The Astrophysical Journal Supplement Series, 228:24 (19pp), 2017 February https://doi.org/10.3847/1538-4365/228/2/24

The Galah Survey: Classification and Diagnostics with t-SNE Reduction of Spectral Information

G. Traven[1], G. Matijevič[2], T. Zwitter[1], M. Žerjal[1], J. Kos[3], M. Asplund[4], J. Bland-Hawthorn[3], A. R. Casey[5], G. De Silva[3,6], K. Freeman[4], J. Lin[4], S. L. Martell[7], K. J. Schlesinger[8], S. Sharma[3], J. D. Simpson[9,10], D. B. Zucker[6,10,11], B. Anguiano[10], G. Da Costa[4], L. Duong[4], J. Horner[12], E. A. Hyde[13,14], P. R. Kafle[15], U. Munari[16], D. Nataf[4,17], C. A. Navin[10,18], W. Reid[18,19], and Y.-S. Ting[20]

[1] Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia; gregor.traven@fmf.uni-lj.si
[2] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, D-14482 Potsdam, Germany
[3] Sydney Institute for Astronomy, School of Physics, A28, The University of Sydney, NSW 2006, Australia
[4] Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia
[5] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[6] Australian Astronomical Observatory, P.O. Box 915, North Ryde, NSW 1670, Australia
[7] School of Physics, University of New South Wales, Sydney, NSW 2052, Australia
[8] Research School of Astronomy & Astrophysics, Mount Stromlo Observatory, Cotter Road, Weston Creek, ACT 2611, Australia
[9] Australian Astronomical Observatory, North Ryde, NSW 2113, Australia
[10] Department of Physics and Astronomy, Macquarie University, North Ryde, NSW 2109, Australia
[11] Research Centre in Astronomy, Astrophysics & Astrophotonics, Macquarie University, North Ryde, NSW 2109, Australia
[12] Computational Engineering and Science Research Centre, University of Southern Queensland, Towoomba QLD 4350, Australia
[13] Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia
[14] Australian Astronomical Observatory, P.O. Box 296 Epping, NSW 1710, Australia
[15] ICRAR, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia
[16] INAF National Institute of Astrophysics, Astronomical Observatory of Padova, I-36012 Asiago (VI), Italy
[17] Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD, USA
[18] Department of Physics and Astronomy, Macquarie University, Sydney, NSW 2109, Australia
[19] Western Sydney University, Penrith South DC, NSW 1797, Australia
[20] Harvard Center for Astrophysics. 60 Garden Street, Cambridge, MA 02138, USA

Received 2016 September 28; revised 2016 December 6; accepted 2016 December 6; published 2017 February 23

Abstract

Galah is an ongoing high-resolution spectroscopic survey with the goal of disentangling the formation history of the Milky Way using the fossil remnants of disrupted star formation sites that are now dispersed around the Galaxy. It is targeting a randomly selected magnitude-limited ($V \leqslant 14$) sample of stars, with the goal of observing one million objects. To date, 300,000 spectra have been obtained. Not all of them are correctly processed by parameter estimation pipelines, and we need to know about them. We present a semi-automated classification scheme that identifies different types of peculiar spectral morphologies in an effort to discover and flag potentially problematic spectra and thus help to preserve the integrity of the survey results. To this end, we employ the recently developed dimensionality reduction technique t-SNE (*t*-distributed stochastic neighbor embedding), which enables us to represent the complex spectral morphology in a two-dimensional projection map while still preserving the properties of the local neighborhoods of spectra. We find that the majority (178,483) of the 209,533 Galah spectra considered in this study represents normal single stars, whereas 31,050 peculiar and problematic spectra with very diverse spectral features pertaining to 28,579 stars are distributed into 10 classification categories: hot stars, cool metal-poor giants, molecular absorption bands, binary stars, Hα/Hβ emission, Hα/Hβ emission superimposed on absorption, Hα/Hβ P-Cygni, Hα/Hβ inverted P-Cygni, lithium absorption, and problematic. Classified spectra with supplementary information are presented in the catalog, indicating candidates for follow-up observations and population studies of the short-lived phases of stellar evolution.

Key words: binaries: general – catalogs – methods: data analysis – stars: activity – stars: peculiar – surveys

Supporting material: machine-readable tables

1. Introduction

In recent decades, the technology of optical fiber-fed spectrographs has enabled very efficient large-scale automated spectroscopic surveys. With the ability to observe up to several hundred stars simultaneously, it is now possible to obtain large numbers of high-quality spectra in a reasonable amount of time (Watson 1987). Surveys such as the RAdial Velocity Experiment (RAVE; Steinmetz et al. 2006), the Apache Point Observatory Galactic Evolution Experiment (Majewski et al. 2015), the LAMOST survey (Large Sky Area Multi-Object Fiber Spectroscopic Telescope; Luo et al. 2015), the ongoing *Gaia*-ESO Survey (Gilmore et al. 2012), the Galah (GALactic Archaeology with Hermes; De Silva et al. 2015), and the *Gaia* mission (Prusti 2012) with its future follow-up projects WEAVE (Dalton et al. 2012) and 4MOST (4 m Multi-Object Spectroscopic Telescope; de Jong et al. 2012) are some of the leading examples of continuous production of overwhelming amounts of data.

To provide a general overview of the observed spectra and learn more about the studied sample of stars in a spectroscopic survey such as Galah, it seems reasonable and necessary to address this task in an unbiased and automated way. A common approach is to employ different numerical dimensionality reduction methods to reveal the complex morphological structure of the data set at hand. By projecting the spectra into a low-dimensional space, it becomes feasible to grasp their intercorrelations and identify diverse morphological groups, thus constructing a classification of the whole data set, and particularly its outstanding features. A plethora of linear and nonlinear mathematical techniques has been developed in the past decades to tackle the problem of classification of complex high-dimensional data such as spectra, and they have been successfully applied in the astronomical community as well.

Many authors have used different techniques in order to classify stellar and other types of spectra (galaxy, quasar) or to discover new classes and unusual objects: Gulati et al. (1994) and von Hippel et al. (1994) were among the first to use artificial neural networks; Ibata & Irwin (1997), Bailer-Jones et al. (1998), and McGurk et al. (2010) demonstrated the use of PCA (principal component analysis) as a very robust classifier; Daniel et al. (2011) and Matijevič et al. (2012) succesfully used the LLE method (locally linear embedding) to identify anomalous or peculiar spectra as well as classify normal spectra. Except for the last two cases, where the authors used the nonlinear dimensionality reduction technique LLE, these studies have mostly dealt with distinguishing between classes of the MKK scheme. However, many have mentioned the potential of identifying and characterizing spectral populations, which is the focus of the present work.

Galah is an ongoing spectroscopic survey that aims to unveil the Milky Way's history by studying the fossil record of ancient star formation and accretion events preserved in stellar light. Detailed knowledge of the chemical information of fossil remnants, which have disrupted and are now dispersed around the Galaxy, is essential to disentangle its formation history and explain its current stellar populations. Recent studies of chemical abundances of stars in individual (undispersed) open clusters show that their abundance distributions are homogeneous to the level at which they can be measured, and their abundances are different from cluster to cluster (e.g., De Silva et al. 2007; conversely, most globulars show inhomogenities, e.g., Na−O anticorrelation, e.g., Carretta et al. 2009; Gratton et al. 2012; furthermore, small abundance variations have been detected in star-to-star studies in open clusters, e.g., Liu et al. 2016a, 2016b). This enables the technique of chemical tagging (Freeman & Bland-Hawthorn 2002) to identify the fossil remnants of old dispersed clusters from their abundance patterns over many chemical elements. Galah will achieve this by measuring up to 29 elemental abundances from 7 independent element groups, each with 5 measurable abundance levels, thereby obtaining enough cells (5^7) in multidimensional chemical abundance space (C-space), in which stars from chemically homogeneous aggregates (e.g., disrupted open clusters) will lie in tight clumps (Freeman 2012; Ting et al. 2012). This level of accuracy and the amount of elemental abundance information by far surpasses any existing single or multiple system stellar studies.

The Galah automatic pipeline is currently running without a classification processing stage. By manually scanning the observed sample, it has become obvious that there is a significant number of peculiar and otherwise problematic spectra. Although the majority belong to single stars and can be properly fit by synthetic spectra, neglecting the outliers can lead to erroneous results in radial velocities, atmospheric parameters, and especially detailed chemical abundances. Finding outliers by comparison to databases of known peculiar spectra might produce useful results, but would fail to give a reliable classification of the whole sample. We therefore aim to diagnose and classify the diverse morphologies in the Galah data set, with the goal of (1) highlighting all problematic spectra with unpredictable effects from either instrumentation or reduction stages, (2) identifying any peculiar spectra that are interesting per se and merit further investigation, and (3) providing a clean sample without any peculiar or problematic spectra, so that further studies, based on the detailed stellar parameters and chemical abundances produced by Galah, can be more reliable. The method that we use to identify patterns or groups in the "feature space" to achieve the stated goals is unsupervised classification with t-SNE (van der Maaten & Hinton 2008) reduction of spectral information. The main advantage of this approach is that we are more likely to detect various unfamiliar morphological features as well as the many known and expected peculiar stars. For a very nice overview of the vast range of classification and data-mining techniques we refer the reader to Sharma & Johnston (2009).

The paper is organized as follows: the data reduction and overview of Galah spectra is described in Section 2, the classification procedure with a description of the employed techniques is detailed in Section 3, and the discovered classes of spectra are examined in Sections 4 and 5. In Section 6 we present the structure of the catalog with final classification results including supplementary information, and in Section 7 we briefly describe a web-based visualization tool nicknamed *Galah Explorer*, which displays the t-SNE projection map featuring various useful functionalities. We conclude with the discussion in Section 8.

2. Data and Reduction Overview

2.1. Galah Spectra

The Galah survey was the main driver for the construction of Hermes (High Efficiency and Resolution Multi-Element Spectrograph), a fiber-fed multi-object spectrograph on the 3.9 m Anglo-Australian Telescope. Its spectral resolving power (R) is about 28,000, and there is also an $R = 45{,}000$ mode using a slit mask. The spectrograph is fed via 400 fibers distributed over π square degrees of sky. Taking into account the Galah magnitude limitation ($V = 14$), up to 392 stars can be observed simultaneously in that relatively small angle up to a Galactic latitude of $|b| \sim 28°$. Hermes has four simultaneous non-contiguous spectral arms centered at 4800, 5761, 6610, and 7740 Å (hereafter blue, green, red, and IR band), covering about 1000 Å in total, including Hα and Hβ lines. The spectrograph is designed to have ~10% efficiency and to achieve a signal-to-noise ratio S/N ~100 per resolution element at $V = 14$ in a 1 hr exposure, resulting in measured RV errors <1 km s^{-1} (Sheinis et al. 2015).

2.2. Reduction Pipeline

All the spectra subject to our analysis are reduced by the pipeline used in the Galah survey to produce fully calibrated spectra for subsequent stellar atmospheric parameter estimation (Kos et al. 2016). The reduction pipeline is based on reliable IRAF routines and other readily available software. After the IRAF-based reduction, a code that provides first estimates of radial velocity and three basic atmospheric parameters is run, and the entire observed spectrum is normalized for each star (see Section 6 in Kos et al. 2016). For some spectra, processing by this code can fail due to various data issues, and such cases are excluded from further consideration. Otherwise, the values of the three parameters (T_{eff}, log g, [Fe/H]) are produced by this code. We refer to them in the text and they are color coded in several figures. These values are preliminary, but they are the best in terms of completeness for the currently analyzed data set. Improved stellar parameters determined by the Galah spectroscopic analysis pipeline are also available, using a combination of the spectral synthesis program Spectroscopy Made Easy (Valenti & Piskunov 1996; Piskunov & Valenti 2016) and the data-driven approach The Cannon (Ness et al. 2015). The procedure is summarized in Martell et al. (2016) and will be detailed M. Asplund et al. (2016, in preparation). We are in the process of analyzing these improved results.

About 300,000 spectra have been taken to date, including various calibration exposures. However, we concentrate on ~210,000 spectra recorded before 2016 January 30 and reduced with the IRAF reduction pipeline version 5.1. In the future, the same study will be extended to include additional spectra once they become available. For more details, we refer the reader to the thorough description of the reduction process (Kos et al. 2016).

3. Classification

We devise a custom classification procedure which is based on two independently developed methods, the novel dimensionality reduction technique t-SNE (van der Maaten & Hinton 2008) and the renowned clustering algorithm DBSCAN (Ester et al. 1996). Both are used more than once in an iterative approach to enable the most efficient classification and overview of our data set. It should be noted that these purely mathematical methods are used extensively in various domains of research for unsupervised classification or clustering, and were not primarily intended for astrophysical purposes. The t-SNE/DBSCAN *dimensionality reduction/clustering* combination was chosen out of the large variety of data-mining techniques because it is able to

1. handle inherently nonlinear data (physics of line formation in stellar photospheres) as opposed to traditional linear dimensionality reduction methods, e.g., PCA, MDS (multidimensional scaling; Young 2013), LDA (linear discriminant analysis; Izenman 2008, p. 237), CCA (canonical correlation analysis; Hotelling 1936), MAF (min/max autocorrelation factors; Switzer & Green 1984)
2. reveal the local as well as the global structure of the high-dimensional data in a single map
3. alleviate the "crowding problem" that hampers many other nonlinear techniques
4. detect clusters in the projection map without a priori knowledge of their number and the form of their distribution function, in contrast to some classical clustering methods, e.g., K-means (Hartigan & Wong 1979) and Gaussian mixture models (Marin et al. 2005, p. 459).

These advantages, supported by previous experience and a detailed performance comparison of t-SNE to other techniques by van der Maaten & Hinton (2008), guided us toward the choice of algorithms for classification. However, we do not claim that these are the optimal methods, and an evaluation of the effectiveness of t-SNE/DBSCAN versus other methods in the context of the Galah data set is beyond the scope of this study. Nevertheless, we are encouraged by our results in that the t-SNE/DBSCAN combination proved to be very useful and efficient in visualizing and distinguishing different morphological groups of Galah spectra.

We first present the two techniques in more detail and then focus on our custom classification procedure.

3.1. t-SNE Reduction of Spectral Information

The widely varying dimensionality and huge amounts of data in different domains of research necessitate some form of reduction and visualization tools in order to efficiently extract information. t-SNE (*t*-distributed Stochastic Neighbor Embedding) can visualize any high-dimensional data set by projecting each data point into a low-dimensional map that reveals the local as well as the global structure of the data at many different scales. This is particularly suitable for high-dimensional data that lie on several different but related low-dimensional manifolds. An example in astronomy is a data set of single-lined binary spectra of multiple spectral types shifted by different radial velocities.

The technique significantly improves the overall performance of its predecessor, the stochastic neighbor embedding (Hinton & Roweis 2002, p. 833), mainly due to easier optimization of the algorithm and a reduction of the tendency to crowd points together in the center of the projection map. For more information on the performance of t-SNE on a wide variety of data sets and comparison with many other nonparametric visualization techniques, including sammon mapping, isomap, and locally linear embedding, we refer the reader to the original paper (van der Maaten & Hinton 2008). We note that t-SNE has been used quite recently in the astronomical community (Lochner et al. 2016; Matijevič 2016; Valentini et al. 2016). A brief introduction to the technique including formulae and text adapted from Pezzotti et al. (2016) is given in the following paragraphs.

A low-dimensional representation of high-dimensional data is achieved by optimally positioning data points in the projection map. For this purpose, t-SNE defines the similarity between N data points in the original high-dimensional space X and in the projection space Y, described by the symmetric joint-probability distributions P and Q, respectively. More precisely, the pairwise similarity between data points is modeled by the probability that one data point would pick another point as its neighbor, which depends on the probability density under a Gaussian in space X, whereas a *Student's t-distribution* is used in space Y:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{1}$$

$$\text{where} \qquad p_{j|i} = \frac{\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2/2\sigma_i^2)}, \tag{2}$$

for space X and

$$q_{ij} = ((1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)Z)^{-1} \tag{3}$$

$$\text{where} \qquad Z = \sum_{k=1}^{N}\sum_{l\neq k}^{N}(1 + \|\boldsymbol{y}_k - \boldsymbol{y}_l\|^2)^{-1}, \tag{4}$$

for space Y.

The σ_i is computed for each data point $\boldsymbol{x}_i$ so that the effective number of its neighbors corresponds to the fixed user-defined parameter μ (perplexity):

$$\mu = 2^{-\sum_j^N p_{j|i} \log_2 p_{j|i}}. \quad (5)$$

In regions of space X with a higher data density, σ_i tends to be smaller than in regions of lower density. The importance of modeling the separations between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ does not depend on their absolute distance in space X as long as they are close to each other relative to σ_i. Likewise, the size of the local neighborhood of $\boldsymbol{x}_i$ depends strongly on σ_i, and the similarity measure $p_{j|i}$ becomes almost infinitesimal for $\boldsymbol{x}_j$ at the distance of several σ_i due to the nature of the Gaussian distribution. These properties effectively define a soft border between the local and the global structure of the data.

The novel element in computing the joint-probability distribution Q in space Y is in using a normalized *Student's t-distribution* kernel with a single degree of freedom. The heavy tails of this distribution put more space between the moderately dissimilar points than a Gaussian would. As a result, there is more projection space available so that $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ can model the local structure of the corresponding $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ more accurately.

Having defined the joint-probability distributions P and Q, t-SNE aims to optimally position the points in space Y by minimizing the non-convex cost function C given by a simple measure of (Kullback–Leibler) divergence between probability distributions:

$$C(P, Q) = \mathrm{KL}(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (6)$$

This is achieved with an iterative gradient descent using a stochastic element to avoid local minima and is governed by the gradient of the above divergence:

$$\frac{\partial C}{\partial \boldsymbol{y}_i} = 4 \sum_{i=1}^{N} (F_i^{\mathrm{attr}} - F_i^{\mathrm{rep}}) \quad (7)$$

$$= 4 \sum_{i=1}^{N} \left(\sum_{j \neq i}^{N} p_{ij} q_{ij} Z(\boldsymbol{y}_i - \boldsymbol{y}_j) - \sum_{j \neq i}^{N} q_{ij}^2 Z(\boldsymbol{y}_i - \boldsymbol{y}_j) \right). \quad (8)$$

The gradient descent can be seen as an N-body simulation, where each data point exerts an attractive and a repulsive force on all the other points (F_i^{attr}, F_i^{rep}). In this respect, t-SNE gradient has the advantage of strongly repelling data points modeled by small pairwise distances in space Y that are otherwise very dissimilar in space X, but these repulsions do not increase to infinity as in some other attempts to address the crowding problem (see Sections 3.2, 3.3, and Figure 1 from van der Maaten & Hinton 2008).

In this work, we are dealing with over 200,000 spectra that can be viewed as data points in space X of dimensionality over 13,000. The high computational complexity introduced by employing t-SNE on our growing data set requires that we make use of the Barnes-Hut t-SNE (van der Maaten 2013), an evolution of the t-SNE algorithm that introduces different approximations to reduce the computational cost from $O(N^2)$ to $O(N \log(N))$ and the memory complexity from $O(N^2)$ to $O(N)$.

When computing the t-SNE embedding, we always (a) project our data set into two-dimensional space, (b) set the Barnes-Hut parameter θ to 0.5, and (c) unless otherwise specified, set the perplexity to 30, a value that has generally proven to be most effective for our purpose. Lower values of perplexity produce sparser projection maps with denser collections of points, and higher values produce more evenly covered projection space, but with less pronounced separations between distinct groups. The choice of two-dimensional space compared to three-dimensional space is pragmatic, since we only use one image (projection map) for visual inspection in the 2D case, whereas in 3D we would need to inspect three projected planes or use some advanced 3D visualization tool. We plan to explore the 3D option in the future, as it potentially preserves more information of the underlying structure of data.

The t-SNE procedure can be summarized in: (1) converting Euclidean distances in space X to pairwise similarities (often computationally the most intensive part), (2) sampling map points randomly from an isotropic Gaussian with small variance that is centered around the origin of projection space Y, and (3) initializing the gradient descent with a fixed number of iterations (usually 1000).

3.2. DBSCAN Clustering

Density-based spatial clustering of applications with noise (DBSCAN) is a data-clustering algorithm relying on a density-based notion of clusters (collections) and is designed to discover any arbitrary shape of collections of points in some space. It defines clusters from densely packed points—those that have many nearby neighbors—and rejects those points that lie alone in low-density regions (outliers). DBSCAN is one of the most common clustering algorithms and also one of the most frequently cited algorithms in scientific literature.

There are two input parameters to the DBSCAN method that have to be set by the user: *minPts,* and ε. Furthermore, the points are classified into three groups: core points, (density-) reachable points, and outliers. They are defined as follows:

1. A point u is a core point if at least *minPts* points are within distance ε of it (including u), and by definition, these are directly reachable from u, whereas no points are directly reachable from a non-core point.
2. A point v is reachable from u if there is a path u_1, ..., u_n with $u_1 = u$ and $u_n = v$, where each u_{i+1} is directly reachable from u_i (all the points on the path must be core points, with the possible exception of v, in which case, v is a border point).
3. All points not reachable from any other point are outliers.

To find a collection of points, DBSCAN starts with an arbitrary point u and retrieves all points density-reachable from u with respect to ε and *minPts*. If u is a core point, this procedure yields a collection. If u is a border point, no points are density-reachable from u, and DBSCAN visits the next point.

In this study, DBSCAN is employed merely as a tool for automatic detection of distinct collections in the projection map produced by t-SNE, without any bias except for the manual selection of ε and *minPts*. By definition of the t-SNE projection map, data points that are similar to each other should be closely packed together, thereby, in DBSCAN's terminology, forming a collection which can be detected and labeled.

3.3. Classification Procedure

In the process of finding the most objective and efficient way of classifying Galah spectra, we have established a procedure

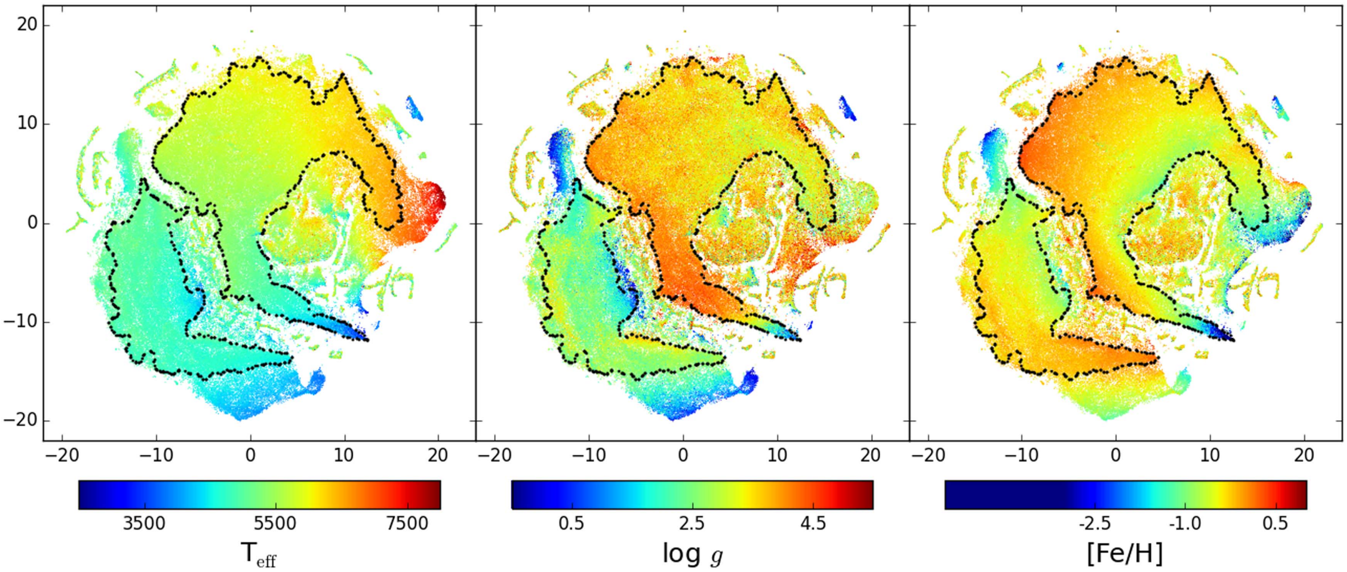

Figure 1. The first t-SNE projection of the whole working set containing 209,533 data points (spectra). The three panels feature $T_{\rm eff}$, log g, and [Fe/H] values for spectra (represented by the color scale) measured by the Galah reduction pipeline (see Section 2.2). The lowest values of [Fe/H] are only a few and probably erroneous. The two areas encircled by black points are the two largest collections of the most appropriate DBSCAN mode ($\varepsilon = 0.2$, *minPts* = 30) for large-scale cluster detection that were removed in the third step of our classification procedure. The outer borders of the two collections are dotted, indicating the approximate shape, which is not smooth and can be patchy on the inside as well. A relatively smooth parameter distribution is clearly visible in the enclosed areas. Altogether, DBSCAN defines 235 collections in this mode, which are not marked here. The axes of the panels have no physical meaning; they merely span the low-dimensional projection space.

that is a combination of automatic and manual processing and inspection of our data. The current Galah reduction pipeline includes several stages (see Section 2), and we only retain those spectra that pass radial velocity determination and normalization, which are in our analysis the two key properties for optimal use of t-SNE dimensionality reduction. In principle, we could use the whole data set of spectra before any kind of reduction, but in that case, the most important features driving our low-dimensional embedding (projection map) would be a result of missing wavelength calibration, flux normalization, radial velocity determination, etc. In essence, we would be concerned with properties that are less scientifically compelling and can be reliably enough accounted for with standard reduction procedures. After this selection, the number of spectra that remain is 209,533, and we join the four spectral bands together to produce the so-called data points of our working set.

Taking the whole practically usable range of red, green, blue, and IR band (excluding the strongest telluric features in the latter), the number of normalized flux values, which are basically pixels or original dimensionality of data points, amounts to 13,600 per data point (spectrum). At such high dimensionality multiplied by 209,533 data points, the computational cost despite the Barnes-Hut implementation of t-SNE can still turn out to be overwhelming or impractical (taking over 10 days to compute on an Intel(R) Xeon(R) 2.60 GHz CPU, using over 80 GB of memory). To facilitate the whole process in terms of memory consumption and computation time and to produce a projection map most suited for the purpose of this work, we make use of the following scheme:

1. *First t-SNE projection*: the first projection is computed for the whole working set (209,533 spectra). We use only the following wavelength ranges: 4850–4880 Å, 5750–5780 Å, 6550–6580 Å, and 7730–7760 Å. These ranges are selected so as to include the more diagnostic parts of each spectral band, with equal contribution from all of them, amounting to 2400 as dimensionality of data points. The map resulting from this first projection can be used as is, for this is the most basic and objective clustering of all data points from our working set. However, a significant portion of spectral information is missing due to our cut in wavelength range. There is also a practical caveat, in that although the projection map is fairly homogeneously populated, it is also very dense, making the smaller scale clustering, the one we are most interested in, more difficult to recognize. The next steps alleviate these issues.
2. *DBSCAN large-scale cluster detection*: DBSCAN input parameters are set in a way that a few large collections of data points are defined across the projection map. This is done in order to select those collections that presumably contain only the "normal" spectra. In our experience, well-behaved spectra are usually clustered together in one or a few large areas of the map where the atmospheric parameters ($T_{\rm eff}$, log g, [Fe/H]) are continuously distributed.
3. *Select and filter out collections of "normal" spectra*: It is here that manual interaction is most important, since we are rejecting (from our analysis) less interesting data points, and even when we are very careful, we can unwittingly discard some of the desired spectra from further consideration. The map of the first t-SNE projection with DBSCAN clustering on a large scale is shown in Figure 1. The two largest collections outlined by the black dotted line, amounting to 137,155 data points, are rejected because they presumably contain only "normal stars," and 76,938 remaining data points are considered in the next steps.
4. *Second t-SNE projection*: the second projection is computed for the subsample of the working set (76,938 spectra), resulting from the previous step. Owing to the

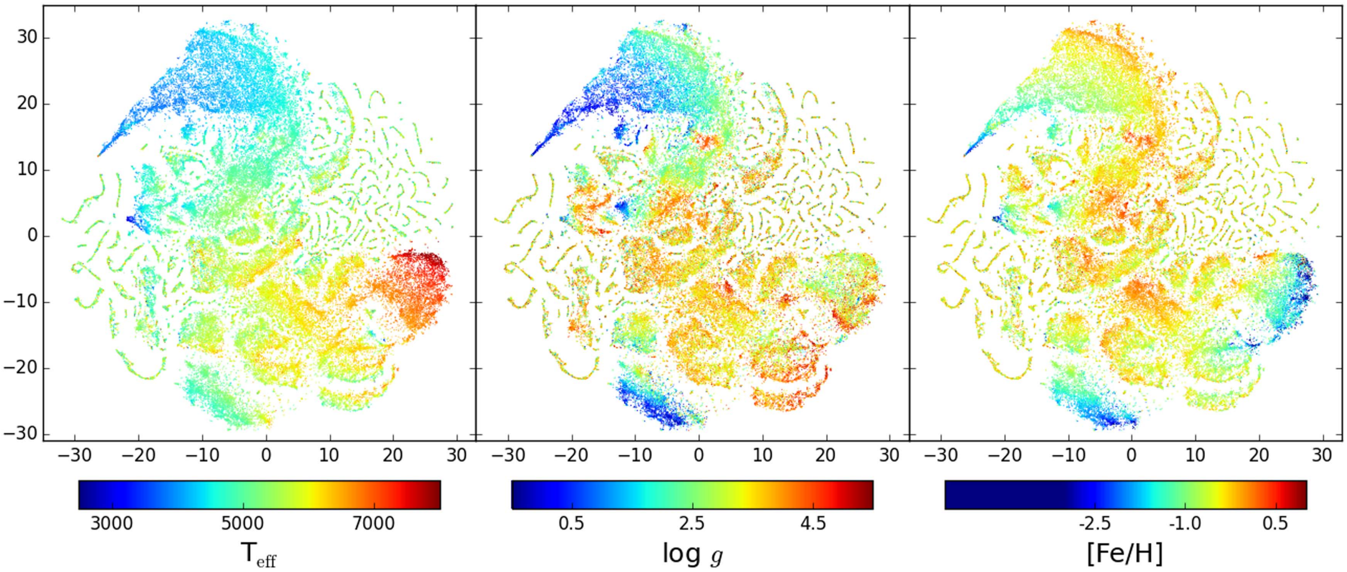

Figure 2. Same as Figure 1, but for the second t-SNE projection map of the filtered working set containing 76,938 data points (spectra), i.e., those that are not part of the two large collections marked in Figure 1.

smaller number of data points compared to the first projection, it is now feasible to operate with all practically usable flux values (13,600) from the four spectral bands 4730–4880 Å, 5670–5850 Å, 6500–6710 Å, and 7725–7865 Å. The projection map is shown in Figure 2 and serves as the final basis for our selection and analysis of peculiar spectra. Some "normal" spectra are still present in this map, but the largest portion should belong to all the peculiar objects that we are interested in. Their small-scale structure, which was hidden in the first t-SNE projection, is now reflected in the large-scale structure, and it is also more easily discernible because there are overall fewer data points in the map, as is evident from comparing Figures 1 and 2.

5. *DBSCAN small-scale cluster detection*: DBSCAN input parameters are set in a way that the defined collections correspond to relatively small and dense regions in the map that represent distinct morphological classes of spectra. In our experience, there is no unique parameter set for DBSCAN that would allow us to properly select the various collections, therefore many sets of parameters are tried and the corresponding DBSCAN results (hereafter DBSCAN modes), producing different sizes, shapes, and numbers of collections, are available for inspection in the next step.
6. *Select relevant/categorical collections and assign classification categories/flags*: The final step involves manual overview of individual spectra in different collections with the help of the visualization tool we presented in Section 7. The goal is to find the best collection from different DBSCAN modes that fully encompasses the manually examined spectra belonging to a distinct category. Some outliers in terms of a chosen category will usually be present, therefore the final results should be regarded as a list of candidate members of a certain category. We do not assign a quantitative probability of membership as it is beyond the scope of this study. Furthermore, the selected collections from different DBSCAN modes might sometimes overlap, so that one spectrum might be assigned to two categories, in which case it can be viewed as a candidate member of both.

4. Morphological Classes of Spectra

Table 1 lists six distinct categories that were defined using the classification procedure described in Section 3.3. This classification is not strictly limited to peculiar objects that have spectra without a counterpart in the library of synthetic spectra, although they remain the principal motivation for this work. It is instead a search for any coherent group in the projection map, from which a category of interest can be selected. They range from larger collections of points (spectra), like the *hot stars* category, to smaller collections that mostly contain problematic spectra with typically one, albeit very prominent, feature (e.g., a strong emission spike).

The projection map that was used to search for and define the six general categories is presented in Figure 2. The overall distribution of parameters $T_{\rm eff}$, log g, and [Fe/H] in the three panels indicates their importance in feature space. $T_{\rm eff}$ clearly is the principal discriminant with a gradient over the whole projection map, followed by [Fe/H] and log g, which influence the distribution of points inside larger and well-separated collections. The collections that represent distinct categories are marked in Figure 3. Some of them are characterized by a certain strong spectral feature, hence the three main stellar parameters can be well mixed within the collection, in addition to being erroneous in cases where such features prevent a reliable estimate of their values.

For all targets corresponding to spectra with an assigned classification category, a search by coordinates inside 1″ radius was performed on the SIMBAD database. The most common main type and other type properties of the matched SIMBAD objects are listed in Table 1. In the following paragraphs, categories are described individually, with several issues related to observations and reduction combined in the category *Problematic*.

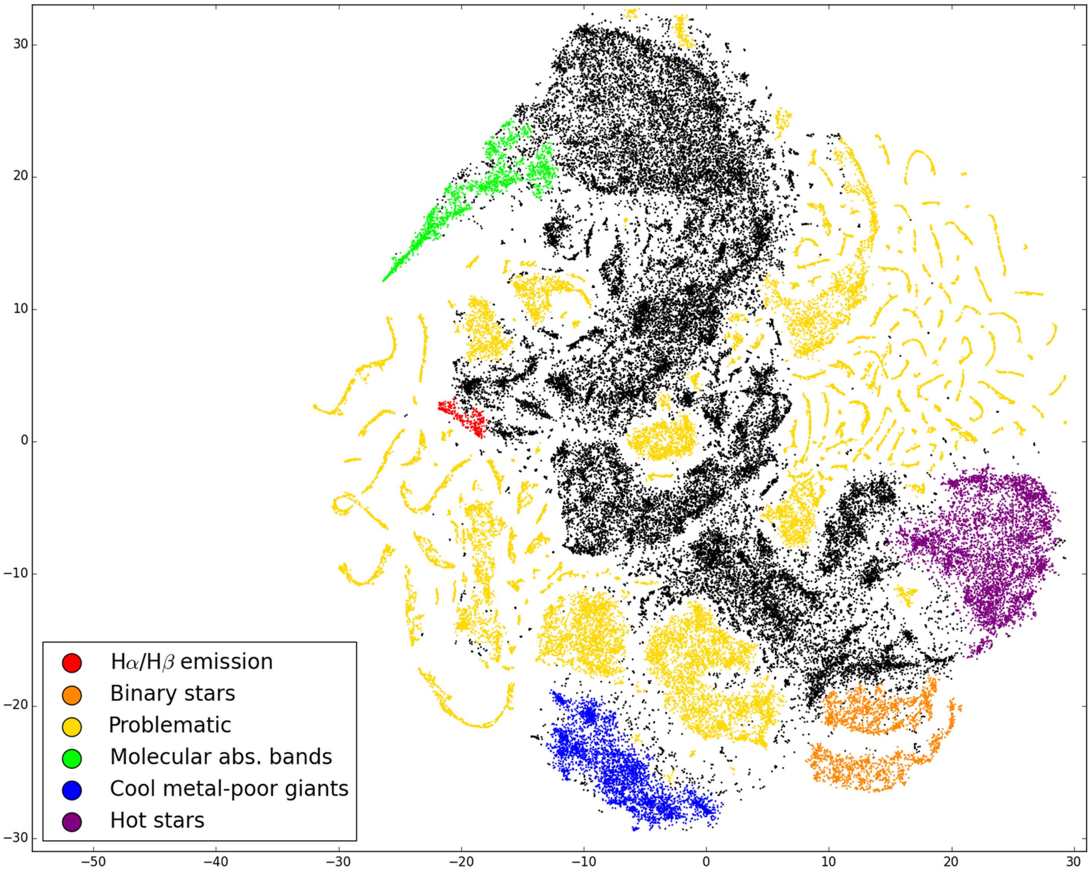

Figure 3. The result of the classification procedure, based on the projection map in Figure 2. Collections of spectra assigned to distinct categories are flagged (color coded), the rest are black. The axes of the panels have no physical meaning, they merely span the low-dimensional projection space.

Table 1
Classification Categories Based on the General Projection Map (see Section 4)

Category	N	Main type	Other types
Hot stars	4130	Star in cluster (20), variable star of RR Lyr type (13), variable star (12), variable star of delta Sct type (12), eclipsing binary of Algol type (detached) (8)	Infrared source (1486), variable star (41), star in cluster (26), rotationally variable star (19), variable star of RR Lyr type (13)
Cool metal-poor giants	2784	Star in cluster (371), red giant branch star (162), possible red giant branch star (26), variable star of RR Lyr type (10), high proper-motion star (8)	Infrared source (695), star in cluster (552), red giant branch star (165), possible red giant branch star (84), variable star (13)
Molecular abs. bands	1274	Variable star (4), long-period variable star (3), star in cluster (1), S star (1), possible red supergiant star (1)	Infrared source (500), variable star (9), long-period variable star (3), star in cluster (2), possible red supergiant star (2)
Binary stars	1817	Rotationally variable star (2), eclipsing binary of Algol type (detached) (1), star in cluster (1), double or multiple star (1), spectroscopic binary (1)	Infrared source (230), rotationally variable star (6), double or multiple star (4), spectroscopic binary (3), variable star (2)
Hα/Hβ emission	215	High proper-motion star (5), rotationally variable star (4), X-ray source (2), double or multiple star (1), infrared source (1)	Infrared source (33), X-ray source (10), high proper-motion star (7), rotationally variable star (5), variable star (5)
Problematic[a]	19095	Star in cluster (218), red giant branch star (47), high proper-motion star (11), variable star (8), variable star of RR Lyr type (7)	Infrared source (1313), star in cluster (253), red giant branch star (48), variable star (17), possible red giant branch star (12)

Note. The columns give the classification category, number of classified spectra, and the most common SIMBAD *main types* and *other types*. SIMBAD defines a main type for each astronomical object in its database, and usually several other types generally inferred from its identifiers. For the last two columns, only the five most common types are listed, excluding the less interesting type *Star*.

[a] A large portion of such spectra are recoverable (see text).

(This table is available in machine-readable form.)

Table 2
List of Spectra Plotted in Figures 4–12 Representing Distinct Classification Categories

Gal. ID	R.A. (J2000) (hh mm ss.s)	Decl. (J2000) (dd mm ss.s)	V (mag)	Figure	sp.
2048539	$12^{\rm h}\,40^{\rm m}\,59\overset{\rm s}{.}46$	$-45^\circ\,27'\,14\overset{\prime\prime}{.}1$	12.2	4	1
1860013	$10^{\rm h}\,59^{\rm m}\,01\overset{\rm s}{.}3$	$-47^\circ\,34'\,19\overset{\prime\prime}{.}9$	…	4	2
2009968	$12^{\rm h}\,59^{\rm m}\,40\overset{\rm s}{.}48$	$-45^\circ\,53'\,04\overset{\prime\prime}{.}4$	…	4	3
3627010	$09^{\rm h}\,00^{\rm m}\,57\overset{\rm s}{.}77$	$-27^\circ\,09'\,39\overset{\prime\prime}{.}8$	…	4	4
3253503	$12^{\rm h}\,09^{\rm m}\,53\overset{\rm s}{.}99$	$-31^\circ\,25'\,10\overset{\prime\prime}{.}5$	…	5	1
2020691	$13^{\rm h}\,01^{\rm m}\,19\overset{\rm s}{.}27$	$-45^\circ\,45'\,51\overset{\prime\prime}{.}8$	12.6	5	2
9514457	$00^{\rm h}\,24^{\rm m}\,59\overset{\rm s}{.}034$	$-72^\circ\,07'48\overset{\prime\prime}{.}33$	…	5	3
3611818	$09^{\rm h}\,03^{\rm m}\,05\overset{\rm s}{.}5$	$-27^\circ\,20'\,36\overset{\prime\prime}{.}9$	12.9	6	1
1581751	$17^{\rm h}\,58^{\rm m}\,33\overset{\rm s}{.}44$	$-50^\circ\,51'\,15\overset{\prime\prime}{.}5$	…	6	2
4935389	$08^{\rm h}\,07^{\rm m}\,58\overset{\rm s}{.}28$	$-10^\circ\,29'\,30\overset{\prime\prime}{.}7$	10.5	6	3
1715870	$13^{\rm h}\,23^{\rm m}\,49\overset{\rm s}{.}52$	$-49^\circ\,14'\,21\overset{\prime\prime}{.}1$	…	6	4
157469	$09^{\rm h}\,43^{\rm m}\,55\overset{\rm s}{.}26$	$-76^\circ\,28'\,55\overset{\prime\prime}{.}5$	13.4	7	1
1345018	$18^{\rm h}\,39^{\rm m}\,12\overset{\rm s}{.}45$	$-53^\circ\,58'\,07\overset{\prime\prime}{.}6$	13.3	7	2
1281061	$06^{\rm h}\,13^{\rm m}\,09\overset{\rm s}{.}26$	$-54^\circ\,49'\,54\overset{\prime\prime}{.}4$	12.6	7	3
2584208	$11^{\rm h}\,34^{\rm m}\,04\overset{\rm s}{.}8$	$-39^\circ\,17'\,39\overset{\prime\prime}{.}5$	12.7	7	4
3217487	$06^{\rm h}\,30^{\rm m}\,58\overset{\rm s}{.}37$	$-31^\circ\,49'\,29\overset{\prime\prime}{.}7$	13.6	7	5
2061842	$14^{\rm h}\,05^{\rm m}\,59\overset{\rm s}{.}68$	$-45^\circ\,18'\,20\overset{\prime\prime}{.}7$	…	7	6
6122038	$21^{\rm h}\,28^{\rm m}\,56\overset{\rm s}{.}42$	$+06^\circ\,06'\,25\overset{\prime\prime}{.}3$	12.4	8	1
203377	$02^{\rm h}\,04^{\rm m}\,32\overset{\rm s}{.}8$	$-74^\circ\,55'\,28\overset{\prime\prime}{.}6$	13.2	8	2
1720551	$12^{\rm h}\,20^{\rm m}\,52\overset{\rm s}{.}77$	$-49^\circ\,11'\,05\overset{\prime\prime}{.}7$	…	8	3
1692074	$12^{\rm h}\,43^{\rm m}\,04\overset{\rm s}{.}53$	$-49^\circ\,31'\,11''$	13.1	8	4
2292604	$10^{\rm h}\,25^{\rm m}\,20\overset{\rm s}{.}92$	$-42^\circ\,41'\,53\overset{\prime\prime}{.}9$	12.7	8	5
1237165	$07^{\rm h}\,41^{\rm m}\,05\overset{\rm s}{.}11$	$-55^\circ\,26'\,32\overset{\prime\prime}{.}54$	13.8	9	1
2076959	$19^{\rm h}\,42^{\rm m}\,39\overset{\rm s}{.}47$	$-45^\circ\,08'\,14\overset{\prime\prime}{.}4$	13.9	9	2
364936	$09^{\rm h}\,56^{\rm m}\,53\overset{\rm s}{.}13$	$-70^\circ\,57'\,57\overset{\prime\prime}{.}8$	…	9	3
598623	$22^{\rm h}\,58^{\rm m}\,39\overset{\rm s}{.}82$	$-67^\circ\,09'\,45\overset{\prime\prime}{.}3$	12.8	9	4
226575	$01^{\rm h}\,13^{\rm m}\,02\overset{\rm s}{.}51$	$-74^\circ\,14'\,22\overset{\prime\prime}{.}9$	13.8	9	5
2791332	$12^{\rm h}\,41^{\rm m}\,51\overset{\rm s}{.}89$	$-36^\circ\,50'\,37\overset{\prime\prime}{.}3$	…	9	6
9520401	$18^{\rm h}\,23^{\rm m}\,16\overset{\rm s}{.}32$	$-34^\circ\,01'\,27\overset{\prime\prime}{.}5$	…	9	7
1230979	$07^{\rm h}\,37^{\rm m}\,19\overset{\rm s}{.}419$	$-55^\circ\,31'44\overset{\prime\prime}{.}49$	14.5	10	1
3039593	$20^{\rm h}\,57^{\rm m}\,51\overset{\rm s}{.}85$	$-33^\circ\,52'\,38\overset{\prime\prime}{.}4$	13.8	10	2
2400420	$20^{\rm h}\,08^{\rm m}\,37\overset{\rm s}{.}69$	$-41^\circ\,26'\,45\overset{\prime\prime}{.}8$	…	10	3
3082604	$11^{\rm h}\,14^{\rm m}\,59\overset{\rm s}{.}85$	$-33^\circ\,22'\,27\overset{\prime\prime}{.}8$	13.7	10	4
2414299	$15^{\rm h}\,24^{\rm m}\,42\overset{\rm s}{.}98$	$-41^\circ\,17'\,09\overset{\prime\prime}{.}9$	…	10	5
2105243	$13^{\rm h}\,04^{\rm m}\,09^{\rm s}$	$-44^\circ\,49'\,18\overset{\prime\prime}{.}7$	14.2	10	6
2292604	$10^{\rm h}\,25^{\rm m}\,20\overset{\rm s}{.}92$	$-42^\circ\,41'\,53\overset{\prime\prime}{.}9$	12.7	10	7
2504307	$21^{\rm h}\,28^{\rm m}\,30\overset{\rm s}{.}21$	$-40^\circ\,14'\,30\overset{\prime\prime}{.}5$	13.9	11	1
1043258	$07^{\rm h}\,50^{\rm m}\,47\overset{\rm s}{.}825$	$-58^\circ\,43'\,17\overset{\prime\prime}{.}31$	13.9	11	2
9518625	$07^{\rm h}\,45^{\rm m}\,32\overset{\rm s}{.}447$	$-58^\circ\,50'\,26\overset{\prime\prime}{.}46$	…	11	3
1264994	$07^{\rm h}\,35^{\rm m}\,59\overset{\rm s}{.}296$	$-55^\circ\,03'\,09\overset{\prime\prime}{.}68$	14.0	11	4
9519860	$18^{\rm h}\,05^{\rm m}\,49\overset{\rm s}{.}37$	$-31^\circ\,44'\,26\overset{\prime\prime}{.}88$	…	11	5
9518173	$07^{\rm h}\,33^{\rm m}\,39\overset{\rm s}{.}016$	$-55^\circ\,34'\,54\overset{\prime\prime}{.}83$	14.7	11	6
9518474	$07^{\rm h}\,41^{\rm m}\,48\overset{\rm s}{.}343$	$-55^\circ\,24'\,54\overset{\prime\prime}{.}13$	…	11	7
9518210	$07^{\rm h}\,34^{\rm m}49\overset{\rm s}{.}249$	$-56^\circ\,02'\,27\overset{\prime\prime}{.}53$	15.0	11	8
9518390	$07^{\rm h}\,39^{\rm m}\,37\overset{\rm s}{.}67$	$-54^\circ\,47'\,05\overset{\prime\prime}{.}14$	14.6	11	9
9518511	$07^{\rm h}\,42^{\rm m}\,50\overset{\rm s}{.}476$	$-55^\circ\,26'\,22\overset{\prime\prime}{.}95$	14.9	11	10
9519416	$14^{\rm h}\,13^{\rm m}\,35\overset{\rm s}{.}84$	$+07^\circ\,15'\,08\overset{\prime\prime}{.}1$	16.1	11	11
9519827	$18^{\rm h}\,05^{\rm m}\,16\overset{\rm s}{.}78$	$-31^\circ\,38'\,24\overset{\prime\prime}{.}72$	…	11	12
9519903	$18^{\rm h}\,06^{\rm m}\,34\overset{\rm s}{.}01$	$-32^\circ\,02'\,07\overset{\prime\prime}{.}08$	…	11	13
9519743	$18^{\rm h}\,03^{\rm m}\,45\overset{\rm s}{.}79$	$-32^\circ\,20'\,08\overset{\prime\prime}{.}52$	…	11	14
9517171	$06^{\rm h}\,43^{\rm m}\,53\overset{\rm s}{.}4$	$+00^\circ\,11'\,10\overset{\prime\prime}{.}34$	14.5	11	15
3316259	$06^{\rm h}\,42^{\rm m}\,36\overset{\rm s}{.}03$	$-30^\circ\,42'\,56\overset{\prime\prime}{.}1$	12.0	12	1
1311297	$04^{\rm h}\,42^{\rm m}\,27\overset{\rm s}{.}14$	$-54^\circ\,25'\,10\overset{\prime\prime}{.}5$	12.5	12	2
1260144	$06^{\rm h}\,22^{\rm m}\,39\overset{\rm s}{.}33$	$-55^\circ\,07'\,13\overset{\prime\prime}{.}1$	13.2	12	3
2532175	$15^{\rm h}\,38^{\rm m}\,51\overset{\rm s}{.}73$	$-39^\circ\,54'\,46\overset{\prime\prime}{.}3$	13.5	12	4
3305889	$12^{\rm h}\,09^{\rm m}\,56\overset{\rm s}{.}1$	$-30^\circ\,50'\,05\overset{\prime\prime}{.}8$	13.0	12	5

Note. The columns give the Galah ID (unique star identifier), coordinates, APASS V magnitude where available, figure number, and sequential spectrum number (bottom to top) in panels for each figure.

4.1. Hot Stars

A large collection in the right part of the map in Figure 2 contains mostly early-type stars, with temperatures well above solar, which are characterized predominantly by widened wings of the Hα and Hβ absorption lines. We observe a smooth transition of temperatures inside this collection, ranging from about 6500 up to 8000 K (upper limit of the grid of synthetic templates, see Kos et al. 2016). The metallicity distribution is also very smooth, along an axis perpendicular to temperature, while the surface gravity is more patchy, with dwarfs more clustered in some parts and giants more dispersed throughout the collection. Examples of spectra in this category with different metallicities and temperatures are shown in Figure 4.

4.2. Cool Metal-poor Giants

The collection in the bottom part of the projection map features mostly late-type stars with a measured metallicity well below solar value ($-4.5 < [\mathrm{Fe/H}] < -0.5$, see Figure 5). The surface gravity distribution is clearly visible, with the majority of stars being giants, and most with temperatures in the range from 4000 K to a little above solar. The available records from SIMBAD support these claims, with 162 stars classified as red giant branch star, 26 as possible red giant branch star, and 10 as variable star of RR Lyr type.

4.3. Stars with Molecular Absorption Bands

The upper left part of the map in Figure 2 contains a region populated by spectra with strong molecular absorption bands. It is well isolated on one end, but still connected to the rest of the late-type stars on the other end. The temperatures in this collection are mostly low, as expected, but not unquestionable, as is nicely demonstrated by the very tip of this area on the left side, where we find the strongest absorption bands, while the temperatures derived for some of these spectra are much too high (above 6500 K). The increase in strength of absorption bands nicely follows the direction from this extreme end to the larger region of late-type spectra (from top to bottom on panels in Figure 6). The surface gravities and metallicities in this collection may also be problematic because producing reliable synthetic templates for these stars is challenging.

4.4. Binary Stars

Multiple stars, of which the majority represents double-lined spectroscopic binaries (SB2), are found in the bottom right part of the projection map, clustered in two well-separated collections. The main and most obvious difference between them is the position of the stronger of the two components in terms of the equivalent width of the absorption lines. For the lower group the stronger component is positioned blueward, while for the upper group it is redward. Although the distinction is not physically significant, it is evidently morphologically important. Following the arc-like shape of the two collections from left to right, the spectra show a progressively larger radial velocity separation of the two components, from almost blended double lines to lines separated by as much as 150 km s^{-1}. Examples of spectra of stars in this category are shown in Figure 7. The top star is a W

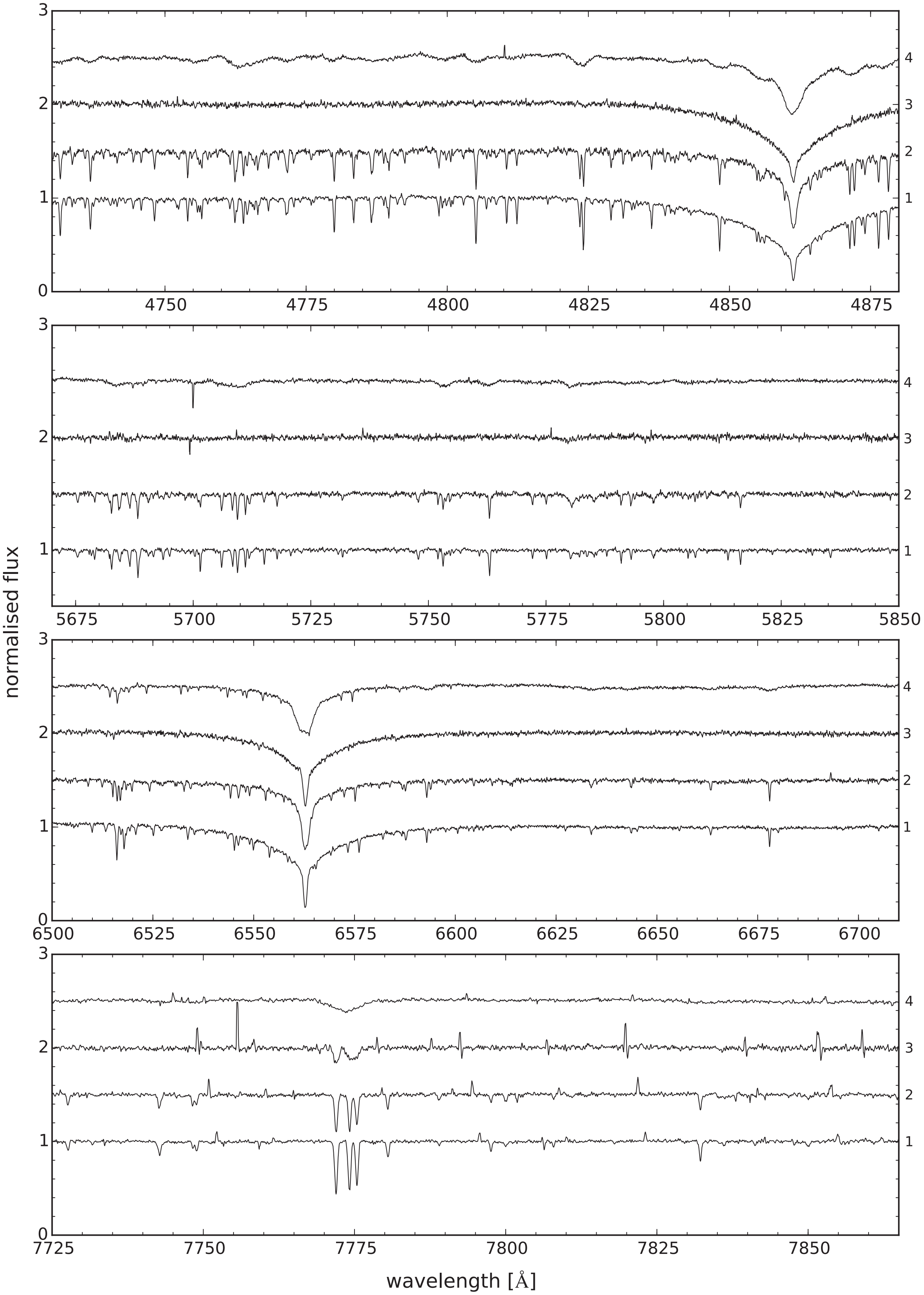

Figure 4. Examples of spectra of the hot stars category. Separate panels represent spectra from different Hermes spectral bands but for the same stars. The vertical spacing between the spectra in each panel is adjusted for clarity. From bottom to top in each panel, the spectra are labeled in sequence on the right axis, corresponding to the spectrum number in Table 2.

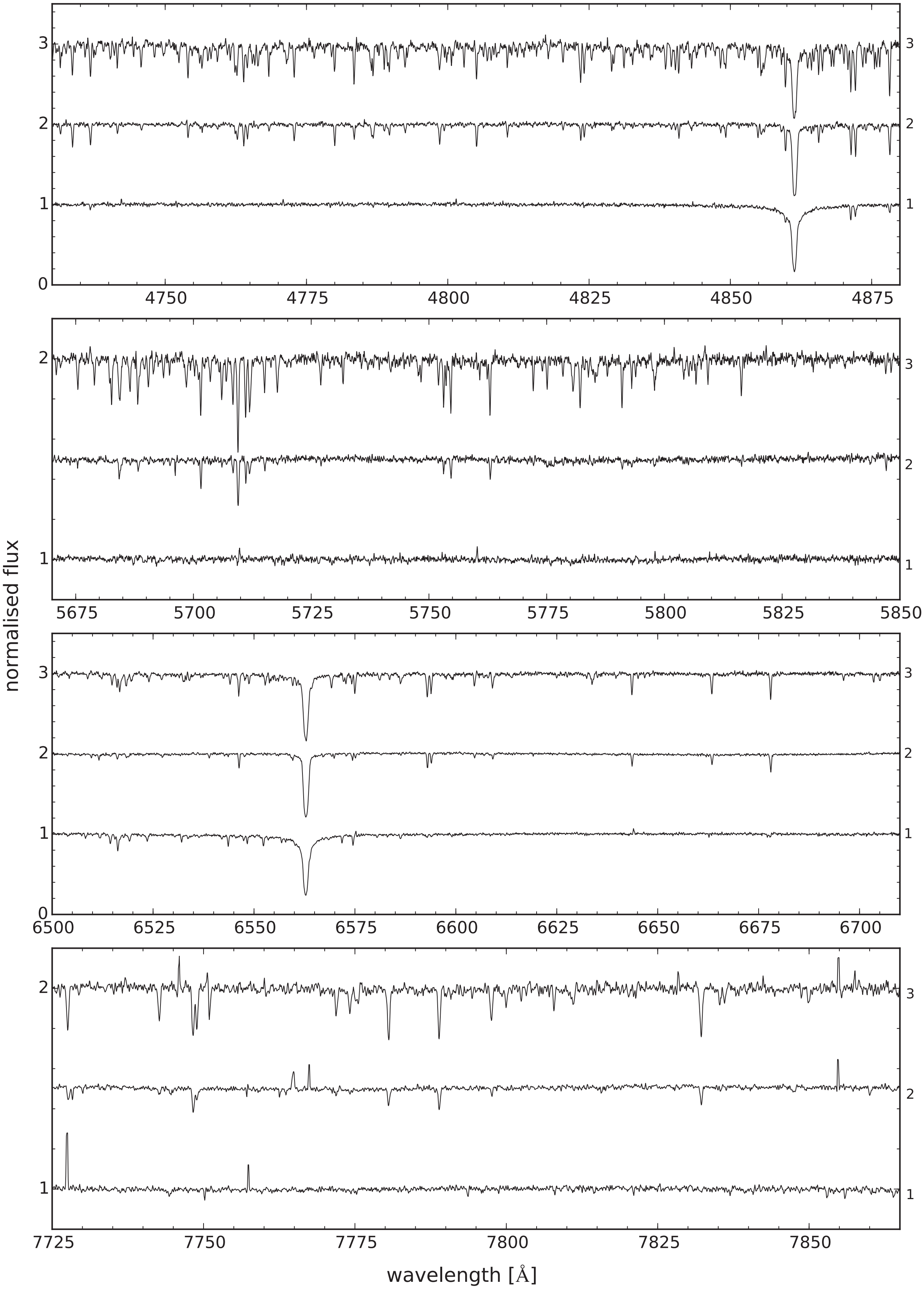

Figure 5. Same as Figure 4, but for the cool metal-poor giants category.

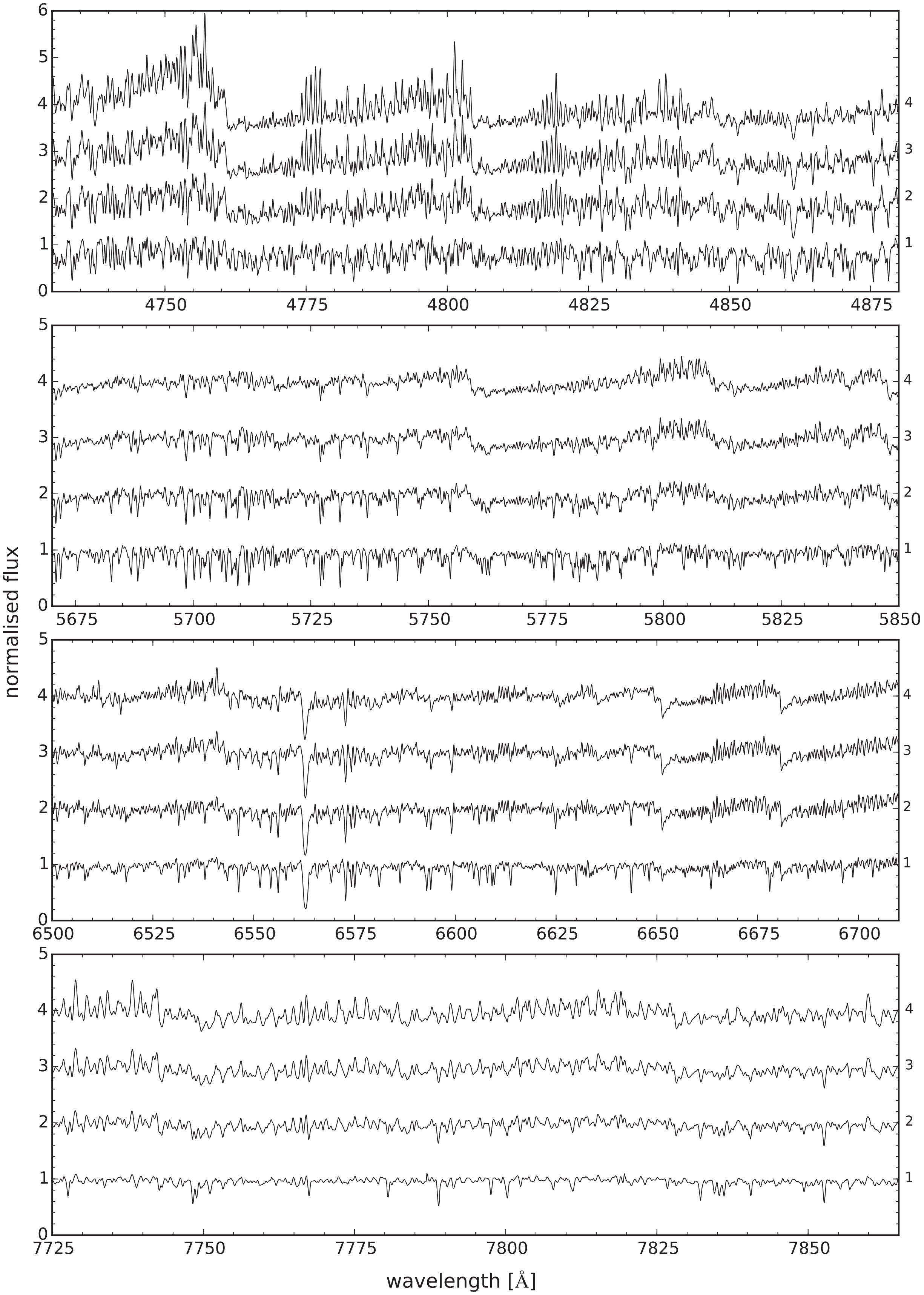

Figure 6. Same as Figure 4, but for the molecular absorption bands category.

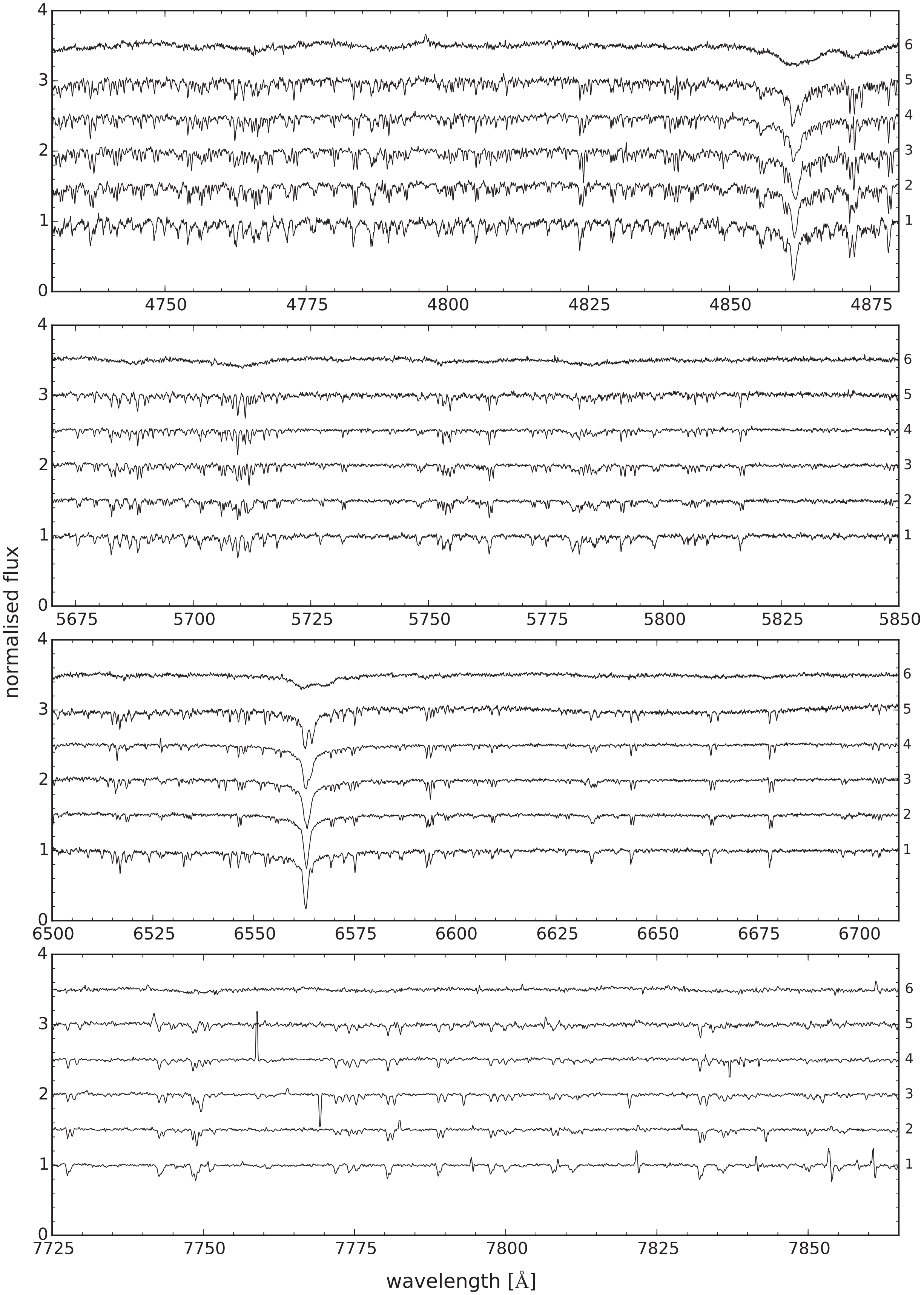

Figure 7. Same as Figure 4, but for the binary stars category.

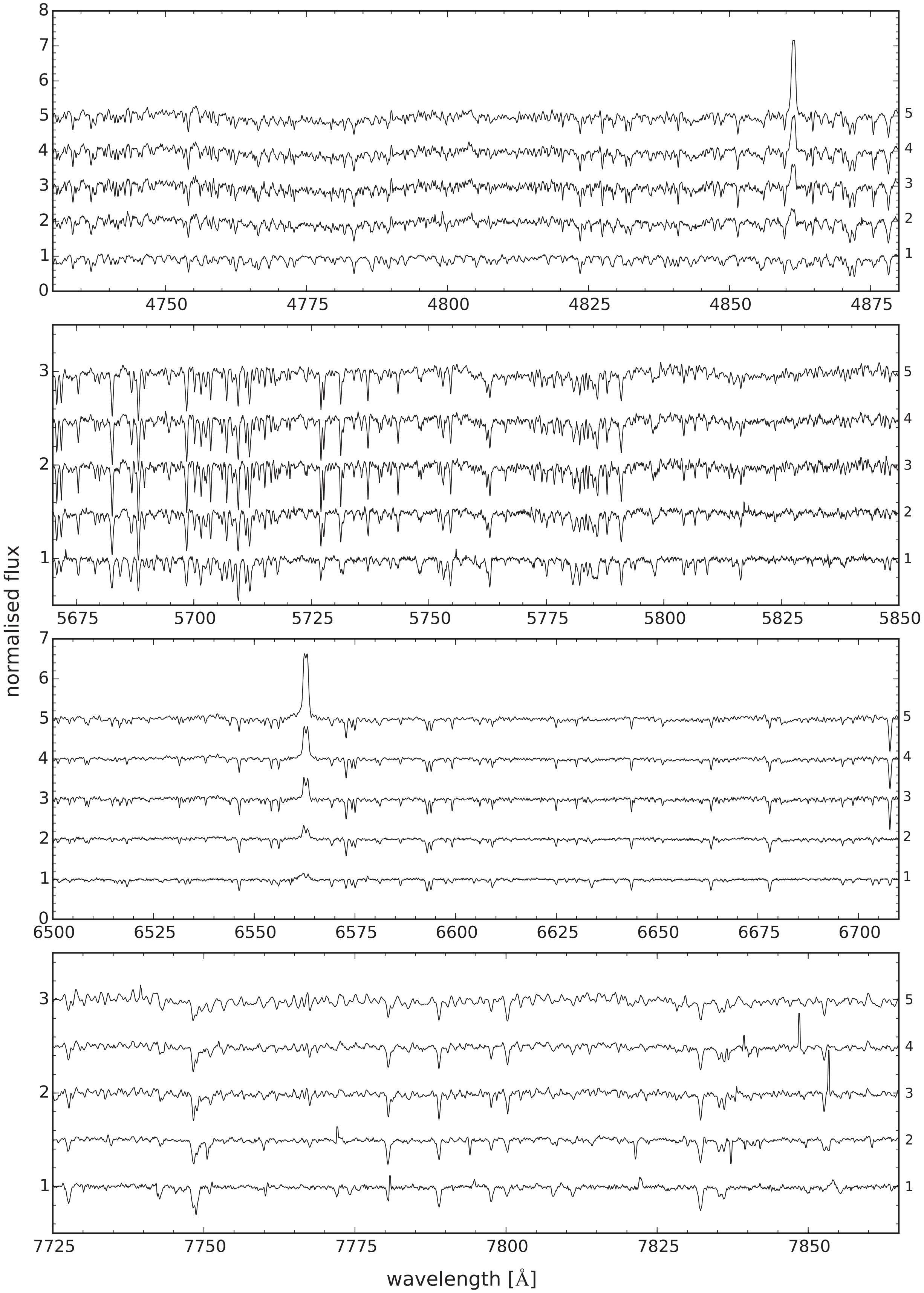

Figure 8. Same as Figure 4, but for the Hα/Hβ emission category.

Uma star, and the spectra are shown with increasing radial velocity separation of the two components from bottom to top. We have some indications of binarity in Table 1, although only for a handful of stars, the other candidates are currently unknown for their binary nature according to the SIMBAD database. The same search by coordinates as performed on SIMBAD reveals no systems in Pourbaix et al. (2004) and two systems in the Mason et al. (2001) catalogs. By visual inspection, some SB3 candidates and W UMa type SB2s are also part of this collection, although they are not isolated enough in the projection map to be labeled separately.

The number of spectra in this collection represents around 1% of the investigated Galah data set. However, the true number of such objects is doubtlessly larger, as there are many factors hindering their detection: SB2 with blended lines, exclusion of potential candidates in the third step of our classification procedure, or classification as problematic because of a stronger spectral feature. The simulation from Matijevič et al. (2010) performed for an SB2 analysis of RAVE spectra found that the detection rate should be fairly high (~80%) for systems with orbital periods shorter than ≈100 days. The limiting line separation $\triangle v_{\mathrm{orb}} \approx 50\ \mathrm{km\,s^{-1}}$ for RAVE (near-IR, S/N ~ 45, $R \sim 7500$) should be smaller for Galah because of the higher resolution and S/N of spectra, and therefore the detection of longer period systems should be greatly improved. Indeed, the smallest separations among the detected binaries in this collection are $\triangle v_{\mathrm{orb}} \approx 15\ \mathrm{km\,s^{-1}}$.

4.5. Hα/Hβ Emission

Emission-type stars often feature diverse profiles in Hα and Hβ emission lines, indicative of young stars, cataclysmic variables, symbiotic stars, stars with massive outflows or inflows, and many other types of active objects. The shapes of emission profiles can be described by meaningful morphological and possibly also physical categories, as demonstrated by Traven et al. (2015). In this collection, the diverse profiles (double peaks, emission superimposed on absorption, P-Cygni, and others) are presented together, as there are relatively few of them and they are also not clearly separated in the projection map. The Hα emission line is mostly present and often accompanied by a similar profile shape of Hβ line. In some cases, molecular absorption bands and the lithium absorption line are clearly visible, all together are indicative of cooler, younger, and active stars. Examples of spectra in this category are shown in Figure 8.

4.6. Problematic

This collection is very diverse as it assembles spectra that are in some peculiar, and generally undesirable fashion, affected in either observations or data reduction. Emission spikes (5352 spectra) are most often present in the IR band and sometimes, but less strongly, in the red band, and are probably due to undersubtracted sky lines. The left and right parts of the map, shaped by low-density snake-like collections, represent spectra with one strong emission spike (9599) in the IR band. A normalization issue in the form of an oscillating continuum (2025) in the red band is also very common. Negative flux (2078) is most often present in the IR band, followed by the blue band and less often in the red band, and might be due to sky oversubtraction. There is one more quite interesting, but less frequent, reduction effect in the IR band. This is in the form of very low continuum (41), which is either at ~0.3 or close to and below zero level, often accompanied by strong oscillating features. These subcategories follow each other in Figure 9 from bottom to top in each panel. Most spectra in this category are well behaved in all aspects except for the described issues, and their automatic detection without manual inspection is very helpful for the iterative development and improvement of our reduction pipeline (Kos et al. 2016).

5. Specific Search for Young/Active Stars

We also present additional classification results based on a more specific projection map, in contrast to the general map presented in Section 4. These results follow the same procedure as explained in Section 3.3, but with different t-SNE input parameters and input spectral ranges. The motivation for this approach is to search for stars in their early evolution phases (Žerjal et al. 2013) for which features in Hα, Hβ, and ^{7}Li spectral lines can be diagnostic of their activity (Soderblom 2010; Jeffries 2014). Perplexity is set to 50, and the spectral ranges 4841–4881 Å(Hβ) and 6543–6583 Å(Hα) are selected for the first t-SNE projection of the whole working set, while a perplexity of 15 and reduced spectral ranges (4859–4863, 6561–6565, and 6706–6710 Å) around the three diagnostic lines are selected for the second t-SNE projection of the filtered working set. Other combinations of perplexity and spectral ranges were tried, but this combination produced the most useful projection map.

Compared to the general classification from the previous section, we find additional candidates in the categories of binary stars (522), problematic spectra with oscillating continuum (665), and Hα/Hβ emission (868). With this projection map, we are able to partition the latter category and identify four distinct morphological subtypes (see Figures 10 and 11): Hα/Hβ emission, Hα/Hβ emission superimposed on absorption, Hα/Hβ P-Cygni, and Hα/Hβ inverted P-Cygni, all indicative of diverse underlying physical processes (Traven et al. 2015 and references therein). The Hα/Hβ emission category is a counterpart to the category presented in the previous section, and contains diverse multicomponent profiles of Hα/Hβ emission lines, that were not clearly separated in the projection map, as in the case of the latter three categories. It is possible that some emission profiles are a consequence of reduction issues instead of intrinsic properties of stars and their environment. We plan to address this possibility in future classification studies.

A new category lithium absorption is defined to account for spectra that display varying equivalent widths of the ^{7}Li line, from weak to very strong absorptions, as shown in Figure 12. Significant ^{7}Li absorption sometimes accompanies spectra in the Hα/Hβ emission categories, as is evident from Figure 10.

The categories in this section are listed in Table 3, along with the SIMBAD classes that indicate the connection between the youth and activity of stars and the observed Hα/Hβ multicomponent profiles and prominent lithium absorption.

6. Catalog

The final classification results are collected in the catalog, whose contents are described in Table 4. Spectra with at least one assigned category from either Sections 4 or 5 are listed by their catalog ID, internal Galah ID of the corresponding target, coordinates of the target, APASS (Henden et al. 2012;

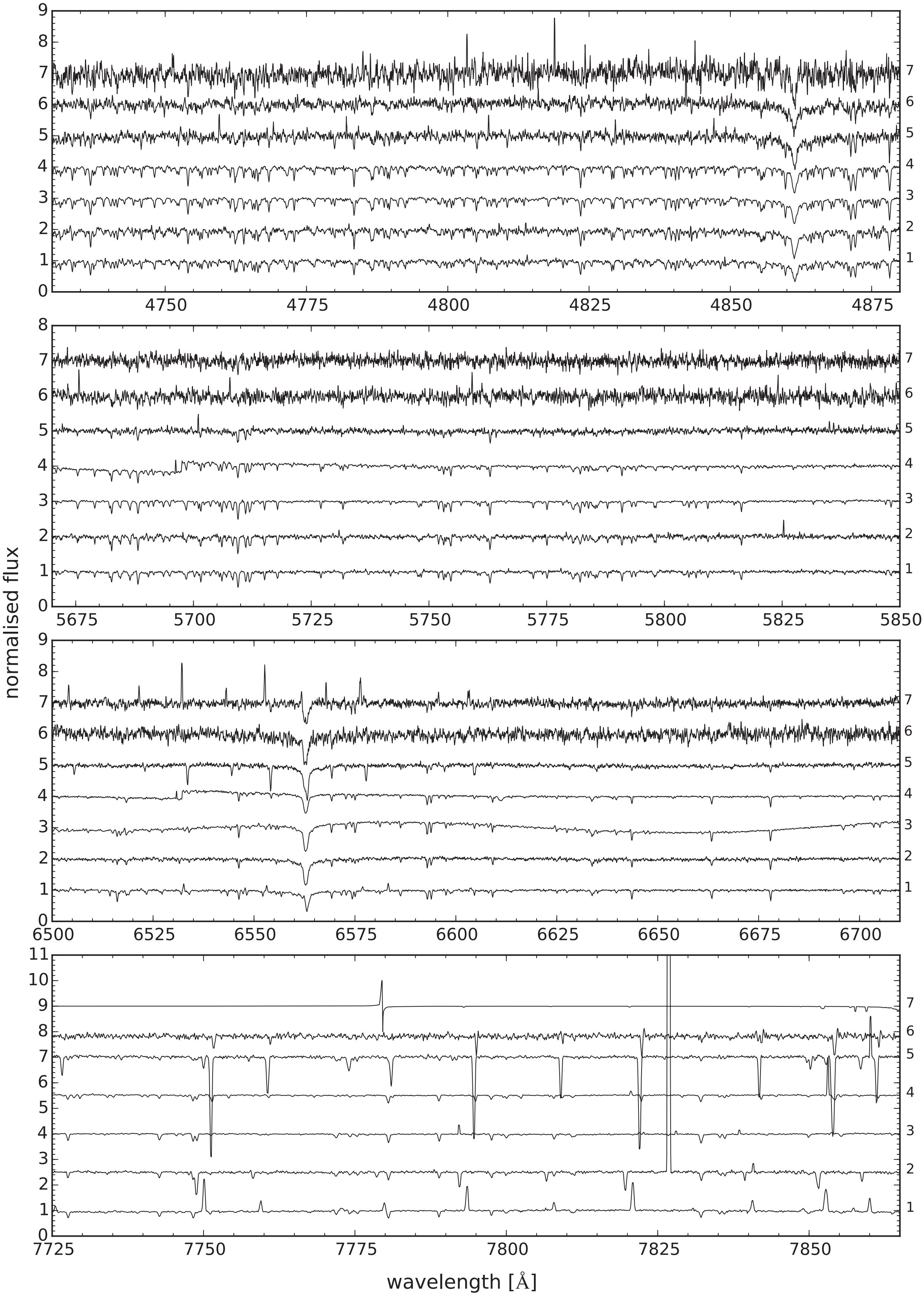

Figure 9. Same as Figure 4, but for the problematic category. Subcategories of spectra in the panels from bottom to top: emission spikes, strong emission spike, oscillating continuum, oscillating continuum, negative flux, low continuum, and low continuum.

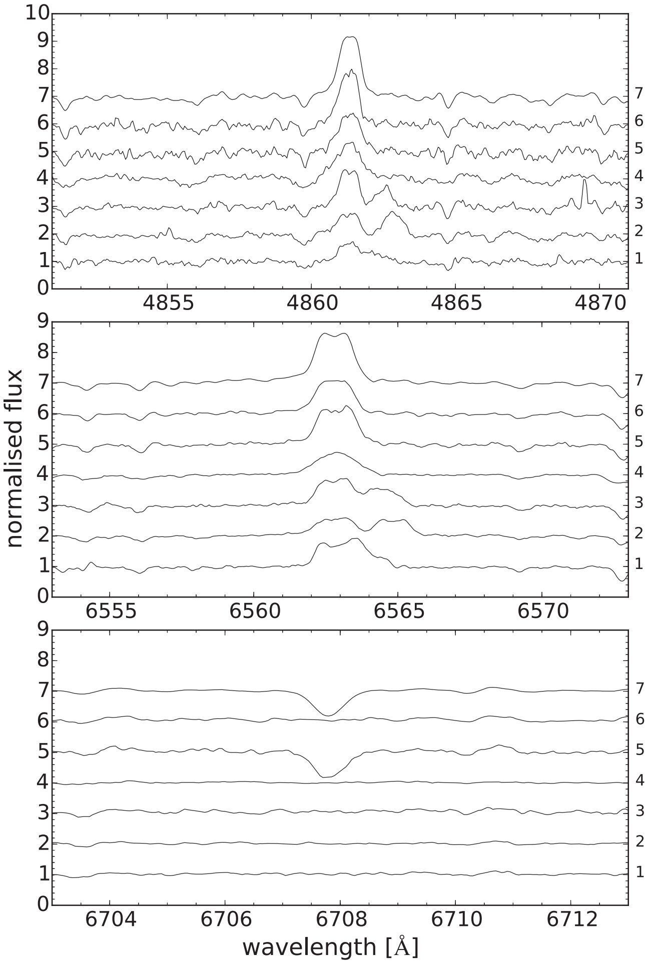

Figure 10. Same as Figure 4, but based on results from the specific search for young/active stars. Examples for Hα/Hβ emission category are displayed.

Munari et al. 2014) V magnitude, classification category, and supplementary information from SIMBAD, VizieR, OGLE, and ADS databases.

We cross matched the coordinates of stars to retrieve information from the SIMBAD, VizieR, OGLE, and ADS online databases. Epoch 2000.0 coordinates of our targets are not identical with those from the catalogs, therefore we adopted a search radius of 1″ where applicable. The results of the search in VizieR catalogs are retrieved in the wavelength ranges gamma-ray, X-ray, EUV, UV, optical, IR, and radio. In the catalog, we list the number of VizieR tables in which a match is found. We also state the type of variability (class) for matched targets in OGLE-III online Catalog of Variable Stars (Soszyński et al. 2013). References from the literature (ADS) should provide additional information about objects of interest, but are not necessarily reliable sources of the characteristics of a certain object.

Some of the 28,579 stars (with 31,050 spectra) in our catalog are known to be peculiar types and have been discussed in the literature or listed in different sources. SIMBAD matched 5956 targets, VizieR finds at least one match in at least one of its catalogs for all unique targets (28,579), OGLE matched 148 targets, and 350 targets are matched successfully with references from the ADS database.

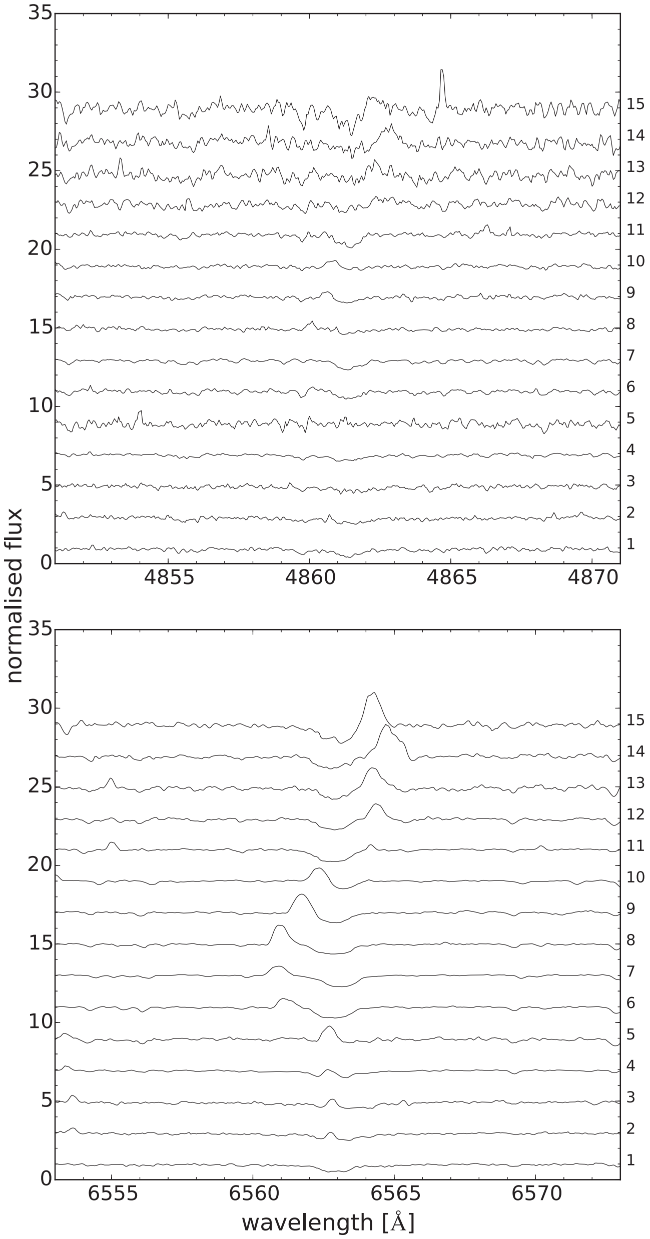

Figure 11. Same as Figure 10, but for the categories Hα/Hβ emission superimposed on absorption, Hα/Hβ inverted P-Cygni, and Hα/Hβ P-Cygni. From bottom to top in panels, each category features five examples of spectra.

The electronic version of the catalog will be made publicly available at the CDS, excluding results for spectra from the problematic category, since these mainly stand out for data reduction reasons and will be recoverable in the upgraded versions of the reduction pipeline.

7. Visualization—t-SNE Explorer.

The t-SNE or Galah Explorer is an interactive web application developed for members of the Galah collaboration that provides a visualization of the feature-based distribution of spectra in the t-SNE projection map. The basic view contains the following:

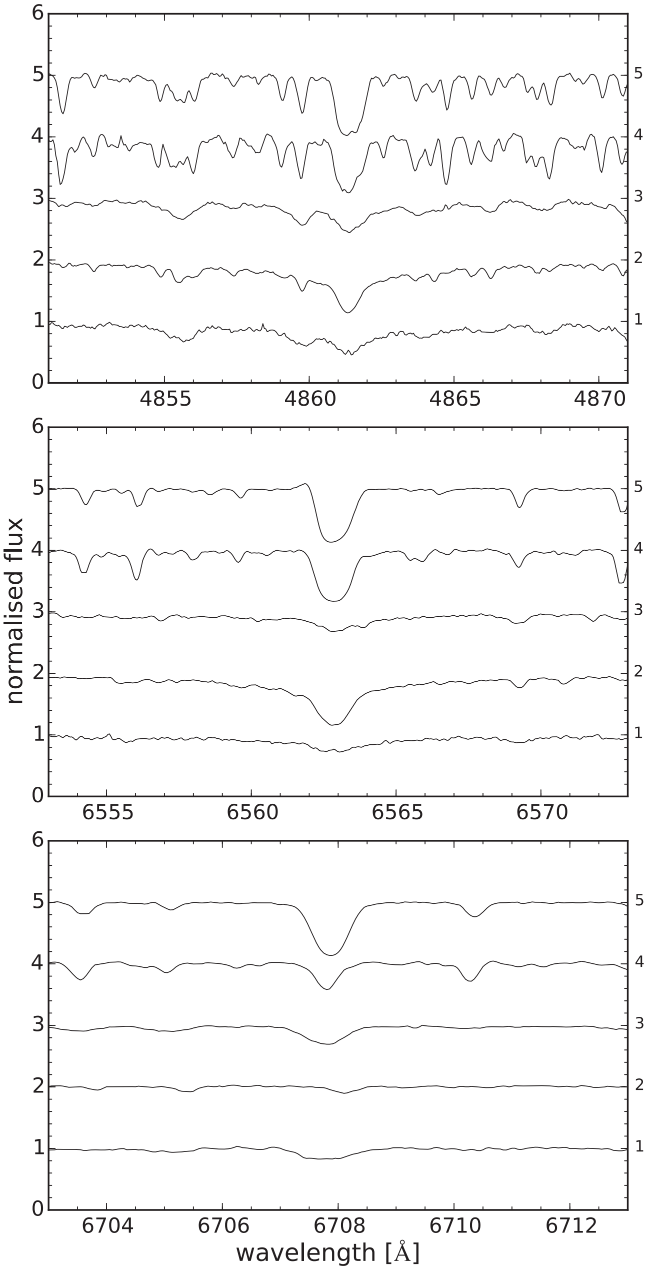

Figure 12. Same as Figure 10, but for the lithium absorption category.

t-SNE map: similar to those in Figures 1 and 2. The map is split into hexagons that are color coded based on average values of parameters of contained data points.
Large hexagonal frame: this displays data points of a selected hexagon from the map, where each data point is color coded depending on selected numerical or descriptive (e.g., classification) parameters.
List of parameters: this can be chosen for color coding the map, while the individual values for all parameters are always displayed for the currently selected data point. Any supplementary information on the corresponding object (star) can also be displayed together with a link to Simbad and Vizier matches.
Plotting area: this has four panels corresponding to the four Galah spectral bands, where the median and the dispersion of normalized fluxes of all spectra of the currently selected hexagon is displayed, overplotted with the currently selected data point (spectrum).
Search fields: here the user can search by Galah identifier or other parameter values or labels available in the Galah database of reduced spectra. This immediately selects and displays the matching object.

The presented segments of the Galah Explorer enable the user to locate specific areas of interest in the map that feature characteristic values of parameters. It is also possible to search for a specific object using e.g., its unique identifier. Once selected, the user can inspect its morphological vicinity (parent hexagon) with the neighboring spectra, using statistical plots to evaluate their similarity.

The (briefly) described functionality offers a very powerful and useful way of reviewing any kind of a data set, locating and exploring its inherent structure in feature space, and detecting outliers. Many different projection maps can be incorporated into the Galah Explorer together with different DBSCAN modes for an efficient selection and identification of distinct morphological collections of spectra. The tool will be available on the Galah official website at http://galah-survey.org.

8. Discussion

We have demonstrated that t-SNE can be used as an efficient tool for discovery of diverse spectral features and classification of stellar spectra. By projecting the Galah data set onto a two-dimensional space, it is able to preserve and visually reveal its complex morphological structure. Although not tested on our sample of spectra, it was shown by van der Maaten & Hinton (2008) that t-SNE is far superior in its domain, placing emphasis on (1) modeling dissimilar data points by means of large pairwise distances, and (2) modeling similar data points by means of small pairwise distances, which is not obvious for other nonlinear dimensionality reduction techniques and even less so for the linear ones.

The complexity of spectral morphologies in principle increases with increasing wavelength range. In this respect, Galah with its four spectral bands surpasses many other spectroscopic surveys. Consequently, this makes the task of classification more difficult because spectral features can appear differently and from different effects in each band (wavelength-dependent markers of physical processes, reduction issues, hardware malfunction, different optical paths, etc.).

Our classification procedure can accept any arbitrary spectral ranges selected by the user, which (1) enables an emphasis on the particular physics we are interested in, and (2) removes possibly unwanted influences from strong features in other parts of the spectrum, which can complicate classification of the desired types of objects. These advantages were demonstrated with the specific projection in Section 5, selecting only narrow regions of Hα, Hβ, and ^{7}Li for the search of young, active stars. Many detections of such objects were not possible with the first projection as the full spectral information along with strong problematic features, e.g., in the IR band, concealed those in other bands. For the same reason, we might miss some interesting morphological categories with weaker characteristic features hidden by stronger ones. The specific projection map yielded new candidates for three already defined categories from the general map, while also providing four new categories, validating the principle of exploiting

Table 3
Same as Table 1, but for the Specific Projection Map Produced in the Search for Young/Active Stars (see Section 5)

Category	N	Main type	Other types
Binary stars	1428	Star in cluster (5), variable star of RR Lyr type (3), eclipsing binary of Algol type (detached) (2), variable star (1), rotationally variable star (1)	Infrared source (178), variable star (6), star in cluster (5), rotationally variable star (4), variable star of RR Lyr type (3)
Hα/Hβ emission	135	Rotationally variable star (2), high proper-motion star (1), double or multiple star (1), X-ray source (1), infrared source (1)	Infrared source (19), X-ray source (7), variable star (4), rotationally variable star (3), high proper-motion star (3)
Hα/Hβ emission superimposed on absorption	479	Star in cluster (5)	Infrared source (8), star in cluster (5)
Hα/Hβ P-cygni	18	Variable star (1)	Variable star (1), rotationally variable star (1), infrared source (1)
Hα/Hβ inv. P-cygni	345		
Lithium absorption	664	Red giant branch star (6), rotationally variable star (5), high proper-motion star (1), spectroscopic binary (1), pre-main sequence star (1)	Infrared source (173), X-ray source (8), rotationally variable star (6), red giant branch star (6), pre-main sequence star (6)
Problematic[a]	1902		Infrared source (174)

Note. The categories binary stars, Hα/Hβ emission, and problematic are defined in Section 4, while the others are described in Section 5.
[a] A large portion of such spectra are recoverable (see the text).

(This table is available in machine-readable form.)

Table 4
Description of the Content for the Catalog Containing Results of Our Classification (see Section 6)

Label	Unit	Description
Catalog_ID		
Galah_ID	⋯	Unique star identifier
DATEOBS		Date and time of the observation
R.A.	°	R.A. (J2000)
Decl.	°	Decl. (J2000)
Class_cat_general	⋯	General classification category as given in Section 4
Class_cat_specific	⋯	Specific classification category as given in Section 5
SIMBAD_main_id	⋯	Main ID of the source in SIMBAD
SIMBAD_angular_distance	″	Angular distance of Galah target to the source in SIMBAD
SIMBAD_main_type	⋯	SIMBAD *main type*
SIMBAD_other_types	⋯	SIMBAD *other types*
VizieR_n_Radio	⋯	Number of VizieR tables for the radio wavelength range in which Galah target has a match
VizieR_n_IR	⋯	As VizieR_n_Radio, but for the IR wavelength range
VizieR_n_optical	⋯	As VizieR_n_Radio, but for the optical wavelength range
VizieR_n_UV	⋯	As VizieR_n_Radio, but for the UV wavelength range
VizieR_n_EUV	⋯	As VizieR_n_Radio, but for the EUV wavelength range
VizieR_n_Xray	⋯	As VizieR_n_Radio, but for the X-ray wavelength range
VizieR_n_Gammaray	⋯	As VizieR_n_Radio, but for the Gamma-ray wavelength range
OGLE_class	⋯	OGLE variable star type (class)
ADS_literature	⋯	A comma-separated list of articles (title and bibcode)

Note. The full table will be available at the CDS.

(This table is available in machine-readable form.)

different t-SNE set-ups to select the best (or several) projection maps for classification purposes.

The search for peculiar objects and the classification presented here is not exhaustive or absolutely representative of the whole data set. It is limited by and reflects our choice of the t-SNE and DBSCAN algorithms that form the basis of our analysis, the selection of their parameters, our iterative approach, and the spectral range used in each of the steps of our classification scheme. In this respect, we recognize only the most prominent features revealed by the selected classification setup, which further define 10 distinct categories listed in Tables 1 and 3, containing a total of 31,050 spectra (28,579 unique targets). We acknowledge the possibility of establishing additional categories of exotic spectra in the Galah data set, and this will be explored in future studies.

When more than one projection map is used, possibly produced by different input selected spectral ranges, or simply using different DBSCAN modes, it can happen that the same spectrum is assigned to more than one category as a result of the previously discussed reasons. Additional factors contributing to such cases are morphologically similar features, such as double lines from binary stars and emission superimposed on absorption. These might be located close in the projection space, with a possible overlap region. Spectra with more than one category can be easily identified in the catalog (Table 4) as having values for both general and specific classification fields.

The novel dimensionality reduction technique t-SNE is capable of representing astronomical spectra in a low-dimensional space where their morphology and hidden features can be efficiently discovered and studied. This was shown with an effective classification of the largest astronomical high-resolution spectroscopic data set so far, comprising 209,533 spectra, each containing 13,600 flux values. All data products along with the t-SNE Explorer will be publicly available in the coming data releases. The source code is freely available online, and our custom procedure for classification is easy to adapt to different spectroscopic or other astronomical data sets. This work will facilitate further investigation and understanding of the still-growing Galah data set and enable focused studies of distinct categories of objects (e.g., binary stars).

Although this work has made use of external sources, it is not dependent on them, and they mostly serve to support this proof-of-concept for the classification of a wide variety of astronomical data.

Based on data acquired through the Australian Astronomical Observatory, via programs 2013B/13, 2014A/25, 2015A/19. G.T. acknowledges the financial support from the Slovenian Research Agency (research core funding No. (P1-0188)). S.L. M. acknowledges the support of the Australian Research Council through grant DE140100598. This research has made use of the SIMBAD database (Wenger et al. 2000) and of the VizieR catalog access tool operated at CDS, Strasbourg, France. The original description of the VizieR service was published in Ochsenbein et al. (2000).

References

Bailer-Jones, C. A. L., Irwin, M., & von Hippel, T. 1998, MNRAS, 298, 361
Carretta, E., Bragaglia, A., Gratton, R. G., et al. 2009, A&A, 505, 117
Dalton, G., Trager, S. C., Abrams, D. C., et al. 2012, Proc. SPIE, 8446, 84460P
Daniel, S. F., Connolly, A., Schneider, J., Vanderplas, J., & Xiong, L. 2011, AJ, 142, 203
de Jong, R. S., Bellido-Tirado, O., Chiappini, C., et al. 2012, Proc. SPIE, 8446, 84460T
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., Jorg, S., & Xu, X. 1996, in Proc. 2nd Int. Conf. on KDD, ed. E. Simoudis, J. Han, and U. Fayyad 226
Freeman, K., & Bland-Hawthorn, J. 2002, ARA&A, 40, 487
Freeman, K. C. 2012, in ASP Conf. Ser. 458, Galactic Archaeology: Near-Field Cosmology and the Formation of the Milky Way, ed. W. Aoki et al. (San Francisco, CA: ASP), 393
Gilmore, G., Randich, S., Asplund, M., et al. 2012, Msngr, 147, 25
Gratton, R. G., Carretta, E., & Bragaglia, A. 2012, A&ARv, 20, 50
Gulati, R. K., Gupta, R., Gothoskar, P., & Khobragade, S. 1994, ApJ, 426, 340
Hartigan, J. A., & Wong, M. A. 1979, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100
Henden, A. A., Levine, S. E., Terrell, D., Smith, T. C., & Welch, D. 2012, JAVSO, 40, 430
Hinton, G., & Roweis, S. 2002, Advances in Neural Information Processing Systems 15 (Cambridge, MA: MIT Press)
Hotelling, H. 1936, Biometrika, 28, 321
Ibata, R. A., & Irwin, M. J. 1997, AJ, 113, 1865
Izenman, A. J. 2008, Linear Discriminant Analysis (New York: Springer)
Jeffries, R. D. 2014, EAS, 65, 289
Kos, J., Lin, J., Zwitter, T., et al. 2016, MNRAS, arXiv:1608.04391
Liu, F., Asplund, M., Yong, D., et al. 2016a, MNRAS, arXiv:1608.03788
Liu, F., Yong, D., Asplund, M., Ramírez, I., & Meléndez, J. 2016b, MNRAS, 457, 3934
Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ApJS, 225, 31
Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA, 15, 1095
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2015, arXiv:1509.05420
Marin, J.-M., Mengersen, K., & Robert, C. P. 2005, in Handbook of Statistics, Vol. 25, ed. D. Dey & C. R. Rao (Amsterdam: North Holland), 459
Martell, S., Sharma, S., Buder, S., et al. 2016, arXiv:1609.02822
Mason, B. D., Wycoff, G. L., Hartkopf, W. I., Douglass, G. G., & Worley, C. E. 2001, AJ, 122, 3466
Matijevič, G. 2016, in IAU Symp. 317, The General Assembly of Galaxy Halos: Structure, Origin and Evolution, ed. A. Bragaglia et al. (Cambridge: Cambridge Univ. Press), 336
Matijevič, G., Zwitter, T., Bienaymé, O., et al. 2012, ApJS, 200, 14
Matijevič, G., Zwitter, T., Munari, U., et al. 2010, AJ, 140, 184
McGurk, R. C., Kimball, A. E., & Ivezić, Ž. 2010, AJ, 139, 1261
Munari, U., Henden, A., Frigo, A., & Dallaporta, S. 2014, JAD, 20, 4
Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Pezzotti, N., Lelieveldt, B., van der Maaten, L., et al. 2016, arXiv:1512.01655
Piskunov, N., & Valenti, J. A. 2016, arXiv:1606.06073
Pourbaix, D., Tokovinin, A. A., Batten, A. H., et al. 2004, A&A, 424, 727
Prusti, T. 2012, AN, 333, 453
Sharma, S., & Johnston, K. V. 2009, ApJ, 703, 1061
Sheinis, A., Anguiano, B., Asplund, M., et al. 2015, JATIS, 1, 035002
Soderblom, D. R. 2010, ARA&A, 48, 581
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2013, AcA, 63, 21
Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645
Switzer, P., & Green, A. A. 1984, Computer Science and Statistics, 13
Ting, Y.-S., Freeman, K. C., Kobayashi, C., De Silva, G. M., & Bland-Hawthorn, J. 2012, MNRAS, 421, 1231
Traven, G., Zwitter, T., Van Eck, S., et al. 2015, A&A, 581, A52
Valenti, J. A., & Piskunov, N. 1996, A&AS, 118, 595
Valentini, M., Chiappini, C., Davies, G. R., et al. 2016, arXiv:1609.03826
van der Maaten, L. 2013, arXiv:1301.3342
van der Maaten, L., & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579
von Hippel, T., Storrie-Lombardi, L. J., Storrie-Lombardi, M. C., & Irwin, M. J. 1994, MNRAS, 269, 97
Watson, F. G. 1987, PhD thesis, Edinburgh Univ.
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Young, F. W. 2013, Multidimensional Scaling: History, Theory, and Applications (Psychology Press)
Žerjal, M., Zwitter, T., Matijevič, G., et al. 2013, ApJ, 776, 127