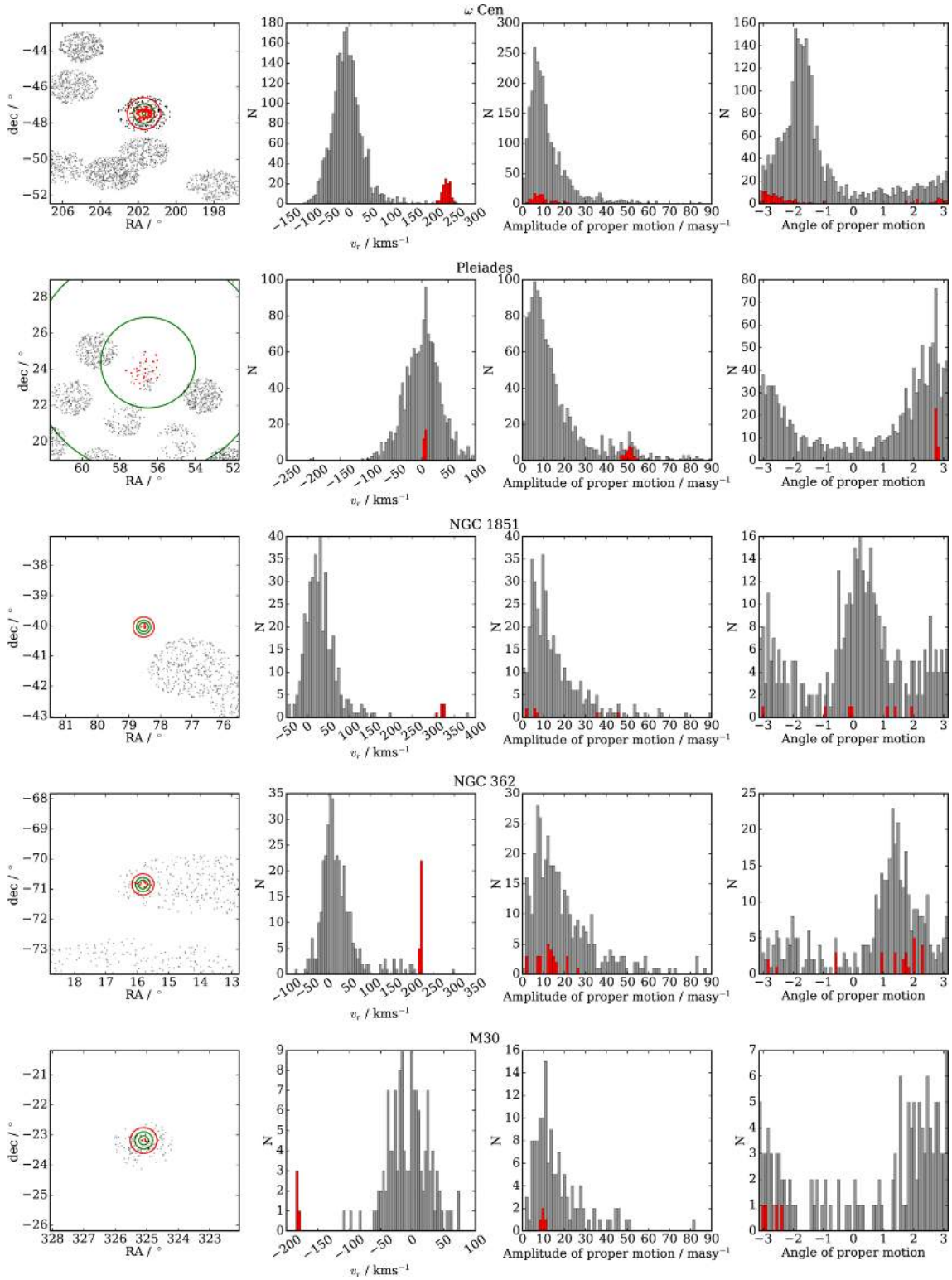

Figure C1. Top to bottom panel: basic information used to determine membership for six clusters, where we did not use sources from the literature. Left- to right-hand panel: position of observed stars (grey) and members (red) in a small region around each cluster centre. Red circle shows the maximum radius at which the members can be. Green circles show values r_1 (size of the cluster centre) and r_2 (size of the cluster) from Kharchenko et al. (2013). Second panel shows distribution of the radial velocities of all stars in the plotted field (grey) and cluster members (red). Last two panels show amplitude and angle of proper motion for all stars in the plotted field (grey) and cluster members (red). There is no panel for NGC 288, because all observed stars made the cut and there are no field stars within 5° of the cluster.

[1]*Sydney Institute for Astronomy, School of Physics, A28, The University of Sydney, NSW 2006, Australia*
[2]*ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO-3D)*
[3]*Research School of Astronomy and Astrophysics, Australian National University, ACT 2611, Australia*
[4]*Max-Planck-Institut for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany*
[5]*Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia*
[6]*Australian Astronomical Observatory, North Ryde, NSW 2133, Australia*
[7]*Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden*
[8]*School of Physics, UNSW, Sydney, NSW 2052, Australia*
[9]*Department of Physics and Astronomy, Macquarie Univeristy, Sydney, NSW 2109, Australia*
[10]*Stellar Astrophysics Centre, Department of Physics and Astronomy, Aarhus University, DK-8000 Aarhus C, Denmark*
[11]*Department of Astronomy, University of Virginia, Charlottesville, VA 22904-4325, USA*
[12]*INAF – Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-35122 Padova, Italy*
[13]*Computational Engineering and Science Research Centre, University of Southern Queensland, Toowoomba, Queensland 4350, Australia*
[14]*International Centre for Radio Astronomy Research (ICRAR), The University of Western Australia (M468), 35 Stirling Highway, Crawley, WA 6009, Australia*
[15]*INAF National Institute of Astrophysics, Astronomical Observatory of Padova, I-36012 Asiago, Italy*
[16]*Center for Astrophysical Sciences and Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218, USA*
[17]*Western Sydney University, Locked Bag 1797, Penrith South DC, NSW 2751, Australia*

This paper has been typeset from a TEX/LATEX file prepared by the author.