UNIVERSITY OF LJUBLJANA
FACULTY OF MATHEMATICS AND PHYSICS
DEPARTMENT OF PHYSICS

Klemen Čotar

Understanding of open stellar clusters and peculiar spectra in sky surveys using machine learning

Doctoral thesis

ADVISER: prof. dr. Tomaž Zwitter

Ljubljana, 2020

UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO
ODDELEK ZA FIZIKO

Klemen Čotar

Razumevanje razsutih zvezdnih kopic in posebnih tipov spektrov v pregledih neba z uporabo strojnega učenja

Doktorska disertacija

MENTOR: prof. dr. Tomaž Zwitter

Ljubljana, 2020

Zahvala

Uvedba v raziskovalni proces in delo morda ni najlažja naloga, zato se na tem mestu najprej zahvaljujem kolegom iz ZRC-SAZU in Vesolje-SI za neomajno podporo in spodbudo pri delu in odločitvah, brez katerih tega zapisa morda celo ne bi bilo. Hvala za vse, tako še vedno z veseljem rad rešim uganke in nejasnosti mojih procedur.

Hvala tudi mentorju ter celotni astronomski skupin na Fakulteti za matematiko in fiziko za sprejem in možnost sodelovanja v različnih mednarodnih kolaboracija.

Del tega so bila tudi opazovanja v Asiagu, ki sem jih izkostiristil kot zanimivo priložnost za oddih, sodelovanje in spoznavanje novih stvari, za kar se zahvaljujem prof. Ulisse Munari, ki me je vedno podučil o čem novem.

Navsezadnje pa je delo tudi odraz podpore domačih v vseh teh letih študija za kar se jim zahvaljujem.

Razumevanje razsutih zvezdnih kopic in posebnih tipov spektrov v pregledih neba z uporabo strojnega učenja

Izvleček

Naraščajoče število popolnoma avtomatiziranih in visoko multipleksiranih teleskopov nam posreduje vedno večje število opazovanj, katerih začetna obdelava in analiza sta vedno bolj in bolj avtomatizirani. Ker omejeno število raziskovalcev ne more ročno obdelati in preveriti vseh zbranih informacij, se bo stopnja avtomatizacije s časom le še povečevala. To ne pomeni, da so uporabljeni procesi brezhibni in bo znanost v bližnji prihodnosti predana računalnikom. Ravno nasprotno, zapletenost teleskop in opazovanj ter znanstvenih vprašanj še vedno zahteva, da operator prilagodi procese, ugotovi prej neopažene težave in razmišlja o naslednjem znanstvenem problemu.

Orodja za strojno učenje niso čarobna orodja, saj le skušajo rešiti težavo, ki si jo je zamislil njihov uporabnik. Če je vprašanje slabo opredeljeno ali so podatki nepravilno pred-pripravljeni in filtrirani, bodo tudi rezultati obremenjeni z vhodnimi odločitvami. To predstavlja pomembno težnjo po poznavanju ne samo razpoložljivih podatkov, ampak tudi kako so bili pripravljeni in označeni z raznimi statusi njihove kvalitete.

Disertacija predstavlja rezultate našega raziskovanja obsežne podatkovne baze GALactic Archaeology with HERMES (GALAH) z različnimi orodji strojnega učenja in poudarja potrebo po nadzoru in razumevanju vhodnih podatkov. Predstavljeni rezultati obravnavajo raziskovanje razsutih zvezdnih kopic in njihov kemični podpis. Pri tem sta glavna problema homogenost in pravilnost izračunanih kemičnih zastopanosti. Pri kemičnem raziskovanju okolice razsutih kopic smo uporabili razumevanje omejitev uporabe metode kemičnih podpisov razsutih zvezdnih kopic, ki naj bi imele homogeno sestavo.

Drugi večji sklop disertacije se ukvarja z iskanjem manjših podskupin posebnih tipov spektrov v podatkovni bazi GALAH. Z različnimi pristopi primerjave med opazovanimi in modeliranimi normalnimi spektri smo iskali spektre z izrazitimi molekularnimi pasovi molekule C_2, emisijskimi črtami in spektroskopsko nerazpoznavne večkratne sisteme.

Ključne besede: metode: analiza podatkov – razsute kopice in asociacije: splošno – Galaksija: zvezdne populacije – zvezde: zastopanosti – zvezde: hitrosti in premikanje – dvojnice: splošno – zvezde: posebni tipi – zvezde: karbonske – zvezde: aktivnost – zvezde: emisijske črte – zvezde: podobne Soncu – črte: profili – katalogi

PACS: 95.10.Eg, 95.75.De, 95.75.Fg, 95.75.Mn, 95.80.+p, 97.10.-q, 97.10.Ri, 97.10.Tk, 97.10.Vm, 97.21.+a, 97.30.Eh, 97.30.Fi, 97.80.-d, 97.80.Fk, 98.58.H

Understanding of open stellar clusters and peculiar spectra in sky surveys using machine learning

Abstract

The increasing number of fully automated and highly multiplexed telescopes produces an ever-increasing number of observations whose reduction and analysis tend to be as automated as possible. As a limited number of researchers is unable to process and check all collected information manually, this degree of automation will steadily increase with time. This does not mean that used processes are flawless, and science will be in the near future handed to machines. Quite the opposite, the complexity of machinery and investigated scientific topics still requires a person to tweak the processes, identify previously unrecognised problems and think about the next scientific problem.

Machine learning tools are not magical solutions as they only try to solve the problem introduced by the operator. Therefore if a question is poorly defined or data improperly prepared and filtered, results will also be burdened by those decisions. This presents an important need of not only knowing the data at hand but also how were they prepared and possibly quality flagged during production.

This thesis presents the results of our exploration of extensive GALactic Archaeology with HERMES (GALAH) data set with various machine learning tools and emphasizes the need for control and understanding of input data. Presented results deal with the exploration of open stellar clusters and their chemical signature, where the main problems are homogeneity and correctness of the determined chemical compositions. By understanding the limitations of chemically tagging open cluster stars with homogeneous composition, the findings were applied to the chemical exploration of their surrounding.

The second large part of this thesis deals with finding smaller subsets of peculiar spectra in the GALAH data set. By various approaches of comparison between observed and modelled normal spectra, we searched for spectra with pronounced molecular bands of C_2 molecule, emission line features and spectroscopically unresolved multiple systems.

Keywords: methods: data analysis – open clusters and associations: general – Galaxy: stellar content – stars: abundances – stars: kinematics and dynamics – binaries: general – stars: peculiar – stars: carbon – stars: activity – stars: emission-line – stars: solar-type – line: profiles – catalogues

PACS: 95.10.Eg, 95.75.De, 95.75.Fg, 95.75.Mn, 95.80.+p, 97.10.-q, 97.10.Ri, 97.10.Tk, 97.10.Vm, 97.21.+a, 97.30.Eh, 97.30.Fi, 97.80.-d, 97.80.Fk, 98.58.Hf

Contents

Chapter 1

Introduction

Observational studies and simulations show that stars in our Galaxy were not formed at the same time, but at multiple epochs in different places of the Galaxy [1, 2]. One of the youngest building blocks of the Galaxy are open stellar clusters whose stars formed from the same molecular cloud of material [3] and therefore retain some properties of the original cloud they were made from. Stellar properties of an individual star can be separated into a kinematic and chemical component. The first is describing the position and movement of a star in the Galaxy. The second gives information about its chemical composition and physical properties. During the evolution of the Galaxy, clusters that have a lower number of members and are therefore gravitationally loosely bound, slowly evaporate because of different effects such as gravitational dynamical friction and ejection of stars during close intra-cluster stellar interactions. The latter is a result of close gravitational interactions among members, where one of them can be ejected out of a cluster at high velocity [4, 5, 6]. Such past members can on the sky be found even several degrees away from their main cluster body [7, 8, 9]. The gravitational dynamical friction happens when a cluster moves trough regions of the Galaxy with a higher density of stars [10]. During such event, the gravity of a cluster starts pulling seemingly fixed stars around it. Considering energy and momentum conservation law, we can conclude that a cluster will be slowed down for the same amount. This slowing subsequently changes the orbit of individual cluster members around the centre of the Galaxy. Small orbital changes lead to a gradual expansion of a cluster volume that can be tracked until its individual stars blend with a general stellar field population, making them unrecognizable as past cluster members. The most prominent transitional features we can observe are compact cluster tidal tails [11, 12, 13, 14] and lose extended halos of evaporated stars. The halos can be observed as a slowly decreasing over-density [15, 16, 17] of stars far from a denser central cluster core.

Born from the same molecular cloud, open clusters are therefore ideal tracers of formation, assembly and evolutionary history of their host galaxies. Being influenced by external and internal processes, their lifetime is limited from about 100 Myr to a few Gyr for the densest structures [18, 19]. Studies have shown that the dissolution time of a cluster depends mainly on its total mass, its radius and its galactic environment (density of the stars around it and on its galactic path). Many papers have so far dealt with the question of determining the precise lifetime of stellar clusters before they completely dissipate into field population. The problem has been tackled using direct N-body simulations [18] where the movement of individual stars

was traced, and from direct observational data [20, 21, 22, 23] using isochrone fits. Age distribution of open clusters within a distance of 1 kpc from the Sun shows a cluster median lifetime of 200 Myr. Their broad range in ages gives us a possibility of observing them at different evolutionary stages [24, 25] before they blend [26] into a field stellar population. On the other hand, such a short lifetime (in comparison with a lifetime of a Sun-like star) gives us a limited number of bounded clusters in the sky that can be studied at a given time.

The processes described above heavily depend on the mass of a cluster, its internal structure and mass distribution of its components. When star formation ceases, we are left with a wide range of stellar masses. Their distribution can be described by an initial mass function (IMF) [27, 28, 29] that appears invariant among clusters and even stars in the field [30]. The initial spatial distribution of stars in young clusters may reflect the structure of parental molecular cloud [31]. In many stellar clusters, the brightest and most massive stars are concentrated towards the centre of a cluster, this state is usually attributed to mass segregation. Whether mass segregation occurs due to an evolutionary effect or it is of primordial origin is not yet entirely clear [32, 33, 34, 35]. In the first case, massive stars are formed all around the cluster volume and eventually sink to its centre through the effect of two-body relaxations. In the second scenario, massive stars form preferentially in the central region of a cluster either by gas accretion due to their favourable location at the bottom of a gravitational potential well or through a coalescence process of less massive stars. The fact that mass segregation is also observed in young clusters might suggest that the second scenario is more likely that the evolutionary mass segregation, but the question is still under investigation [36].

Similarly to the IMF, we can also define the initial binary population (IBP) of the observed sample of stars. Characterizing fraction of binaries in stellar clusters is of great importance for many fields of astrophysics. Since binaries are on average more massive than single stars, they are thought to be useful tracers of dynamical mass segregation [37]. Comparing observations with a theoretical model for the radial distribution of binary systems and their properties (mass and luminosity ratio, orbital period) distribution can be used to assess the dynamical state of a cluster [38].

1.1 Open clusters in the *Gaia* era

Since *Gaia* Data Processing and Analysis Consortium (DPAC) centres responsible for analysis of the *Gaia* observations started publishing publicly available ready-to-use stellar properties, many new discovers and insights have been made. For the first time in history, we reliably know the position, distance, complete spatial velocity (proper motion and radial velocity) and luminosity for millions of bright and dim stars all across the observable sky. Their accuracy heavily depends on the source brightness. For example, published *Gaia* parallax uncertainties are in the range from 0.04 to 0.7 mas for the faintest sources. Similarly, radial velocities are precisely known in the range from 300 $\mathrm{m\,s^{-1}}$to 3 $\mathrm{km\,s^{-1}}$(further details are given in Section 2.1).

One of the fields in astronomy that massively benefited from those new and improved measurements was a research of open stellar clusters.

A traditional historical way to look for open stellar clusters was through counting

stars as they are seen on the sky from Earths' location and finding over-densities in those counts [39, 40]. To perform a more robust selection, members were additionally filtered based on their apparent distance from the cluster centre and their motion vectors [41]. Using the latest second release of *Gaia* data [DR2, 42], we can go beyond that and build upon the results achieved by the methods mentioned above. Complete *Gaia* information on stellar distance, kinematics and photometric measurements enable us to go beyond simple methodologies, to unravel even the faintest and sparsest components of open clusters. So far, many works have been published trying to refine parameters, and membership information of long known open clusters [41, 43, 44] and find new, less numerous or fainter clusters [45, 46, 47, 48, 49]. Such thorough and the improved investigation uncovered that many of the clusters listed in modern catalogues, initially discovered as apparent stellar overdensities, are no more than chance alignments of stars and not true physically bound clusters [50, 51, 52, 53, 54].

Precisely determined open cluster membership also enables accurate age determination of a cluster and its stars [55]. When observing stars in any random stellar field, it is notoriously hard to determine their ages as they remain unchanged for the majority of their lifetime. While it is possible to obtain precise age estimates for some specific classes of stars [56], a uniform methodology does not exist for all. Age determination is much easier for stars in a cluster as they are all of a very similar age. Stellar evolution and its cycle are determined mainly by the initial mass of a star. Observing stars of different masses in a cluster reveals its evolutionary stage and consequently also its age by comparing them to theoretical evolutionary models. The latest *Gaia* data enabled researchers to compute stellar absolute magnitudes and colours that are commonly used in those comparisons [22, 23, 57].

Another advantage of the newest *Gaia* dataset, with accurately determined distance to objects, is the possibility of determining the actual shape of a cluster, mass segregation of member stars and their gravitational potential from the first principles. Until recently that was possible only through n-body simulations [36, 58, 59] that were compared with available observational datasets. For distant stellar associations, the distance error bars are still larger than a cluster itself, but improvements are expected when *Gaia* EDR3 will be released in the second half of the year 2020 or later. Internal cluster dynamics can in those cases be inferred using precisely determined radial velocities and proper motion vectors whose uncertainty is not affected by stellar distance, but their apparent brightness. Deviations (overtaking or lagging) from the mean cluster velocity can be therefore be used to assess in which direction, away from the mass centre, stars are moving. This velocity deviation consequently also indicates its position in a cluster. Reliable determination of stellar mass distribution in a cluster also depends on accurate classification of binary or higher multiplicity stars and their masses. By using pure *Gaia* information, they can be identified as stars with higher radial velocity uncertainty [60], by its temporal variation in future data releases [61], using precise absolute magnitudes [62, 63, 64] or by exploring astrometric uncertainties [65] that indicate movement of their combined photocentre.

1.2 Chemical tagging

Majority of previously described works in Section 1.1 relies on a complete 6D positional and kinematics information to discover clusters and their sub-structures. Advances in observational techniques and data analysis enables us to go beyond kinematics information and include a multidimensional chemical signature of stars – the procedure known as chemical tagging [66, 67]. So far a blind chemical tagging (without kinematics) that would delineate between cluster and field stars has not yet been demonstrated with great success unless the observed structure has obviously different chemical composition [68]. A trait that it is not common to open stellar clusters formed at about the same time [69], but to galactic components formed at vastly different epochs [70] of Galaxy formation.

The measured chemical signature of a star gives us its composition in the outermost layer, where it remains unchanged for young stars. Its composition is, therefore, a direct reflection of a medium from which it was built. Galaxy started evolving from dust made predominantly from hydrogen and helium which slowly got polluted by evolved stars during their final evolutionary stages. This gradual enrichment is nowadays observable as gradients of chemical abundances in radial and vertical directions of Galaxy. Internal gravitational mixing causes blending of those signatures as stars move to different orbits. Already complicated chemical structure of Galaxy is in some regions interrupted by newly created dense stellar structures – open clusters – which we would like to discover by the above-described procedure. Because of multiple gradual processes going on in Galaxy, the chemical tagging so far does not have clearly defined cuts to separate galactic components.

The latest research showed that many, even unexpected, observational and data reduction issues still have to be thoroughly investigated and resolved, especially if data from different surveys are to be combined [71]. Studies suggest that open clusters might not be as homogeneous as thought before [8, 72], and abundances show traits of stellar evolution [73, 74, 75]. On top of that, the main concerns of chemical tagging are abundance trends assumingly induced by spectral analysis [72, 76, 77]. Observed trends depend on determined stellar physical parameters (i.e. $T_{\rm eff}$ and $v \sin i$) and might be results of inadequate stellar models or actual stellar processes. To cope with this complexity and uncertainties, complex Bayesian models are being developed [72, 78] in order to uncover and cluster abundance patterns. With this in mind, many work and validation still have to be done until large surveys are fully ready for blind chemical tagging experiments.

1.3 Peculiar stellar spectra

Observational difficulties and data reduction issues are not the only obstacles that have to be controlled in order to perform successful blind chemical tagging. In every larger observational sample of stars, we can find a small sub-sample of stars with properties that deviate from the majority of the randomly selected stellar population. Such stars can be identified by having kinematic properties vastly different than their local volume of space [79, 80, 81], having distinctly unusual chemical abundances [82] or have spectral features that are seen in a small percentage of stars [83, 84]. Among them, we also count sources on the sky whose observables show that they are a composite of two or more stellar components [85, 86, 87]. During the early stages

of stellar evolution, stars could be found surrounded by the optically thin material from which they formed [88]. Such peculiar stars, commonly found in young open clusters [89, 90], exhibit additional emission features in their spectra. In the opposite context, we commonly refer to the majority of spectra as normal, but the delineation between classes is not strict and may depend on a considered scientific question.

Every peculiarity in observed spectra could potentially influence derived stellar physical parameters and consequently also chemical abundances as they all depend on the comparison between observed spectra and computed synthetic spectra that rely on physical models of stellar interior. As the exploration of galactic history, also known as the galactic archaeology, relies on normal stars for which their parameters can be reliably measured, we would like our set of analysed stars to be as clean as possible. The sets are usually delineated by the procedure known as classification. This refers to a procedure that sorts objects into categories and labels them according to their properties. Those labels can reflect actual physical properties and nature of stars or uncover unexpected features in observed data. Anyhow, both normal and peculiar stars are essential for understanding large-scale galactic dynamics, the internal structure of stars and their life cycle.

1.4 Machine learning in large sky surveys

In the past decades, we witnessed a significant change in many areas of scientific research, including the filed of astronomy. New observational methods and advance scientific experiments are serving us ever-increasing amounts of data that have to be analysed by researchers. For example, astronomy has seen a shift from painstaking and time-consuming observations of a single star and its spectra to fibre-fed spectrographs that are able to simultaneously acquire spectra of hundreds to thousands of stars in a comparable time. Given the vast amounts of acquired data that can not in reasonable time be individually inspected and analysed, we are increasingly relying on computer algorithms to ease and speed-up those steps for us.

Among the studies dedicated to the exploration of stellar objects, we can choose spectroscopic observations among the following ongoing surveys: *Gaia* spacecraft [91], GALactic Archaeology with HERMES (GALAH [92]), Apache Point Observatory Galactic Evolution Experiment (APOGEE [93]), and Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST [94]). In addition to them, we can also obtain data from the past large surveys Sloan Extension for Galactic Understanding and Exploration (SEGUE [95]), RAdial Velocity Experiment (RAVE [96]), and Gaia-ESO Survey (GES [97]). The technological development is not stopping, bringing us even more complex telescopes and spectrographs that will acquire even more spectra in a single exposure. We, therefore, look forward to the observation start of surveys 4-metre Multi-Object Spectrograph Telescope (4MOST [98]), WHT Enhanced Area Velocity Explorer (WEAVE [99]), and Funnel Web.

All mentioned surveys differ in the target selection methodology (magnitude and/or colour thresholds), properties of observed spectra (resolution, wavelength coverage etc.), and sky coverage, but all produce normal and peculiar spectra that have to identified and analysed. Because of a large number of observations, it is not feasible or justified to manually reduce and analyse all those spectra, therefore automatic pipelines have to be used. The first step in understanding observed spectroscopic data is a reduction pipeline, tailored to a specific telescopic and spectrograph

setup [100, 101, 102], that will homogeneously prepare all spectra for further use and potentially add quality flags, warning a user that something might be wrong.

After the data have been prepared, machine learning approaches have been used in many different ways to extract further information from the observed spectra. The most important stellar properties in galactic archaeology are iron content [Fe/H] and individual chemical abundances. The first gives information about the amount of iron in the observable outer layers of the star compared to the amount of hydrogen. It is easily measurable as iron atoms cause numerous absorption lines in a spectrum. Similarly, individual abundances give the abundance of analysed element X compared to iron, usually denoted as [X/Fe]. Both of them are measured on the logarithm scale and compared to abundances found in the Sun. The iron abundance of our Sun is in the given system defined as [Fe/H] = 0.

When having access to a small set of observations with correctly determined parameters, different approaches have been used to project them onto the whole dataset. Currently, the most widely used approaches are *The Cannon* [103, 104] and Payne [105], which fall into the category of generative approaches. Such an approach, in order to infer parameters of the analysed spectrum, internally generates a model spectrum that is compared to observations. Another large group of parameter determination machine learning approaches are regression algorithms that directly infer parameters defined by the training set. Of many existing regression approaches, currently, the most explored is the use of various neural network architectures [106, 107, 108, 109]. The usage ranges from deep fully connected networks and convolutional networks to autoencoder structures that first extract latent spectral features which are later used in the regression procedure. As both methodologies do not include any knowledge about stellar physics and are trained on a limited set of data, usually much smaller than the size of the parameters space in Galaxy, their results could also be misleading. Questionable results can be returned in the case of extrapolation when observed spectrum has parameters outside of the training grid or when an algorithm is trained on an element that is not present in the used spectral range. Of course, an algorithm could find some average correlations of an unobservable feature with other features, but this has no underlying basis with observations or physics and could, therefore, give us wrong results.

Moving to chemical tagging experiments, that builds on top of previously determined stellar parameters and abundances, a shift from supervised to unsupervised machine learning algorithms is sensible as we rarely know the desired result of a conducted experiment. In the literature, chemical tagging experiments often rely on existing unsupervised clustering methods such as K-means [68, 73, 110], DBSCAN [111], t-SNE [8, 70], hierarchical clustering [112, 113], and other custom methodologies [78]. Results of those algorithms are not some numerical values, but groups of data points with similar input parameters, such as abundances in the case of chemical tagging.

For some other surveys, such as *Gaia*, users do not have full access to raw observations but have to trust published stellar properties and their warning signs – such as uncertainties and quality flags – to filter out published data before performing any machine learning operations. Understanding of used data and proper treatment of warning flags of uncertain data is therefore essential before blindly applying any machine learning algorithm to an unfamiliar data set. Failing to do so, we can get stuck in a process that is in computer science commonly referred to as “garbage

in, garbage out". The phrase emphases use of clean input training data as it is directly transferred upon investigated test set. Many attempts have been made to filter out *Gaia* kinematic and astrometric parameters effectively, but so far the most commonly used and recommended parameter is Renormalised Unit Weight Error (RUWE) [114] that is used in numerous explorations of the *Gaia* DR2.

Various machine learning approaches have been so far applied to the latest *Gaia* DR2 dataset, among other trying to classify variable stars [115], determine stellar effective temperature [116], catalogue young stellar objects [117], study streams and moving groups in the galactic halo [118, 119], identify accreated stars and structures [120, 121], track hypervelocity stars [122], delineate between stellar and extragalactic sources [123, 124], determine interstellar extinction rates [125], and define new open stellar clusters [126].

1.5 Our exploration of observed data

Our research into open stellar clusters and spectroscopically peculiar stars consists of three related subtopics that are further explained in the following chapters: analysis of possible runaway stars that were ejected from their birth clusters (based on *Gaia* kinematic and positional measurements and the GALAH spectroscopic parameters and abundances), identification of chemically peculiar and emission stellar spectra (based on the GALAH spectroscopic data), and investigation of spectroscopic solar twin stars and their multiplicity (based on multiple photometric surveys, *Gaia* distances, and the GALAH spectroscopic data).

Investigated subtopics share the observational data, but have vastly different physical question with the common goal of understanding current chemical composition and structure of Galaxy. Open clusters that are thought to be chemically the most homogeneous structures can finally be analysed in dept to uncover whether this is true. Additionally, we could find possible sources of intra chemical enrichment by evolved stars or engulfed planets. Determination of precise chemical composition is demanding, especially when observing stars that do not behave like the majority of the population. As we want to clean the GALAH dataset, we focused on finding carbon enriched and emission stars that could endanger the chemical analysis. With limited observing time and capability, the whole sky can not be thoroughly observed by only one survey. Therefore we need to combine and equalise results from multiple surveys. Standard way for inter-calibration is using solar-like stars as Sun is the best observed and studied star with precisely known parameters and composition.

1.5.1 Open clusters and ejected stars

To study ejected stars, we built upon available research that already defined membership possibilities and cluster parameters for more than 1000 open clusters in the Galaxy [43] using the latest *Gaia* DR2 data. The refined membership probabilities of previously known open clusters [127] were redefined using improved positional and kinematic information. To further define the best cluster members, we additionally used radial velocities to sift out outlying members and multiple stars. With the initial members of a cluster in place, we can continue with the analysis on ejected stars located far away from their cluster centre. Observed 6D positional and kinematics vector for every star enables simulating their position in the Galaxy

at different epochs. By integrating those properties, we can model the evolution of stellar clusters and their dissolution [18]. It has already been shown that it is possible to retrace some of the nearest runaway stars back to their original clusters using older Hipparcos astrometric observations [128].

To perform a similar task, we used the latest *Gaia* astrometric measurements supplemented with radial velocities measured by the GALAH spectroscopic survey. Their observations were done for fainter sources that are inaccessible for the spectrograph onboard the spacecraft. We integrate the movement of stars inside and outside a cluster backwards in time to determine potential points in time where their orbits around the centre of the Galaxy intersect. Those intersections give us a list of candidates that were once members of a cluster.

The prime focus of this open cluster exploration was the determination if machine learning approaches can be readily used for blind chemical tagging of known stellar structures. With this task in mind, we first determined homogeneity and trends of individual chemical abundances for clusters observed in the GALAH survey. As every cluster is surrounded with numerous unrelated field stars, we explored if the chemical signature of a cluster is any different from its surrounding. More significant the difference among both chemical signatures, easier the chemical tagging is. In the case of very similar chemical compositions, kinematic information of stars gives us and additional information, and sometimes the only one, that can help with the delineation among field, cluster and ejected stars. An in-depth explanation of our work and results is given in Chapter 3.

1.5.2 Spectroscopically peculiar stars

Every large, unbiased spectroscopic observational set is prone to target some peculiar stars whose spectrum does not resemble the majority of so-called normal spectra. One of the ways to produce those spectra is using simulations that try to reproduce the complete physics of the stellar interior [129]. Given the complexity of those computations, different approximations and simplifications are made. As this might introduce unwanted differences towards typically observed spectra, we prefer to relay on observations itself to produce a set of most common normal-looking spectra. The GALAH and other similarly vast spectral sets have enough diversity and number of of observations to produce normal-looking spectra in such manner.

During our initial search trough the GALAH spectral dataset [84], we confirmed that peculiar spectra could also be found in our acquired data. For a more consistent search of those peculiarities, we employed a multitude of supervised and unsupervised machine learning techniques to broaden the search onto a complete set of spectra. In it, we browsed for chemically peculiar spectra and spectra with pronounced emission lines. All of those special spectral types need to be identified as thoroughly as possible because they can be interesting for further studies on the one hand and might be difficult to correctly determine their physical parameters and abundances using automated pipelines on the other hand.

The supervised search for peculiar stars was based on the generation of normal reference spectra without any peculiarities that were compared with observed spectra. Any mismatch at the predetermined expected wavelengths was thoroughly measured and analysed in order to extract some physically meaningful explanations about the observed peculiarity. Reference spectra were in our case constructed using

two different methodologies. During the supervised construction, we compared observed spectra towards the median spectrum of stars with similar physical properties. The more complicated unsupervised methodology performed spectrum generation using neural network autoencoder that extracts the most important latent features from a spectrum and reconstructs it using those few latent features. A result of this spectral compression and un-compression was a peculiarity free reference spectrum.

For the same task of classifying peculiar stars, we also applied unsupervised clustering technique t-SNE [130] to normalised spectra. The algorithm treats individual spectra as vectors of multiple features of very high dimensionality whose complexity can not be perceived or visualised by humans. To convert this multitude of dimensions into a visually manageable form, the algorithm first computes similarities between all those vectors and groups spectra based on their similarity. The final result is a 2D or 3D map of points. In such a map, it is easy to select denser groups of data points and investigate if all selected spectra have a peculiarity that we were looking for. Further explanations of our search for peculiar spectra and results are given in Chapters 4 and 5.

1.5.3 Spectroscopic solar twins and their multiplicity

Among all, the best-studied chemically peculiar star is our own Sun. When compared to spectroscopically and/or photometrically similar stars, also named solar twins (as defined by Adibekyan *et al.* [131]), it shows signs of under-abundance of volatile chemical elements [132] that may hint to a formation of a solar system around Sun. When searching for solar twins, we do not consider only the chemical composition of stars, but also their physical parameters. For us, it was interesting to see where in the Galaxy we could find solar twins observed by the GALAH survey. Especially interesting are solar twins embedded in open clusters for which we can study their possible differences in composition towards their birth open cluster. Studies revealed that old open cluster M67 (also observed in GALAH) seems to contain several possible solar twins [133, 134].

Solar twins are also interesting for the following reasons. From their observed luminosity and known luminosity of the Sun, we can accurately determine their distance. Underabundace of volatile chemical elements in solar twins can be correlated with the presence of planetary systems around selected stars. The fraction of studied stars in the GALAH survey was studied by finished K2 [135] and ongoing TESS [136] spaceborne missions that are searching for planets around other stars. Matching known stars with planetary systems with our GALAH sample might reveal abundance patterns that can hint at the presence of rocky or gaseous planets.

To search for solar twins, we used raw spectra and compared them to the reference solar spectrum, that was constructed by averaging multiple acquired twilight flats. The selection of the best candidates was based on similarity metrics, where we selected only a few percents of the best matching spectra. Close inspection of absolute magnitudes (observed magnitude corrected for stellar distance) showed that some of the stars in the selection looks too bright for their spectral type and distance. To investigate possible physical scenarios that would reproduce the observations, we build a spectroscopic and a photometric model of a single star. Combined flux and spectrum of multiple single stars revealed that observed objects might contain multiple stellar sources that are slowly revolving around their common mass centre.

Further details about the search for solar twins and their multiplicity modelling is given in Chapter 6.

Chapter 2

Spectroscopic, photometric, and astrometric surveys

In the last few decades, we are witnessing fast and numerous shifts from dedicated single object observations to massive all-sky surveys producing hundreds or even thousands of unbiased observations in a single telescope pointing. Along with the complexity of data acquisition and storage, new challenges and problems involving data reduction arose, requiring dedicated computer power to reduce acquired data. The reduction challenges span from timely, almost real-time reduction requirements, to complex, computationally demanding processes that try to take into account as many telescopic- and observations-induced biases as possible. Some of those processes will be discussed in the following sections discussing specific surveys.

This thesis shows a few cases of synergies of such vastly different data sets and at the same time points to a necessity of having knowledge about the automatic processing pipelines that produced final products – something that is quite often blindly overlooked by users.

The primary sources of data for our studies are the following three stellar surveys producing information about the stars' brightness, composition, distance, kinematics and many additional parameters that can be inferred from the observed quantities.

2.1 *Gaia* space mission

Gaia is the one-billion-star surveyor of the European Space Agency (ESA). It has been continuously scanning the sky since July 2014 from its location close to the second Lagrange point of the Sun-Earth/Moon system. *Gaia* aims to map the entire sky, down to magnitude ~ 20.7, and to collect micro-arcsecond-level astrometry and milli-magnitude-level photometry for the brightest billion stars as well as medium-resolution spectroscopy for mainly radial-velocity determination of the brightest subset of ~ 150 million objects.

The *Gaia* scanning of the sky is composed of two independent, superimposed motions: rotation around the spacecraft spin axis with a period of 6 hours plus a slow, 63 day period precession of the spin axis around the Solar direction at a fixed Solar-aspect angle of 45°. Over the nominal five-year mission, *Gaia* has completed 29 of these precession periods, leading to an optimally uniform sky coverage with, on average, ~ 70 astrometric and photometric transits across the focal plane (and ~ 40 for the spectroscopic instrument). In the extended mission phase that started

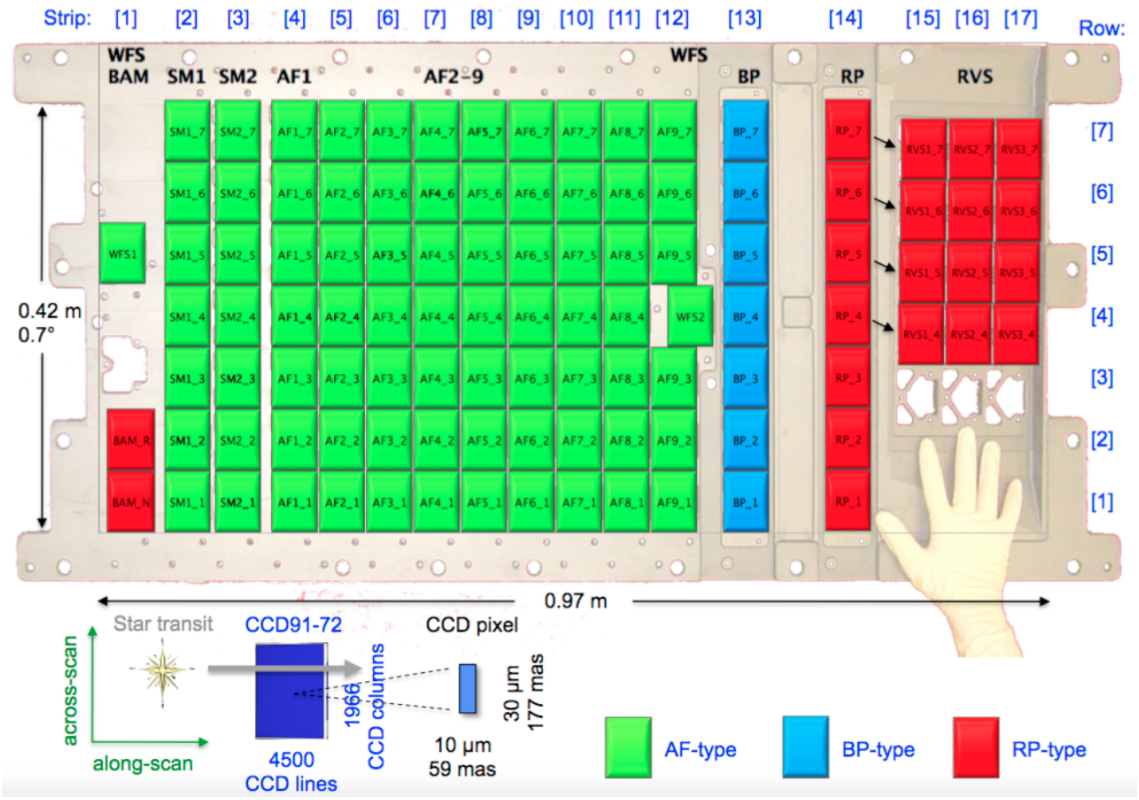

Figure 2.1: Size comparison and spatial arrangement of the CCD detectors in the *Gaia* focal plane. Image credit: Gaia Collaboration *et al.* [91].

Table 2.1: Past and future predicted release dates of *Gaia* data and products.

Release designation	Date
Gaia DR1	14 September 2016
Gaia DR2	25 April 2018
Gaia EDR3	2^{nd} half of 2020 or later
Gaia DR3	postponed into late 2021 or later
Final release	not yet determined

in July 2019, a similar scanning law is being employed but with a reversed precession direction during the first year. Passage of a star through the focal plane is called a field-of-view transit. During each transit, *Gaia* collects instantaneous, so-called epoch data of each object. Publication of all epoch data is scheduled (see Table 2.1 for the final data release.

As the spacecraft slowly rotates, observed stars traverse the *Gaia* focal plane equipped with 106 CCD detectors (shown in Figure 2.1). Every star that gets observed therefore passes through a sequence of detectors which analyse a star and determine its properties in the given order: precise position on the sky, spectral energy distribution, and its medium-resolution spectrum in a narrow pass-band.

2.1.1 Photometry and astrometry

The first array of CCDs that collects light from stars is a Sky Mapper (SM) that autonomously detects objects. Stars brighter than magnitude ~ 3 are too bright to be detected automatically. The faint detection threshold is set at 20.7 magnitude in its broadband photometric system. The detection threshold is not sharp due to on-the-fly magnitude estimation errors of the simplified on-board software.

After the source detection, stars pass into the most extensive array of CCDs that is attributed to the Astrometric Field (AF). It collects the instantaneous positions and fluxes of all objects detected by the Sky Mapper as they traverse along the field. Astrometric measurements are made in a white-light bandpass that coves the range from 3300 to 10,500 Å. The band is also referred to as the *Gaia* G band.

The last spectrophotometric measurements are then performed by two low dispersion detectors that are measuring precise fluxes in a number of narrow-pass subbands of previously mentioned wide-pass G band. The Blue Photometer (BP) collects low-resolution spectra of all objects over the wavelength range from 3300 to 6800 Å. The integrated magnitude is referred to as the G_{BP} or BP magnitude. The Red Photometer (RP) collects low-resolution spectra of all objects over the wavelength range from 6300 to 10,500 Å. The integrated magnitude is referred to as the G_{RP} or RP magnitude. Both, BP and RP low-resolution spectra, are on-board binned on-chip over 12 pixels to form one-dimensional spectra that are planed to be available to users in the third data release.

2.1.2 Spectroscopy

A final measurement performed by the spacecraft is spectroscopy over the whole observed field of the sky. The integral-field Radial Velocity Spectrometer (RVS) [137] collects medium-resolution spectra (spectral resolving power (R) $\sim 11,700$) over the wavelength range from 8450 to 8720 Å, for all objects brighter than magnitude ~ 16 in this bandpass. The location of the pass-band is selected to cover the single ionised calcium (CaII) triplet with prominent absorption features over a large temperature range of stars. Therefore, it can be used to determine radial velocity of spectroscopically diverse stars. The integrated magnitude in the RVS bandpass is referred to as the G_{RVS} magnitude. The RVS has a reduced field of view across the scan direction such that fewer observations, up to the RVS limiting magnitude, are collected compared to photometric fields in a ratio of 4:7.

2.2 *Gaia* DR2

The second data release of *Gaia* data (*Gaia* DR2) was heavily used during the production of this thesis; therefore, it requires a detailed description of the provided tables, its use and potential problems. The DR2 set is far from complete or similar to the final release, but on the other hand, provides an unprecedented set of homogeneously acquired and reduced stellar information newer seen before. Visual and numerical representation of the specific stellar products is given in Figure 2.2. *Gaia* DR2 is based on data collected by the spacecraft between 25^{th} July 2014 and 23^{rd} May 2016, the span of 22 months.

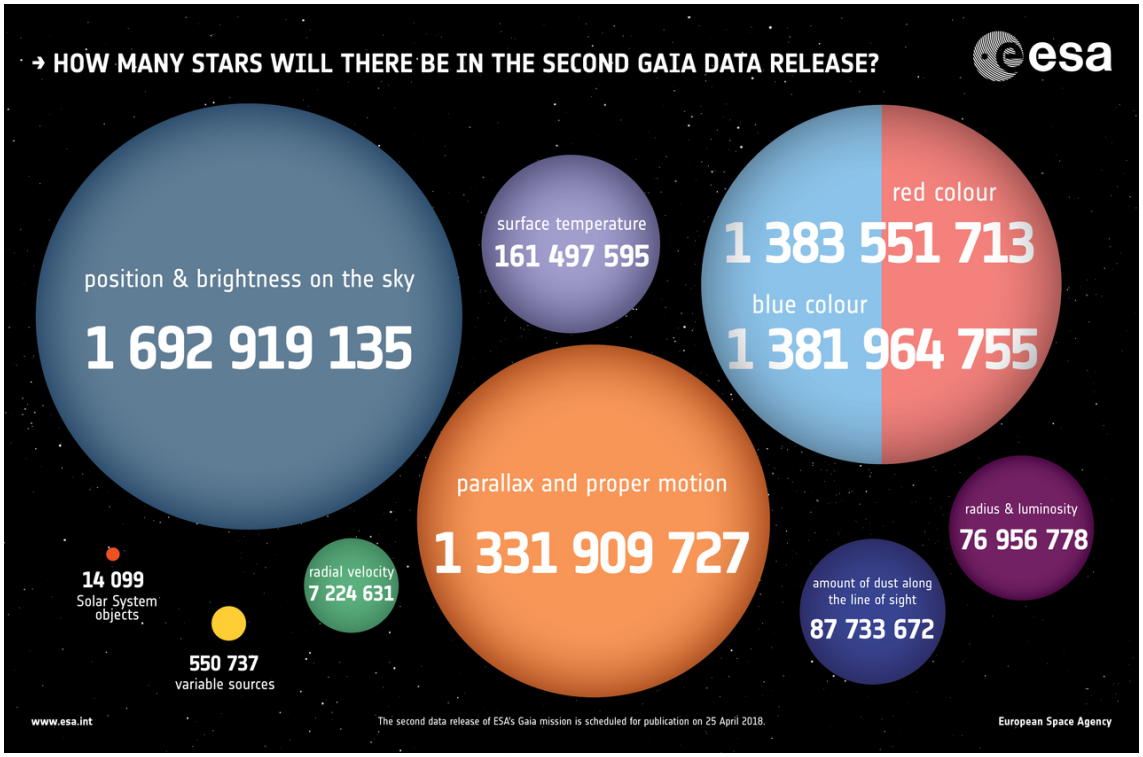

Figure 2.2: Visual and numerical representation of *Gaia* DR2 stellar content. Image credit: ESA, CC BY-SA 3.0 IGO.

Along with the calibrated raw measurements, Gaia Data Processing and Analysis Consortium (DPAC) provides numerous parameters and properties of stars, pre-selected Solar system bodies and quasars that can be used and explored by users. Among them we can find:

- **Astrometric set** consisting of partial 2-parameter (limited to celestial positions α and δ) and full 5-parameter astrometric solution with an addition of parallax, and proper motion in both directions. The 2-parameter sources are typically faint (with about half of them at magnitude G $>$ 20.6), have very few observations (less than five, as required for full solution), or very poorly fit five-parameter astrometric model. All sources fainter than magnitude G $=$ 21 have only positional information. In the current data release, all stars are still treated as single during the astrometric fit, which could significantly influence solutions in a case of multiple fast-moving sources contributing to the position of an observed photocentre. The *Gaia* coordinate reference frame is aligned with the International Celestial Reference Frame (ICRF). The aliment is performed by using positions of extragalactic sources (such as quasars) as common fixed points on the coordinate grid [138]. After the initial release, it was determined that the quality of an astrometric fit is best described by the astrometric re-normalised unit weight error (RUWE) [114]. The astrometric quality indicator RUWE is computed from the astrometric chi-squared goodness of fit and number of used observations that are both given in the published data table. For the metric to be comparable among stars of different absolute magnitude and colour, it was normalised by a factor given in the additional gridded data set. Normalising factors were produced by fitting a

quadratic spline function to its non-normalised unit weight error (UWE) that depend on the stellar type and observed brightnesses. The most typical value of RUWE is 1 and can be as large as 50 for extreme cases. Further details about astrometric processing and validation are given in Lindegren *et al.* [139] and Luri *et al.* [140].

- **Photometric data set** contains the broadband photometry in the G, G_{BP}, and G_{RP} bands, giving us colour information for *Gaia* DR2 sources that were observed at least twice. The mean value of the G-band fluxes is reported for all sources while colour information (BP and RP) is available for about 80% of them. The integrated colour information suffers from strong systematic effects at the faint end of the survey ($G > 19$), in crowded regions, and near bright stars. In the case when measured fluxes are inconsistent between the G and the G_{BP} and G_{RP} bands (sum of the latter two is significantly larger than G measurement) a warning is raised. A quantitative index of this effect is provided in the numerical form as *flux excess factor*. Further details about processing and photometry validation are given in Evans *et al.* [141] and Riello *et al.* [142].

- **Radial velocity measurements** indicate stellar median radial velocity, averaged over the first 22 months of the observations. Distinct double-lined spectroscopic binaries were taken out from the list and will be published in later data releases. This filtering, to some degree, limits the possibility of finding multiple stars among observations. Some indication that a star could be part of a multiple system is unusually large velocity uncertainty. *Gaia* radial velocities are provided for sources which are brighter than magnitude 12 in the G_{RVS} photometric band. A limited temperature range of the spectral template set which was used for the radial velocity determination implies that velocities are reported only for stars with effective temperatures in the range between 3550 and 6900 K (referring to the effective temperature of a used template and not an actual effective temperature of a star). By the RVS pipeline design, the determined absolute value of radial velocity is limited to within 1000 $\mathrm{km\,s^{-1}}$. The precision of the radial velocities at the faint end depend on stellar effective temperature and range from 1.4 $\mathrm{km\,s^{-1}}$ for cooler to 3.6 $\mathrm{km\,s^{-1}}$ for hotter stars. Those values indicate the typical spread of individual measurements during the observing period. The zero-point of the RVS velocities was determined using a comprehensive set of 4813 standard stars and asteroids with numerous dedicated, precise and temporally spread radial velocity measurements made at different observatories [143]. Further processing details of RVS data are given in Sartoretti *et al.* [144].

- **Stellar variability data set** consists of sources that were identified as photometrically variable (based on at least two photometric observations of the two *Gaia* telescopes). The final number still represents only a small subset of the total amount of variables expected in the *Gaia* survey. The sources were classified into the following nine categories based on their light curves: RR Lyrae (anomalous RRd, RRd, RRab, RRc), long-period variables (Mira type and Semi-Regulars), Cepheids (anomalous Cepheids, classical Cepheids, type-II Cepheids), δ Scuti and SX Phoenicis stars. If a star had 12 or more ob-

servations, its light curve was analysed in greater detail. They are designated as specific object studies (SOS) and consist of variables of the type Cepheid and RR Lyrae, long-period variables, short time scale variables, and rotational modulation variables. Full details on the variable star processing, results and their validation are given in Holl *et al.* [145], Mowlavi *et al.* [146], Molnár *et al.* [147], and Clementini *et al.* [148].

- **Astrophysical parameters** derived by the astrophysical parameter inference system in the *Gaia* data processing (Apsis) include estimates of effective temperature T_{eff}, extinction A_G and reddening $E(G_{BP} - G_{RP})$, radius, and luminosity for stars brighter than magnitude G = 17. Values of T_{eff} are reported only in the temperature range between 3000 and 10,000 K Limits are induced by the training set of the algorithm responsible for the T_{eff} estimation. Estimates of the other astrophysical parameters are published for about half of the sources with determined T_{eff}. As the processing pipeline was run individually for every object and with a limited set of input data (three *Gaia* photometric bands and parallax), some errors are expected because of high degeneracy between determined parameters. If a star is located far from expected isochrones used in the processing, extinction becomes overestimated and less usable. Full details of the astrophysical parameter processing and result validation are described in Andrae *et al.* [149].

- **Solar system objects** (SSO) data set provides epoch astrometry and unfiltered G photometry for a pre-selected list of 14,099 known minor bodies in the solar system that are numbered in the Minor Planet Center repository. Each time a given SSO enters the field of view of *Gaia* telescopes, celestial positions are recorded as seen from the spacecraft. The data set and its production are thoroughly described in Gaia Collaboration [150].

The above sections are partially adapted and summarized from Gaia Collaboration *et al.* [42, 91], and Gaia Helpdesk [151].

2.3 The GALAH survey

The GALactic Archaeology with HERMES (GALAH, [152]) is an ongoing spectroscopic survey that aims to unveil the Milky Way's formation history by studying the detailed chemical composition of observed stars. As already explained in Chapter 1, the complete Galaxy did not form at once, but gradually over time. This formation history can be traced back by precisely determining the chemical composition of stars in its different regions. Remnants of initial building blocks, which have been disrupted during the formation and evolution, are now dispersed throughout the Galaxy. Chemical tagging theory [66] shows that those individual galactic components should have conserved their original chemical signature over time. It is therefore essential to disentangle their formation location and migration history in order to explain the current mixture of stellar populations. This component disentanglement can be achieved through the technique of chemical tagging that promises identification of old dispersed fossil remnants based on their unique abundance patterns over numerous chemical elements. The GALAH aims to achieve this by measuring up to 31 elemental abundances in every acquired spectrum. The

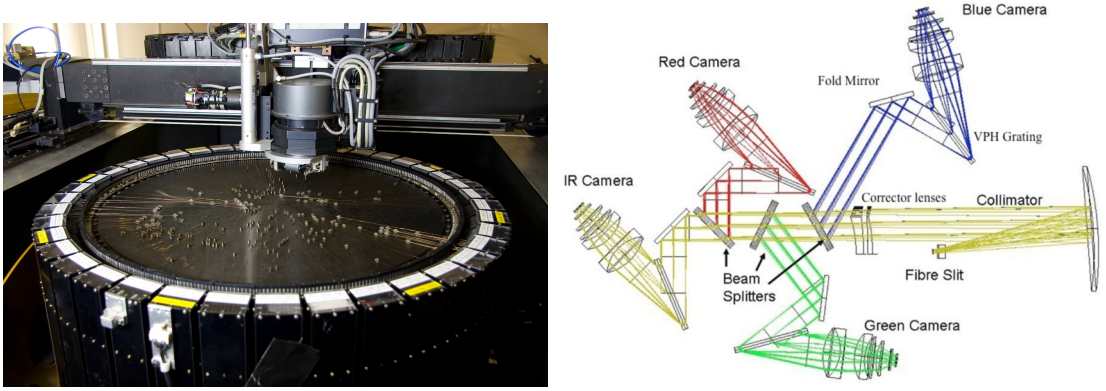

Figure 2.3: Crucial parts of the HERMES spectrograph are 2dF fiber positioner (left image) and four individual spectrograph arms.

observed elements come from seven independent major element groups with the different physical nucleosynthetic origin, which include light proton-capture elements Li, C, O; α-elements Mg, Si, Ca, Ti; odd-Z elements Na, Al, K; iron-peak elements Sc,V, Cr, Mn, Fe, Co, Ni, Cu, Zn; light and heavy slow neutron capture elements Rb, Sr, Y, Zr, Ba, La; and rapid neutron capture elements Ru, Ce, Nd and Eu [152].

The GALAH survey was the main driver for the construction of the High Efficiency and Resolution Multi-Element Spectrograph (HERMES, [153, 154]), a multi-fibre spectrograph mounted on the 3.9-metre Anglo-Australian Telescope (AAT) at the Siding Spring Observatory, Australia. The spectrograph has a resolving power of R $\sim 28,000$ (or R $\sim 45,000$ when slit mask is used) and covers four wavelength ranges (4713 – 4903 Å, 5648 – 5873 Å, 6478 – 6737 Å, and 7585 – 7887 Å), together covering approximately 1000 Å, including the Hα and Hβ lines. The ranges are frequently referred to as blue, green, red and near-infrared spectral arms (see left image of Figure 2.3). This configuration can simultaneously record spectra from up to 392 fibres distributed over a 2° diameter field of the night sky, with an additional 8 fibre bundles used for telescope guiding. The fiber positioner has two identical plates that are used to precisely position fibers at designated stellar locations. During the exposure with the first plate, the robotic positioner is placing fibers on the second plate as shown in Figure 2.3. The complete fiber allocation process takes 45 min per plate. The spectrograph can typically achieve a signal to noise ratio (SNR) ~ 100 per resolution element at magnitude V=14 in the red arm during an 1-hour long exposure. To achieve as high SNR as possible and minimise atmospheric diffraction, all observations are ideally done when an observed field is passing through the meridian. With a fibers' finite field of view of 2″, we also want the object to be as high as possible on the sky. Observing at low altitudes could diffract different wavelengths over an angle that is larger than the fiber entrance point. Something that is not desired as we would loose input flux in one of the spectrograph arms, further reducing quality of spectra in that wavelength region.

2.3.1 Acquired spectra and target selection

The spectroscopic data used during the production of this thesis were taken from the pilot survey, the main GALAH survey [152], the K2-HERMES survey [155], the TESS-HERMES survey [156], and specially dedicated HEMRES open clusters

program (De Silva et al. in preparation) and the HERMES Orion star-forming region program (Kos et al. in preparation) surveys. Together they form a dataset of 669,845 successfully reduced stellar spectra, of which a small fraction belongs to repeated observations. All acquired spectra are homogeneously reduced to one-dimensional spectrum, normalised and shifted to the stellar velocity frame (detailed description in Kos *et al.* [100]). Combination of those surveys produces an increased number of spectra compared to the main GALAH survey. However, at the same time, this breaks the rule of a simple, unbiased selection function (Sharma et al. in preparation) that is desired for population studies and easier comparison with synthetic galactic models.

The original selection function of the main GALAH survey is separated into two magnitude limited field selections - bright ($10<V<12$) and regular ($12<V<14$) fields whose target selection is colour independent. Used V magnitude is inferred from magnitudes measured by the Two Micron All-Sky Survey (2MASS, [157]) whose photometric bands are shifted into the infra-red spectral region. Because of that, some, especially peculiar and variable stars, might have an erroneous estimation of V magnitude leading to underexposure or excessive spectral crosstalk. Because of expected crowding problems (projected diameter of used optical fibre on the sky is equal to $2''$) observed stars are located at higher Galactic latitudes ($|b| > 10°$) where the density of stars is lower. Additional surveys sometimes break those rules by selecting fainter/dimmer stars, going closer to the Galactic plane, employ colour cuts, or favour interesting preselected stars such as K2 [135] targets, TESS [158] targets and cluster members. Therefore some care is needed when trying to infer global stellar or galactic properties based on such inhomogeneous selection criteria.

2.3.2 Spectral reduction and parameters determination

The first step after recording spectra in the form of a 2D image is their extraction to multiple 1D spectra. The procedure, extensively documented by Kos *et al.* [100], consist of the following steps: raw image cosmetic corrections, spectral tracing, optical aberrations correction, scattered light and apertures cross-talk removal, wavelength calibration, sky subtraction, and telluric absorption removal. After reduction, spectra are normalised and shifted into their rest-frame by cross-correlating them with a set of 15 AMBRE model spectra [159]. This procedure produces initial velocities that are usable for parameter determination, but not accurate enough for possible inter-cluster analysis. The more precise methodology uses the GALAH observations itself to compute a much denser set of reference spectra for cross-correlation. Its details are described by Zwitter *et al.* [160]. In the last step, the methodology also accounts for gravitational redshift and convective blueshift to determine actual stellar radial velocity.

Stellar atmospheric parameters and individual elemental abundances are in a uniform way derived from all normalised spectra. The applied parameter analysis pipeline slowly evolved and improved during the GALAH survey. The three most important milestones in an ongoing process are:

- The initial stellar parameters (T_{eff}, $\log g$, and [Fe/H]), accompanying the first GALAH data release (**GALAH DR1**), were derived as a global fit (all arms at the same time) of the observed spectra to a grid of 16,783 AMBRE spectra [159], which were convolved down to the average resolution of individual

HERMES CCD. The aim of this procedure was to provide indicative stellar parameters, which could be used as a first initial guess to help speed up more complex stellar parameters and abundance pipelines.

- The parameters (with extension to $v \sin i$, v_{mic}, and A_{K_S}) and up to 23 elemental abundances, released as part of the **GALAH DR2** [104], were produced using the multi-step data-driven approach. The complete analysis depended on a set of 10,605 spectra that were selected in such way to span a large portion of the parameter space and did not contain any peculiar star, especially binary and emission line spectra. Selected spectra were analysed using a physics-driven spectrum synthesis code Spectroscopy Made Easy (*SME*, [161, 162]) that performs spectrum synthesis for 1D stellar atmosphere models. In the case of DR2, it consists of MARCS theoretical 1D hydrostatic models [129] under the assumption of local thermodynamic equilibrium (LTE) for the majority of elements. The procure is computationally efficient but does not completely describe the physics inside stars. Therefore several vital elements (Li, O, Na, Mg, Al, Si, and Fe) were analysed using more realistic non-LTE line formation. A common approach to compute non-LTE abundances is to use the same pipeline as for LTE but internally account for the non-LTE departure coefficients that were in advanced computed using complex non-LTE computations that account for realistic atom collisions. Such departure coefficients are usually computed for a limited grid of stellar parameters and afterwards interpolated in between when need [163, 164, 165].

  To propagate the parameter and abundance results of the training set to the whole survey, *The Cannon* [103] generative data-driven approach was used. Here a crucial step is the preparation of the training set as it defines the quality of the employed machine learning procedure and consequently, quality and precision of the results. The training set was selected to span the stellar parameter ranges covered by stellar spectra completely. Therefore it mostly consists of well analysed and explored stars, with many published results. Additionally, the set was cleared of any peculiarities (binary stars, fast rotators, reduction issues, etc.) that would introduce uncertainties into the learned model. Internally, *The Cannon* procedure adopts a simple quadratic model which uses stellar parameters to describe the observed flux of a given spectrum. Independent quadratic model is built for every spectrum wavelength bin. Flux at that bin is computed as a weighted linear combination of all possible two-parameter product combinations of input stellar parameters. To train *The Cannon* model, all the GALAH spectra in the training set were interpolated to a common wavelength grid. After the training was performed, the model was inverted to produce parameters and abundances for every observed spectrum. When a spectrum is introduced to an inverted *The Cannon* model, the latter tries to find the best matching stellar parameters. At every fitting step, *The Cannon* generates a learnt synthetic spectrum that is compared to the input spectrum. By minimising the difference between the generated and input spectrum, it tries to find the best matching parameters. Further details of the described process are given in Buder *et al.* [104].

- Not relying on the GALAH spectral information alone, but including *Gaia* parallax, colour, and absolute magnitude, additional constraints and priors can be

used to infer spectroscopic stellar parameters. That additional information can be used to confine information about $\log g$ and stellar age. Adaption of *SME* software, thoroughly described in Buder *et al.* [166], was used to accommodate additional *Gaia* information in order to produce the latest **GALAH DR3** data set. Unlike in DR2, no data-driven methodology was used to produce stellar parameters and abundances as *SME* was run for every individual spectrum. Reference spectra were generated using MARCS 1D stellar models. For eleven abundances (out of 30 measured) non-LTE corrections were performed. The published catalogue contains several additional Value-Added-Catalogues defining stellar ages and their galactic dynamics. To compute stellar masses and ages, the best matching PARSEC+COLIBRI isochrone was used.

With the addition of so many auxiliary information with their own uncertainties, that are sometimes not even known or unable to be estimated, the question arises if they also impact the quality of the published results. As the precise parameters and abundances are the driving focus of the ongoing survey, we tried to estimate the possible trend and offsets by analysing open clusters in Chapter 3.

2.4 Asiago spectroscopic observations

Vastly different from the previous two massive all-sky surveys, telescopes at the Asiago site are mainly used for dedicated observations or monitoring of pre-selected targets whose observational and astrophysical potential was identified from all-sky surveys. During our stay at the Asiago Observatory, that usually lasted for four consecutive bright nights (close to full Moon) every month, we used the 1.82 m Copernico telescope located on top of the nearby hill Mount Ekar (Asiago, Italy - the altitude of $1,366$ m).

Our observations were performed by the Echelle spectrograph that is during bright near-full Moon nights mounted on the telescope. At that time, the quality and deepness of photometric observations are heavily reduced and therefore not performed. Design of the Echelle instrument and its slit length enables the observer to observe only one star a time. Obtained spectra have a resolving power of R$\sim$ $20,000$ and a wide span of wavelengths between 3600 and 7400 Å. They are divided into 30 interference orders which partially overlap with succeeding and preceding order, providing an undisturbed coverage of observed wavelengths. Acquired spectra were recorded by the Andor DW436-BV CCD camera. The camera uses a back-illuminated CCD detector with a size of 2048×2048 pixels. This setup enables us to capture spectra of stars with magnitudes $V < 10$ at high SNR with reasonable exposure time (less than 1 hour per spectrum). Because of the mechanical limitation, observed stars must be positioned at least 15° above the local horizon. At those low altitudes, only the brightest stars are reasonable to be observed because of strong atmospheric attenuation and diffraction differences between red and blue wavelengths. The latter effect can be reduced by rotating the slit of the spectrograph into the paralactic angle.

Combining the location of the observatory and above observational limitations with the fact that our exciting stars were in advance selected from the GALAH survey, highly reduces the number of potentially observable objects. To reduce the

Figure 2.4: Structure and dome above the Copernico telescope on top the Mount Ekar hill, Asiago, Italy.

atmospheric effects, we observed only stars which rose at least 30° above the local horizon. This limit enables observation of stars with $\delta > -15°$. As described in more detail below (see Chapters 4 and 6), we used additional Asiago observations to inspect spectroscopic features not accessible by the GALAH spectra and to extend radial velocity time series of possible multiple stars who could show signs of radial velocity changes not detectable by a single epoch GALAH spectrum.

Additionally, to our program observations, we also contributed spectroscopic observations that resulted in published astronomer's telegrams [167] and scientific papers [168].

Chapter 3

Chemo-dynamic tracing of open cluster stars

One of the main goals of the GALAH survey is to explore possibilities of chemical tagging for random field and known open cluster stars. A task that sounds easy in theory, but its working applications are far from ready for large spectroscopic surveys. The road to getting precise stellar chemical abundances leads around many different obstacles, which all have an impact on the final determined abundance values whose precision and accuracy dictates the possibility and success of implementing chemical tagging.

In this chapter, we present our exploration of abundances for a few open clusters that were observed as part of multiple different surveys served by the HERMES spectrograph. In Section 3.1, we briefly describe the history of open cluster membership, the evolution of clusters, and means to discover these ongoing processes using stellar abundance information only. Of multiple observed open cluster in the GALAH, we focus only on a small subset of them that have the highest number of members (see Section 3.2.1 for the complete list). Section 3.3 describes the selection of clusters and integration of orbits for stars inside and around the clusters. Chemical signature of field and cluster stars is analysed in Section 3.4.2 and the results are summarised and discussed at the end of this chapter.

3.1 Introduction

The latest second release of *Gaia* data (*Gaia* DR2, [42]) revolutionized numerous fields of astronomy, including research of galactic open clusters. Its combined information of stellar distance, kinematics, and photometric measurements enables us to go beyond simple methodologies, such as star density counts, to unravel even the faintest and sparsest components of open clusters. So far, many works have been published trying to refine parameters, and membership information of long known open clusters [41, 43, 44] and find new, less numerous or fainter clusters [46, 47, 48, 49]. Such thorough and the improved investigation uncovered that many of the clusters listed in modern catalogues, initially discovered as apparent stellar overdensities, are no more than chance alignments of stars and not true physical clusters [50, 51, 52, 53, 54].

Born from the same molecular cloud, open clusters are ideal test structures for different astrophysical principles. Being influenced by external and internal pro-

cesses, such as tidal stripping and close stellar interactions, their lifetime is limited from about 100 Myr to a few Gyr for the densest structures [18, 19]. This gives us a possibility of observing them at different evolutionary stages [24, 25] before they blend [26] into field stellar population. The most prominent transitional features we can observe are compact cluster tidal tails [11, 12, 13, 14] and lose extended halos of evaporated stars. They are observed as a slowly decreasing over-density [15, 16, 17] of stars far from a denser cluster core. Due to close gravitational interactions among members, they can be ejected out of a cluster at high velocities [4, 5, 6]. Such cluster members can on the sky be found several degrees or even further away from their main cluster body [7, 8, 9]. Identification of such stars could be done using a chemical tagging procedure, whose importance and potential problems have already been discuses in Section 1.2.

3.2 Additional data specifics

3.2.1 The GALAH and cluster stars

Among the dedicated HERMES cluster observations and other HERMES surveys, such as the main GALAH survey, we detected members of known open clusters, whose stellar membership was taken from results published by Cantat-Gaudin *et al.* [43]. Their clustering methodology relies on the unsupervised membership assignment code called supervised Photometric Membership Assignment in Stellar Clusters (UPMASK, [169]). Initially, the methodology was designed to find overdensities using only stellar position and photometry. As the methodology is unsupervised and has zero knowledge about physics or input data, Krone-Martins and Moitinho [169] easily applied it to work with astrometric and positional data. Internally UPMASK creates many incarnations of random values from the input data and their uncertainties, selects the four most important principal components, and runs the clustering algorithm on the components. At the end, detected overdensities are compared with random stellar fields and overdensities of the last iteration. Membership probabilities depend on how often a star was inside the relevant overdensity.

As some of the clusters were not targeted intentionally by the surveys or only their cores, the number of observed members and surrounding field stars of interest can vary substantially. The clusters analysed in this paper, having the most significant number of spectroscopic observations are Berkeley 32, NGC 2516, NGC 2112, NGC 6253, Blanco 1, Ruprecht 147, NGC 2632, NGC 2682, Melotte 22, and Collinder 261. To supplement their selection, we added members of Melotte 25 cluster whose membership selection was performed by us as it was missing in the mentioned published paper [43]. The first step of our methodology was comparable to UPMASK. Similarly, we generated many incarnations of the data to determine the kinematics of a cluster. It was computed as a median proper motion of the overdensity that was closest to the previously known centre at every iteration. After selection in proper motion space, we used the same stars to determine clusters centre in position, distance, and radial velocity. A multivariate Gaussian distribution was fitted at the determined parameter to assess membership probabilities.

3.2.2 Gaia

For a complete 6D positional and kinematics stellar information, we augmented the GALAH data with proper motion, parallax and radial velocity from the *Gaia* DR2 data-set. As all investigated open cluster stars are located close to the Sun, their distances can be inferred by the inversion of a parallax value ($1/\varpi$). Of course, we could, at this point, use a bit more accurate distances that were determined using distance priors based on a Galaxy model [170]. The second approach does give more reliable and symmetric distance uncertainties, but does not reduce the elongated shape of stellar clusters (in radial direction away from our location in the Galaxy) as the distance to every star is determined individually.

The current release of the *Gaia* data contains magnitude limited range of recovered radial velocities, that are, whenever possible, supplemented or substituted with the GALAH measurements of higher accuracy [160]. Supplemented are mostly stars fainter than currently adopted *Gaia* RVS [137] analysis threshold as the GALAH targets are much fainter stars than the RVS limit. The synergy, therefore, increases the set of useful stars in our case.

3.3 Cluster and field members

The first step in our analysis was the acquisition of data relevant for each cluster identified among the GALAH observations. Identification of observed clusters was made by matching observed stars with known cluster members published by Cantat-Gaudin *et al.* [43]. As some of the clusters had a low number of stars or were proved to be chance alignments of stars [53], they were not considered in the analysis. Sky coordinates and distances of selected open cluster members (see Section 3.2.1 for the list of considered clusters) were taken from Cantat-Gaudin *et al.* [43] and served us as anchors around which we queried the *Gaia* DR2 data. A cone query with a radius of 6° and distance limit of ±900 pc around a cluster centre was performed to download a subset of the whole dataset. The elongated shape of queried dataset volume is a result of clusters' apparent elongated shape. A bit different volume was used for a nearby and visually extensive cluster Melotte 25, for which we used radius of 12° and distance limit of ±200 pc around its centre.

The downloaded subset included stars with an incomplete set of *Gaia* parameters. To complement and improve quality of radial velocity measurements, all available GALAH velocity estimates in a subset were used to override or supplement *Gaia* measurements. In the case of multiple GALAH observations, a median velocity per star was used.

The initial open cluster memberships were taken from Cantat-Gaudin *et al.* [43] but needed some refinement before it was suitable for us. To select as many possible cluster members, the employed membership algorithm did not relay on magnitude limited radial velocity information to assign cluster membership. To make cluster volume more compact and retain only the most probable core members, we discarded all member stars whose radial velocity deviated for more than 5 $\mathrm{km\,s^{-1}}$ from the cluster median value of all retrieved members. The used threshold was determined empirically by observing velocity distributions to discard only a few of the most dissimilar stars. The reason behind this velocity limitation will be evident in Section 3.3.1 where we try to keep the cluster volume as compact as possible during its

integration. This thresholding prevents unwanted pollution of a cluster by field stars during comparison of their chemical signatures in Section 3.4.

3.3.1 Stellar tracing

After the selection of open cluster members, we proceeded with the analysis of stellar movements inside and outside the cluster. In order to get the most reliable motion information, only stars with a complete 6D kinematic information (proper motion + radial velocity + sky coordinate + parallax) were considered. No additional *Gaia* quality flagging was used to remove stars with potentially wrong parameter estimates as we would like to show in the following steps that they could be discovered and eliminated based solely on their chemical composition.

By knowing members of the observed clusters, their current position, and complete motion vectors, we can trace the path of a volume constrained by the cluster stars backwards or forwards in time. This integration procedure was performed by individual integration of cluster stars in axisymmetric gravitational potential (*MWPotential2014* potential described by Bovy [171]) using *galpy* software library version 1.5.0. [171]. Being interested in the past ejected members of a cluster, we integrated orbits of cluster stars for 120 Myr (comparable to ages of the youngest open clusters in our set) into their past and saved their location after every step of 20 kyr. As the integration process relies only on the present uncertain measurements of their velocities and distance, longer integration is not precise or reliable. This uncertainty is observed by the fact that the cluster volume gets larger during backwards integration instead of staying approximately constant as it would in the case of gravitationally bound stellar components. The volume could also keep shrinking during backwards integration if cluster is loosely bound and is already slowly dissipating at the present time. At every integration step, the cluster volume was described by a minimum convex hull defined by its outer-most members. They serve as vertex points of the constructed geometric body. Such a geometrical shape presents the smallest bounding volume with partially flat boundaries which encompasses all considered members.

The next step of our analysis consisted of finding stars in the clusters' immediate neighbourhood that could be traced back to having origin inside the considered open cluster. To filter-out field stars that travel into a completely different direction than the cluster, we discarded all stars whose galactic velocity vector difference towards present cluster velocity vector was >50 $\mathrm{km\,s^{-1}}$. This threshold, therefore, defines the fastest possible speed at which stars could have been ejected. Orbits of the remaining set of stars (usually more than half of the queried stars) were integrated using the same configuration as cluster stars described above. At this point, we could investigate which orbit of the field stars crosses clusters' volume at any given integration step.

To get a more descriptive crossing probability, we created 250 incarnations of every field star. Initial kinematic properties of each incarnation were drawn from the Gaussian distributions of parallax, radial velocity, and proper motion defined by their reported values and uncertainties. After analysing all 250 orbits of each star, we described its cluster crossing probability by the longest stay inside the cluster volume and by the percentage of crossing events. For a crossing to be counted as confirmed, a star had to be located inside the cluster volume for at least 0.4 Myr –

Table 3.1: Clusters statistics. Only stars with a complete 6D positional and kinematic information were considered for this statistics and orbit integration analysis. Numbers in columns successively present: number of all stars with complete information, number of analysed stars that meet initial criteria of having the galactic velocity similar to a cluster ($\Delta v < 50$ km s^{-1}), number of stars that do not cross cluster volume during integration, number of possibly ejected stars, and number of cluster members that defined volume of a cluster.

Cluster	Queried from *Gaia* DR2	Analysed stars	Field stars	Possibly ejected	Known members
Berkeley 32	11322	2659	2047	125	23
Blanco 1	5043	2734	2687	15	81
IC 4665	15022	10155	9823	26	34
Mamajek 4	21776	11623	10513	85	48
Melotte 22	9097	6335	5944	105	239
Melotte 25	6836	3782	3511	165	126
NGC 1817	12826	4489	4060	74	54
NGC 1901	12666	7323	7204	19	30
NGC 2112	13866	6665	6323	38	49
NGC 2204	4314	1777	1170	180	59
NGC 2516	17383	11906	11030	315	182
NGC 2548	14371	9212	8842	60	34
NGC 2632	9951	5290	4991	170	222
NGC 2682	10947	5244	4776	226	287
NGC 6253	62975	30114	17267	1362	64
Ruprecht 147	17749	5062	4850	23	103

time that is equivalent to 20 integration steps. The final selection of probable ejected stars consist of stars, whose integration procedure revealed that they were crossing a cluster volume in at least 68% of the incarnations and their longest stay there was at least 1 Myr. Remaining volume crossing stars, that did not meet the criteria, were considered as random field stars. They were also discarded from further analysis as they might, in the case they were real past cluster members, additionally pollute chemical signature of the field population. Summary of investigated and discovered stars for every cluster is given in Table 3.1.

Table 3.2: The number of acquired GALAH spectra among different cluster components that were considered during the chemical comparison. No parameter or abundance flagging was yet used at this point. Stars represented by these statistics are a subset of stars given in Table 3.1. Because of a small percentage of repeated observations, few clusters (especially NGC 2682 that was used as the calibration and verification field) can have a higher number of spectra than the number of member stars.

Cluster	Field	Ejected	Members
Berkeley 32	42	2	36
Blanco 1	151	10	109
IC 4665	1114	6	26
Mamajek 4	2907	26	13
Melotte 22	1016	21	67
Melotte 25	653	27	64
NGC 1817	569	11	36
NGC 1901	6963	19	28
NGC 2112	401	4	66
NGC 2204	129	23	58
NGC 2516	4604	101	48
NGC 2548	767	6	18
NGC 2632	2266	77	116
NGC 2682	1710	147	1000
NGC 6253	825	34	86
Ruprecht 147	798	14	175

3.4 Chemical signature of clusters

After defining potential members of different cluster components (field, ejected, and members), we can look into the abundance signatures of an individual component and their overlap. Of all 30 possible GALAH chemical abundances, we initially excluded only Li because of its intrinsic variability that depends on the stellar evolutionary stage. Scatters plots of all considered abundances and [Fe/H] as a function of $T_{\rm eff}$ for clusters most populated by the GALAH data are shown in Figures 3.1, 3.2, 3.3, and 3.4. Not all plots for the same cluster have the equal number of points as reporting of abundance values depends on the estimation of their reliable detectability that is based on equivalent widths of elemental absorption lines (thoroughly described in Buder *et al.* [166]). The plots show only stars with unflagged (`flag_sp` = 0, other flag values are described in Buder *et al.* [166]) stellar parameters, that are presumably of the highest quality.

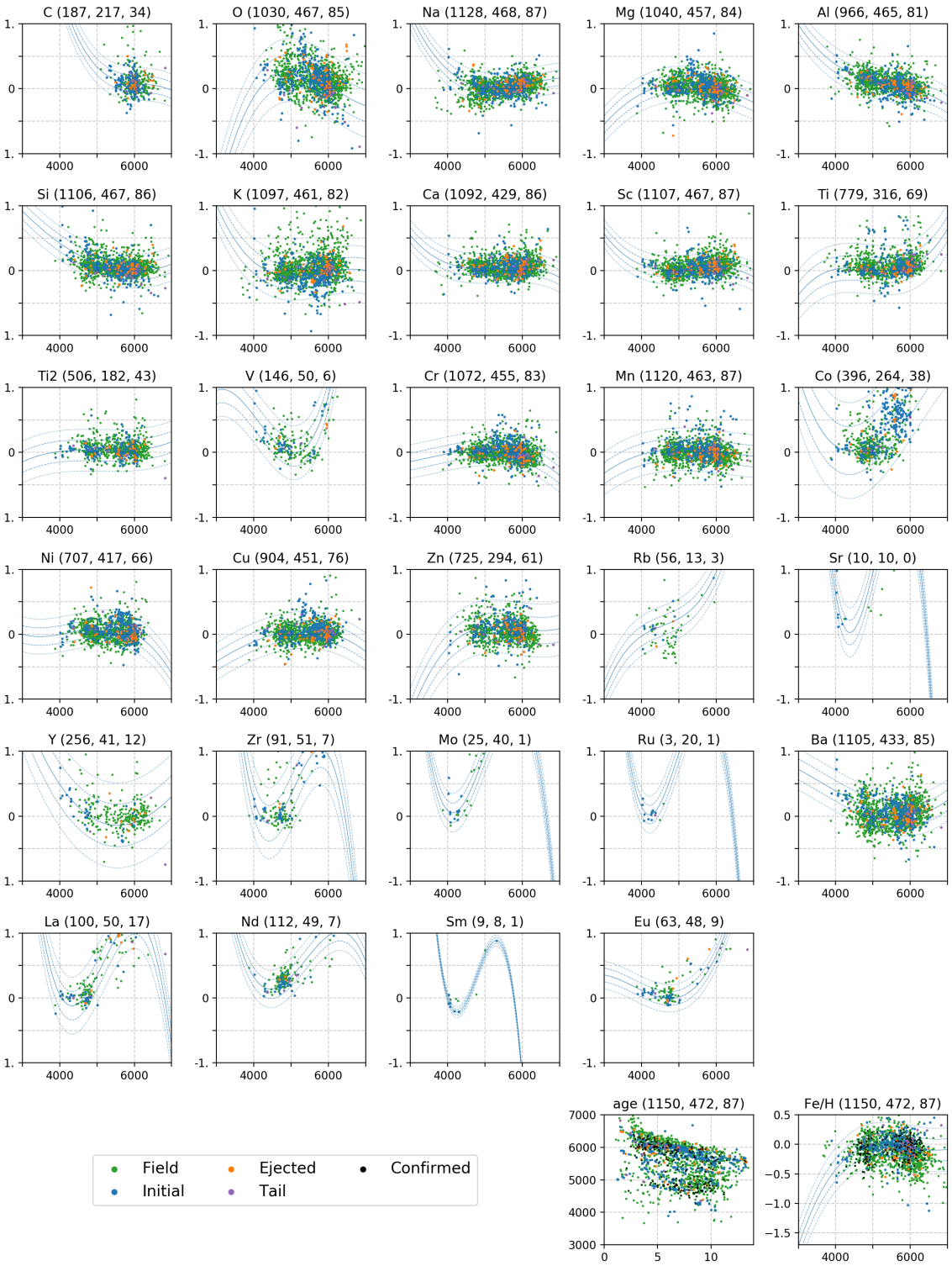

Figure 3.1: NGC 2632 abundance scatter plots as a function of effective stellar temperature. The solid blue line represents the best fit on the cluster population. The 1σ and 2σ abundance deviations from the fit are given by dashed blue lines of decreasing intensity. Coloured dots represent field (green), members (blue), and possibly ejected (orange) stars. Their numbers are given above every panel, following the elements' name. Purple dots preset known tails (described later in Section 3.5) of slowly evaporating stars in some clusters. The last two panels present $T_{\rm eff}$ of stars at different ages (in Gyr), and dependence of [Fe/H] on their $T_{\rm eff}$. The black dots in those two panels indicate possibly ejected stars whose chemistry was matched to cluster stars.

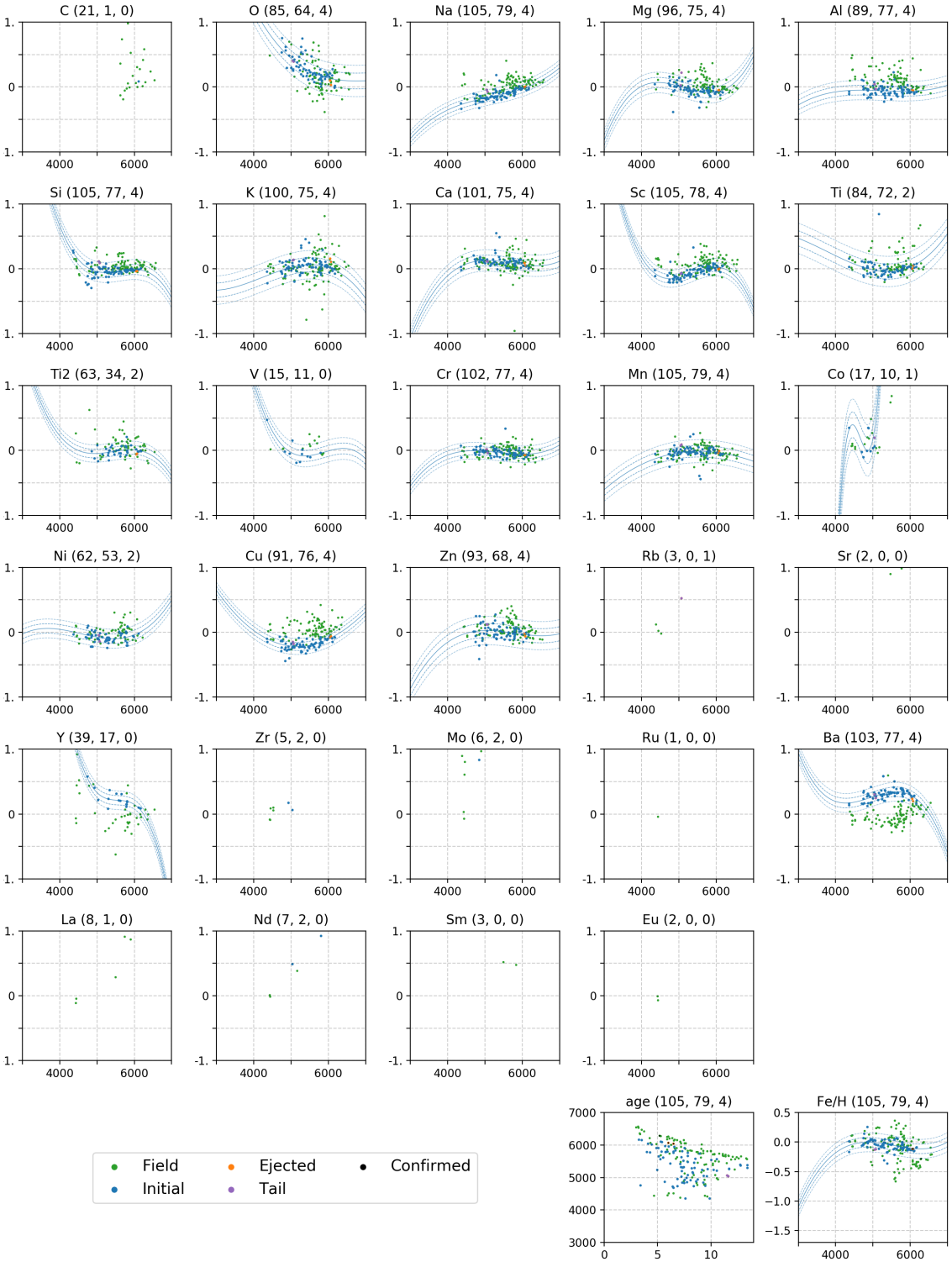

Figure 3.2: Same plots as in Figure 3.1 but for open cluster Blanco 1.

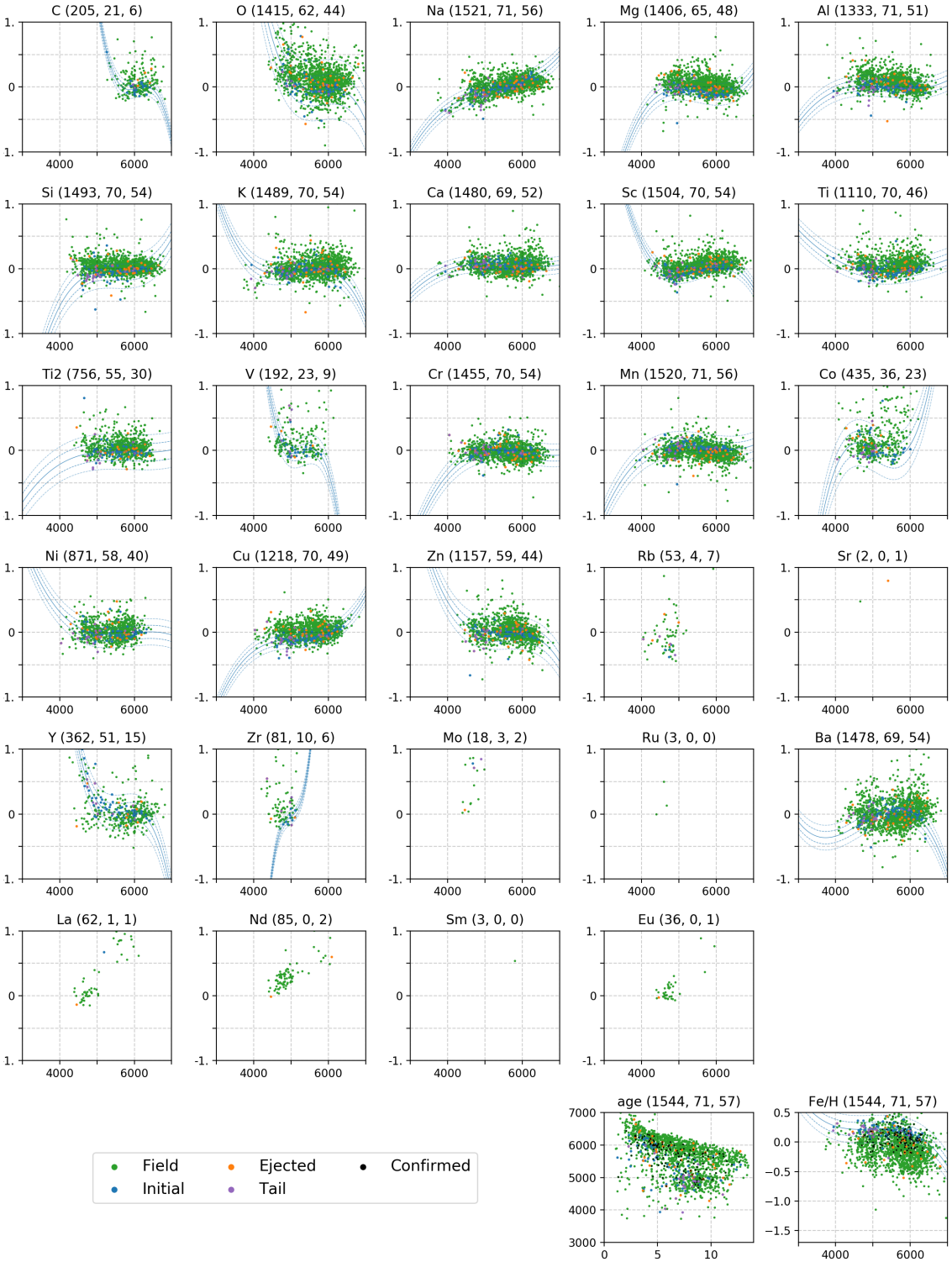

Figure 3.3: Same plots as in Figure 3.1 but for open cluster NGC 2632.

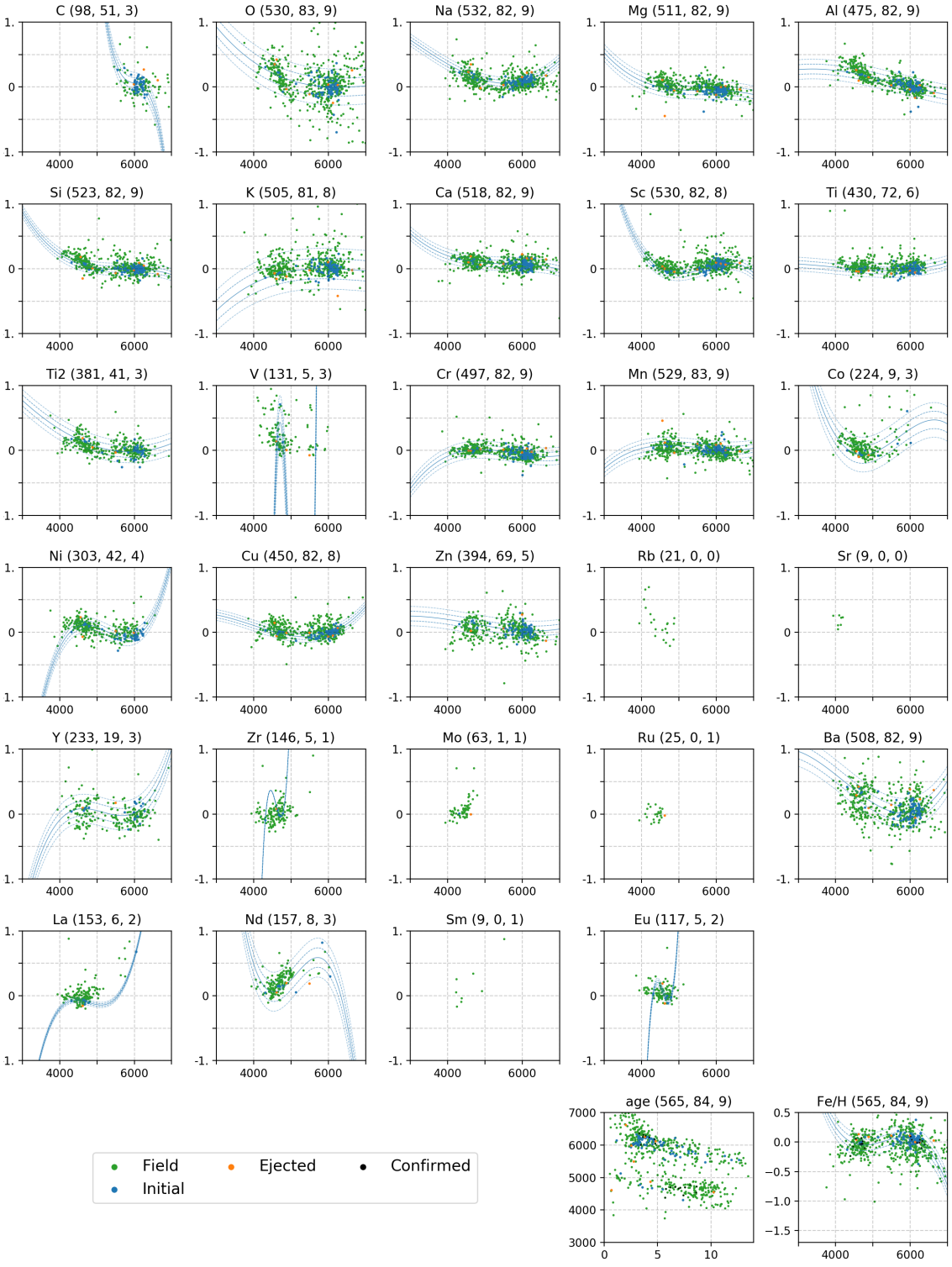

Figure 3.4: Same plots as in Figure 3.1 but for open cluster Ruprecht 147.

3.4.1 Abundance and age trends

One of the first things we noticed on the shown abundance scatter plots is their strong dependence on physical parameters, especially $T_{\rm eff}$. As this is not the first or unique observation of those trends [76, 172, 173, 174, 175], they are most likely products of insufficient/inaccurate stellar models and/or actual abundance patterns, and not induced solemnly by the employed *SME* spectrum analysis pipeline that was used by Buder *et al.* [166] to determine stellar abundances.

If we presume that observed trends are artificially induced and considered cluster members should have a homogeneous chemical composition that is independent of stellar type, a differential chemical tagging analysis can be used [76]. Such analysis considers only comparisons among stars with a similar set of stellar parameters. To describe observed trends, we independently fitted a 3^{rd} degree polynomial function using 2.5σ clipping algorithm in 2 steps to every abundance versus $T_{\rm eff}$ diagram. By subtracting fitted trends, we estimated degree of intracluster scatter for every element. Because of limitations of measuring certain abundances, the fit was not performed if the number of valid abundance measurements was lower or equal to a used polynomial degree +1.

In the ideal case, all of the trends for the same element observed in different open clusters would have the same shape and would be distinguished only by their abundance offsets that result from their distinct chemical pattern. Looking at our fitted trends, everything is not so simple and easy as the fitted curves can have significantly different shapes. For example, trends of elements Ba and Cu are in general U-shaped. Therefore the highest or lowest abundance values are present in the middle of the $T_{\rm eff}$ region. Those results imply that using the latest GALAH abundances, we can not easily find the global behaviour of measured elements. Blindly using the wrong trend line would, therefore completely change results in our case.

Additional identification that there might something be wrong with the parameters lies in the determined ages of individual stars in the clusters. The ages are determined by the best fitting isochrone with appropriate stellar [Fe/H] that lies closest to the stars' position on H-R diagram (further details in Buder *et al.* [166]). As the procedure is run independently for every star in the survey, ages among cluster stars (see plots in Figures 3.1, 3.2, 3.3, and 3.4) are strewn over a range of 5 Myr or more. Fortunately the abundance computation does not require information about the stellar age, but they both share the prior for an accurate information about the basic stellar parameters.

3.4.2 Determining chemical similarity

Having an analytical description of an individual abundance behaviour for every cluster, we can estimate how many and how accurately do the identified ejected stars match with cluster abundance patterns and trends. The most straightforward way to perform this is to count how often does abundance value of an investigated star fall inside a 1σ (or 2σ for a more relaxed selection) region around an abundance trend. Both regions and fitted trends are visualised in Figures 3.1, 3.2, 3.3, and 3.4. Before performing such counting, we additionally omitted abundance trends of the following chemical elements: V, Rb, Sr, Y, Zr, Mo, Ru, La, and Sm. Their low number of successful measurements per cluster and uncertain trends were not beneficial to the whole chemical tagging experiment and influenced only a small

fraction of stars. In general it is not advised to reduce the dimensionality of chemical space for greater differentiation between chemical signatures. As this was one of the first experiments analysing whether the latest GALAH DR3 abundances could be used for cluster and global blind tagging, we tried to remove as many additional systematic effects of uncertain measurements as possible. In our case, an individual star was counted as chemically similar if it matched (e.g. fell inside selected σ region around the fitted trend line) to a cluster in at least 68% of the considered abundances with valid trend fit. Chemically similarity of ejected stars, given as percentage of tagged stars, is presented in Table 3.3.

3.4.3 Tagging remaining field stars

The same principle can also be applied to remaining nearby field stars. As clearly evident from Figures 3.1, 3.2, 3.3, and 3.4, the cluster abundances are mostly similar to field stars and lie close to their densest regions in shown abundance scatter plots. Therefore, we were interested in the probability of a random field star being chemically similar to a nearby open cluster. In contrast, some of the investigated clusters, especially Blanco 1, show evident signs of being chemically separable from neighbouring stellar populations. Elements crucial for the populations' separation are commonly used as chemical tracers of galactic evolution and stellar age [176, 177, 178]. Therefore a young cluster, such as Blanco 1, that is located at high galactic latitudes, far from the main galactic plane, can locally be separated from its neighbouring stars. Something that is not common for typical open clusters that can currently be observed in the sky as they mainly reside close to the Galactic plane.

For the field chemical tagging procedure, we used the same selection principle as previously described in Section 3.4.2. The results of both tagging experiments are together presented in Table 3.3. Even if the abundance distributions of cluster and field stars are intertwined, the results of the tagging experiment are encouraging, as it was more likely that kinematically tracked stars were chemically similar to open cluster than a random nearby field star. The percentage of tagged ejected stars for almost all clusters in Table 3.3 is higher than percentage of tagged field stars. Clusters with zero tagging success suffer the curse of having low number statistics as they have only few possible candidates. For those clusters, if only one ejected star would be tagged, the probability would again be much higher than for the field stars.

3.5 Comparison with known tidal structures

In the previous section, we analysed stars whose integrated orbits indicate that they could be ejected from neighbouring open cluster sometime in the last 120 Myr. Depending on the mass of involved stars and their proximity during the slingshot mechanism, stars could be thrown out of a cluster into the interstellar space at random velocities and directions. This is true in the case when we presume that stars have no preferential way of moving inside a cluster. Such a process would therefore form a sphere of candidates around the main cluster. The density of candidates would homogeneously decrease in all radial directions away from the centre of a cluster. This effect is also visible in Figure 3.5, where ejection candidates

Table 3.3: Number and percentage of all considered and chemically similar (tagged) stars in the spatial neighbourhood around the analysed open clusters. Percentages indicate the number of stars in different components that are similar to cluster abundance pattern and scatter. Tagging algorithm is detailed in Section 3.4.2. Only spectra with unflagged stellar parameters were used to produce shown statistics.

Cluster	Ejected stars		Field stars	
	All	Tagged	All	Tagged
Berkeley 32	2	0 (0.0%)	39	2 (5.1%)
Blanco 1	4	2 (50.0%)	150	1 (1.0%)
IC 4665	4	0 (0.0%)	919	0 (0.0%)
Mamajek 4	25	6 (24.0%)	2300	82 (3.6%)
Melotte 22	10	1 (10.0%)	724	33 (4.6%)
Melotte 25	15	3 (20.0%)	389	71 (18.3%)
NGC 1817	6	0 (0.0%)	432	4 (0.9%)
NGC 1901	19	4 (21.1%)	5582	5824 (14.8%)
NGC 2112	4	0 (0.0%)	268	7 (2.6%)
NGC 2204	18	2 (16.7%)	113	14 (12.4%)
NGC 2516	71	4 (5.6%)	3330	90 (2.7%)
NGC 2548	5	0 (0.0%)	605	5 (0.8%)
NGC 2632	56	13 (22.8%)	1544	126 (8.2%)
NGC 2682	87	33 (37.6%)	1150	218 (19.0%)
NGC 6253	27	3 (11.1%)	653	21 (3.2%)
Ruprecht 147	9	0 (0.0%)	565	19 (3.4%)

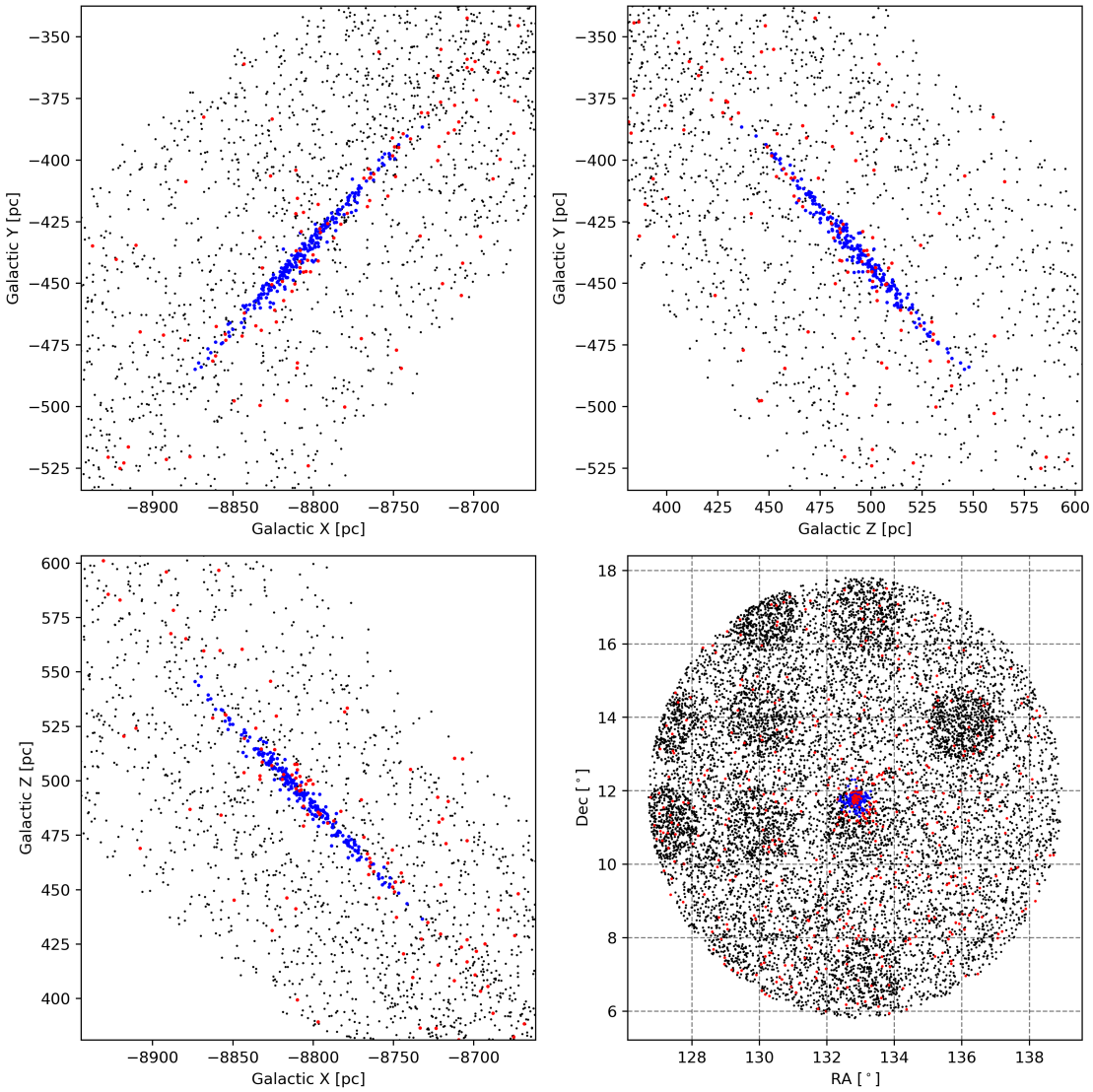

Figure 3.5: Plots show the spatial distribution of the ejected candidates in red around the definite cluster members in blue. All other stars considered in the analysis are shown as black dots. First three panel shows the Galactic position of described cluster components. The denser circular patches of stars in the bottom right panel are not real stellar overdensities, but indicate regions where combined radial velocities of *Gaia* and GALAH are used. The example is given for the open cluster NGC 2682, which has the highest number of potentially ejected stars. Uncertainty of determined stellar distance is clearly evident as elongation of the main cluster body in the radial direction, away from the observers' position.

are scattered all around the confirmed cluster members. A bit denser population of candidates is evident alongside the cluster volume, which probably corresponds to nearby initial cluster members that were discarded due to their mismatching radial velocity as described in Section 3.3.

A much less energetic and more gradual mechanism that also influences the lifetime of an open cluster is tidal stripping of stars [179]. It happens during clusters' journey through a more densely populated region such as galactic spiral arms. Tails

reaching out of a cluster would, therefore, have elongated, probably slightly bent shape (not to be confused by the elongated cluster shape in Figure 3.5), and not a spherical distribution as considered for ejection mechanism. In this section, we compare the chemical composition of known tidal structures (discovered in *Gaia* DR2 data by Röser and Schilbach [11], Meingast and Alves [13], Zhang *et al.* [14], Carrera *et al.* [17]) of a few the GALAH open clusters with other previously defined and analysed structures. All of the aforementioned tidal structures were identified by different clustering methods of the same *Gaia* data and should, therefore, be in a sense similar to results of our orbital classification procedure. From the published tidal structures [11, 13, 14, 17], we first removed all cluster members that were already identified by Cantat-Gaudin *et al.* [43] and consequently used in our procedure for the definition of clusters' chemical signature. In that way, our possibly ejected stars and tidal structures can be compared directly. Their overlap is shown in the fourth and fifth column of Table 3.4. As the overlap between them is not zero, we can reliably say that we are all detecting similar kinematic structures using completely different approaches (but the same *Gaia* data). The identified structures therefore may or might not be related to neighbouring open cluster.

Among randomly acquired The GALAH spectra, we observed some stars that where identified to belong to the kinematically discovered tidal structures around open clusters. If we consider only stars with unflagged the GALAH stellar parameters, the overlap between the sets is quite significant as more than 50% of unflagged tail stars were also identified as possible ejected for every studied cluster. This significant overlap reduces probability of the tails having a completely different chemical signature than our selections, but at the same time confirms validity of our orbit tracing approach.

For a tidal tail chemical tagging procedure, we used the same selection principle as previously described in Section 3.4.2. The results of the tagging experiment are presented in the last column of Table 3.4. Majority of the tails have quite significant probability of being chemically similar to a nearby open cluster.

3.6 Summary and conclusions

Open clusters, as long known and widely studied stellar structures, still provide many opportunities for their exploration using new and improved information about their stellar components and neighbouring environment. In this chapter, we explored the latest multidimensional abundance data determined for stars observed in the scope of the GALAH survey, whose targets also consisted of stars in multiple open clusters. Combined with the *Gaia* kinematic information, a precise position, shape, and velocity of the cluster volume can be defined. We used the method of backward orbit integration to determine if any of the neighbouring stars could be kinematically traced back to the cluster and have the same chemical signature.

To verify that our orbital integration methodology gives sensible results, they were compared towards identified cluster tidal structures around the same open clusters. Given non zero overlap in all cases, we are confident that our structures are also related to the main cluster body.

Deepened analysis of field and cluster abundance patterns showed improvement in the quality of determined abundances (in comparison towards older GALAH releases), but at the same time revealed the pitfalls of blindly using massively deter-

Table 3.4: Statistics of detected stellar tidal structures around some of the clusters that were observed by the GALAH. The columns shown in the following order give information about: name of the analysed cluster, number of all stars found in the surrounding tidal structure, number of stars without cluster members, size of overlap between ejected and tail, number of unflagged GALAH observations among tidal structure, chemical similarity with parent open cluster, and reference source of the used data.

Cluster	Stars in reference	Without used cluster members	Common with ejected (% of ejected)	GALAH unflagged parameters (% in common with ejected)	Chemically tagged (% of valid)	Membership reference
Blanco 1	644	276	3 (50%)	7 (42.9%)	0 (0.0%)	Zhang *et al.* [14]
NGC 2632	1393	738	17 (22.1%)	25 (68%)	7 (28.0%)	Röser and Schilbach [11]
NGC 2682	952	241	13 (14.4%)	18 (72.2%)	6 (33.3%)	Carrera *et al.* [17]
Melotte 25	238	238	8 (30.8%)	57 (14.0%)	1 (1.8%)	Meingast and Alves [13]

mined abundances for blind chemical tagging experiments that are highly desired in the community of galactic archaeology. Identified problems can successfully be overcome by applying differential analysis, which in our case showed some encouraging success of using the GALAH abundance data as tracers of stellar birth signatures of individual stars. The tagging experiment showed that it was more likely that we could chemically tag a kinematically pre-selected star than a random nearby star. For a much firmer confirmation, we would require more observations as some of the cluster components were sparsely observed by the GALAH.

The explored methodology depends on the quality and correctness of the determined abundance trends. From the current data, it seems that identified abundance trends are valid only for an individual cluster as they can vary among them. This decreases possibility of preforming differential chemical tagging on larger, potentially galactic, scales.

To improve the parameters used in this chapter, we would have to analyse cluster members as a homogeneous structure that was formed at the same time. This adoption of formation time would force the use of a single isochrone and stellar age for all members, improving their determined $\log g$, [Fe/H], and consequently also stellar surface chemical composition. Our first experiments with the mentioned analysis upgrade already show significant improvements in the stability of the trends and reduction of abundance scatter among cluster members.

Chapter 4

Chemically peculiar stars

This chapter has been adapted from the published paper titled *The GALAH survey: a catalogue of carbon-enhanced stars and CEMP candidates* [180] whose first author is author of this Doctoral thesis. The used computer code is published on GitHub platform [1] and results of the analysis as a catalogue on the VizieR service [2].

In the previous chapter, we saw that the chemical tagging of open cluster stars is far from a straightforward process in large spectroscopic studies and relies on many assumptions. One of them is that analysed spectra are of normal, non-active stars whose outer layer was not polluted by other stars or anyhow modified during their lifetime. To improve the accuracy of chemical tagging, such peculiarities in acquired spectra must be detected and derived parameters effectively filtered out before any analysis is performed. One of such peculiarities are chemical enrichments that are usually a result of pollution by nearby evolved stars, such as novae and supernovae. This chapter deals with the identification of carbon-enhanced stars using supervised and unsupervised classification algorithms. The chapter begins with a brief introduction of carbon-enhanced stars, and their historical and present importance in Section 4.1. The section is followed by the description of the used algorithms for the detection of carbon-enhanced stars in Section 4.2. Properties of the classified objects are investigated in Section 4.3, CEMP candidates are a focus of Section 4.4, with Section 4.5 describing a follow-up study for one of them. Final remarks are given in Section 4.6.

4.1 Introduction

Chemically peculiar stars whose spectra are dominated by carbon molecular bands were first identified by Secchi [181]. Their spectra are characterised by enhanced carbon absorption bands of CH, CN, SiC_2, and C_2 molecules, also known as Swan bands. Possible sources of enhancement are dredge-up events in evolved stars [182], enrichment by carbon-rich stellar winds from a pulsating asymptotic giant branch (AGB) star, which settles on a main sequence companion [183], or it can be the result of a primordial enrichment [184]. Historically, high latitude carbon stars, presumed to be giants, were used as probes to measure the Galactic rotation curve [185], velocity dispersion in the Galactic halo [186], and to trace the gravitational potential of the Galaxy.

[1] `https://github.com/kcotar/GALAH-survey-Carbon-enhanced-stars`

[2] `http://vizier.u-strasbg.fr/viz-bin/VizieR?-source=J/MNRAS/483/3196`

Because of their strong spectral features, the most prominent candidates can easily be identified from large photometric surveys [187, 188]. Specific photometric systems [189, 190, 191] were defined in the past to discover and further classify stars with enhanced carbon features in their spectra. Specifics of those systems were catalogued, compared, and homogenised by Moro and Munari [192] and Fiorucci and Munari [193].

Other useful data come from low-resolution spectroscopic surveys, whose classification identified from a few hundred to a few thousand of those objects [194, 195, 196, 197, 198]. High-resolution spectroscopy is required to search for candidates with less pronounced molecular absorption features or to determine their stellar chemical composition. Multiple studies have been carried out to determine accurate abundances of metal-poor stars [199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210]. Such detailed abundance information is especially important for the analysis and classification of chemically peculiar objects [211].

Today, the most sought after, of all carbon-enhanced stars, are the carbon-enhanced metal-poor (CEMP) ones whose fraction, among metal-poor stars, increases with decreasing metallicity [M/H] [196, 199, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222]. Amongst these, those near the main-sequence turn-off are expected to be of particular importance, as they may have accreted enough material from their AGB companion to produce an observable change in their atmospheric chemical composition [209, 223, 224]. The accreted material could ideide insight into the production efficiency of neutron capture elements in AGB stars [204]. Multiple studies show that a peculiar observed abundance pattern and carbon enrichment in a certain type of CEMP stars could be explained by the supernova explosions of first-generation stars that enriched the interstellar medium [225, 226, 227, 228]. The exact origin and underlying physical processes governing multiple classes of CEMP stars are not yet fully understood and are a topic of ongoing research [184, 229, 230]. Classification into multiple sub-classes is performed using the abundance information of neutron-capture elements [184, 231, 232, 233] that are thought to originate from different astrophysical phenomena responsible for the synthesis of those elements.

4.2 Detection procedure

To search for carbon-enhanced stars in the GALAH data set, we focused on spectral features that can be distinguished and are known markers of carbon enhancement. Instead of using one very weak atomic carbon absorption line (at 6587.61 Å), used by *The Cannon* to determine [C/Fe] abundance, we focused on a region between 4718 and 4760 Å observed in the blue arm that covers 4718–4903 Å in its rest-frame. In this range we can, depending on the radial velocity of the star, observe at least four Swan band features [234] with their band heads located at approximately 4715, 4722, 4737, and 4745 Å. The bands are caused by molecules that contain carbon. Therefore their shape is not a simple single absorption line, but a wedge-like absorption feature - a sequence of molecular vibrational bands. Historically, the bands have also widely been used for the exploration of cometary spectra.

Carbon enhancement is observable in spectra as a strong additional absorption feature (Figure 4.2) that is the strongest at the wavelength of the band's head. After that its power gradually decreases with decreasing wavelength. The most prominent and accessible for all of the spectra is a feature located at 4737 Å, pro-

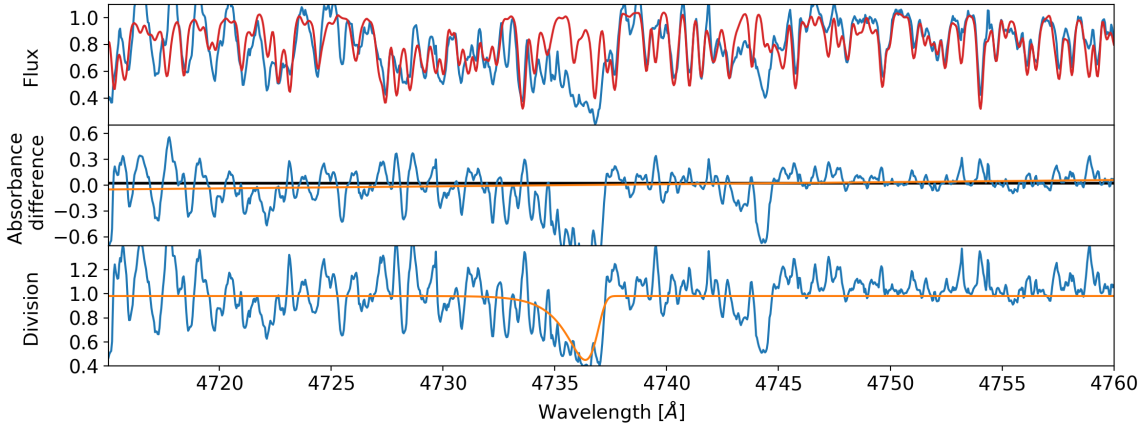

Figure 4.1: Equivalent plot as in the Figure 4.2 but presenting an example of a metal-rich star with multiple strong Swan features around 4737 and 4745 Å. The latter is not fitted by the orange line in the bottom panel as it is rarely present in the analysed spectra. Presented star has a 2MASS identifier J13121354-3533120 and is known Galactic carbon star [235].

duced by a $^{12}C^{12}C$ molecule. If other carbon features, like the one produced by a $^{13}C^{12}C$ molecule at 4745 Å (shown in Figure 4.1) are present in the spectrum, the carbon isotope ratio $^{12}C/^{13}C$ in a star can be determined. Its determination was not attempted in the scope of this chapter.

Detection of spectral features was tackled using two different classification procedures. First, a supervised procedure was used to identify the most prominent spectra with carbon enhancement. The supervised selection was based on the following two assumptions: where in the spectrum those features are located and their shape. This assumption was augmented with an unsupervised dimensionality reduction algorithm that had no prior knowledge about the desired outcome. The goal of a dimensionality reduction was to transform n-dimensional spectra onto a 2D plane where differences between them are easier to analyse. The unsupervised algorithm was able to discern the majority of carbon-enhanced spectra from the rest of the data set and enabled us to discover spectra with less prominent carbon enhancement features.

4.2.1 Supervised classification

To search for additional absorption features that are usually not found in spectra of chemically normal stars, we first built a spectral library of median spectra based on an initial rough estimate of stellar physical parameters that are derived by the automatic reduction pipeline, described in detail by Kos *et al.* [100]. The median spectrum for every observed spectrum in our data set was computed from physically similar spectra. Their effective temperature $T_{\rm eff}$ had to be in the range of $\Delta T_{\rm eff} = \pm 75$ K, surface gravity $\log g$ in the range of $\Delta \log g = \pm 0.125$ dex, and iron abundance [Fe/H] in the range of Δ[Fe/H] $= \pm 0.05$ dex around the stellar parameters of the investigated spectrum. The median spectrum was calculated only for observed targets with at least five similar spectra in the defined parameter range and with minimal noise level of SNR = 15 per resolution element, as determined for the blue spectral arm. All considered spectra were resampled to a common wavelength

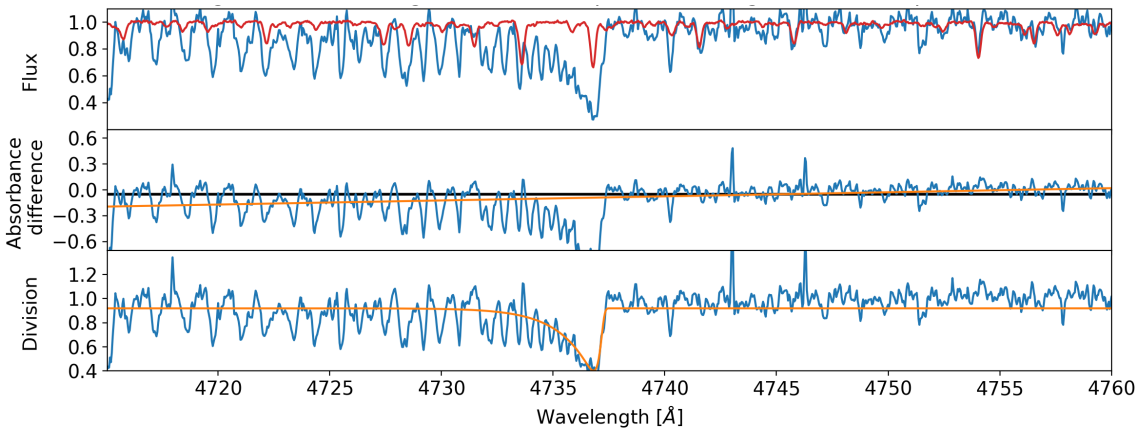

Figure 4.2: Example of a metal-poor carbon-enhanced candidate with strong Swan absorption feature at 4737 Å, caused by the carbon C_2 molecules. The first panel shows the stellar spectrum in blue and a corresponding reference median spectrum in red. The reference median spectrum was computed as the per-pixel median flux value of spectra with similar stellar parameters (selection defined in Section 4.2.1) as the spectrum shown in blue. The spectral difference (second panel) and division (third panel) were computed from those two spectra. The middle panel shows in orange a linear fit to the spectral difference that was used to identify spectra with the reduction problems on the blue border of the spectral range. The black horizontal line gives the median value of the spectral difference. The orange curve in the last panel shows a fit that was used to determine the strength of the observed carbon feature. The shown spectrum belongs to a star with a Two Micron All-Sky Survey (2MASS, [157]) identifier J11494247+0051023 and iron abundance [Fe/H] of −1.17, as determined by *The Cannon*.

grid with 0.04 Å wide bins and then median combined. The automatic reduction pipeline [100] performed the normalisation of the spectra along with the radial velocity determination and the corresponding shift to the rest frame. We checked that spectral normalisation and radial velocity determination are adequate also for carbon-enhanced stars. The normalisation procedure was done using a polynomial of low-order that is not strongly affected by the Swan band features or similar spectral structures. The radial velocity of a star was determined as an average of radial velocities that were independently determined for the blue, green, and red spectral arm. If one of the arms has a radial velocity deviating for more than two times the difference between the other two, it was excluded from the average (further details in Kos *et al.* [100]). Therefore the final velocity should be correct even if one of those arms contains features that are not found in the set of reference spectra used in the cross-correlation procedure.

With the limitation tat least five spectra must be used for the computation of the median spectrum, some possibly carbon-enhanced stars could be excluded from the supervised classification. The final number of spectra analysed by this method was 558,053.

Spectra, for which we were able to determine the median spectrum of physically similar objects, were analysed further. In the next step, we tried to determine possible carbon enhancement by calculating a flux difference and flux division between the observed stellar and median spectra, as shown in Figure 4.2.

In order to describe the position, shape, and amplitude of the Swan feature with its head at 4737 Å, we fitted a function that is based on a Log Gamma ($\log\Gamma$) distribution. The distribution, with three free parameters, was fitted to the division curve, where the Swan feature is most pronounced. Division curve, shown in the bottom panel of Figure 4.2, was computed by dividing the observed spectrum with its corresponding median spectrum. The fitted function f can be written as:

$$f(\lambda) = f_0 - \log\Gamma(\lambda, S, \lambda_0, A). \tag{4.1}$$

The shape of the curve is defined by an additional offset f_0, shape parameter S, constant centre wavelength λ_0 that was not fitted, and amplitude A of $\log\Gamma$ distribution. λ represents rest wavelengths of the observed spectrum. This function was selected because of its sharp rise followed by the gradual descent that matches well with the shape of a residual absorption observed in the Swan regions. The steepness of the rising part is determined by the parameter S (lower value indicates steeper raise) and its vertical scaling by the parameter A. We are not aware of any other profile shapes used for fitting Swan bands in the literature.

To narrow down possible solutions for the best fitting curve, we used the following priors and limits. The initial value for the parameter f_0 was set to a median of all pixel values in the division curve and allowed to vary between 0.5 and 1.5. The limiting values are, however, never reached. The centre of the $\log\Gamma$ distribution λ_0 was set to 4737 Å and was allowed to vary by 2 Å. Wavelength limits were set to minimise the number of misfitted solutions, where the best fit would describe the nearby spectral absorption lines not present in the median spectra or problematic spectral feature caused by the spectral data reduction as shown by Figure 4.21. We did not set any limits on parameters A and S in order to catch fitted solutions describing a spectrum difference that is different from the expected shape of the molecular absorption band.

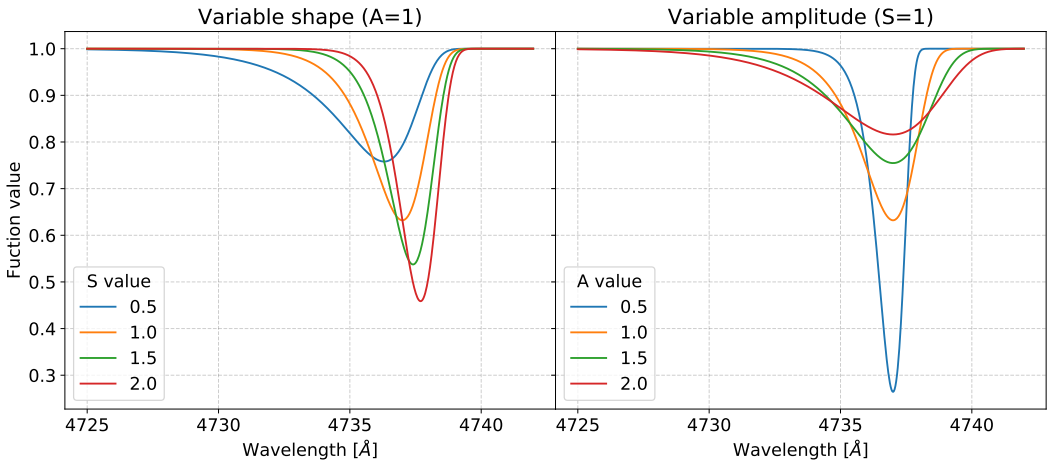

Figure 4.3: Visualization of selecting appropriate values of shape and amplitude parameters during the function fitting explained in Section 4.2.1. The left panel shows effect of varying shape parameter S and the right panel varying amplitude parameter A in Equation 4.1.

By integrating the surface between the offset f_0 and the fitted curve we calculated the strength of the Swan band. The integral (`swan_integ` in Table 4.1) is derived between 4730 and 4738 Å. It should not be used as a substitute for a carbon abundance measurement, but only to sort the detections of carbon-enhanced stars by their perceivable strength of the Swan band.

With so many spectra in our data set, unexpected reduction and analysis problems can hinder the selection of carbon-enhanced stars. In the first iteration, the results were ordered only by the value of the integrated Swan band region, but this proved to select too many spectra with reduction problems. Most of the problematic detections were caused by the incorrect normalisation of spectra with strong, non-carbon molecular bands. This normalisation issue is best observable at the border of thr spectral range, where Swan bands are located in the case of HERMES spectra. There, normalisation can be poorly defined in the case of numerous nearby absorption lines. In order to prevent miss-detections, additional limits on the shape ($S <= 1$) and amplitude ($A <= 1$) of the $\log\Gamma$ distribution were used to filter out faulty fitting solutions. Impact of varying A and S on the shape of the function is shown in Figure 4.3. Selecting too large parameter S causes function to be too narrow as shown in Figure 4.21. Similarly, too large amplitude A results in function with reduced skewness as fitted to the spectrum in Figure 4.22.

Figure 4.22 represents misfitted example where the function $f(\lambda)$ was fitted to the absorption lines of a double-lined spectroscopic binary, producing a shape of the function that is not characteristic for the analysed molecular band head. To remove spectra with reduction problems or peculiarity that would result in wrongly determined strength of the Swan band, we are also analysing the slope of the spectral difference and its integration in the limits of the Swan bands. One of the spectral trends that we are trying to catch with those indicators is shown in Figure 4.22, where spectral difference and its linear fit are steeply rising at the border of the spectrum.

By visual inspection of the algorithm, diagnostic plots are shown in Figure 4.2,

Table 4.1: List and description of the fields in the published catalogue of detected objects and objects matched with multiple literature sources.

Field	Unit	Description
`source_id`		*Gaia* DR2 source identifier
`sobject_id`		Unique internal per-observation star ID
`ra`	deg	Right ascension from 2MASS, J2000
`dec`	deg	Declination from 2MASS, J2000
`det_sup`	bool	Detected by the supervised fitting method
`det_usup`	bool	Detected by the t-SNE method
`swan_integ`	Å	Swan band strength if determined
`teff`	K	*The Cannon* effective temperature T_{eff}
`e_teff`	K	Uncertainty of determined T_{eff}
`logg`	$\log(\mathrm{cm/s^2})$	*The Cannon* surface gravity $\log g$
`e_logg`	$\log(\mathrm{cm/s^2})$	Uncertainty of determined $\log g$
`feh`	dex	*The Cannon* iron abundance [Fe/H]
`e_feh`	dex	Uncertainty of determined [Fe/H]
`flag_cannon`	int	*The Cannon* flags in a bit mask format
`type`		G for giants and D for dwarfs
`rv_var`	bool	Is radial velocity variable.
		Minimal radial velocity change of 0.5 $\mathrm{km\,s^{-1}}$
`li_strong`	bool	Shows strong lithium absorption
		The Cannon [Li/Fe] had to be > 1.0
`cemp_cand`	bool	Is star CEMP candidate
		The Cannon [Fe/H] had to be > -1.0
`bib_code`		ADS bibcode of the matched literature source

we limited a final selection to 400 spectra with the strongest carbon enhancement that was still visually recognisable. The last selected spectrum is shown in the Figure 4.20. Selection of spectra with lower enhancement would introduce possibly wrong classification of stars whose enhancement is driven by spectral noise levels, data reduction or any other process that has a subtle effect on the spectral shape.

4.2.2 Unsupervised classification

With numerous spectra of different stellar types, chemical composition, and degree of carbon enhancement, some might show different carbon features or be insufficiently distinctive to be picked out by the supervised algorithm described above.

Another analysis technique, which is becoming increasingly popular is a dimensionality reduction procedure named t-distributed Stochastic Neighbor Embedding (t-SNE, [236]) that has already proved to be beneficial in comparison and sorting of unknown spectral features of the same data set [84]. With the method, we aim to reduce the number of dimensions while preserving the structure of analysed data. In our case of a stellar spectrum, we have numerous dimensions (one per wavelength bin), among which we want to look for groups of data points. This search for groups could be done in the original space with many dimension, but visualisation of such dataset is problematic. As the relations between wavelength bins are not linear, methodologies such as Principal Component Analysis (PCA), have to be supplemented with non-linear methodologies; t-SNE for example. An additional problem, we are dealing with is a density of investigated spectra. As peculiar spectra are usually in a minority, the algorithm must recover global and local structures alike. t-SNE is able to map data sets with high variations in density, so it is able to resolve small, local details, as well as the global picture. This feature is achieved by the internal use of a Student's distribution. The algorithm is fine-tuned using the property perplexity that controls whether t-SNE is more sensitive to global or local structures. This property is comparable to setting the number of nearest neighbours taken into consideration. The role of the perplexity can be tested in an online interactive tool produced by Wattenberg *et al.* [237]. After computing similarities between all pairs of the investigated spectra, the algorithm tries to find the best transformation that optimally arranges spectra in a 2D plane. As the transformation is variable and non-linear, the actual distance between two objects in a final 2D plane does not linearly depend on the spectral similarity measure. This property of the t-SNE algorithm additionally ensures more homogeneous coverage of the 2D plane in comparison to other dimensionality reduction methods.

The t-SNE projection shown in Figure 4.4 was computed from normalised spectra between 4720 and 4890 Å. To maximise the number of analysed spectra, no other limiting cuts than the validity of the wavelength solution (bit 1 in `red_flag` set to 0 by reduction pipeline [100]) in this arm was used. This resulted in 588,681 individual spectra being analysed by the automatic unsupervised algorithm. This is $\sim$ 30k more spectra than in the case of supervised classification, where we applied more strict criteria for the selection of analysed spectra (Section 4.2.1).

Without any prior knowledge about the location of objects of interest in the obtained projection, we would have to visually analyse every significant clump of stars in order to discover whether the carbon-enhanced population is located in one of them. This can be simplified by adding the results of the supervised classifica-

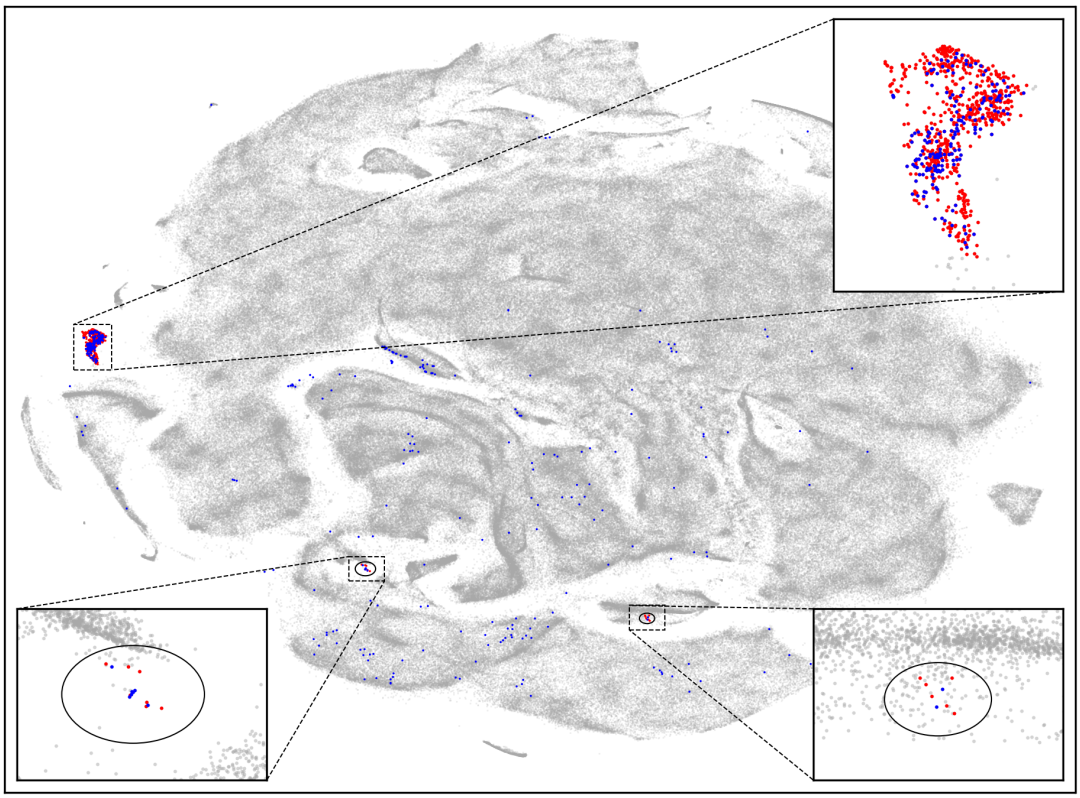

Figure 4.4: t-SNE projection of 588,681 observed spectra ranging between 4720 and 4890 Å. Red dots (756 spectra) mark a clump in the projection that was manually selected to contain carbon-enhanced spectra. Superimposed blue dots represent carbon-enhanced spectra determined by the supervised algorithm. Outside the t-SNE selected clump, we have 224 spectra that were determined to be carbon-enhanced only by the supervised method. All other analysed spectra are shown in grey shades, depending on their density in the 2D projection. Two ellipses indicate regions where the majority of CEMP candidates are located in the projection.

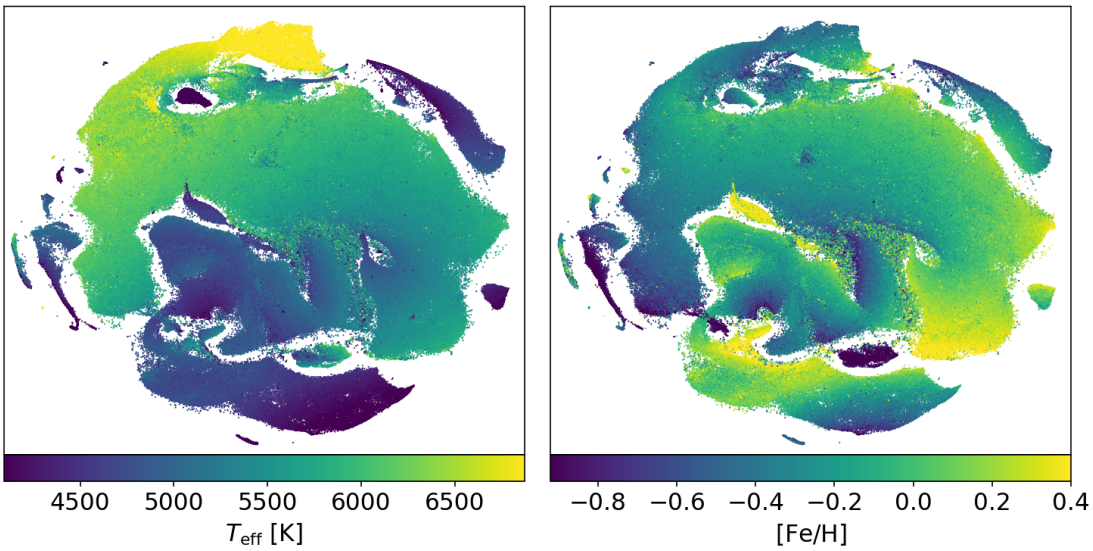

Figure 4.5: Spatial distribution of all available measurements of $T_{\rm eff}$ (left panel) and [Fe/H] (right panel) as determined by *The Cannon*. Dots, representing analysed spectra in the t-SNE projection, are colour coded by their parameter values. Colours and their corresponding values are explained by a colourbar under the graph.

tion into this new projection. In Figure 4.4, the stars identified by the supervised classification are shown as blue dots plotted over grey dots representing all spectra that went into the analysis. The majority of blue dots are located in a clump on the left side of the projection. A high concentration of objects detected by a supervised method leads us to believe that this isolated clump represents carbon-enhanced objects in the t-SNE projection. To select stars inside the clump, we manually have drawn a polygon around it.

Inspection of other blue labelled spectra outside the main clump revealed that their slight carbon enhancement could not be identified by the t-SNE similarity metric as another spectral feature might have dominated the spectral comparison.

Additional exploration of the t-SNE projection revealed two smaller groups of metal-poor carbon-enhanced spectra located inside ellipses shown in Figure 4.4. A confirmation that those regions are populated with metal-poor stars can be found in Figure 4.5 where the dots representing spectra in the projection are colour coded by [Fe/H] and $T_{\rm eff}$. To maximise the number of those objects in the published catalogue, we manually checked all undetected spectra in the vicinity of the detected ones. This produced an additional 13 CEMP detections.

t-SNE limitations

While checking the local neighbourhood of some of the blue dots in Figure 4.4 that are strewn across the t-SNE projection we identified a possible limitation of our approach for the automatic detection of specific peculiar spectra if their number is very small compared to the complete data set. Figure 4.6 shows a collection of a few carbon-enhanced spectra embedded between other normal spectra that were taken out of the right ellipsoidal region in Figure 4.4. As they are quite different from the others, they were pushed against the edge of a larger cluster in the projection, but their number is not sufficient to form a distinctive group of points in the projection.

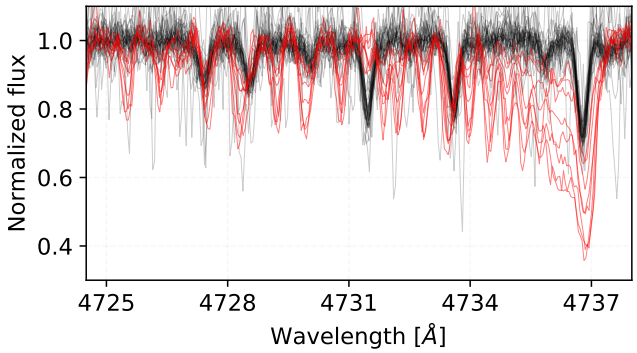

Figure 4.6: A collection of spectra that were determined to be mutually very similar by the t-SNE algorithm. Out of 46 spectra inside the right black ellipse in Figure 4.4 we identified 8 carbon-enhanced spectra with visually very different and distinctive spectrum in the region from 4734 to 4737 Å that is also depicted in this figure. For easier visual recognizability, they are coloured in red.

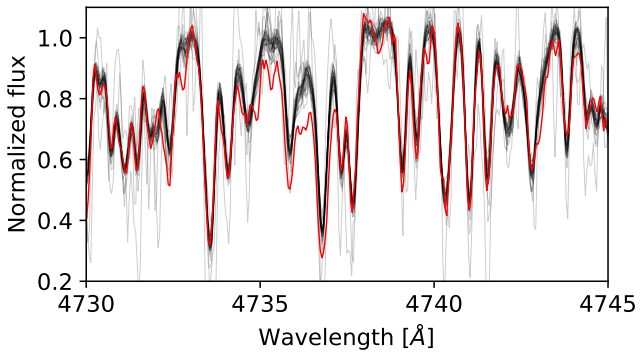

Figure 4.7: Spectral comparison between one of the detected carbon-enhanced stars in red and its 30 closest neighbours in the t-SNE projection shown as black curves. Enhancement in the spectrum was probably not sufficiently distinct and was dominated by the spectral noise. Therefore the spectrum was placed among other physically similar spectra without visible enhancement.

Therefore any automatic algorithm that would try to distinguish those objects based solely on a local density of points would most probably fail.

Another specific of the t-SNE projection that we must be aware of is how it computes the similarity between analysed spectra. Combined similarity, which is computed as a sum of per pixel differences, has zero knowledge about the locations where in the spectrum those differences occur. The red spectrum in Figure 4.7 with a slight signature of carbon enhancement in the range between 4734 and 4737 Å has been placed among spectra with similar physical properties. Its slight carbon enhancement and comparable spectral noise to other spectra in its vicinity are probably the reason why it was placed in such a region of the t-SNE projection. This unrecognizability could be solved by using a smaller portion of the spectrum in a dimensionality reduction, which could, at the same time, lead to a loss of other vital information about a star.

4.3 Characteristics of candidates

The final list of detected carbon-enhanced stars consists of 918 stars, corresponding to 993 spectra detected by at least one of the described methods. Among them, 63 stars were observed and identified at least twice and up to a maximum of four times. Those identifications belong to repeated observations that were performed at different epochs. Because not all of the observed spectra were considered in the classification procedure (due to the limitations described in Section 4.2) this is not the final number of stars with repeated observations. By searching among the complete observational data set, the number of carbon-enhanced stars with repeated observations increases to 90.

Out of those 90 stars, every repeated observation of 56 stars was classified as being carbon-enhanced. In total, we detected 76.5 % of the carbon-enhanced spectra among repeated observations where at least one of the repeats have been classified as having enhanced carbon features in its spectrum. The unclassified instances usually have a low SNR value that could decrease their similarity value towards other carbon-enhanced stars in the t-SNE analysis or have incorrect stellar parameters and were therefore compared to an incorrect median spectrum during the supervised analysis.

4.3.1 Radial velocity variations

With repeated observations in the complete observational data set, we can look into measured radial velocities and investigate a fraction of possible variables among detected stars. For certain types of carbon-enhanced objects, the fraction of variables should be high or even close to 100 % Sperauskas *et al.* [238]. Taking into account all of the repeated observations in our data set and not just the repeats among the identified spectra, 52 out of 90 stars show a minimum velocity change of 0.5 $\mathrm{km\,s^{-1}}$(70 stars with minimum change of 0.25 $\mathrm{km\,s^{-1}}$) and a maximum of 45 $\mathrm{km\,s^{-1}}$in different time spans ranging from days to years. The detailed distribution is presented by Figure 4.8. The threshold was selected in a such way to be significantly larger than the estimated typical accuracy of ~ 0.1 $\mathrm{km\,s^{-1}}$as determined for the GALAH radial velocities by Zwitter *et al.* [160].

Any radial velocity change can hint at the presence of a secondary member or at intrinsic stellar pulsation [239, 240, 241], as carbon-enhanced stars are found among all long-period variable classes (Mira, SRa, and SRb [242, 243]). Follow-up observations are needed to determine their carbon sub-class and subsequently, the reason behind variations of radial velocity.

Visual inspection of variable candidates revealed that none of them shows obvious multiplications of spectral absorption lines, a characteristic of a double-lined binary system. Therefore we can conclude that none of them is a binary member in which both components are of comparable luminosity, and a difference between their projected radial velocities is high enough to form a double-lined spectrum. From our simulations with median spectra, such line splitting becomes visually evident at the velocity difference of $\sim$14 $\mathrm{km\,s^{-1}}$. If the components do not contribute the same amount of flux, the minimal difference increases to $\sim$20 $\mathrm{km\,s^{-1}}$.

The chemical peculiarity of a dwarf carbon-enhanced star (dC) that exhibits enhancement of C_2 in its spectra could be explained by its interaction with a primary star in a binary system [244]. Chemically enhanced material is thought to be ac-

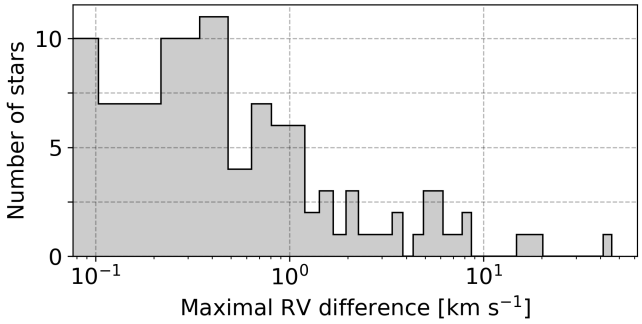

Figure 4.8: Distribution of maximal velocity change between repeated observations of the stars that were classified as carbon-enhanced.

creted from the evolved AGB companion. Less than thirty of such systems that show signs of the existence of an invisible evolved companion who might have enriched a dC by the carbon have been identified spectroscopically to date [244, 245, 246]. This low number of confirmations gave us the possibility to greatly increase the list with every additional confirmed object. The only detected dC star (for criteria see Section 4.3.2) with repeated observations shows that its radial velocity is unchanged on the order of 0.1 $\mathrm{km\,s^{-1}}$ during the two years between consecutive observations. Hence, it cannot be classified as a possible binary system from those two observations alone. The lack of clear evidence for binarity among dC stars, especially among the most metal-poor, can also be explained by another enrichment mechanism. Farihi *et al.* [247] showed that a substantial fraction of those stars belongs to the halo population based on their kinematics information. We performed similar analysis in Section 4.4 that is supported by Figures 4.16 and 4.17. Combined with the results of Yoon *et al.* [184] that classified the prototype dC star G 77-61 as a CEMP-no star, that is known to have intrinsically low binarity fraction [224, 248], their carbon-enhancement may be of a primordial origin.

4.3.2 Stellar parameters

For the analysis of stellar parameters, we used values determined by *The Cannon* data interpolation method that was trained on actual observed HERMES spectra. To exclude any potentially erroneous parameter, we applied a strict flagging rule of `flag_cannon`=0 (extensive description of the flagging procedure can be found in Buder *et al.* [104]), thus obtaining a set of 347 objects with trustworthy stellar parameters. Such a large percentage of flagged objects could be attributed to their nature as an additional elemental enhancement that we are looking for might not be a part of the training set. A raised quality flag would hint that the spectrum is different from any other in the training set or that the fit is uncertain and has a large χ^2. Therefore flagged parameters have to be used with care, especially on the border of, and outside the training set.

The majority (338) of the unflagged detected objects are giants, and only nine are confirmed to be dwarf stars based on their spectroscopic stellar parameters (Figure 4.9).

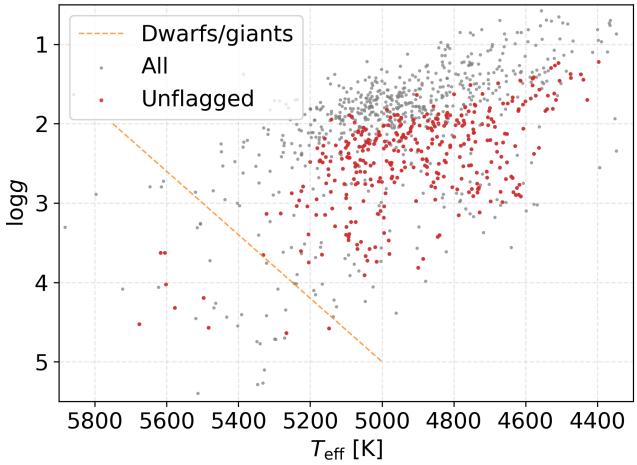

Figure 4.9: Kiel diagram for a subset of 338 detected carbon-enhanced stars with valid stellar parameters in red. Uncertain positions of flagged stars are shown with grey dots. Dashed orange line illustrates the border between giants and dwarfs.

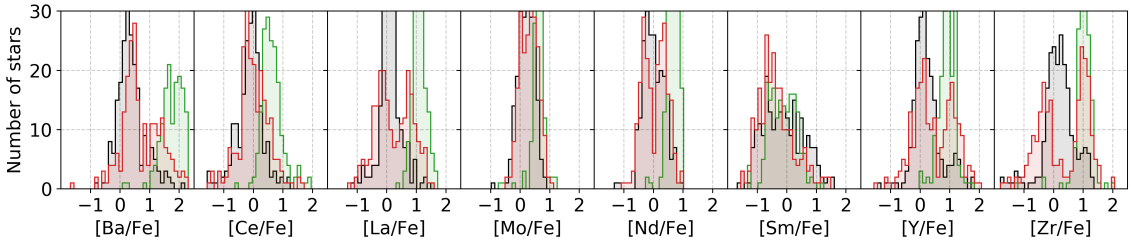

Figure 4.10: Distribution of s-process element abundances for stars in three different groups. The most enhanced group in green represent a carbon-enhanced stars located in the t-SNE selected clump of stars. The red distribution presents all other detections that are placed around the projection, and outside the clump. As a control group, the same distribution in black is shown for their closest t-SNE neighbours, therefore the black and red distribution contain an equal number of objects. No abundance quality flags were used to limit abundance measurements.

4.3.3 S-process elements

Focusing on a spectral signature of the detected objects inside and outside the t-SNE selected clump (Figure 4.4) we can further investigate which spectral feature might have contributed to their separation. The distributions of their abundances in Figure 4.10 and strength of spectral features corresponding to the same elements in Figure 4.11 hints to an enhancement of s-process elements among stars inside the selected clump. The abundance difference is slightly different for every elements. In general the abundance peaks are separated for about 1 dex This additional enhancement might be another reason, besides the carbon enhancement, for the algorithm to cluster all of those stars as being different from the majority of spectra.

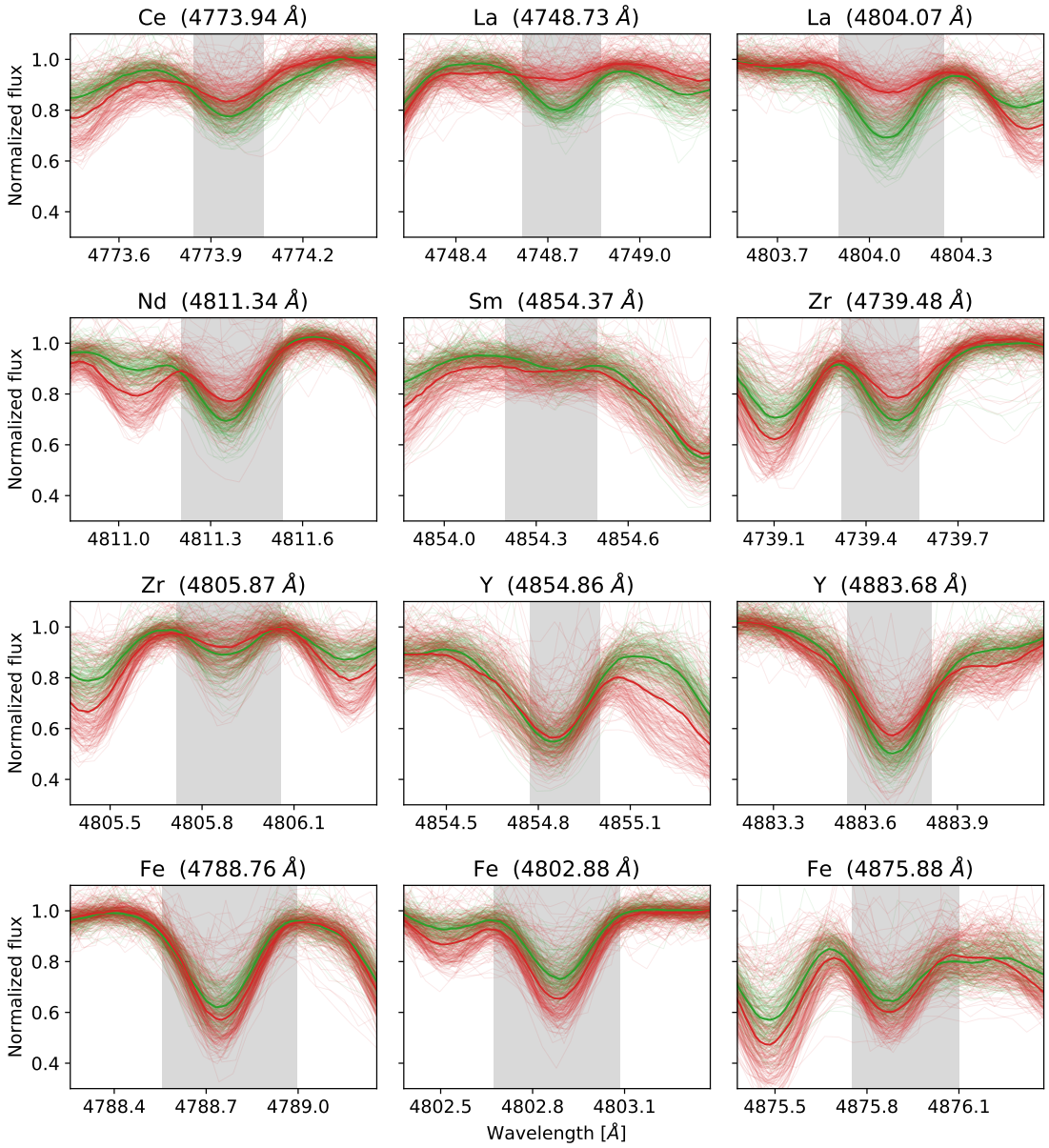

Figure 4.11: Spectral subset around the absorption features in the blue arm that were used to determine abundances of Fe and s-process elements. Same colour coding is used as in Figure 4.10. Spectra inside the t-SNE determined clump are shown in red, and outside it in green. Median of all spectra is shown with a bold line of the same colour. The shaded area gives the wavelength range considered in the computation of abundances. The abundance difference is the result of green spectra having lower Fe enhancement and higher enhancement (deeper absorption lines) of s-process elements.

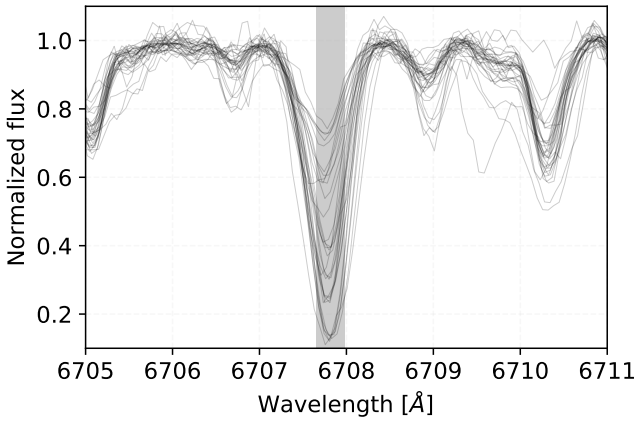

Figure 4.12: Spectral subset of 32 lithium-rich carbon-enhanced stars among the identified stars. The highlighted wavelength region is used by *The Cannon* to determine the lithium abundance of a star.

4.3.4 Lithium abundance

The derivation of elemental abundances for known carbon-enhanced stars has shown that some of them can exhibit strongly enhanced levels of Li in their atmosphere [249]. Lithium is thought to be produced by hot-bottom burning [250] and brought to the surface from the stellar interior. Investigation of the Li line at 6707 Å revealed 32 such stars. Their spectra, centred around the Li feature, show a greatly varying degree of Li abundance in Figure 4.12.

4.3.5 Sub-classes

Following a revision of the original MK classification [251] introduced by Barnbaum *et al.* [252], carbon stars are separated into five different classes named C-H, C-R, C-J, C-N, and Barium stars. Of all the spectral indices proposed for the spectral classification, we are only able to measure a small part of Swan C_2 bands and Ba II line at 6496 Å. For a more detailed classification of detected objects into proposed classes, we would need to carry out additional observations with a different spectroscopic setup to cover all the significant features.

Additionally, the features caused by the $^{13}C^{12}C$ molecule are strongly enhanced only for a handful of spectra in our data set, therefore we did not perform any isotopic ratio analysis or identification of possible C-J objects, which are characterised by strong Swan bands produced by the heavier isotopes.

According to the abundance trends presented in Section 4.3.3 and the classification criteria defined by Barnbaum *et al.* [252], we could argue that the stars selected from the t-SNE projection belong to the C-N sub-class. Their s-process elements are clearly enhanced over solar values (see Figure 4.10), but the actual values should be treated with care as they are mostly flagged by *The Cannon*(having quality flag `flag_cannon` > 0). This uncertainty might come from the fact that the training set does not cover carbon-enhanced stars and/or stars with such enhancement of s-process elements.

4.3.6 Match with other catalogues

In the literature we can find numerous published catalogues of carbon-enhanced (CH) stars [194, 197, 235] and CEMP stars [221, 253, 254, 255, 256, 257] observed by different telescopes and analysed in inhomogeneous ways. Most of those analyses were also performed on spectra of lower resolving power than the HERMES, therefore some visual differences are expected for wide molecular bands. By matching published catalogues with the GALAH observations that were analysed by our procedures, we identified 44 stars that matched with at least one of the catalogues. Of these, 28 were found in CH catalogues and 16 in CEMP catalogues.

From the stars recognised as CEMPs in the literature, we were able to detect only 1 star using the described methods. Visual assessment of the diagnostic plots provided by our analysis pipeline proved that the remaining 15 CEMP matches do not expresses any observable carbon enhancement in Swan bands and were therefore impossible to detect with the combination of our algorithms. The reason for this difference between our and literature results might be in the CEMP selection procedure employed by the aforementioned literature. Every considered study selects their set of interesting stars from one or multiple literature sources based on values of [M/H] and [C/Fe] that were measured from the atomic spectral lines and not molecular lines.

The match is larger in the case of CH matches, where we were able to confirm 11 out of 33 possible matched carbon-enhanced stars. As the observed molecular bands are prominent features in the spectra, we explored possible reasons for our low detection rate. Visual inspection of spectra for the remaining undetected matched stars proved that they also show no or barely noticeable carbon enhancement in the spectral region of Swan bands, therefore reason must lie in the detection procedures used in the cited literature. Christlieb *et al.* [194] used low-resolution spectra to evaluate enhancement of C_2 and CN bands. The results are also summarised in their electronic table. In here, all of our undetected stars are marked to contain enhanced CN bands but no C_2 bands. Combining this with Figure 4.13 we speculate that those stars occupy a narrow range of parameter space where C_2 is not expressed and therefore undetectable in the HERMES spectra.

Number of successfully detected stars matched between the surveys could also be influenced by different excitation temperatures of analysed carbon-rich molecules. Frequently studied photometric G-band (centred at 4640 Å and FWHM of 1200 Å), that is not present in our spectra, covers a spectral region rich in CH molecule features whose temperature dependence is different than for a C_2 molecule. Presence of those bands is identified by classifying a carbon-enhanced star into C-H sub-class (see Section 4.3.5). As we detected all C-H stars identified by Ji *et al.* [197], that are also present in the GALAH data set, we are unable to discuss about the selection effect in the $T_{\rm eff}$ range between ∼5100 and ∼5300 K where those three stars were found.

The position of all stars matched with the literature is also visualised on the t-SNE projection in Figure 4.19, where it can be clearly seen that they lie outside the selected clump with identified carbon enhancement and are strewn across the projection. Close inspection of spectra that are spatially near the aggregation of CEMP stars from the literature, revealed no visible carbon enhancement. The enhancement is present neither in form of molecular bands nor expressed as stronger atomic carbon line. They therefore are indistinguishable from other metal-poor stars

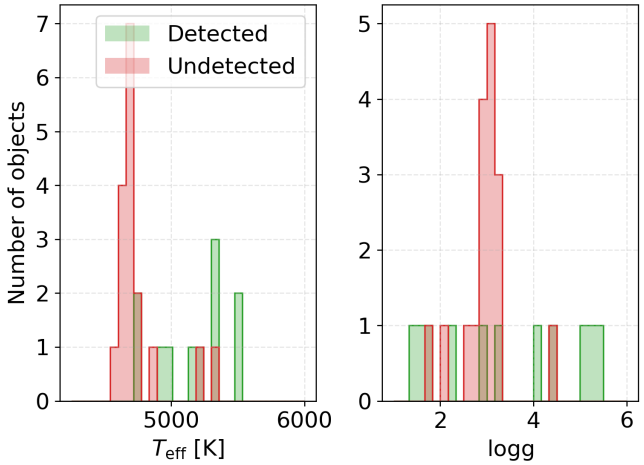

Figure 4.13: Comparison between the stellar parameters of detected (green histogram) and undetected (red histogram) carbon-enhanced stars found in literature.

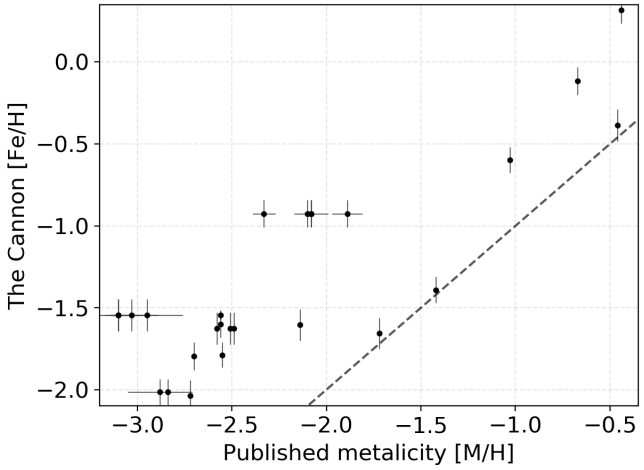

Figure 4.14: Correlation between published metallicities and *The Cannon* iron abundance for the stars that were classified as CEMPs in the literature. As some of those stars were taken from multiple literature sources, we also have multiple determinations of [M/H] for them. This can be identified as horizontal clusters of dots at different [M/H], but with the same [Fe/H]. Where available, uncertainties of parameters are shown. The dashed line follows a 1:1 relation.

with similar physical parameters.

4.4 Metal-poor candidates

CEMP stars are defined in the literature as having low metallicity [M/H]<-1 and strong carbon enrichment [C/Fe]$>+1$. In the scope of this analysis, we assume that our measurement of [Fe/H] is a good approximation for the metallicity. To be sure about this we compared [M/H] values of CEMP stars found in the literature and [Fe/H] derived by *The Cannon* for the same stars. The relation between them is

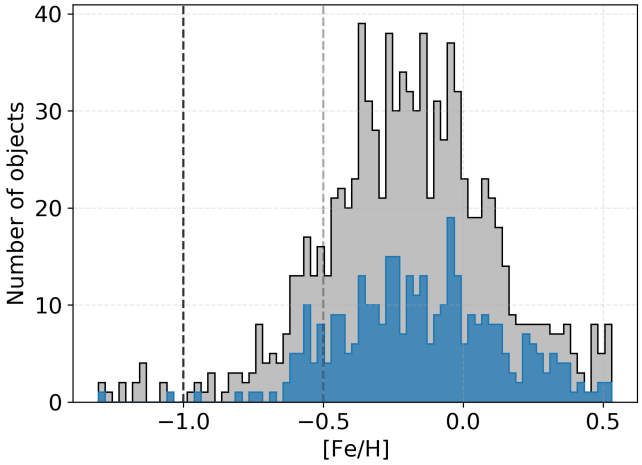

Figure 4.15: Histogram of [Fe/H] for detected carbon-enhanced stars with valid *The Cannon* stellar parameters in blue and for every detected carbon-enhanced star in grey. Two vertical lines are located at iron abundances of −1.0 and −0.5.

shown in Figure 4.14. We see that our values start deviating from the published values at metallicities bellow −1.5. Bellow that threshold the differences are in the range of $\sim$ 1 dex, but the same trend is obvious for both data sets. General offset between [Fe/H] and [M/H] is expected as the later gives abundance information of all elem and not just iron. Uncertainty of the published [M/H], derived from multiple sources, can reach up to 0.5.

Taking unflagged *The Cannon* parameters and abundances of the detected objects we can determine possible CEMP candidates among our sample. As also shown by Figure 4.15 our set of carbon-enhanced stars consists of 41 objects with [Fe/H]<-0.5 and 2 objects with [Fe/H]<-1.0. If we also include potentially incorrect parameters, the number of objects with [Fe/H]<-1.0 increases to 28, which is equal to 2.8 % of detected carbon-enhanced spectra. In any case, none of them has a valid determination of carbon abundance. Analysing HERMES spectra in order to determine carbon abundance is difficult because the automatic analysis is based on only one very weak atomic absorption line that is believed to be free of any blended lines. Consequently, we are also not able to measure the [C/O] abundance ratio, as a majority of determined [C/Fe] abundances is flagged as unreliable. Complementary observations are needed to determine the abundance and confirm suggested CEMP candidates.

A low number of metal-poor candidates could also be explained by the specification of the HERMES spectrograph as its spectral bands were not selected in a way to search for and confirm the most metal-poor stars with [Fe/H]<-3.0. With the release of *Gaia* DR2 data [42], stars low/high-metallicity could also be compared with their Galactic orbits. To determine the distribution of detected stars among different Galactic components, we performed an orbital integration in `MWPotential2014` Galactic potential using the `galpy` package [171]. In order to construct a complete 6D kinematics information, *Gaia* parallax and proper motion measurements were supplemented with the GALAH radial velocities. Results shown in Figure 4.17 suggest that our CEMP candidates could belong to two different components of the Galaxy. Stars with maximal $z < 4$ kpc most probably belong to the thick disk and

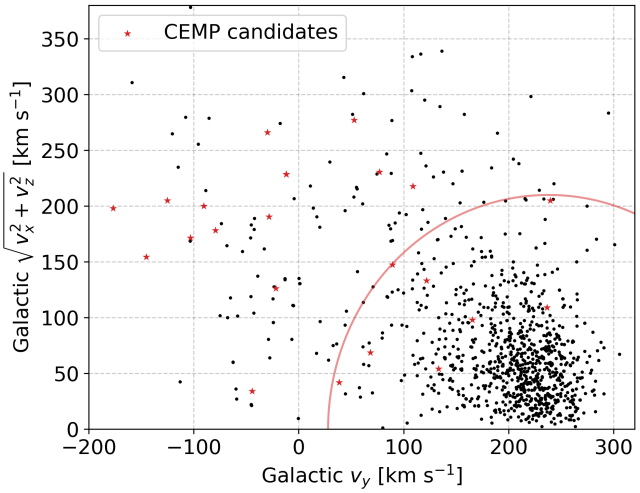

Figure 4.16: Toomre diagram used to identify possible local halo stars among our detected carbon-enhanced stars, especially CEMP candidates. Halo stars in this diagram are located above the red circular line, satisfying the velocity condition $|\mathbf{v} - \mathbf{v_{LSR}}| > 210$ km s^{-1} (the threshold taken from Koppelman *et al.* [258]). CEMP candidates are marked with star symbols.

stars with $z > 5$ kpc to the halo population that is inherently metal-poor. This is also supported by their angular momentum in Figure 4.17 and their Galactic velocities shown in Figure 4.16.

When looking at the distribution of [Fe/H] for the complete set of observed stars, we find a comparable distribution as for carbon-enhanced stars. Similarly, about 1.8 % of stars are found to be metal-poor with [Fe/H]<-1.0.

4.5 Follow-up observation

To further classify and analyse one of the detected objects, a star with 2MASS identifier J11333341-0043060 was selected for a follow-up observation. We acquired its high-resolution Echelle spectrum (with the resolving power R $\sim 20,000$), using a spectrograph mounted on the 1.82 m Copernico telescope located at Cima Ekar (Asiago, Italy). Because only a few of our detected candidates are observable from the Asiago observatory, we selected the best observable CEMP candidate, whose [Fe/H] was determined by *The Cannon* to be -0.96. The selected star, with V $=$ 12.79, was on the dark limit of the used telescope, therefore low SNR was expected. The one-hour long exposure of the selected object was fully reduced, normalised order by order, and shifted to the rest frame.

Although the acquired spectrum covers a much wider and continuous spectral range (from 3900 to 7200 Å) than the HERMES spectra, only subsets, relevant for the classification of carbon-enhanced stars are presented in Figure 4.18. They were identified by visually matching our observed spectrum with the published moderate-resolution spectral atlas [252] of peculiar carbon stars. Where available, the GALAH spectrum is shown alongside the Asiago spectrum. Carbon enhancement is not expected to vary over a period of several years, therefore both spectra should show similar features. The second and fourth panel in Figure 4.18 confirm that both

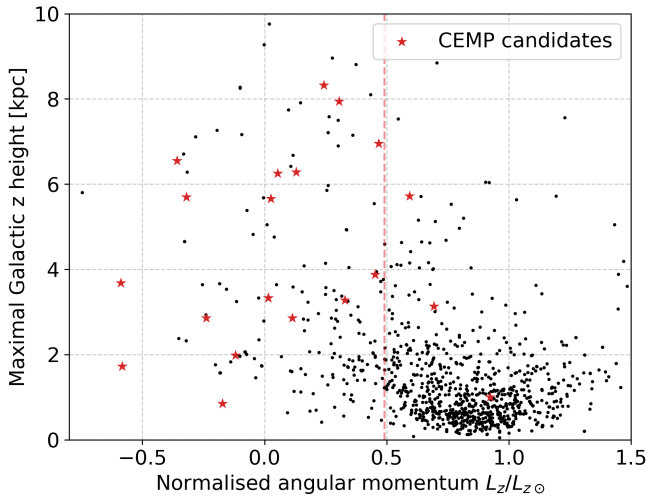

Figure 4.17: Distributions of maximal height above/below the Galactic plane reached by the detected stars on their orbit around the centre of the Galaxy. Heights are compared towards their normalised angular momentum L_z, where $L_\odot = 2033.6$ km s^{-1} kpc. Vertical dashed line at $L_z = 1000$ km s^{-1} kpc highlights the transition from the halo to the disk population, where a majority of the halo stars is located below this threshold (the threshold was visually estimated from similar plots in Koppelman *et al.* [258]). CEMP candidates are marked with red star symbols.

observations indicate a similar degree of carbon enhancement.

Following the classification criteria of carbon stars, we determined that the star belongs to the C-H sub-class. The definitive features for this class are strong molecular CH bands, prominent secondary P-branch head near 4342 Å (top panel in Figure 4.18), and noticeable Ba II lines at 4554 and 6496 Å [198], which are all present in the spectrum. The star definitely does not have a high ratio between ^{13}C and ^{12}C isotopes as the Swan features corresponding to ^{13}C are clearly not present, therefore it can not be of a C-J sub-class.

Following the current state of knowledge [238, 259, 260] that most, if not all, C-H stars show clear evidence for binarity, we compared the radial velocity between both observations. They hint at the variability of the object as the follow-up radial velocity (126.75 ± 1.63 km s^{-1}) deviates by more than 3 km s^{-1} from the velocity (123.43 ± 0.08 km s^{-1}) observed as part of the GALAH survey. The time span between the two observations is more than 2.5 years, where the exact JD of the observation is 2458090.702 for the Asiago spectrum, and 2457122.095 for the GALAH spectrum. Further observations along the variability period would be needed to confirm whether it is a multiple stellar system.

4.6 Conclusions

This work explores stellar spectra acquired by the HERMES spectrograph in order to discover peculiar carbon-enhanced stars, which were observed in the scope of multiple observing programmes conducted with the same spectrograph.

We show that the spectra of such stars are sufficiently different from other stellar

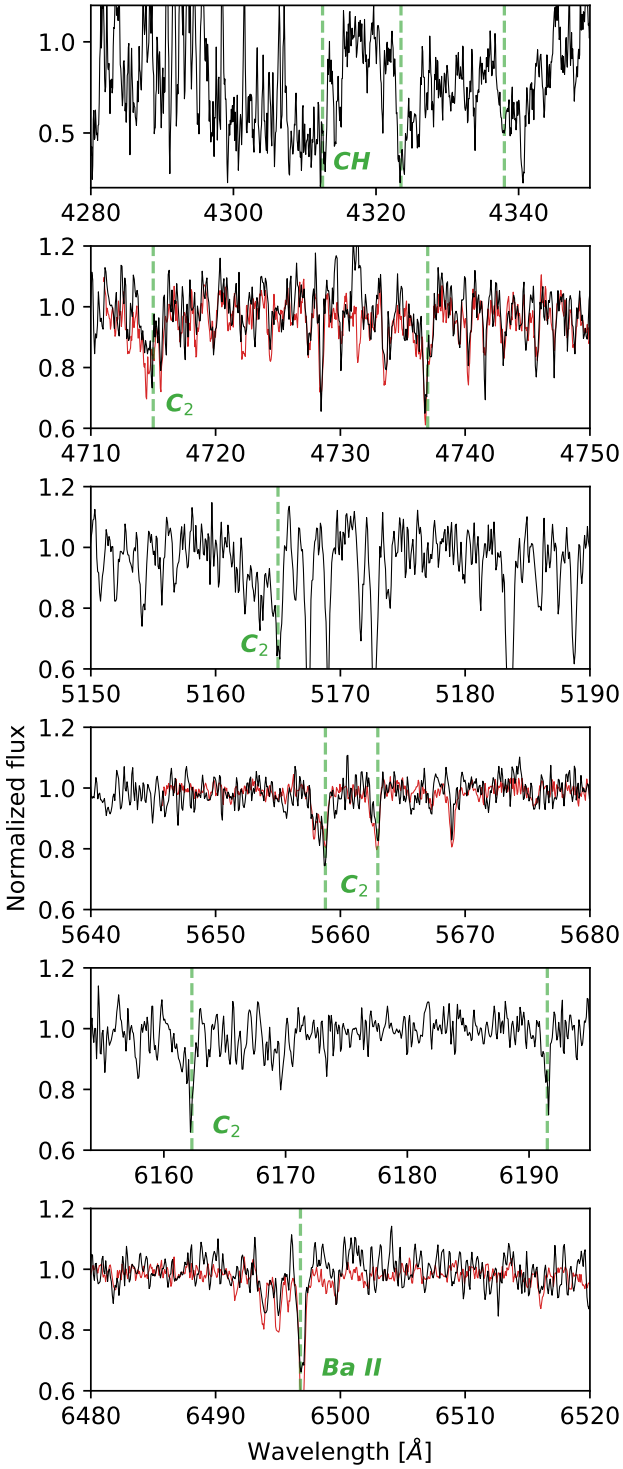

Figure 4.18: Subsets from the follow-up Asiago spectrum with resolving power comparable, but not identical, to the HERMES spectrum. It contains multiple spectral features used to evaluate carbon enhancement in a star and its carbon sub-class. Relevant spectral features (single line or tip of a molecular band) are marked with vertical dashed green lines and labels that represent a molecule or an element that is responsible for the features shown in the individual panel. The 2MASS identifier of the observed star is J11333341-0043060. In the wavelength ranges where it is available, the GALAH spectrum of the same star is shown in red.

types to be recognisable in high-resolution spectra with limited wavelength ranges. This can be done using a supervised or unsupervised method. The latter was used to identify observed stars solely on the basis of acquired spectra. By combining both methodologies we identified 918 unique stars with evident signs of carbon enhancement of which 12 were already reported in the literature. Out of all matched objects from the literature, we were unable to detect and confirm 16 (57 %) CH and 15 (93 %) CEMP stars with our procedures. As some of those objects were proven to contain carbon enhancement detectable outside the HERMES wavelength ranges, this would have to be taken into account to say more about the underlying population of carbon-enhanced stars. In addition to a detection bias imposed by the analysis of C_2 bands and exclusion of CN, and CH molecular bands that might be excitated in different temperature ranges, varying degree of carbon-enhancement also has to be accounted for accurate population studies. As shown by Yoon *et al.* [184], CEMP stars can be found within a wide range of absolute carbon abundances. When an object selection is performed with a pre-defined threshold, as in the case of our supervised methodology, this may reduce the number of objects in only one of the sub-classes. In the case of CEMP stars, this selection may influence a number CEMP-no stars that are known to have lower absolute carbon abundance [184].

The identified objects were separated into dwarf and giant populations using their stellar atmospheric parameters that were also used to select possible CEMP candidates. All of the detections, with multiple observations at different epochs, were investigated for signs of variability. More than half of the repeats show signs of variability in their measured radial velocities. This could be an indicator that we are looking at a pulsating object or a multiple stellar system.

With a follow-up observation of one of the identified stars, we were able to confirm the existence of carbon-rich molecules in its atmosphere in a wider wavelength range. The acquired spectrum was also used to determine its sub-class. Variation in radial velocity points to a possible variable nature of the star or binarity that is common for C-H stars.

Follow-up observations are required to confirm variability of radial velocities observed for some of the detected carbon-enhanced stars and further investigate their nature. Careful spectral analysis, with the inclusion of carbon enhancement in models, is needed to confirm the metallicity levels of the metal-poor candidates.

The list of detected stars presented in this chapter is accessible as electronic table through the CDS. Detailed structure is presented in Table 4.1. The list also includes stars from the literature, matched with our observations, for which we were unable to confirm their carbon enhancement. The list could be used to plan further observations, allowing a better understanding of these objects.

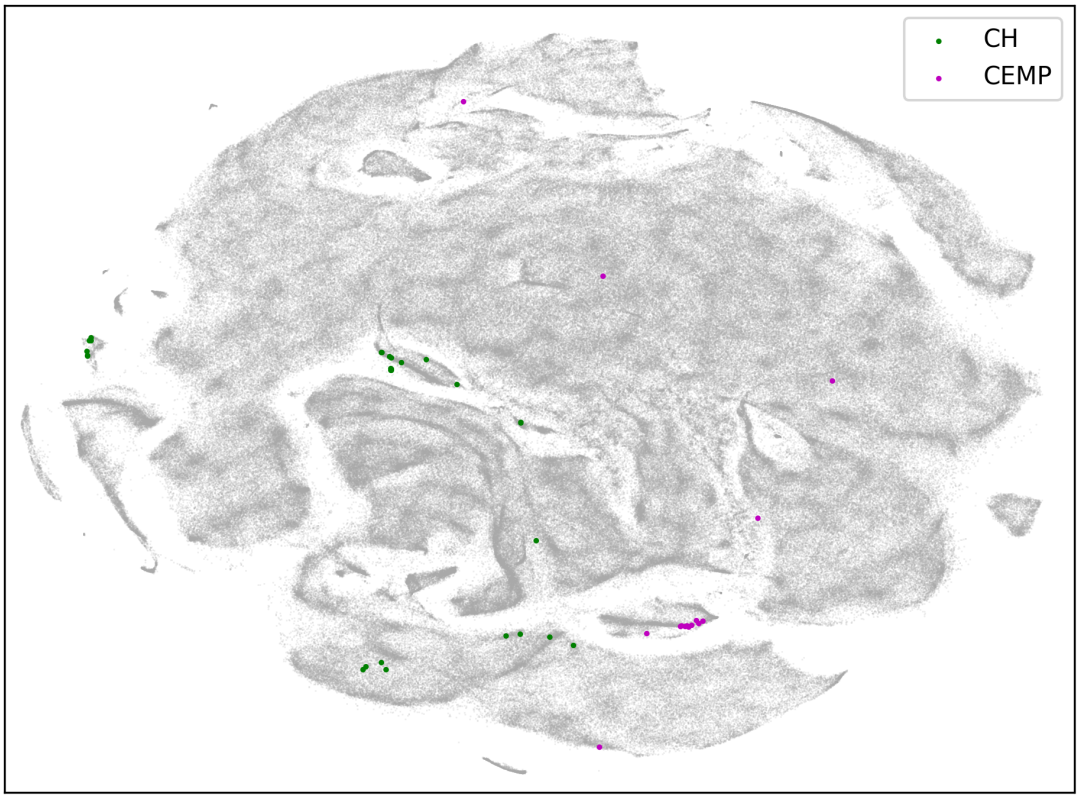

Figure 4.19: t-SNE projection with marked known carbon-enhanced and CEMP objects from multiple different catalogues found in the literature that are also part of our analysed set of spectra. The dense clump of known CEMP stars is located close to our region of detected CEMP stars.

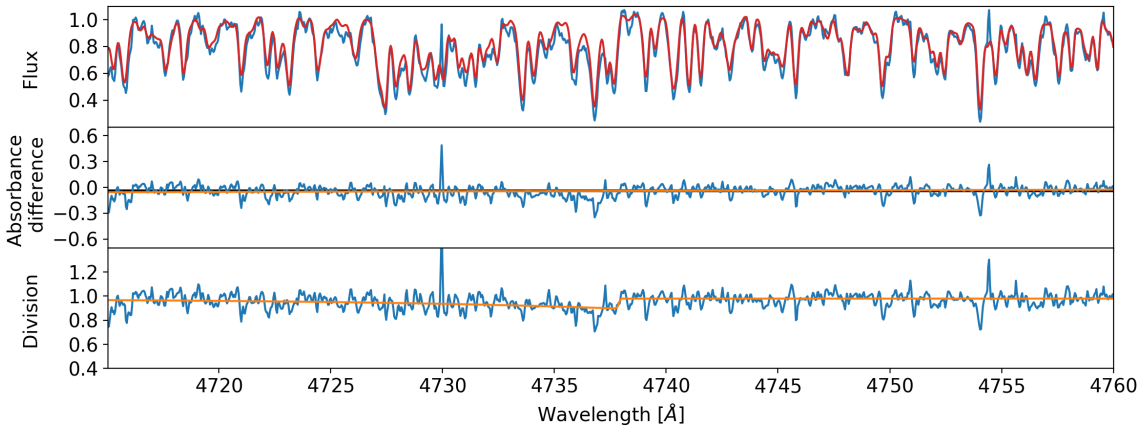

Figure 4.20: Equivalent plot as in the Figure 4.2 showing the last of 400 spectra, ordered by their degree of carbon enhancement, selected by the supervised methodology.

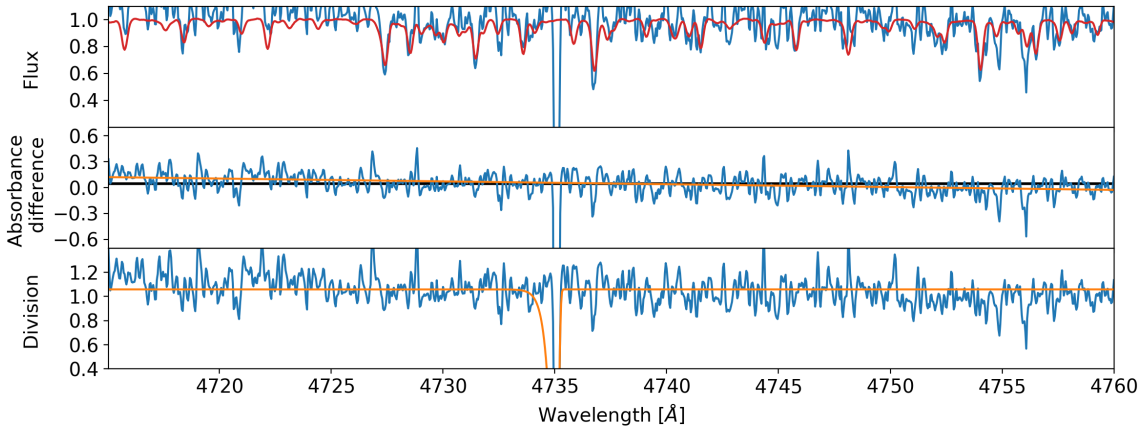

Figure 4.21: Equivalent plot as in the Figure 4.2 but representing grossly over exaggerated carbon enhancement by a fit that describes a reduction problem (a cosmic ray in a subtracted sky spectrum).

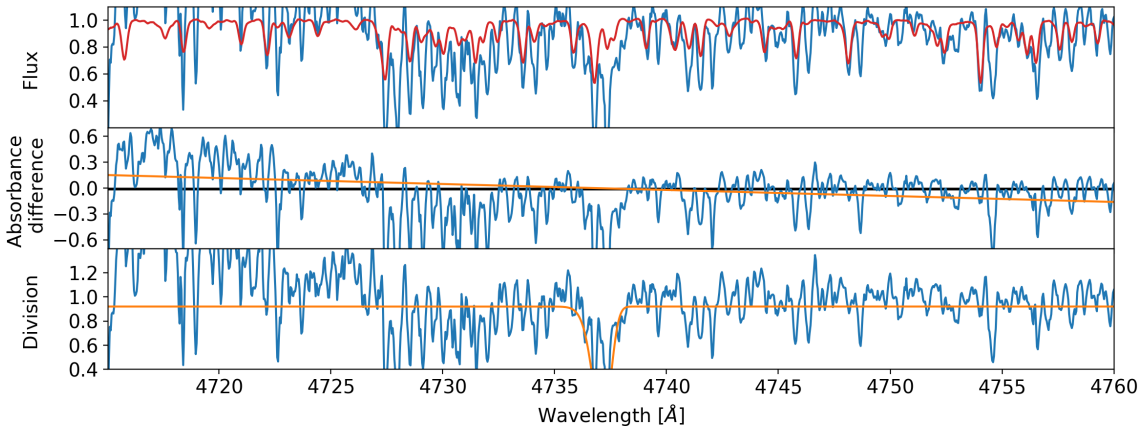

Figure 4.22: Equivalent plot as in the Figure 4.2 but representing a fit to absorption lines of a double-lined spectroscopic binary. Final fit is not skewed as would be expected in the case of carbon enhancement.

Chapter 5

Peculiar emission stars

This chapter has been adapted from the paper in preparation titled *The GALAH survey: Characterization of emission-line stars with spectral modelling using autoencoders* [261] whose first author is author of this Doctoral thesis. The used computer code is published on GitHub platform [1] and results of the analysis as a catalogue on the VizieR service [2].

Among all machine learning approaches, neural network structures are receiving the highest interest in all fields of big data analysis in recent years. Astronomy is no exception in this regard. In previous Chapter 4 we saw that peculiar stars are commonly detected by comparing observed spectra with reference spectra of stars that do not show any peculiarities in them. In this chapter we explore the autoencoder structure that provides us with the reference spectra for the comparison. Whenever we are interested in the precise elemental abundances (as used in Chapter 3), the first step in their determination of stellar physical parameters that could be based on the strongest hydrogen Balmer lines in stellar spectra. As any deviations in their shape might endanger the analysis, we tried to identify all stars whose spectrum shows emission-lines in the Balmer region. The chapter begins with Section 5.1 that gives a detailed description of the problem and presents multiple scenarios of why and how emission-lines can be used and identified. In Section 5.2 we explain our analysis pipeline whose main components are the generation of reference spectra (Section 5.2.1) and identification of multiple emission features (Sections 5.2.3 and 5.2.4). The temporal variability of detected emissions is analysed in Section 5.3. The results are summarised and discussed in Section 5.4.

5.1 Introduction

The identification of peculiar stars, whose spectra contain emission lines, is of interest to a wide field of stellar research. Spectral complexity of such stars brings insight into the ongoing physical processes on and around the star. Presence of emission lines hints to an optically thin material that surrounds a star. Such optically thin structures can be present at different evolutionary stages of the star. As those stages are temporally short compared to the stellar life span, they are regarded as peculiar at that time.

[1] `https://github.com/kcotar/GALAH-suervey-Emission-lines-and-autoencoders`
[2] link will be available after the paper is published

Emission features in stellar spectra might adversely impacted the quality of stellar parameters and abundances determined by automatic data analysis pipelines that are only configured to produce the best results for most common stellar types. Examples of how these features might compromise spectroscopic measurements when we assume that a star is not peculiar include the determination of effective temperature [262, 263, 264], computation of stellar mass [265, 266], and the effects of self broadening on line wing formation [267, 268]. Highly accurate measurement of the hydrogen absorption profiles are needed in those cases. Any deviations in the line shapes from model predictions would produce misleading results. We would therefore like to know if the investigated line is modified by additional, unmodelled physical process or spectral reduction process.

Stars with evident emission lines populate a wide variety of regions on the HR diagram. Because of possible overlaps between different stellar types, detailed photometric (especially in the infrared region where warm circumstellar dust disc can be identified) and spectroscopic observations are needed for an accurate physical explanation of the observed features. An examples of such work is presented in Munari *et al.* [168], who performed detailed a multi-band photometric study of an emission-line star, originally discovered on objective prism plates. The detailed photometric time-series study described in that work, together with observations of the star's infrared excess, led to the star VES 263 being identified as a massive pre-main-sequence star and not a semi-regular AGB cool giant as classified previously. In a similar way, Lancaster *et al.* [269] performed an analysis of the stellar object V* CN Cha, which had previously been identified as an emission star. By studying a long photometric time-series of the star, they concluded that the object was most likely a symbiotic binary star system whose emission was lined to a long-duration, low-luminosity nova phase.

Numerous different physical processes that can contribute to the complex shapes of the Hα emission profile are discussed by Reipurth *et al.* [270], Jones *et al.* [271], Silaj *et al.* [272], Ignace *et al.* [273], who compare observations with expected physical models. Following the classification scheme introduced by Kogure and Leung [274], emission-line stars are predominately observed in close binaries, earliest-type, latest-type, and pre-main sequence stars. For systems in which mass accretion is occurring, the examination of emission lines can allow the mass accretion rate onto the central star to be estimated [275, 276]. The procedure involves measuring simple indices (such as the equivalent width and broadening velocity) of the emission lines in the stars's spectrum.

In recent years, multiple dedicated photometric and spectroscopic surveys (e.g. [89, 277, 278, 279, 280, 281, 282]), and exploratory spectral classifications of large unbiased all-sky spectroscopic observational datasets (e.g. [84, 90, 282, 283, 284, 285]) have been performed, each finding from hundreds to tens of thousands of interesting emission-line stars. Some of these surveys provide a basic physical classification in addition to an emission detection. Therefore they can be used as source lists for further in-depth studies of individual stars.

If a star is engulfed in a hot rarefied interstellar medium or stellar envelope, emission features of the forbidden lines (the most commonly studied of which are the [NII] and [SII] lines) could be observed in its spectrum, providing an insight into the temperature, density, intrinsic movement, and structure of its surrounding interstellar environment [286, 287, 288, 289, 290].

Focusing on spectroscopic data, procedures for the detection of emission lines can roughly be separated into two categories. Simpler procedures searching for obvious emitters above the global continuum [90, 282, 282, 285] and more complex procedures, where the observed spectrum is compared to an expected stellar spectrum of a normal star [291]. The reference spectra in the latter case can be generated using exact physics-based stellar modelling or data-driven approaches. Of these, the data-driven approaches can be separated into supervised and unsupervised generative models, where, for the later, it is not required to provide an estimate of the stellar parameters for a given spectrum in advance. To predict a reliable model using supervised models, we must determine the correct stellar labels of an emission star in advance. This can pose a serious limitation if the strongest lines in the acquired spectrum can be populated by an emission feature, which happens for *Gaia* and RAVE spectra [291]. In light of the future publication of Gaia RVS spectra as part of Gaia DR3 for several million of stars, it is thus important to develop tools to identify emission-line stars, as we aim to do in this study via GALAH spectra.

5.2 Detection and characterization

The first attempts to discover Hα/Hβ emission spectra in GALAH survey observations were performed by Traven *et al.* [84], who also detected emission line stars in the Gaia-ESO [97] dataset using the Gaussian fitting and arbitrary thresholding [90]. Traven *et al.* [84] used the unsupervised dimensionality reduction technique t-SNE [130] to group morphologically similar spectra. As the amplitude and shape of the observed emission can vary substantially depending on the astrophysical source, Traven *et al.* [84] presumably detected only a portion of the strongest emitters. One of the reasons for this is the manual classification of data clumps determined by the clustering algorithm. In the case of weak emissions in an investigated clump (performed manually by the operator), an expressed emission feature must be strong enough to be visually perceived when looking at a spectrum. To broaden the range of detectability to include spectra with marginal levels of emission as well, a more sophisticated and partially supervised procedure must be employed.

To expand the search, our methodology uses additional prior knowledge about the expected wavelength locations of interesting emission spectral lines. The prior wavelengths are used to narrow down the interesting wavelength regions during the comparison between the spectrum of a possibly peculiar star and an expected (reference) spectrum of a star with similar physical parameters and composition.

5.2.1 Spectral modelling using autoencoders

A reference or a synthetic spectrum of a normal star without emission lines can be produced by a multitude of physics-based computational stellar models [292, 293, 294] or supervised generative data-driven approaches [105, 295], whose common weakness is the need for prior knowledge of at least an approximate stellar parameters of an analysed stars used by the data-driven algorithm.

As some of our spectra do not have determined stellar parameters or they are flagged with warning signs that indicate different reduction and analysis problems (missing infrared arm, various reduction issues, bad astrometric solutions, *SME* did not converge etc.), we focused on an unsupervised spectral modelling to produce

our set of reference spectra. Given the large size of available training data set, we chose to use an autoencoder type of an artificial neural network (ANN) that is rarely used to analyse astronomical data. Its current use ranges from data denoising [296, 297, 298] to unsupervised feature extraction and feature based classification [106, 299, 300, 301, 302, 303, 304].

An autoencoder is a special kind of ANN, shaped like an hourglass, that takes input data (a stellar spectrum in our case), reduces it to a selected number of latent features (a procedure known as encoding) and tries to recover the original data from the extracted latent features (decoding process). Our dense, fully connected autoencoder consists of the data input layer, four encoding layers, a middle feature layer, four decoding layers and the output layer. The number of nodes (or latent spectral features) in the encoding part slowly decreases in the following arbitrary selected order: 75%, 50%, 25%, and 10% of input spectral wavelength samples (4500 in the case of the red spectral arm). The exact numbers of nodes at each layer are shown in Figure 5.1. At the middle feature layer, the autoencoder structure reduces to only 5 relevant extracted features. Selecting a higher number of extracted features would also mean that the ANN structure could extract more uncommon spectral peculiarities which is not what we want. In our case, the goal is the reconstruction of a normal non-peculiar spectrum by extraction of a few relevant spectral features. Additionally, because of the low number of extracted features, our decoded output spectrum is a smoothed and denoised version of an input spectrum.

A visual representation of the described architecture is shown in Figure 5.1. The shape of the decoding structure of the autoencoder is the same, except in a reverse order. The Parametric Rectified Linear Unit (PReLU, [305]) activation function defined as

$$f(x) = \begin{cases} x, & \text{if} x > 0 \\ ax, & \text{if} x \leq 0 \end{cases} \tag{5.1}$$

is used for all nodes of the network, with the exception of the final output layer that uses a linear (i.e. identity) activation function. The x denotes one spectrum flux value in the first layer and one latent feature in the remaining layers. The free parameter a in Equation 5.1 is optimised during the training phase.

If the network learns a physics-based generative model of a stellar spectrum, information contained in the extracted features should be related to real physical parameters, such as $T_{\rm eff}$, $\log g$, [Fe/H], and $v \sin i$, or their mathematical combinations.

To train our autoencoder, we created a set of presumably normal spectra (with no emission features), resampled to a common wavelength grid ($\delta\lambda$ equal to 0.04 and 0.06 Å for the blue and red arm). Coverage of the grid is slightly wider than the range of an individual HERMES arm to account for variations in wavelength span because of stellar radial velocity and field curvature which slightly shifts wavelength span of every fiber on a CCD. Observations that did not completely fill the selected range were padded with continuum value of 1. To be classified as normal, spectra must suffice the following selection rules: signal to noise ratio (SNR) in the green arm must be greater than 30, a spectrum must not contain any known reduction issues (`red_flag` $= 0$ in Kos *et al.* [100]) and have valid spectral parameters (`flag_sp` < 16 in Buder *et al.* [166]). Although choosing `flag_sp` $= 0$ returns the spectra with the most trustworthy parameters, we choose to use this higher cutoff in `flag_sp` to filter out only the strangest spectra and not to produce a set of spectra with well defined

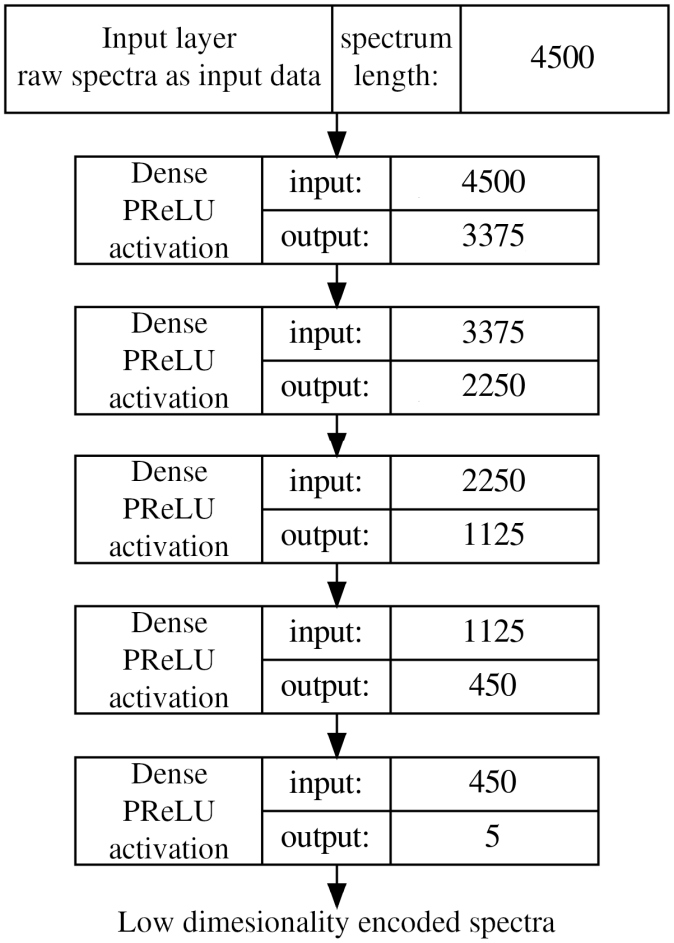

Figure 5.1: Visual representation of an encoder part of the used autoencoder structure for the red spectral arm. After the input spectra are encoded, they are passed through the same inverted architecture to produce modeled low-noise spectra. The value in the right most column indicates a number of input and output connections to neighboring layers. The number of nodes in a layer is equal to the output value. The input spectrum length is given as the number of wavelength bins in a spectrum.

parameters. Spectra with $0 <$ `flag_sp` < 16 include objects with bad astrometric solution, unreliable broadening, and low SNR that are still useful for our training process. From Traven *et al.* [84], Buder *et al.* [104] and Čotar *et al.* [180], we know that some GALAH spectra display peculiar chemical composition or consist of multiple stellar components. Therefore we removed all identified classes of peculiar spectra with the exception of stars classified as hot or cold that are actually treated as normal spectra in our case. Even such a rigorous filtering approach can miss some strange spectra.

After we applied these quality cuts, we were left with $482,900$ spectra, of which last 10% were used as an independent validation set during the training process. Before the training, normalised spectra were inverted (1 − normalised flux), which sets the continuum level to a value of 0. The inversion improved the model stability and decreased the required number of training epochs.

The described autoencoder was trained with the Adam optimisation algorithm [306] for 350 epochs. At every epoch all training spectra were divided into multiple batches of $40,000$ spectra, whose content is randomised at every epoch. A batch is a subset of data that is independently used during a training process. Such splitting and randomisation of training spectra into batches decreases the probability of model over-fitting. To enable the selection of the best network model, it was saved after the end of every training epoch.

The loss score minimised by the Adam optimiser, shown in Figure 5.2, was computed as a mean absolute error (MAE) between the input observed and decoded spectra defined as:

$$loss_{\mathrm{MAE}} = \frac{1}{N n_\lambda} \sum_{n=1}^{N} \sum_{i=1}^{n_\lambda} \left| f_{\mathrm{ae},n,i} - f_{\mathrm{obs},n,i} \right| , \tag{5.2}$$

where N is the number of all spectra, n_λ the number of wavelength bins in each spectrum, $f_{\mathrm{ae},n,i}$ the flux value of a decoded spectrum at one of the training epochs, and $f_{\mathrm{obs},n,i}$ the flux value of a normalised observed spectrum. Such a loss function gives lower weight to gross outliers in comparison to the mean squared error (MSE). At the same time, outputs are closer to a median spectrum of spectra with a similar appearance and less affected by remaining peculiar spectra in the training set.

After examining the decoded outputs at different epochs in comparison with known normal and peculiar spectra, we decided to use the model produced after 150 training epochs. After that, overall improvements of the model are minor, which increases the model opportunity to over-fit on a low number of peculiar spectra. After closer inspection of the last epoch, we found indications of over-fitting on known emission stars, which further confirms the validity of choosing a model with shorter training period (with greater prediction loss) and rejects the need for a longer model training.

To decrease the complexity of a dense neural network and reduce the required training time, two independent autoencoders were trained, separately for the blue and red HERMES spectral arms.

After the training and model selection were completed, all available $669,845$ spectra were run through the same autoencoder to produce their high SNR reference spectra. An example of four such spectra is shown in Figures 5.3 and 5.4.

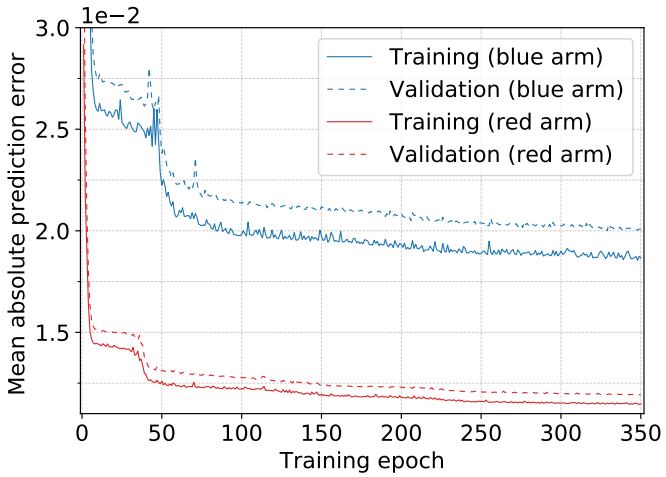

Figure 5.2: Prediction accuracy of the blue and red arm autoencoders at different training epochs. The prediction error is computed as a sum of all absolute differences between the input and output data set (see Equation 5.2). Shown are training (solid line) and validation curves (dashed line) which do not show any strong model overfitting on the training set. The curves indicate that both autoencoders learned in a similar way because the same optimiser was used. The blue arm model has a bit higher loss and shows slower learning because of a greater spectral complexity and lower signal to noise ratio in that wavelength region.

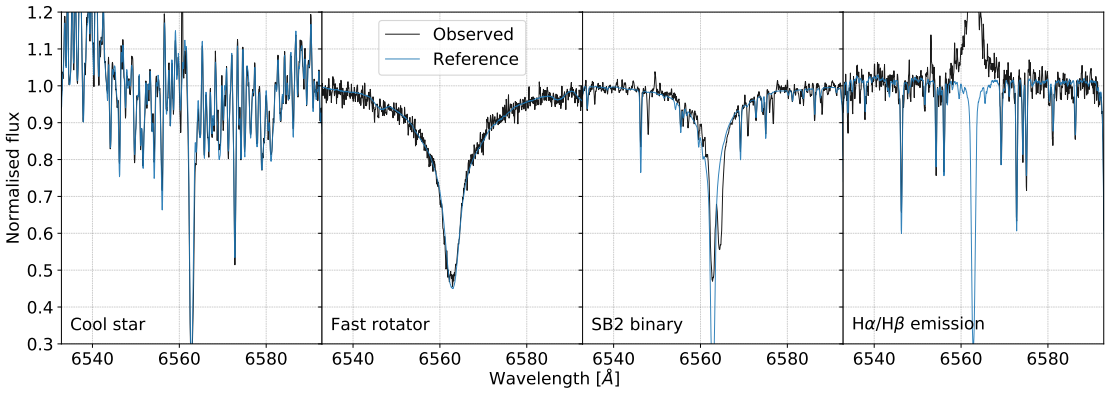

Figure 5.3: The diversity of spectra that must be processed by our reference spectrum generation scheme. Panels show spectra of the following normal and peculiar stars: cool, hot fast-rotating, spectroscopic binary, and Hα/Hβ emission star. All examples show that the autoencoder network did reproduce the observed shapes of the normal spectra (first two) and not the peculiar spectra (last two) as desired from the reference spectrum generator. The original spectra are shown in black and reconstructed in blue.

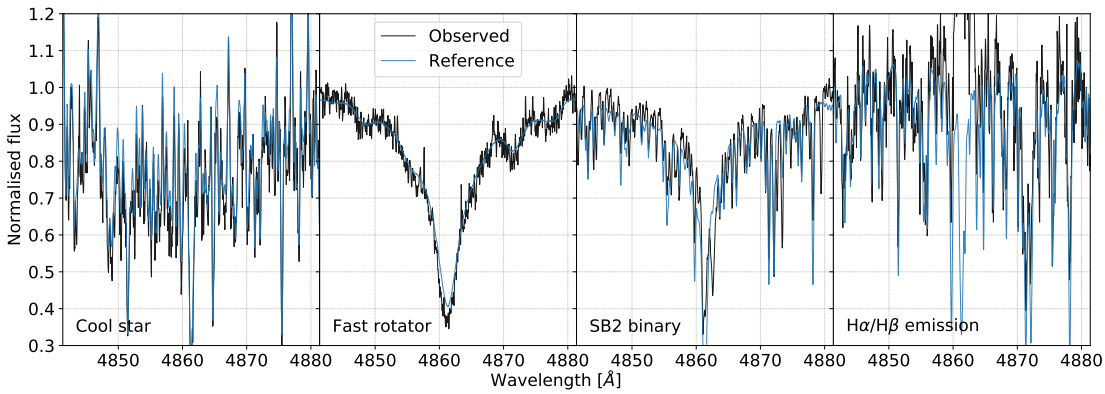

Figure 5.4: Same plots and objects as in Figure 5.3 but for the blue spectral arm.

5.2.2 Latent features

To test the idea of extracted scalar latent features being connected to physical parameters, and to inspect how an autoencoder structure actually orders spectra, we colour coded values of latent features by unflagged physical parameters of input the GALAH spectra. Latent feature scatter plots, colour coded by a different combination of stellar parameters, are presented in Figures 5.5 (with $T_{\rm eff}$ and $\log g$ for the red arm) as well as 5.6 (with $T_{\rm eff}$ and $\log g$ for the blue arm) and 5.7 (with $T_{\rm eff}$ and [Fe/H] for the blue arm).

As expected, all plots show continuous colour changes induced by the changing value of investigated physical parameter. This gives us a confirmation that the derived stellar physical parameters are spectroscopically meaningful and have the strongest influence on the appearance of acquired spectra. Rough physical parameters of previously unanalysed or peculiar spectra can therefore be acquired by averaging the parameter values of their neighborhood in the latent space. Similar procedures for parameter estimation have already been successfully explored by Yang and Li [106], Li *et al.* [299], Pan and Li [300].

5.2.3 Hα and Hβ emission characterization

The detection of emission components in spectra is based on a spectral difference $f_{\rm diff}$, computed as:

$$f_{\rm diff} = f_{\rm obs} - f_{\rm ref}, \tag{5.3}$$

where $f_{\rm obs}$ and $f_{\rm ref}$ are the observed spectrum and the generated reference spectrum respectively. The result of a computed difference $f_{\rm diff}$ for an emission spectrum is shown in the top panel of Figure 5.8. Ideally, this computation would enhance only mismatch between both spectra, with inclusion of spectral noise, if both represent a star with the same physical stellar parameters. During the initial processing, we found out that some observed spectra have slight normalisation problems, therefore we re-normalised them prior to difference computation. As the reference spectrum $f_{\rm ref}$ is known and has a continuum level close to a median value of similar stars in the training set, we first compute a spectral ratio $f_{\rm div}$, defined as:

$$f_{\rm div} = \frac{f_{\rm obs}}{f_{\rm ref}}. \tag{5.4}$$

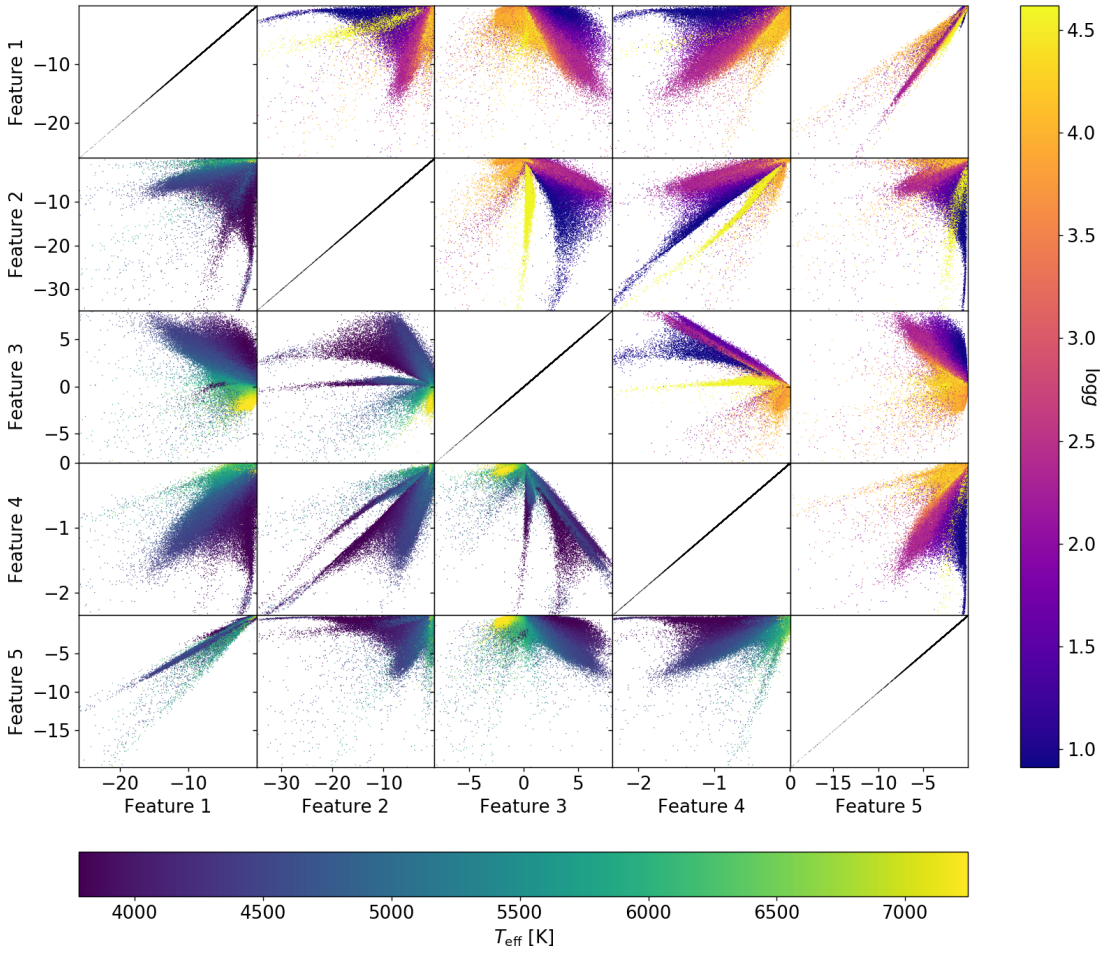

Figure 5.5: Correlation between extracted latent features and physical parameters. Scatter plots between different features are colored by the GALAH physical parameters of original spectra. Points in the lower triangle are colored by their T_{eff} and in upper triangle by their $\log g$. Associated colour mappings are given below the figure (for the lower triangle) and on its right side (for the upper triangle). Presented are results for the red arm autoencoder.

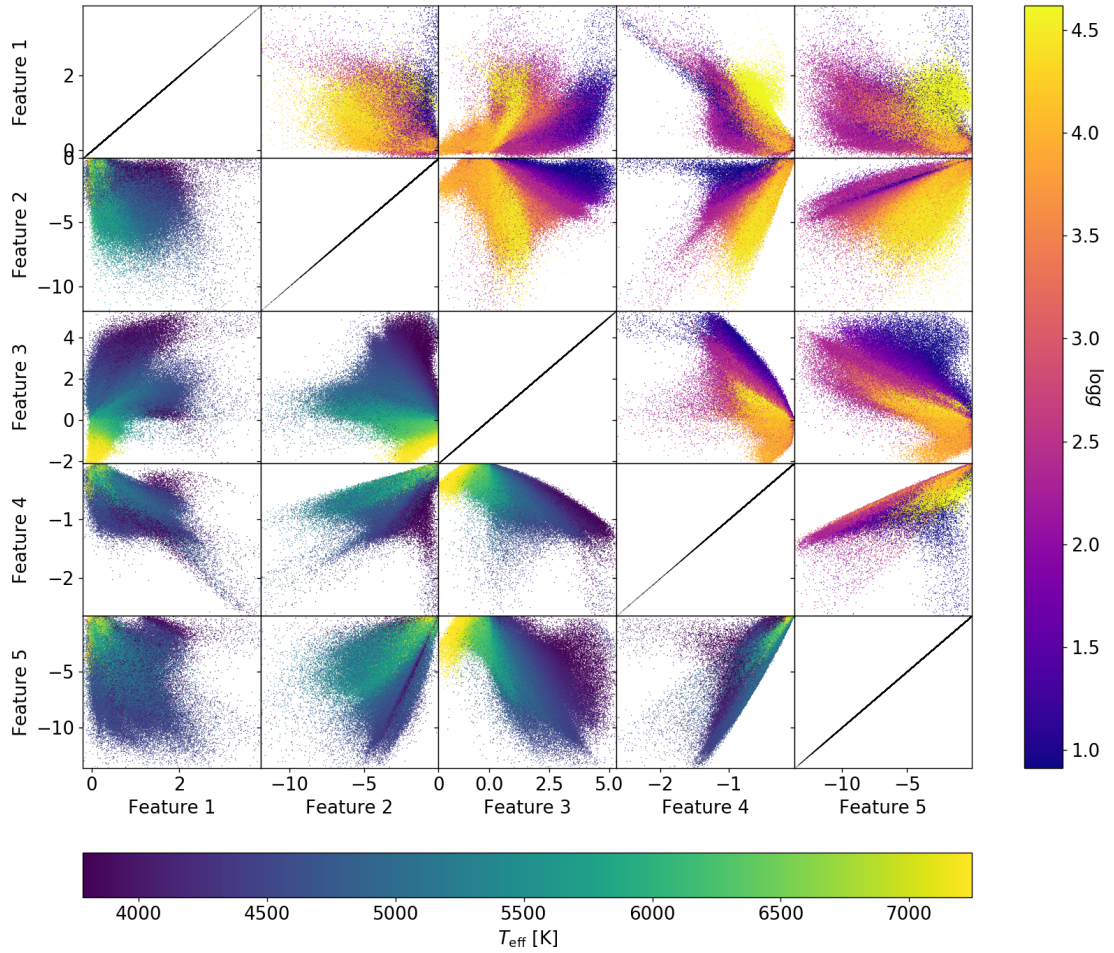

Figure 5.6: Same plots as shown in Figure 5.5, but for the latent features of the blue HERMES band, coloured by parameter $T_{\rm eff}$ on lower triangle and by $\log g$ on the upper triangle.

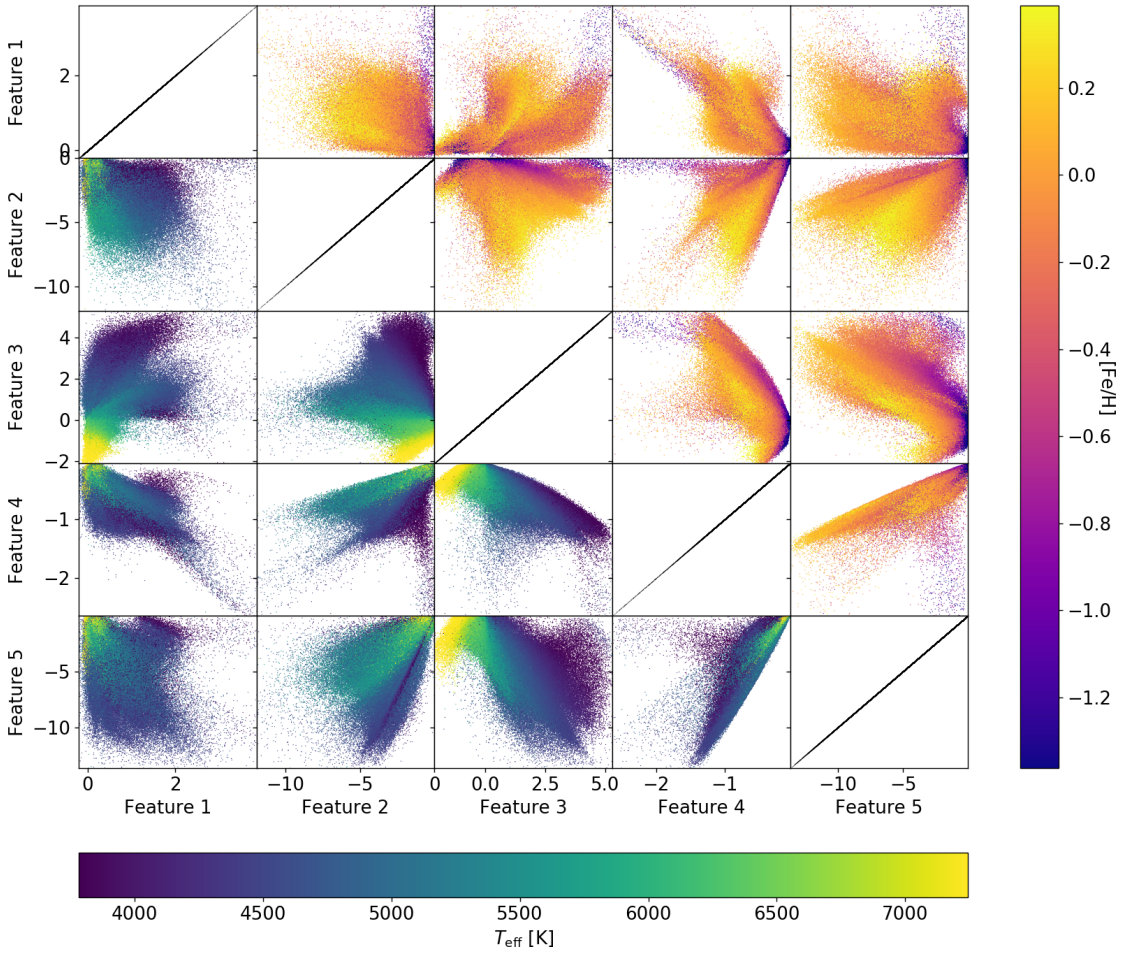

Figure 5.7: Same plots as shown in Figure 5.5, but for the latent features of the blue HERMES band, coloured by parameter $T_{\rm eff}$ on lower triangle and by [Fe/H] on the upper triangle.

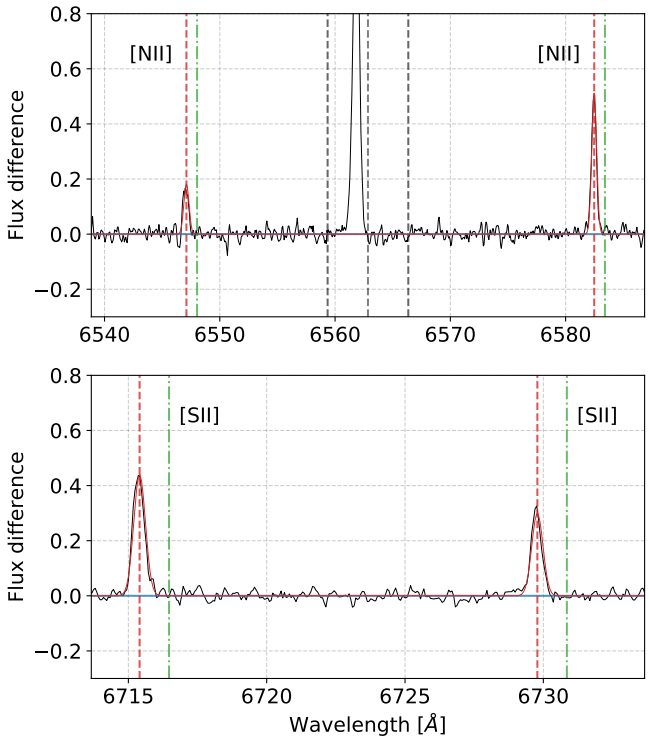

Figure 5.8: Panels show two different wavelength regions of f_{diff} for the same star. The top panel is focused on the Hα and [NII] nebular lines, while the second panel focuses on [SII] lines. Rest wavelengths of both nebular doublets are given by the green dash-dotted vertical lines. Their fitted locations, affected by a gas cloud movement, are given by the red dashed vertical lines that both have the same radial velocity. The constant integration range around EW(Hα) is bounded by the left and right black dashed vertical lines on the top panel. The middle black dashed vertical line represents Hα rest wavelength. All wavelengths are given in the stellar rest frame.

The resulting ratio can be viewed as a proxy for a renormalisation curve that would bring f_{obs} to the same continuum level as f_{ref}, but would at the same time cancel out any spectral differences between them. To avoid the latter, we fitted f_{div} with a 3^{rd} degree polynomial with a symmetrical 2-sigma clipping, ran for five iterations. We used the polynomial fit to renormalise f_{obs}.

To get the first identification of an emission features, we calculate the equivalent width (EW) of the spectral difference in a ± 3.5 Å range around the investigated Balmer Hα and Hβ lines. The selected range (shown in Figure 5.8) is wide enough to encompass emission profile of all spectra, with the exception of a few, which have very broad and structured profiles. We kept the width narrow to reduce the effect of spectral noise and nearby sky emission lines (see Section 5.2.5). The correlation between measurements of both equivalent widths is shown in Figure 5.9 from which it is evident that the Hβ emission feature is weaker than the Hα feature. This gradual intensity reduction is a well known relation also known as the Balmer decrement. As the decrement depends on many physical parameters of stars, absorbing medium, and type of the observed object, Bloom [307] collected multitude of known measurements. In our case the measured ratio EW(Hα)/EW(Hβ) has a value close to 3/2.

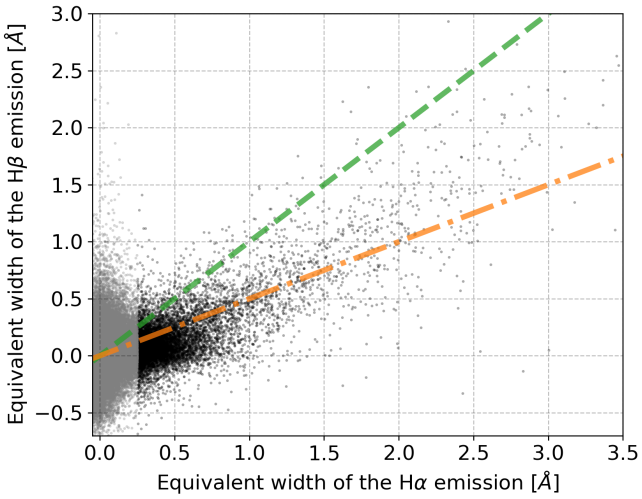

Figure 5.9: Correlation between equivalent widths of the Hα and Hβ emission components for our set of detected stars (defined as having `Ha_EW` > 0.25 Å) as black points. The remaining set of objects is shown with gray dots. All flagged objects and possible spectroscopic binaries are taken out for this plot. The green dashed linear line represent the one-to-one relation and the orange dash-dotted line identicates cases where the equivalent width of the Hβ is half of the Hα line.

The ratio seems to be independent of the Balmer line strength if the emission was detected. Low EW values in Figure 5.9 should not be taken in this approximation as they are burdened by the reference model uncertainty, spectral noise, and precise continuum levels of both spectra.

Alongside the equivalent widths of the residual components (EW(Hα) and EW(Hβ)), we also measured two additional properties of these lines, which give some insight into physical understanding of emission source. The broadening velocity of a line is described by its width at the 10% of the line peak (W10%(Hα) and W10%(Hβ)) expressed in $\mathrm{km\,s^{-1}}$. The automatic measurement procedure first finds the highest point inside the integration wavelength range and then slides down on both sides of the peak until the flux drops below 10% of the peak value. The broadening velocity is defined as a width between those two limiting cuts. As the computation is done for every object in an unsupervised way, the results are meaningful only for the spectra with evident emission lines. In the case when a low broadening velocity is estimated (equivalent to a very narrow peak), the highest peak could be a residual sky emission line or a cosmic ray streak. By combining EW(Hα) and W10%(Hα), mass accretion could be estimated if emission is of a chromospheric origin [308].

The second emission line index measured in the f_{diff} spectrum, that roughly describes the shape and location of an emission feature, is the asymmetry index defined as:

$$Asymmetry = \frac{|EW_{\mathrm{red}}| - |EW_{\mathrm{blue}}|}{|EW_{\mathrm{red}}| + |EW_{\mathrm{blue}}|}, \tag{5.5}$$

where $|EW_x|$ is the equivalent width of the absolute difference $|f_{\mathrm{diff}}|$ on the red and blue side of the rest wavelength of the investigated Balmer line. By this definition, a line that is, as a whole, moved to the redward side would have this index equal to 1, whilst if it was moved to the blueward side, the index would instead equal -1. The

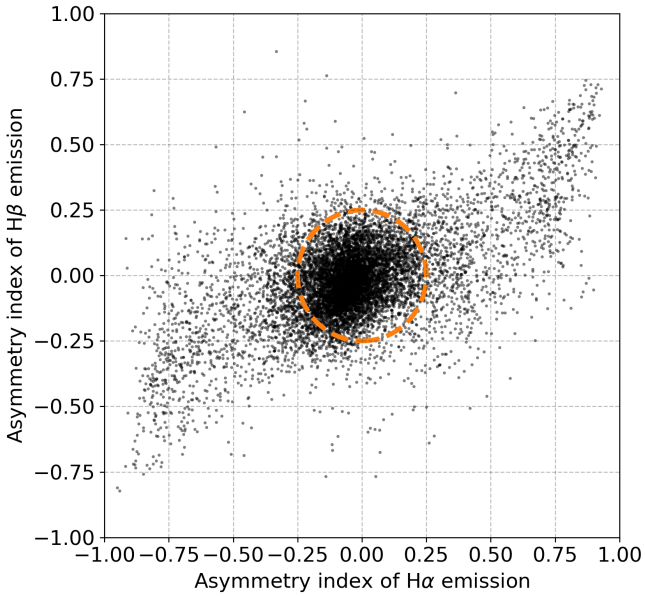

Figure 5.10: Asymmetry index of objects with prominent emission lines in the integration range around investigated Hydrogen Balmer lines. Objects with index inside the green dashed circle are considered to have a symmetric emission contribution, which can be attributed to a chromospheric activity. Central circular region has a radius of asymmetric index 0.25.

distribution of the asymmetry index values for the most prominent and unflagged (see Section 5.2.8) emitters is shown in Figure 5.10, where a strong correlation between the asymmetry of Hα and Hβ lines is evident. As the Hβ line in most cases produces a much weaker or even no emission feature, its asymmetry is much harder to measure. That is evident in Figure 5.10 where its index is scattered a around 0, except for the most asymmetric cases. The distribution of the Hα asymmetry is much more uniform outside the central symmetric region. From this index, we can roughly classify the source of the emitting component as a chromospheric origin would produce a centered component with an asymmetry index close to 0. Everything outside the central region in Figure 5.10, defined by the circle with a radius of 0.25, could be thought to be of an extra-stellar origin as lines are not perfectly aligned. The used thresholding radius value of 0.25 was defined by observing Figure 5.10 to encircle the main over-density of almost symmetric emission profiles.

5.2.4 Detection of nebular contributions

Due to the multiple possible origins of H emission lines [274], we also attempted to detect the extra-stellar nebular contributions of nearby optically thin gas. Its presence is expressed as forbidden emission lines in addition to the H emission. The spectral coverage of the HERMES red arm enables us to observe doublets of [SII] (6548.03 and 6583.41 Å), and [NII] (6716.47 and 6730.85 Å). Having usually a weak emission contribution that could possibly be blended with nearby absorption lines, they are most easily detected when we remove the expected reference spectrum from the observed one (resulting in f_{diff}). To automatically detect the emission strength and position of both doublets, we independently fitted two Gaussian functions with

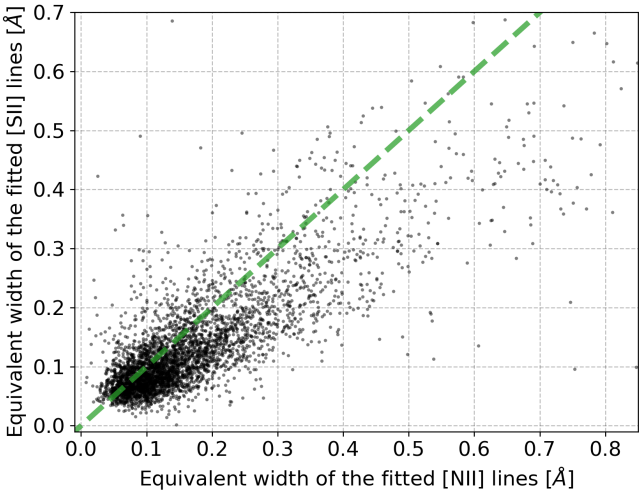

Figure 5.11: Correlation between the strengths of the nebular contributions from both elements. Shown are only cases with a small difference in the determined radial velocities as shown in Figure 5.12.

the same radial velocity shift for each element to f_{diff}. Because the contributing medium is not necessarily physically related to the observed object, its radial velocity could be different, therefore it was treated as a free parameter in our fit. Two independent velocities, one for each of the two doublets, give us an indication of a spurious or unreliable fit component if their difference is large. To filter out outliers, we adopted a threshold of 15 $\mathrm{km\,s^{-1}}$ on their velocity difference. Some of the discarded outliers might be correct detections because few of the spectra show two or more peaks for each nebular line which might point to a contribution of multiple clouds with different radial velocities. Such cases are not fully accounted for by the fitting algorithm that only identifies the strongest emission.

In the absence of additional fitting constraints, the routine might also find two noise peaks and lock onto them. Therefore, we put an arbitrarily selected detection threshold (0.05 of relative flux) on a minimum amplitude of the fitted forbidden lines to be counted as detected. The result from this fitting and analysis procedure is a number of successfully detected peaks per element and their combined equivalent widths (EW([NII]) and EW([SII])), reported in the final published table (Table 5.1). To filter out some possible miss-detection, we count a spectrum as having nebular lines when at least three nebular lines above the threshold were detected. The correlations for measured radial velocities and equivalent widths of identified objects with nebular emission are given in Figure 5.11 and 5.12 respectively.

The radial velocities of both doublets shown in Figure 5.12 give us a first impression that the gas dynamics of the elements in all observed clouds is nearly coincident, but elements are moving at slightly different velocities. This velocity offset, but in the opposite direction, was also observed by Damiani *et al.* [289, 290] who attributed it to the uncertainties in their adopted line wavelengths, that are slightly different to ours (less than 0.05 Å), causing the velocity points to be located either above or under the identity line in Figure 5.12. Additionally, the plot reveals that the majority of the gas clouds have a different radial velocity than stars behind or inside a cloud.

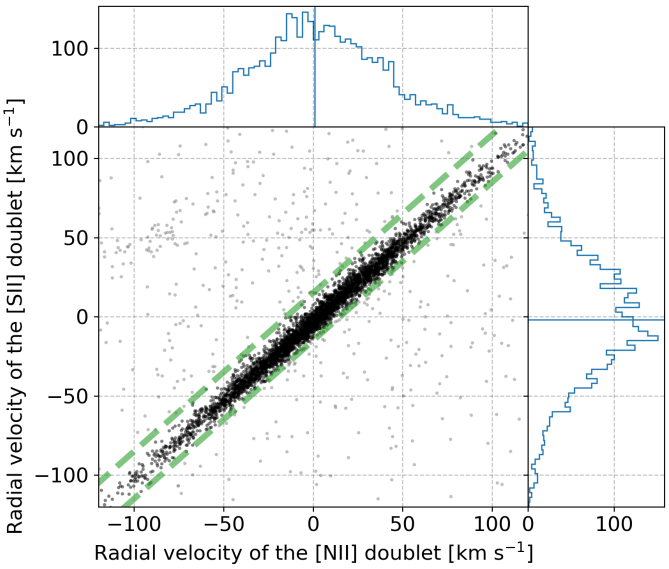

Figure 5.12: Correlation between radial velocity of both assessed nebular contributions that are observable in the red arm of the HERMES spectrum. Shown are only cases with at least three detected forbidden lines. The grey dots were further discarded as their absolute difference between velocities is more than 15 km s^{-1}. The limiting thresholds are visualized by dashed linear lines. Plotted velocities are measured in the stellar rest frame and therefore grouped towards zero velocity, meaning they are moving together with the star. Velocity distributions of selected measurements are presented around the main scatter plot.

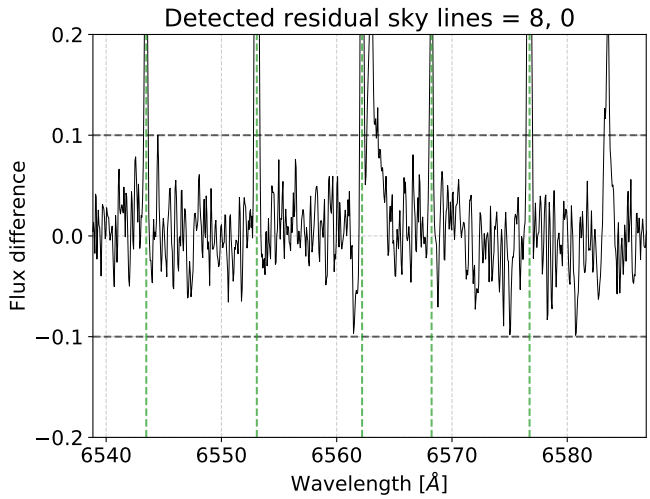

Figure 5.13: Sky emission lines are most evident after spectra subtraction in f_{diff}. Green vertical dashed lines represent expected locations of emission lines in the rest frame of an observed star. The middle sky line in this plot falls inside the actual Hα emission feature and changes its shape from single- to double-peaked and consequently modifies the measured equivalent width. Upper and lower thresholds for detection are given by the bold horizontal dashed lines. The number of detected under- and over-corrected sky lines in this order is given above the plot.

As we are working with fully reduced normalised spectra, with inclusion of sky background removal, the detection procedure would, in the case of an ideal background removal, not detect emission due to nebular clouds. As the measured flux of the nebular contribution is very unlikely the same for object and because of the physical separation of the sky fibres (see next Section 5.2.5 and Kos *et al.* [100]), the ideal cases are very rare. Similarly, the densities and the temperatures of such nebular clouds, extracted from corrected spectra could be influenced by the extraction pipeline and were therefore not performed in our case.

The strength of the identified lines, measured by their equivalent widths, is shown in Figure 5.11. This shows a high degree of correlation, where on average [SII] lines have lower strength than [NII] lines. Rough estimation of ratio between their measured equivalent widths EW([NII])/EW([SII]) is close to the value of 4/3.

5.2.5 Identification of sky emission lines

Attributing a limited and relatively low number of the HERMES fibres to monitor the sky in hopefully star and galaxy free regions, imposes limitations to a quality of the sky background removal in the GALAH reduction pipeline [100]. As the sky spectrum is sampled at 25 distinct locations over the whole 2° diameter field, it must be interpolated for all other fibre locations that are pointing towards stellar sources. Depending on the temporal and spatial variability of weather conditions, and possible nebular contributions, interpolation may produce an incorrect sky spectrum that is thereafter removed from the observed stellar spectra.

In most cases, this does not influence the spectral analysis, unless one of the strongest sky emission lines falls in a range of the analysed stellar line. For us, the

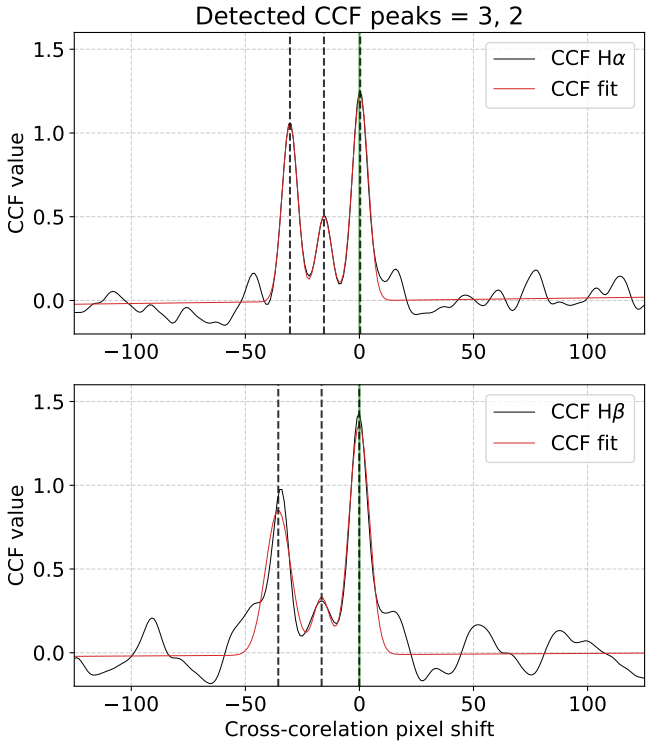

Figure 5.14: Detection of a spectroscopic binary candidate by cross-correlating observed spectrum with its reference spectrum. Three Gaussian functions that are fitted to the resulting CCF (black solid curve) are depicted by their means (dashed vertical lines) and their best fitting sum in red. Presented are CCFs for the red arm in the top panel and for the blue arm in the bottom panel. Number of detected peaks for both arms is given above the figure. The middle marked peak in the bottom plot was not detected due to its low fitted amplitude that could hint to a feature caused by a spectral noise. Peaks of similar amplitude, that could cause wrong binarity determination, are seen left and right of central CCF region.

most problematic sky emission line, which can alter the shape of the Hα profile, is located at 6562.7598 Å (used list of sky emission lines was taken from Hanuschik [309]). Being close to the adopted wavelength of the Hα (6562.8518 Å), it can get blended with a real emission feature or simulate its presence. We tried to estimate the impact of the sky residual in the spectrum from multiple nearby emission lines. First, we select only the strongest sky emitters (with parameter `Flux` ≥ 0.9 in Hanuschik [309]) and shift their reference wavelength into a stellar rest frame. After that, we use a simple thresholding (see Figure 5.13) to estimate their number. By the thresholding procedure, we want to simultaneously catch over- and under-corrected stellar spectra.

When a sufficient number (≥ 4) of strong residual sky lines with a normalised flux above 10% is detected, a quality flag (see Section 5.2.8) is raised, warning a user that the equivalent width of the Hα emission could be affected by uncorrected sky emission. As this potential contamination is present only in the red HERMES arm, we do not check for spurious strong emitters in the region around the Hβ line.

5.2.6 Determination of spectral binarity

During the inspection of our initial results, we noticed that the spectra of spectroscopically resolved binary stars (SB2) produce a mismatch between observed and reference spectra whose f_{diff} have a profile similar to the P Cygni or inverted P Cygni profile [310] that is often observed in emission-line objects. To detect SB2 candidates, we performed cross-correlation between the reference and the observed spectra, disregarding the wavelength range of ± 10Å around the centre of the Balmer lines to avoid broadening of the cross-correlation function (CCF) peak. Cross-correlation was performed independently for both (the blue and red) HERMES spectral arms. The resulting CCF, shown as the black curve in Figure 5.14, was fitted by three Gaussian functions, centred at three strongest peaks, to describe its shape. The location, amplitude and width of those peaks were assessed to determine the number of stellar components in the spectrum. When fitting three peaks, there is a possibility of finding triple stars and distinguishing them from binaries. Every spectral arm with more than one prominent peak was marked as potential SB2 detection in the final results (see Table 5.1), where binarity indication is given independently for both arms. Nevertheless, the results of the blue arm (column `SB2_c1`) are more trustworthy because of the higher number of absorption lines in the red arm (column `SB2_c3`). For even greater completeness of detected SB2 candidates, a list of analyzed binaries, compiled by Traven *et al.* [311] can be used. They combined unsupervised spectral dimensionality reduction algorithm t-SNE and semi-supervised CCF analysis [312] to compile their list of SB2 binaries. After their analysis, they discarded spectra that were falsely identified as SB2 by their detection procedures.

An unexpected result of this binarity search was the realization that some reduced spectra show duplicated lines only in the red arm or even stranger, only in a smaller subsection of it. After a thorough investigation, we uncovered that this effect is caused by improper treatment of fibre cross-talk while extracting spectra from the original 2D image [100]. A partial culprit of this is also a poorer focus in the red arm. Therefore if only flag `SB2_c3` is set, and not `SB2_c1`, this can be used as an indication of the above reduction effect.

Additionally, the highest peak of our CCF function is used to determine the correctness of the wavelength calibration during the reduction of the spectra [100]. If the peak is shifted by more than five correlation steps (maximum shift equals to about 13 $\mathrm{km\,s^{-1}}$) from the rest wavelength of the reference spectrum, the quality flag (see Section 5.2.8) is raised, warning the user that the derived radial velocity, equivalent width, and asymmetry index might be wrong in the respective arm as both spectra were not aligned ideally.

5.2.7 Resulting table

The emission indices and other computed parameters are collected in Table 5.1. The complete table is available in electronic form at the CDS. An excerpt of the published results, containing a subset of 30 rows and 11 most interesting columns for the strongest unflagged emitters is given in Table 5.2.

As we do not perform any quality cuts on our results, a suggested set of limiting parameter thresholds and quality flags is provided in Section 5.2.8. Their use depends on user-specific requirements and the analysed science case.

Table 5.1: List and description of the fields in the published catalogue of analysed the GALAH spectra.

Column	Unit	Description
`source_id`		*Gaia* DR2 star identifier
`sobject_id`		GALAH internal per-spectrum unique id
`ra`	deg	Right ascension coordinate from Two Micron All-Sky Survey (2MASS, [15
`dec`	deg	Declination coordinate from 2MASS
`Ha_EW`	Å	Equivalent width of a difference between observed and template spectrum in the range of ±3.5 Å around the Hα line
`Hb_EW`	Å	Same as the `Ha_EW`, but for the Hβ line
`Ha_EW_abs`	Å	Equivalent width of an absolute difference between observed and template spectrum in the range of ±3.5 Å around the Hα line
`Hb_EW_abs`	Å	Same as the `Ha_EW_abs`, but for the Hβ line
`Ha_W10`	$\mathrm{km\,s^{-1}}$	Width (in $\mathrm{km\,s^{-1}}$) of the Hα emission feature at 10% of its peak flux amplitude
`Ha_EW_asym`		Value of asymmetry index for the Hα line
`Hb_EW_asym`		Value of asymmetry index for the Hβ line
`SB2_c3`		Was binarity detected in the red arm
`SB2_c1`		Was binarity detected in the blue arm
`NII`		Number of detected [NII] peaks in the doublet
`SII`		Number of detected [SII] peaks in the doublet
`NII_EW`	Å	Combined equivalent width of a fitted Gaussian profiles to both studied [NII] emission features
`SII_EW`	Å	Same as the `NII_EW`, but for the [SII] doublet
`rv_NII`	$\mathrm{km\,s^{-1}}$	Intrinsic radial velocity of the [NII] doublet, corrected for the barycentric and stellar velocity
`rv_SII`	$\mathrm{km\,s^{-1}}$	Same as `rv_NII`, but for the [SII] doublet
`nebular`		Is spectrum considered to have an additional nebular component
`emiss`		Is spectrum considered to have an additional Hα emission component
`flag`		Sum of all bitwise flags raised for a spectrum

Table 5.2: Excerpt of 30 strongest unflagged emitters from the published table presented in detail by Table 5.1. The rest of the table can be downloaded in electronic form CDS service and publishers' website.

source_id	Ha_EW	Ha_EW_abs	Ha_W10	Ha_EW_asym	NII	SII	NII_EW	rv_NII	rv_SII	flag
3337923100687567872	5.37	5.37	373.63	0.36	1	0	0.05	-30.22	23.54	0
3217769470732793856	5.06	5.06	252.48	0.07	0	1	0.01	-11.32	35.71	0
4660266122976778240	4.51	4.51	199.66	0.15	0	0	0.02	-310.50	-231.18	0
3340892714091577856	4.35	4.35	205.94	-0.06	2	2	0.20	-24.73	-27.14	0
3336365097008009216	4.14	4.14	188.62	-0.08	0	0	0.07	-76.99	-43.80	0
3217804483306125824	4.02	4.02	152.24	-0.03	0	1	0.01	-85.06	28.70	0
3214742618300312064	3.97	3.98	277.85	-0.08	0	0	0.01	-63.06	-32.82	0
6243142063220661248	3.87	3.87	120.97	-0.07	2	1	0.08	7.21	15.98	0
2967553747040825856	3.80	3.80	330.27	0.07	0	1	-0.00	-66.11	-10.92	0
5948023586013872128	3.72	3.72	277.77	0.17	0	0	0.00	-72.55	-127.22	0
5416221633076680704	3.65	3.65	126.59	-0.08	0	0	-0.01	69.68	19.93	0
3222267297922229248	3.64	3.64	265.38	-0.07	0	0	0.04	-60.66	13.30	0
6245775565362814976	3.59	3.59	133.82	-0.08	0	0	-0.07	194.28	59.86	0
3235905365276381696	3.52	3.82	211.08	-0.43	1	0	0.05	-4.52	57.50	0
3236272877038986240	3.47	3.47	141.67	-0.06	0	0	0.06	-50.01	10.89	0
5200035927402217472	3.46	3.46	148.65	-0.03	1	0	0.04	-89.98	15.81	0
5820283738165246976	3.45	3.45	380.58	-0.08	0	0	-0.02	55.83	78.12	0
3222024374573501952	3.37	3.37	224.71	-0.11	0	1	0.00	-0.04	17.76	0
3221019798902558720	3.37	3.37	142.94	-0.07	0	0	0.03	-64.74	48.73	0
6235172592479759360	3.32	3.32	147.53	0.02	0	0	0.01	-2.29	46.61	0

5.2.8 Flagging, quality control and results selection

The above described pipeline runs blindly on every successfully reduced spectrum (`guess_flag` = 0, for details see Kos *et al.* [100]), and could therefore produce wrong or misleading results for some spectra. To have the ability to filter out such possible occurrences, we created a set of warning flags for different pipeline steps that are listed and described in detail in Table 5.3. An interested user can base their selection of results according to the desired confidence level and a physical question of interest. The cleanest set of 10,364 Hα emission stars can be produced by selecting unflagged stars that do not show any signs of possible binarity, defined such that parameter `emiss` in the published Table 5.1 is set to one (the equivalent of true). To be included among the cleanest set of detections, we considered only spectra whose `Ha_EW` > 0.25 Å. Below this limit, we are less confident in marking an object as having an emission feature because visual inspection showed that this strength could be mimicked by spectral noise, the uncertainty of the reference spectrum, or induced by the reduction pipeline. This selection criteria at the same time discards the weakest chromospheric components, which might be of great interest for specific studies. If the user is interested only in stronger emitters, the threshold should be raised to `Ha_EW` > 0.5 Å or above.

The published Table 5.1 also contains a flag that describes whether the spectrum is considered to contain an additional nebular contribution. Such spectra can be filtered out by choosing the parameter `nebular` to be equal to 1. To compile this less restrictive list of 4004 spectra, we selected entries with at least three prominent forbidden emission lines (`NII` + `SII` ≥ 3) and a small difference in their measured radial velocities ($|$`rv_NII` $-$ `rv_SII`$| \leq 15$ km s^{-1}).

5.3 Temporal variability

The strategy of the GALAH survey is to observe as many objects as possible, and as a result, not many repeated observations were made. The repeated fields were mostly observed to assess the stability of the instrument. Time spans between observations are therefor on the orders of days or years. This greatly limits the possibility of finding a variable object, but still enables us to discover potential interesting objects and diagnose analysis issues.

To find possible emission stars with repeated observations, we selected stars with repeats, among which at least one spectrum was identified to harbour a stronger (`Ha_EW` > 0.5 Å) unflagged emission feature. This selection produced 621 stars, having between 2 and 9 observations. To be confident about the observed variability, we visually inspected the observed and the reference spectra of 208 stars with at least three observations. A subset of these spectra are shown in Figure 5.15, where we present typical types of variability discovered by visual inspection. The types can roughly be described as shape transformation (e.g. change from single- to double-peak or P Cygni emission profile), peak location shift, intensity change, and possible reduction issue.

In the sample of 208 stars, whose spectra were visually inspected, we found that $\sim 20\%$ of the inspected spectra display a stable Hα profile. Noticeable profile shape transformation was observed in $\sim 10\%$ of the cases, and peak location change in $\sim 5\%$ of the cases. Some degree of emission intensity change was noticed for $\sim 40\%$

Table 5.3: Quality binary flags produced during different steps of our detection and analysis pipeline. Lower value of the flag represents lower significance to the quality of detection and classification. The final reported `flag` value in Table 5.1 is a sum of all raised binary quality flags.

Flag	Description
128	Reference spectrum for the Hα range does not exist.
64	Reference spectrum for the Hβ range does not exist.
32	Large difference between reference and observed spectrum in the red arm of a spectrum. Median squared error (MSE) between them was ≥ 0.002
16	Large difference between reference and observed spectrum in the blue arm of a spectrum. MSE was ≥ 0.008.
8	The spectrum most likely contains duplicated spectral absorption lines of a resolved SB2 binary. Binarity was detected in both arms.
4	Possible strong contamination by sky emission features. 4 or more residual sky lines were detected. Could be a result of an under- or over-correction.
2	Wavelength solution (or determined radial velocity) might be wrong in the red arm of the spectrum. Determined from cross-correlation peak between observed and reference spectra.
1	Wavelength solution (or determined radial velocity) might be wrong in the blue arm of the spectrum.

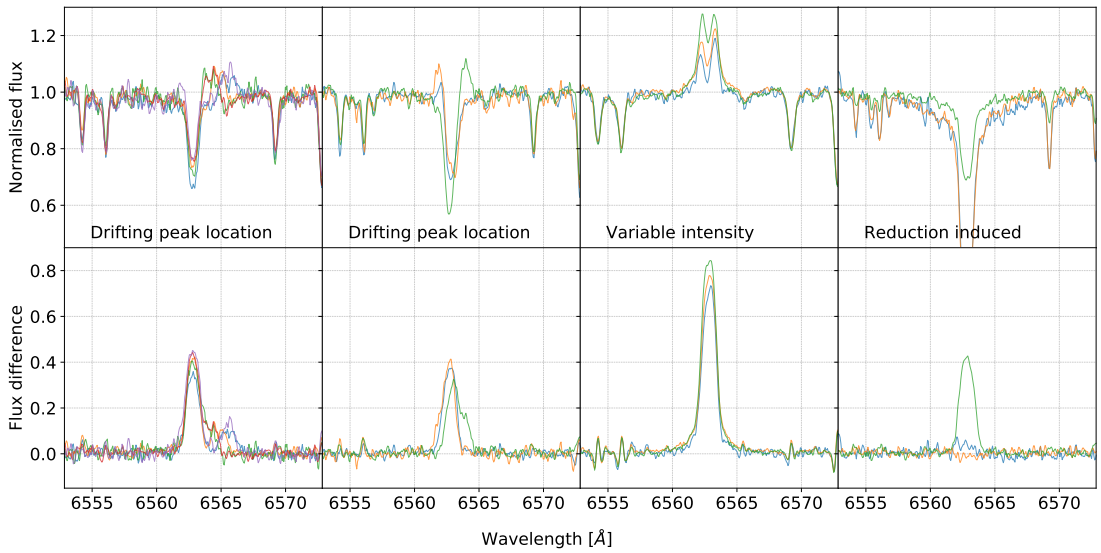

Figure 5.15: A sample of objects with repeated observations, where at least one of the normalised spectra (top row) contains strong Hα emission detected by comparison towards reference spectrum (bottom row). The first two objects (or columns) show shifting location of an additional emission component peak, and the last two varying degree of its strength. The last example is most likely a result of a miss-reduction as not only Hα, but also other absorption lines show reduced strength. The existence of this problem is confirmed by other objects in the same field as majority of them show the same tendency of having weaker absorption lines across the spectrum.

of the cases. Visually similar is reduction induced variability (see the rightmost panel in Figure 5.15), observed for $\sim$ 25% of all inspected repeated observations. In the case of multiple observations of the same star, we can distinguish between the last two profile changes (intrinsic and reduction induced intensity change) by looking at the whole spectrum to inspect whether variability is also exhibited in other absorption lines as shown by the last example in Figure 5.15. That kind of reduction induced variability is limited to a few observed fields.

5.4 Discussion and conclusions

In this chapter, we describe the development and application of a neural network autoencoder structure that is able to extract the most relevant latent features from the spectrum. Low feature dimensionality contains only the most basic spectral informations that are used to reconstruct a non-peculiar spectrum with the same physical parameters as the input spectrum.

Our method of differential spectroscopy is one of the most widely used approaches to find peculiar spectral features that are not found in normal stars. As a part of this chapter, we showed that a dense autoencoder neural network structure can be reliably used for generation of non-peculiar reference spectra if trained on a large set of normal spectra. With the additional exclusion of our detected emission-line stars, the training set could iteratively be further cleaned of peculiar stars before training the network. As all the information about the spectral look is contained in the real flux values, there is no need to add additional convolutional layers for the

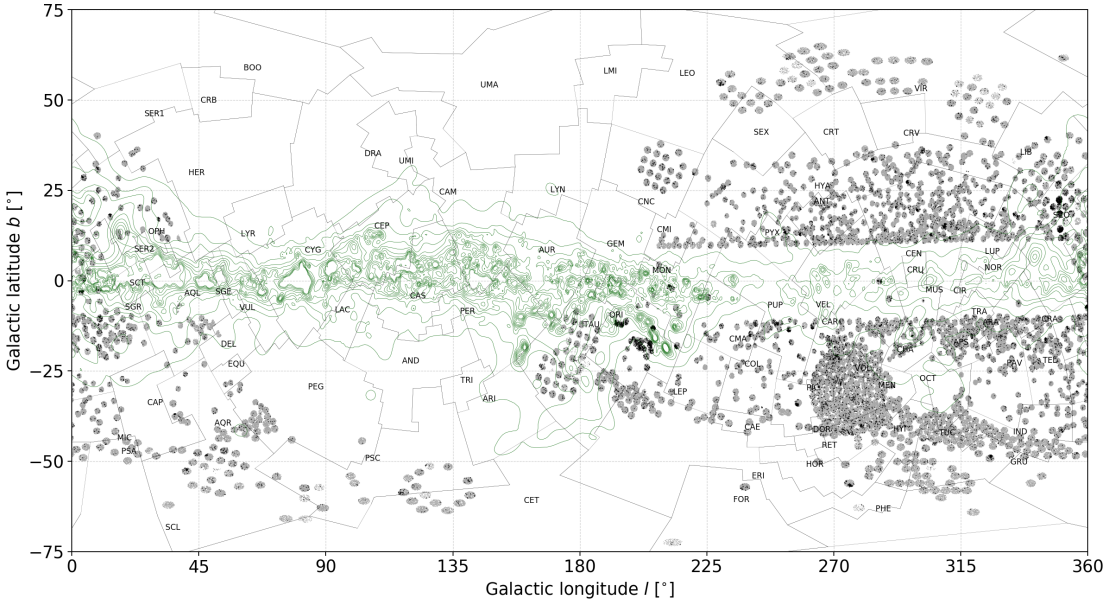

Figure 5.16: Spatial distribution of stars with detected Balmer emission profiles. Grey areas represent regions that were observed and analysed in this chapter. The green lines represent location of equal reddening in steps of 0.1 magnitude at the distance of 2 kpc. Reddening data were taken from results published by Capitanio *et al.* [313]. For readability, no isoline is shown above the reddening of 1 magnitude. Constellation boundaries were taken from Davenhall and Leggett [314]. Locations of their designations are defined by median values of constellation polygon vertices.

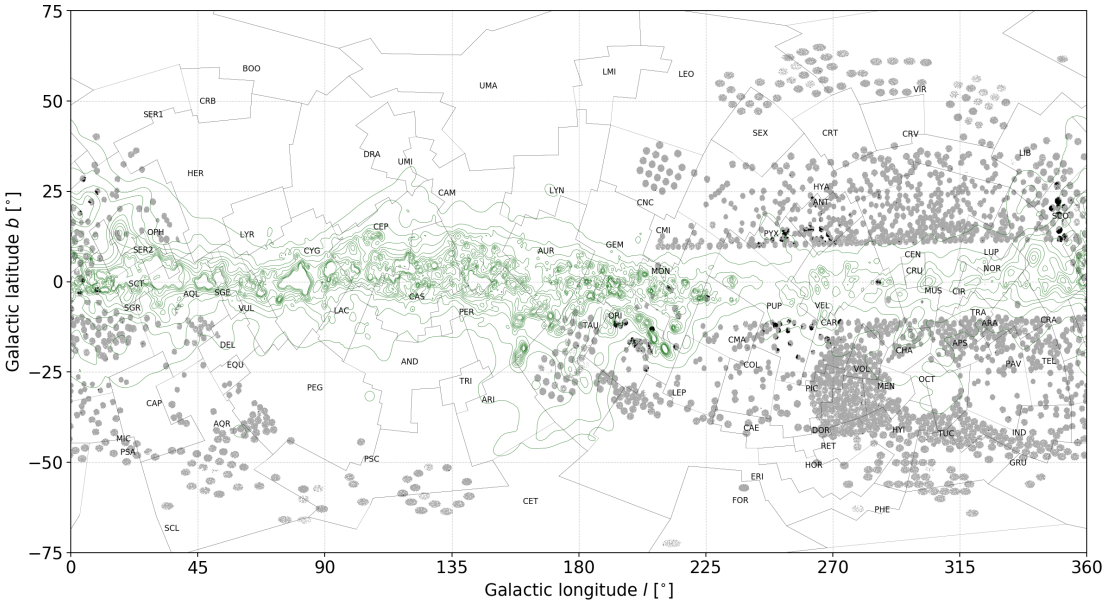

Figure 5.17: Same as Figure 5.17 but showing stars with at least three detected nebula emission lines, shown with black dots.

extraction of more complex spectral shapes.

By identifying significant residuals after subtracting the generated reference spectra from the observed spectra, we detected emission star candidates in the GALAH fields all over the sky. Figure 5.16 shows that we can identify few locations with a higher density of detected emission-line objects. The position of emission-line objects coincides with regions of young stars such as the Orion complex, Blanco 1, Pleiades, and other possibly random over-densities of interstellar gas and dust. Detected nebular emission in stellar spectra, shown in Figure 5.17, coincide with large visually-identified nebular clouds (by comparing detected locations with the red all-sky photographic composite of The Second Digitized Sky Survey, described by McLean *et al.* [315]) such as the Antares Emission nebulae, clouds around π Sco and δ Sco, Barnard's loop, Carina Nebula, nebulae around λ Ori, nebular veils in the constellations of Puppis, Pyxis and Antlia, and other less prominent features.

By combining our detections with additional auxiliary data sets, we can start exploring more detailed physical explanations of the observed emissions and their structure. Among them are two specific photometric surveys, VPHAS [280] and IPHAS [277] which were designed to detect and study emission-line sources close to the Galactic plane. Because of their positional selection function, their combined photometric data are available only for 4431 GALAH spectra. Of these, the spectroscopically confirmed emission stars are shown in Figure 5.18, whose color-color diagram can be used to infer accreting objects.

Our detected emission spectra have a broad range of emission components - these range from very strong to barely detectable chromospheric emission component whose identification can be mimicked or masked at multiple steps of the analysis and data preparation. To limit the number of false-positive classifications due to reduction and analysis limitations, we focused on stronger components (`Ha_EW` > 0.25 Å) whose existence can be confirmed visually. Because that kind of process would be slow for the whole sample, we introduced quality flags that can be used to filter

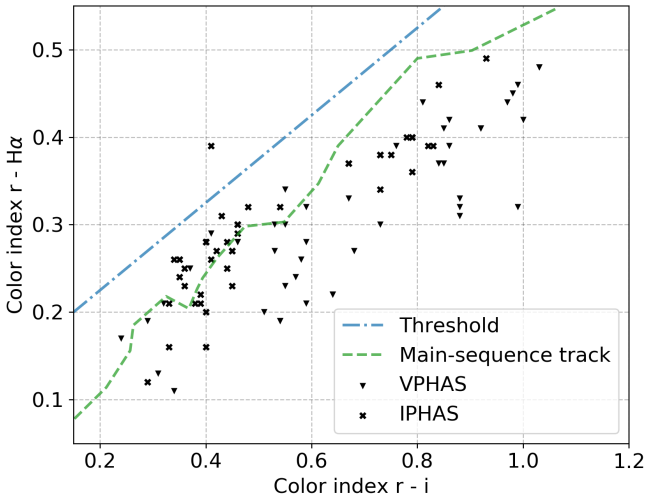

Figure 5.18: r - Hα versus r - i color-color diagram using combined IPHAS and VPHAS photometry for detected emission stars. The dashed green line represents the unreddened main-sequence track tabulated by Drew *et al.* [280]. An empirical threshold, shown by dash-dotted blue line, can be used to distinguish non-accreting and accreting objects above the line [316].

out unwanted or specific cases. Additionally, the stability of the spectra and emission features was investigated by repeated observations of the same objects. Among them, we observed different variability types, of which one could be attributed to the data reduction pipeline, limiting the confidence of finding weak emission profiles in the spectra.

To reliably detect even the weakest chromospheric emissions, uncertainty of the used reference spectra must be well known as well. By showing that the proposed neural network structure can be used as intended, we are looking into possibilities to improve our methodology using variational autoencoder. Its advantage lies in the possibility of simultaneous determination of a reference spectrum and its uncertainty which would enable uncertainty estimation of the measured emission-line indices.

Chapter 6

Peculiar solar-like multiple stars

This chapter has been partially adapted from the published paper titled *The GALAH survey: unresolved triple Sun-like stars discovered by the Gaia mission* [64] whose first author is author of this Doctoral thesis. The used computer code is published on GitHub platform [1] and results of the analysis as a catalogue on the VizieR service [2].

So far, we dealt with spectral peculiarities that are directly observable from the spectra. In the case of multiple stellar systems, that might not always be true. As all of the stars in a multiple system contribute to light received by the observer, the object must be treated as multiple to determine physical properties and abundances of individual components properly. However, to get an indication that an object must be treated as multiple, a single spectrum might not be enough as it does not reveal the dynamics of the components. In this chapter, we show a possible way to identify and characterise multiple objects using data-driven machine learning approaches. We first focused on solar twin stars whose importance is given in Section 6.1 and their selection from GALAH spectra described in Section 6.2. Their homogeneity of chemical composition is presented in Section 6.3. As some of the sources seem to be too bright to consist of only one star, their photometric and spectroscopic signature was further analysed using single star models explained in Section 6.8. Modelling revealed that some of the stars in our selection could consist of two or more very similar stars, whose orbital and physical properties are summarised in Section 6.13.

6.1 Introduction

The Sun, being the closest and most analysed star, is widely used as a reference for the calibration of many fundamental stellar astronomical parameters [317, 318]. This usage implies a desire to find and catalogue as many stars similar to the Sun as possible. They can be used for self-calibration of observations done with the same setup or inter-calibration of multiple surveys that share observations of the same objects. Stars resembling the Sun in all physical parameters such as luminosity, mass, radius, rotation period, and chemical composition should also have an identical spectrum to our Sun. The definition of such stars, also known as solar twins, was first introduced by Cayrel de Strobel *et al.* [319]. However, the definition is not fixed

[1] `https://github.com/kcotar/GALAH-survey-Solar-like-triple-stars` and `https://github.com/kcotar/Solar-spectral-siblings`

[2] `http://vizier.u-strasbg.fr/viz-bin/VizieR?-source=J/MNRAS/487/2474`

as it follows the evolution of astronomical instrumentation, observational techniques [131], and precision of determining stellar parameters and chemical abundances.

Before the release of *Gaia* parallaxes, one of the approaches to determine distances of twin stars or its hosting stellar cluster was based on the assumption that solar twins should have the same intrinsic luminosity as the Sun. Therefore its extinction corrected apparent brightness on the sky is directly related to their distance [320, 321, 322].

Another widely debated topic is the chemical composition of the Sun and its evolution in time. Ramírez *et al.* [323] and Nissen [324] addressed the question whether the composition of the Sun is peculiar when compared to other solar twins. Works like these benefit from using spectra with high signal to noise ratio and very high spectral resolution that are especially needed when trying to disentangle if the observed solar chemical composition is a consequence of the planetary system formation [325] or intrinsic galactic chemical composition during the formation of the Sun [324, 326, 327]. However, even the best solar twin candidate (currently thought to be the star 18 Sco [328]), when observed with the highest possible precision, still shows minute differences in the abundance pattern which will make it even harder to disentangle the place where the Sun and its siblings were formed. At the same time, this unique chemical composition will make it easier to be confident about discovering a real solar sibling when we find one. Among the siblings, we count all stars that were formed at the same time and from the same molecular cloud as Sun, which implies that they can have the same chemical composition (see Section 1.2).

6.2 Selection of the best solar-like spectra

Our search for spectra that best resemble spectrum of the Sun uses data acquired by the expanded GALAH survey that combines multiple surveys acquired by the same telescope and spectrograph (see Section 2.3 for their list). The search is based on the direct comparison between acquired stellar spectra and the reference solar spectrum, where both were observed with the same setup of the HERMES spectrograph.

6.2.1 Reference solar spectrum

Before finding the best solar-like spectra in our data set, the reference spectrum had to be constructed. For its construction, we identified 3708 twilight flats with the SNR per resolution element in the green HERMES arm greater than 210. Observing the sky at twilight enables us to observe the solar spectrum by every fibre of the spectrograph simultaneously. This is true even if our intuition says that the colour of the sky is, because of the scattering in the atmosphere, different than during daytime. The atmospheric effect does not affect our analysis as scattering only modifies the distribution of spectral strength and does not alter relative strengths of absorption lines.

The acquired flux of selected twilight spectra was first normalised using the 7^{th} degree Chebyshev polynomial with an asymmetric sigma clipping (2σ low and 3σ high threshold) that was run for 11 steps. Normalisation procedure was tailored explicitly for solar-like objects as it excluded wavelength ranges with wide absorption features and regions with no near-continuum levels that were determined from the high-resolution spectrum of the Sun created by Kurucz [329]. Especially problematic

for the normalisation process is the blue arm that is packed with blended absorption lines, topped by the Hβ line. Because of high SNR and presence of additional "wiggles" in observed twilight flats, we used a higher degree of polynomial function than in the case of a stellar spectrum normalisation (see Section 6.2.2). Source of identified features that make spectrum to look wiggly is not known and might be contributed by a reflection in optics when observing bright targets or by a residual background flux that was not properly removed by the GALAH reduction pipeline [100].

The "radial velocity" of normalised twilight spectra was determined by correlating an observed spectrum with the high-resolution solar spectrum [329]. The peak of the resulting correlation function was determined by fitting a Gaussian function to it. Before correlation, the high-resolution solar spectrum was convolved with a Gaussian kernel to match its resolving power with the HERMES spectrograph and to appear as close as an observed spectra as visually possible.

Normalised rest-frame twilight spectra were then median stacked to generate the final reference solar spectrum. Before it could be used, it was analysed for possible residual flux offsets that are impossible to remove by the reduction pipeline as we can not acquire a background spectrum at the same time as a twilight spectrum. Any residual flux would render too shallow absorption lines in the median reference solar spectrum. To equalise strengths of absorption lines in our reference solar spectrum and the high-resolution one, a small linearly changing offset was subtracted from the normalised spectrum that was re-normalised after subtraction.

6.2.2 Stellar spectra preprocessing

Preprocessing of observed stellar spectra began with combined and fluxed spectra that were prepared by the pipeline as an intermediate reduction result. First they were normalised using a 5^{th} degree Chebyshev polynomial and asymmetric sigma clipping (with 2σ low and 3σ high thresholds) that was run for 15 steps. To determine the level of a continuum as accurately as possible, we masked spectral ranges without near-continuum pixels (Hα and Hβ) during the normalisation process. Normalised spectra were then shifted for the barycentric and radial velocity that was determined by the GALAH pipeline and linearly resampled to the same wavelength step as used by the generated reference solar spectrum. Performing both steps (radial velocity shift and resampling) at the same time retains a quality of the spectrum and reduces the required number of resampling steps compared to using already shifted and normalised spectra prepared by the reduction pipeline.

6.2.3 Candidate selection

As the number of GALAH spectra in our set is quite large (more than $600,000$) and some of them have entirely different spectrum than the Sun (e.g. hot stars, metal-poor giants, cool stars), we performed an initial selection of possible solar twins based purely on stellar parameters. They were determined by *The Cannon* interpolation pipeline [104], that was part of the GALAH data release 2. For this, we first had to determine the possible offset of our interpolated parameters towards the parameters of well known solar-like benchmark stars [330, 331] and the Sun. Table 6.1 gives reference values of physical parameters for the Sun alongside median

Table 6.1: Stellar physical parameters of reference objects used to determine possible offsets between published benchmark values and values determined by *The Cannon*. For every object, the first line in table gives reference values and the second line of parameters is determined from the GALAH spectra.

Object	$T_{\rm eff}$ [K]	$\log g$ [dex]	[Fe/H] [dex]
Sun	5771	4.44	0.02
Twilight flats	5605 ± 40	4.21 ± 0.06	-0.14 ± 0.03
18 Sco	5810 ± 80	4.44 ± 0.03	0.03 ± 0.01
	5750 ± 20	4.37 ± 0.1	0.05 ± 0.02
β Hyi	5873 ± 45	3.98 ± 0.02	-0.04 ± 0.01
	5784 ± 5	3.93 ± 0.01	-0.11 ± 0.01
μ Ara	5902 ± 66	4.3 ± 0.03	0.35 ± 0.01
	5657 ± 20	4.15 ± 0.01	0.28 ± 0.01

parameters of selected twilight flats and their standard deviations.

Table 6.1 also shows parameters for three benchmark stars with solar-like properties that were observed and reduced as any other object in our data set. Parameter values in the first line were taken from the published set of *Gaia* FGK benchmark stars [330, 331] and the second line gives *The Cannon* parameters for the same object. From the comparison, it is evident that twilight flats have slightly more under-estimated parameters as apposed to other stars used in the comparison. *The Cannon* trend of stellar parameters being a bit too low is observable for all stars. Parameters are given in Table 6.1. Taking this into consideration, we initially selected stars within a broader range of $T_{\rm eff}$ $\pm$250 K, $\log g$ $\pm$0.4 dex and [Fe/H] $\pm$0.3 dex around median parameters of twilight flats. Applying parameter cuts to our data set, we were left with 92, 284 spectra that were analysed further.

6.2.4 Spectral similarity

The similarity between observed spectra and generated reference solar spectrum was calculated using multiple distance metrics, where identical spectra are described by the similarity value of 0. Figure 6.1 shows that all considered similarity metric values are strongly correlated with only few outliers. Some degree of the correlation was expected as all of the shown metrics are based on differently weighted per pixel difference between two spectra. For all the following comparisons we choose to use the Canberra distance metric [332] that is less sensitive to large, unnatural outlying flux values than the Euclidean distance and is defined as

$$d_{Canberra}(f_{\odot}, f_{obs}) = \sum_{\lambda} \frac{|f_{\odot,\lambda} - f_{obs,\lambda}|}{|f_{\odot,\lambda}| + |f_{obs,\lambda}|}, \tag{6.1}$$

where $f_{\odot}$ is the flux at a reference solar spectrum, f_{obs} an observed spectrum and λ wavelength bins of both spectra. Similarity value determined by the given function

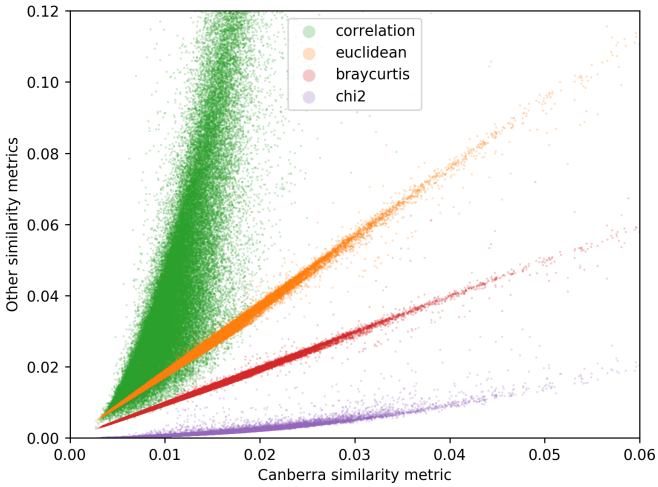

Figure 6.1: Correlation between the Canberra similarity metric and four other similarity metrics for the objects considered in the rough spectral comparison step.

heavily depends on the SNR of the evaluated spectrum, which effectively degrades its spectral similarity at lower SNR values. The dependency was analysed by computing the same similarity metric between the original reference solar spectrum and noisy reference that was generated by adding Gaussian noise to the original spectrum to represent spectra with different SNR values. To determine the uncertainty of similarities, the test was repeated 1500 times for every noise level. Smooth dependence between SNR and similarity metric can be described by the following composite of linear and a power-law function defined as

$$sim(SNR) = A\Big(\frac{SNR - SNR_{off}}{SNR_0}\Big)^{-\alpha} + B(SNR - SNR_{off}) + C, \qquad (6.2)$$

that was fitted to the resulting simulated points shown in Figure 6.2. Parameters A, B, C, SNR_0, SNR_{off}, and α in Equation 6.2 are free and were fitted to the simulated similarity measurements at different SNR levels.

We used the above relation between similarity distance and SNR (Equation 6.1) to compare observed spectra along with the following ranges of HERMES spectrum: 4725 - 4895 Å in the blue arm, 5665 - 5865 Å in the green arm, 6485 - 6725 Å in the red arm, and 7700 - 7875 Å in the near-infrared spectral arm. Selected ranges have non-zero flux values for all observed spectra. They do not pose any limitations on the number of usable spectra that might have undefined flux values at the range borders because of large radial velocities. The similarity between the reference and the observed spectra was not computed for the whole spectral range specified above, but was limited to known and isolated (not blended) spectral absorption lines, representing multiple different chemical elements. Number of lines per element is given in Table 6.2 and is the same as used by *The Cannon* abundance determination procedure described in Buder *et al.* [104]. Around the central wavelength of each absorption line, wavelength bins around 0.5 Å wide are considered for the comparison. The bin size differs for every line and is not necessarily centred on a given line centre [104] to reduce the possibility of contamination by nearby lines.

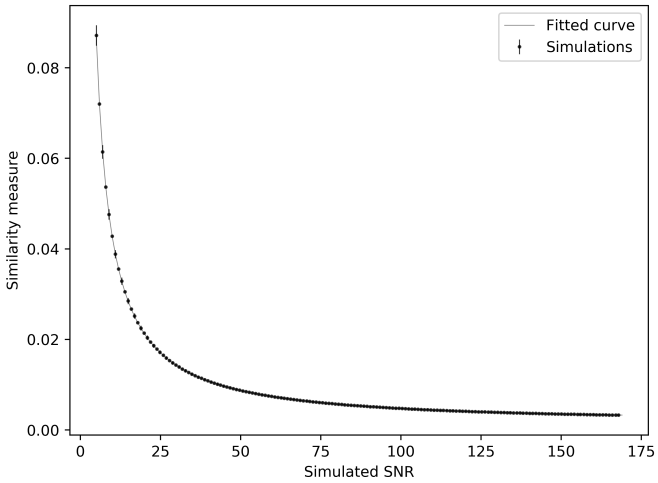

Figure 6.2: Equation 6.2 fitted to the simulated data points describing a relation between the spectral SNR and Canberra distance of a noisy solar spectrum in comparison with the original noise-free spectrum of the Sun.

Table 6.2: Number of absorption lines for different chemical elements that were used to measure similarity between an observed spectra and a reference spectrum. Elements marked with a * are problematic for precise abundance determination in solar twins because of their shallow or nonexistent absorption. Total number of used absorption lines is 180.

Element	Lines	Element	Lines
Al	4	Na	3
Ba	2	Nd*	5
C*	1	Ni	7
Ca	5	O	3
Co*	3	Rb*	1
Cr	9	Ru*	1
Cu	2	Sc	10
Eu*	2	Si	4
Fe	52	Sm*	2
K*	1	Sr*	1
La*	4	Ti	20
Li*	1	V	17
Mg	2	Y	4
Mn	4	Zn	2
Mo*	2	Zr*	4

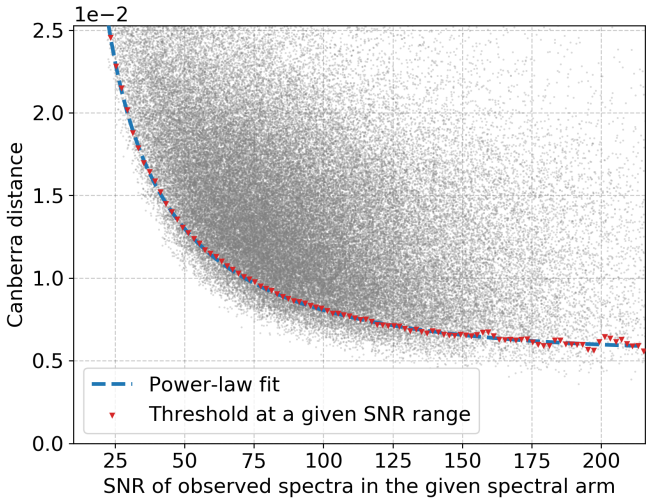

Figure 6.3: Equation 6.2 was fitted to the thresholding values, denoted with red downwards arrows, that delineated 7% of the most similar spectra in every narrow SNR bin. The resulting fit on the thresholding values is represented by the blue dashed line. Shown plot is made for the red spectral arm.

Best candidates

The spectral comparison described in Section 6.2.4 was independently computed for every HERMES arm. With four similarity values per spectrum, we selected the set of best matching spectra in the following way. Because the similarity value heavily depends on the SNR value of a spectrum, we first determined thresholding similarity values for the best 7 % of spectra in each narrow SNR bin. The thresholds are in Figure 6.3 visualized as red downward arrows. Width of the SNR bins was set to 4 units and separation between them to 2 units. After fitting the function described by Equation 6.2 to those thresholds, the best matching spectra were determined to be all spectra whose spectral similarity fell below the fitted line. By combining all four bands, we determined 329 objects whose spectral similarities fell bellow the fitted thresholding line in all four HERMES bands. Considering every spectral arm individually, we effectively removed objects with possible reduction problems in any of them. Portion of the best solar twin spectra are shown in Figure 6.4.

6.3 Physical properties and chemical composition of our candidates

By selecting very similar spectra, we would also expect them to have very similar abundances and parameters to solar values. Distribution of their physical parameters and abundances are shown in Figures 6.5 and 6.6. In general, physical parameters have a slight offset from the published solar values. The most obvious difference is the distribution of $\log g$ for solar twins in the middle panel of Figure 6.5. The tail towards lower $\log g$ values suggests existence of stars with different radii. Those stars are exciting and will be analysed in the following sections.

Even more interesting, especially for the purpose of chemical tagging experi-

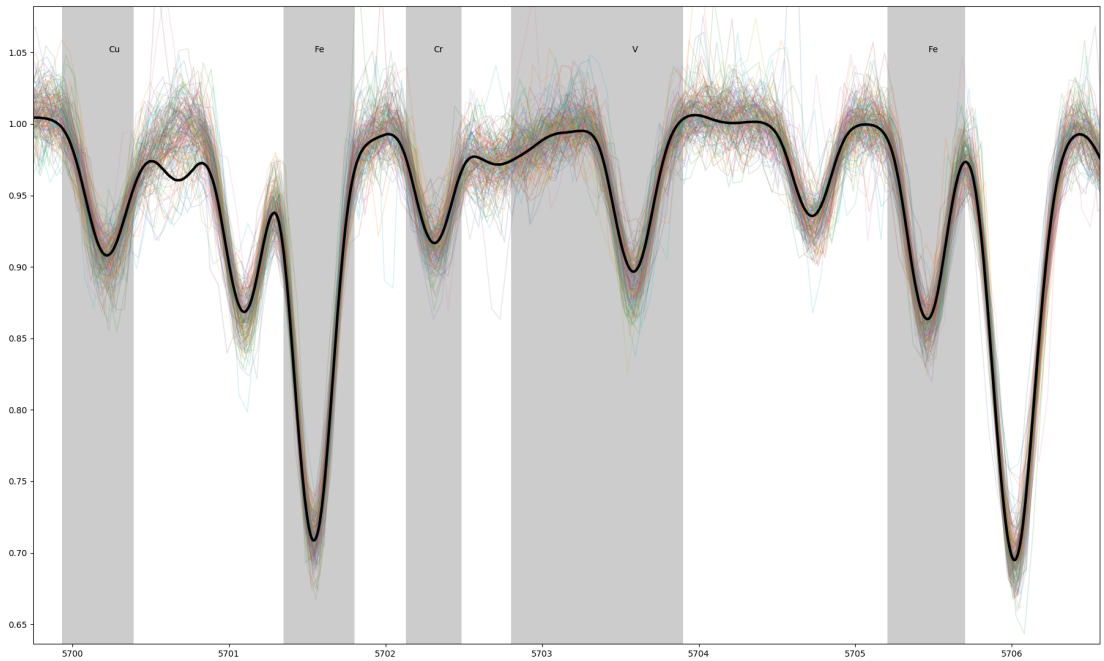

Figure 6.4: Visual comparison between a set of the most likely solar twins spectra and the solar reference spectrum shown in black. Grayed out regions represent wavelength ranges used in the similarity computation described in Section 6.2.4. Chemical element responsible for the observed line is also indicated above the spectra for every absorption line.

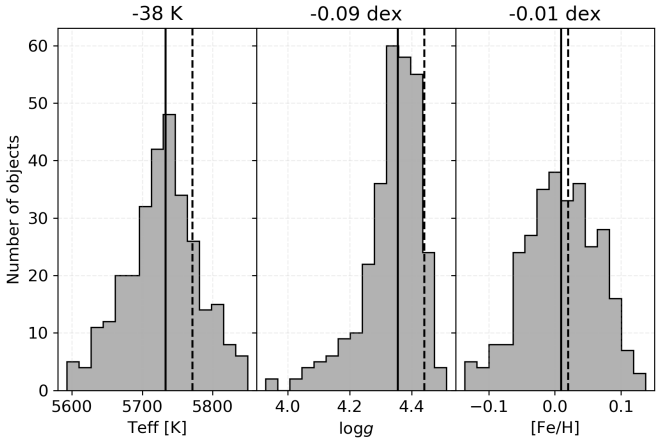

Figure 6.5: Distribution of physical parameters [166] for discovered solar twin candidates. Values above the plots represent the difference between median of the distribution (solid vertical line) and actual solar values (dashed vertical line).

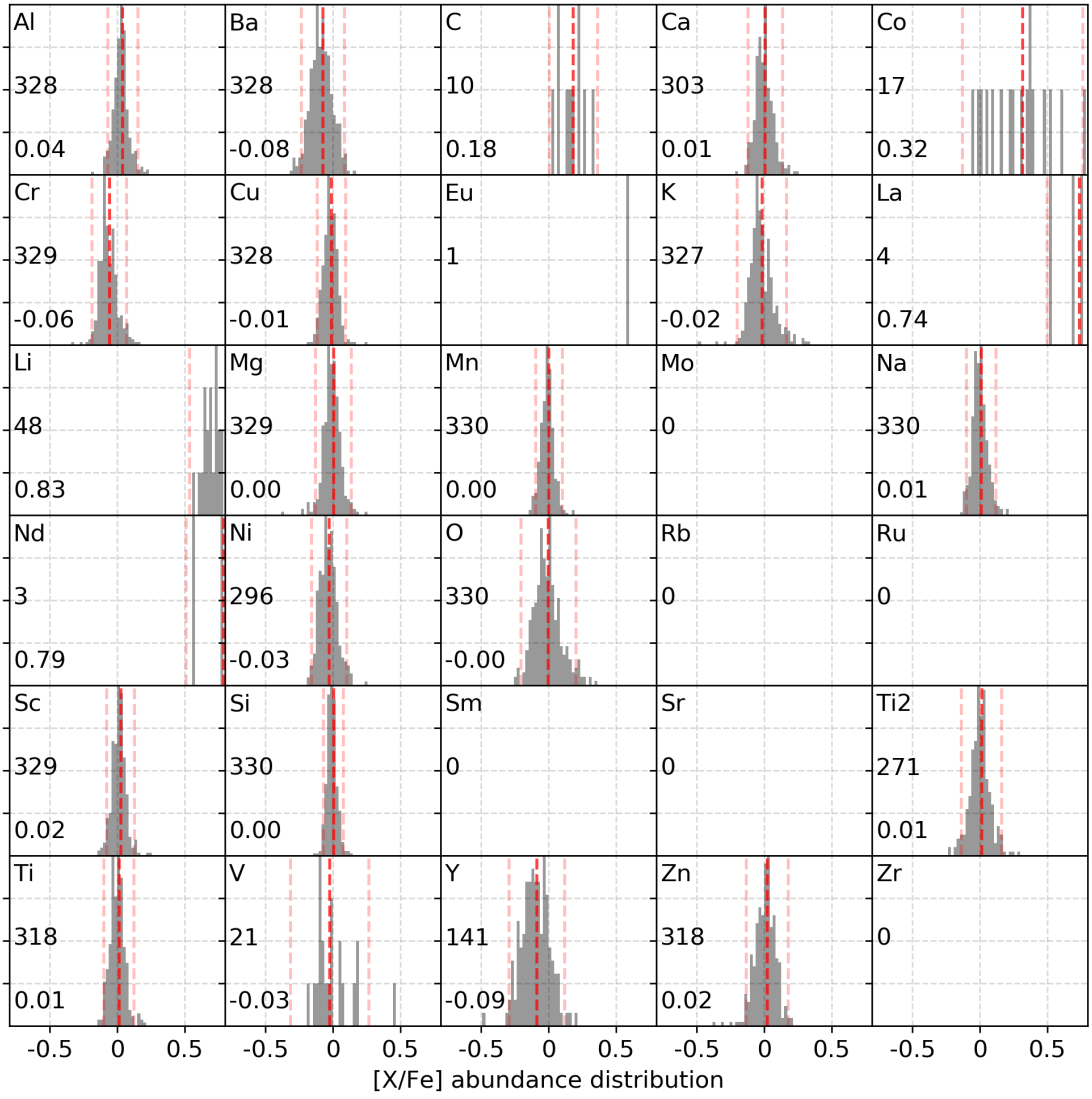

Figure 6.6: Normalised distributions of the measured abundance values determined during production of the third GALAH data release [166]. Plotted are abundances values for our complete set of solar twin candidates. Median values and one standard deviations are in plots marked by strong and transparent red vertical lines respectively. Values on the left, under element name, represent number of used abundance values for the construction of histogram and deviation of its median value from the zero abundance value that should represent solar abundance level. The solar abundance calibration and offset determination are thoroughly explained by Buder *et al.* [166].

ments, are the distributions of abundances in Figure 6.6. From the shown median values, we can determine possible offset with stellar abundances and abundances determined by other large surveys. For the well-behaved elements, that are not marked with a * in Table 6.2, the median values are within 0.1 dex of solar values. A bit more worrying in their standard deviation that can be as large as ~ 0.2 dex - something that we do not want when searching for almost identical stellar spectra. Elements marked as problematic in Table 6.2 are as expected having a very low number or even no abundance measurements.

6.4 Absolute magnitudes of solar twin candidates

The latest *Gaia* data release enables us to accurately determine absolute magnitudes for a large fraction of its stars. Given the fact that all our solar-like stars have similar spectra, we expected their absolute magnitudes to be almost the same. When we plotted their apparent magnitudes against their measured parallaxes, shown in Figure 6.7, it became evident that this assumption does not hold for all stars in our set. A fraction of them appeared to be over-luminous and could be unresolved stellar multiple systems with two or more near-identical stars. As the multiplicity is not uncommon among solar-type stars (see Section 6.5), we used additional photometric data (presented in Section 6.6) to develop a data-driven model (see Section 6.8) that was used to model observed spectrum and absolute magnitudes of our solar twin candidates.

6.5 Solar-type stars and their multiplicity

The investigation of solar-type stars in the solar neighbourhood has revealed that around half of them are found in binary or more complex stellar systems [333, 334, 335]. Of all multiple systems, about 13 % are part of higher-order hierarchical systems [333, 336]. Beyond the solar neighbourhood, the angular separation between members of such multiple systems becomes too small for their components to be spatially resolvable in the sky. Spectroscopic, photometric, and astrometric surveys observe those sources as single light points. It has been suggested that the population of binaries in the field could be even higher than in the solar neighbourhood [337]. Therefore a combination of multiple complementary approaches must be used to detect and analyse multiple stellar systems with different properties [335].

If the orbital period of such a system is relatively short, with high orbital velocities, it can be spectroscopically identified as a multiple system in two ways. When the components are of comparable luminosities, the effect of multiple absorption lines can be observed in the system's spectrum. Such an object is also known as a double-lined binary (SB2) system [85, 312, 338, 339]. By contrast, a single-lined binary (SB1) system does not show the same effect, as the secondary component is too faint to contribute significant flux. Short period SB1 systems can be identified from the periodic radial velocity variations if multi-epoch spectroscopy is available [340, 341, 342, 343, 344]. Other extrema are very wide binaries [345, 346, 347, 348, 349, 350, 351], and co-moving pairs [352, 353] that can only be identified by their spatial proximity and common velocity vector.

Duchêne and Kraus [334] summarized that the majority of Solar-type stars are

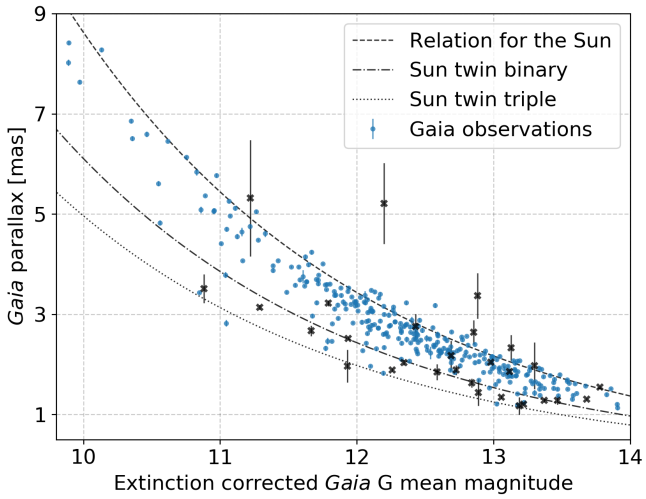

Figure 6.7: Parallax versus measured apparent *Gaia* G magnitude for our solar twin candidates. Stars marked with black crosses have significant re-normalised astrometric uncertainties (RUWE $>$ 1.4) which may lead to wrongly determined distance and consequently their absolute magnitude. The dashed line represents an absolute theoretical magnitude of the Sun G = 4.68 as it would be observed at different parallaxes. The solar absolute magnitude was computed from the relations given by Evans *et al.* [141]. Similarly, the relations for a binary and a triple system composed of multiple solar twins are plotted with the dash-dotted and dotted line.

part of binary systems with periods of hundreds to thousands of years, whose period distribution reaches a maximum at $\log(P) \approx 5$, for P measured in days. Because of the wide separation and long orbital periods of the components in such a scenario, the radial velocity variations will have both low amplitude and long period. This long periodicity makes them challenging to detect in large-scale spectroscopic surveys, which typically last for less than a decade, and have a low number of revisits. A spectrum of such a binary or triple still contains a spectroscopic signature of all members, and those contributions can be disentangled into individual components [86, 87]. Such a decomposition is easier when a binary consists of spectrally different stars [354, 355, 356, 357, 358]. However, it becomes much harder or near-impossible when the composite spectrum consists of contribution from near-identical stars whose individual radial velocities are almost identical [359]. In that case, additional photometric and distance measurement have to be used to constrain possible combinations [62]. If spectroscopic data are not available, determination of multiples can be attempted purely based on photometric information [62, 360, 361, 362], but such approaches are limited to certain stellar types, and yield results that might be polluted with young pre-main-sequence stars [363].

6.6 Additional photometric data

Photometric magnitudes provided by the *Gaia* DR2 release [42, 91] as shown in Figure 6.7 for our set of solar twins cover only a small wavelength region of the light produced by those stars. For a broader wavelength coverage, additional photometric data were taken from three large all-sky surveys. In the visual part of the spectrum

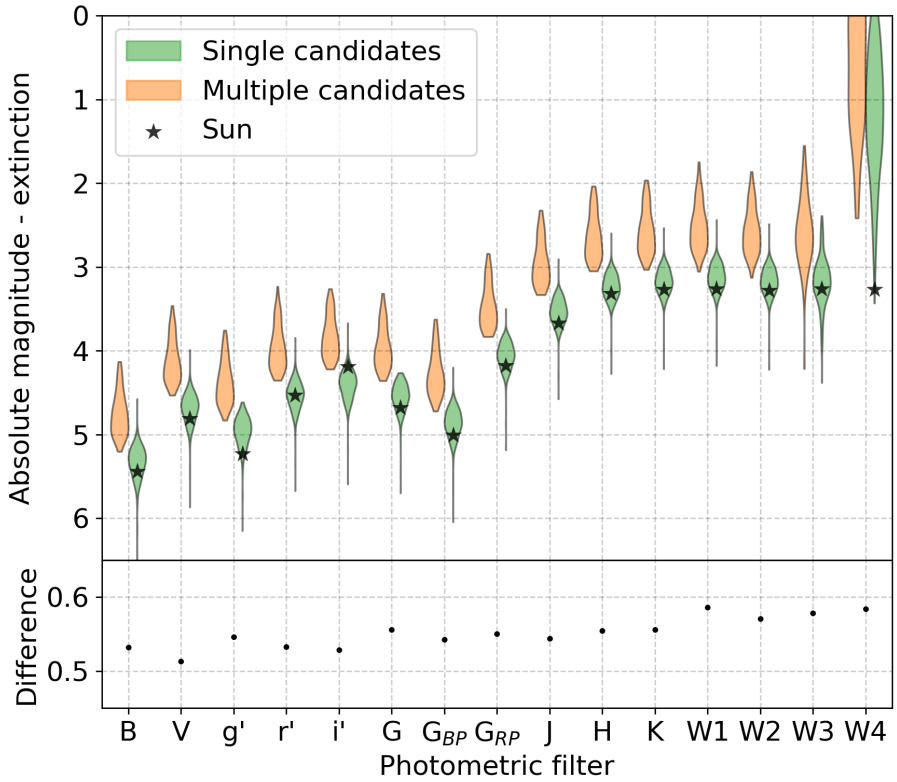

Figure 6.8: Violin plots showing the distribution of extinction-corrected absolute magnitudes in multiple photometric systems. Separate distributions are given for stars that we considered as single and multiple in our analysis. Star symbols indicate absolute magnitude of the Sun [366]. The bottom panel shows a difference between the median magnitudes of both distributions.

we rely on the AAVSO Photometric All-Sky Survey (APASS, [364]) B, V, g', r', i' bands that are supplemented by the Two Micron All-Sky Survey (2MASS, [157]) J, H, K_S bands, and the Wide-field Infrared Survey Explorer (WISE, [365]) W1, W2 bands. All of those surveys were cross-matched with the GALAH targets, which resulted in up to 13 photometric observations per star. Photometric values and their uncertainties were taken as published in these catalogues, ignoring any specific quality flags. During the initial investigation of their usefulness, WISE W3, and W4 bands proved to be unreliable for our application and were therefore removed from further use. The main reason for their removal is a large scatter in magnitude measurements of similar stars and a substantial overlap between single and multiple stars evident in Figure 6.8.

6.7 Solar-like spectra

6.7.1 Candidate multiple systems

Among the solar twin candidates, we noticed a photometric trend that is inconsistent with the distance of an object that resembles the Sun. Solar twins, mimicking the observed solar spectrum, should also be similar to it in all other observables such as luminosity, effective temperature, surface gravity, chemical composition and absolute

magnitude. Plotting their apparent magnitude against *Gaia* parallax measurement (Figure 6.7), all of the detected stars should lie near or on the theoretical line, describing the same relation for the Sun observed at different parallaxes. As the magnitude of the Sun is not directly measured by the *Gaia* or determined by the *Gaia* team, we computed its absolute magnitude using the relations published by Evans *et al.* [141] that connect the *Gaia* photometric system with other photometric systems. The reference solar magnitudes (in multiple filters) that were used in the computation were taken from Willmer [366]. The resulting absolute G magnitude of the Sun is 4.68 ± 0.02, where the uncertainty comes from the use of multiple relations. This value also coincides with the synthetic *Gaia* photometry produced by Casagrande and VandenBerg [367], who determined absolute magnitude of the Sun to be $M_{G,\odot} = 4.67$.

Within our sample of the probable solar twins, we identified 64 stars that show signs of being too bright at a given parallax. In Figure 6.7 they are noticeable as a sequence of data points that lie below the theoretical line and are parallel to it. Another even more obvious indication of their excess luminosity is given by the colour-magnitude diagram in Figure 6.9 where the same group of stars is brighter by ~ 0.7 magnitude. As both groups of stars are visually separable, the multiple stellar candidates can easily be isolated by selecting objects with extinction corrected absolute G magnitude above the binary limit line shown in Figure 6.9. To compute the absolute magnitudes, we used the distance to stars inferred by the Bayesian approach that takes into account the distribution of stars in the Galaxy [170]. As the reddening published along the *Gaia* DR2 [149] could be wrong for stars located away from the used set of isochrones, we took the information about the reddening at specific sky locations and distances from the all-sky three-dimensional dust map produced by Capitanio *et al.* [313]. To infer a band dependent extinction, a total to selective extinction ratio (R) was used. The values of R, considering the extinction law $R_V = 3.1$, were taken from the tabulated results published in Schlafly and Finkbeiner [368].

To determine the limiting threshold between single and multiple candidates, we first fit a linear representation of the main sequence to the median of the distribution of the absolute magnitudes in the 0.02 mag wide colour bins. The lower limit for the binaries was placed 0.25 mag above the fitted line.

Confirmation that the unseen companion could contribute this extra flux also comes from other photometric systems, where the distribution of absolute magnitudes for both groups is shown in Figure 6.8. On average, multiple candidates are brighter by $\sim$0.55 magnitude in every considered band. For an identical binary system, a measured magnitude excess would be 0.75, and 1.2 for a triple system. As the observed difference is not constant in every band, as would be expected for a system composed of identical stars, we expect some differences in parameters between the components of the system. This comparison can be argued under the assumption that all considered photometric measurements were performed at the same time. Of course, that is not exactly true in our case as the acquisition time between different photometric surveys can vary by a few to 10 years. As the Sun-like stars are normal, low activity stars, this effect is most probably negligible, but events like occultations between the stars in the system can still occur.

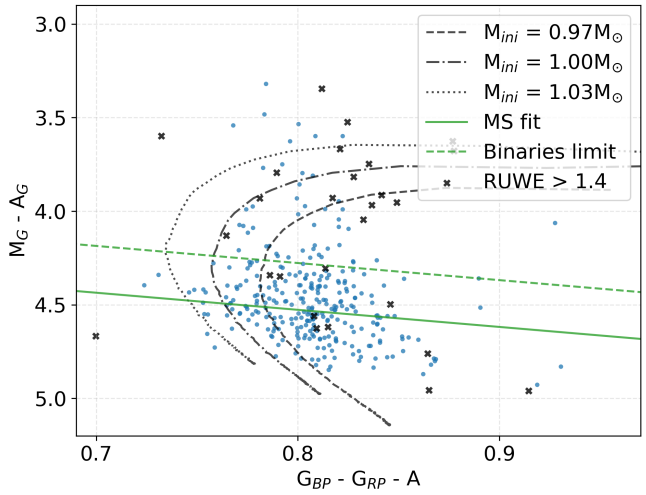

Figure 6.9: *Gaia* extinction corrected absolute G magnitude and colour index computed from G_{BP}, and G_{RP} bands. Stars marked with black crosses have large normalised astrometric uncertainty (RUWE $>$ 1.4). The green dashed line represents a threshold that was used as a delimiter between objects treated as multiple and single stars. Overlaid evolutionary tracks, constructed from the PARSEC isochrones [369], represent an evolution of stars with solar-like initial mass M_{ini} and metallicity [M/H] $=$ 0 for stellar ages between 0.1 and 12 Gyr. The ages are given in respect to the curve of Zero Age Mean Sequence (ZAMS).

6.8 Single star models

Once the selection of interesting stars was performed, we began with the analysis of their possible multiplicity. Our procedure for the analysis of suspected multiple stellar systems is based on spectroscopic and photometric data-driven single star models that were constructed from observations taken from multiple large sky surveys. With this approach, we exclude assumptions about stellar properties and populations that are usually used to generate synthetic data. In this section, we describe approaches that were used to create those models.

6.8.1 Spectroscopic model

Every stellar spectrum can be largely described using four basic physical stellar parameters: $T_{\rm eff}$, $\log g$, metallicity, and $v \sin i$. To construct a model that would be able to recreate a spectrum corresponding to any conceivable combination of those parameters, we used a data-driven approach named *The Cannon*. The model was trained on a set of normalised GALAH spectra that meet the following criteria: the spectrum must not be flagged as peculiar [84], have a signal to noise ratio (SNR) per resolution element in the green arm $>$ 20, does not contain any monitored reduction problems and have valid *The Cannon* stellar parameters. Additionally, we limited our set to main sequence dwarf stars (below the arbitrarily defined line shown in Figure 6.10) as giants are not relevant for our analysis. Additionally, the decision not to consider giants was taken as a result of the fact that accurate modelling of their spectra requires information about their luminosity. It should be noted that the

application of these limits does not ensure that our training set is entirely free from spectra of unresolved (or even clearly resolved SB2 binaries), as would be ideally desired.

In order to train the model, all spectra were first shifted to the rest frame by the reduction pipeline [100], and then linearly interpolated onto a common wavelength grid. The training procedure consists of minimising a loss function between an internal model of *The Cannon* and observations for every pixel of a spectrum [103].

The result of this training procedure are quadratic relations that take desired stellar parameters $T_{\rm eff}$, $\log g$, [Fe/H], and $v \sin i$ to reconstruct a target spectrum. Spectra generated in this manner are trustworthy only within the parameter space defined by the training set, where the main limitation is the effective temperature which ranges from ∼4600 to ∼6700 K on the main sequence. Spectra of hotter stars are easy to reproduce, but they lack elemental absorption lines that we would like to analyse. On the other hand, spectra of colder stars are packed with molecular absorption lines and therefore harder to reproduce and analyse. The model itself can be used to extrapolate spectra outside the initial training set, but as they can not be verified, they were not considered to be useful for the analysis.

6.8.2 Photometric model

With the use of a model that produces normalised spectra for every given set of stellar parameters, we lose all information about the stars' colour, luminosity, and spectral energy distribution. This can be overcome using another model that generates the photometric signature of the star in question. To create this kind of a model, we first collected up to 13 apparent magnitudes from the selected photometric surveys (*Gaia*, APASS, 2MASS, and WISE) for every star in the GALAH survey. Whenever possible, these values were converted to absolute magnitudes using the distance to stars inferred by the Bayesian approach [170]. Before using the pre-computed published distances, we removed all sources whose computed RUWE was greater than 1.4. The magnitudes of every individual star were also corrected for the reddening effect, except for the WISE photometric bands W1 and W2 that were considered to be extinction free as their extinction is more that 10-times smaller than as measured for the V band.

Using the valid *The Cannon* stellar parameters, the inferred and corrected absolute magnitudes in multiple photometric bands were grouped into bins that contain stars within $\Delta T_{\rm eff} = \pm 80$ K, $\Delta \log g = \pm 0.05$ dex, and Δ[Fe/H] $= \pm 0.1$ dex around the bin centre. As spectroscopically unresolved multiple stellar systems could still be present in this bin, median photometric values are computed per bin to minimise their effect. Extrapolation outside this grid that covers the entire observational stellar parameter space is not desired nor implemented as it may produce erroneous values. When a photometric signature of a star with parameters between the grid points is requested, it is recovered by linear interpolation between the neighbouring grid points. The median photometric signature for the requested stellar parameters could also be computed on-the-fly using the same binning, but we found out that this produced insignificant difference and increased the processing time by more than a factor of two.

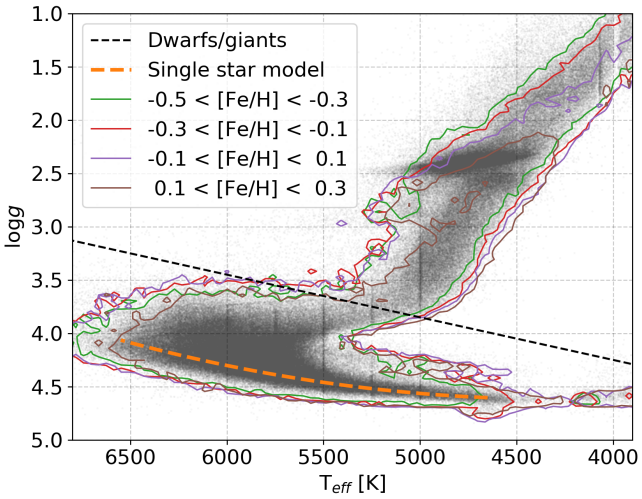

Figure 6.10: Complete observational set of valid stellar parameters shown as a varied density of grey dots. Contours around the diagram illustrate a coverage of parameter space in different [Fe/H] bins. Overlaid orange dashed curve represents a relation on the main sequence from where single stars considered in the fit are drawn. Additionally, black dashed linear line represents arbitrarily defined limit between giant and dwarf stars that were used in the process of training a spectroscopic single star model.

6.8.3 Limitations in the parameter space

As already emphasized, spectroscopic and photometric models were built on real observations and are therefore limited by the training set coverage. The limitations are visually illustrated by Figure 6.10, where the dashed curve, from which single stars are drawn, is plotted over the observations. This arbitrary quadratic function is defined as:

$$\log g = 2.576 + 9.48 \cdot 10^{-4}\ T_{\rm eff} - 1.10 \cdot 10^{-7}\ T_{\rm eff}^2, \tag{6.3}$$

where values of $T_{\rm eff}$ and $\log g$ are given in units of K and $\mathrm{cm\,s^{-2}}$ respectively. The polynomial coefficients were determined by fitting a quadratic function to manually defined points that represent regions with the highest density of stars on the shown Kiel diagram. From Equation 6.3, it follows that the $\log g$ of a selected single star is not varied freely, but computed from the selected $T_{\rm eff}$ whose range is limited within the values $6550 > T_{\rm eff} > 4650$ K.

Focusing on solar-like stars gives us an advantage in their modelling as the whole observational diagram in Figure 6.10 is sufficiently populated with stars of solar-like iron abundance. When going towards more extreme [Fe/H] values (high or low), coverage of the main sequence starts to decrease. For cooler stars, this happens at low iron abundance ([Fe/H] < -0.3) and for hot stars at high abundance ([Fe/H] > 0.3). Those limits pose no problems for our analysis, unless the wrong stellar configuration is used to describe the observations. At that point, the spectroscopic fitting procedure would try to compensate for too deep or too shallow spectral lines (effect of wrongly selected $T_{\rm eff}$) by decreasing or increasing [Fe/H] beyond values reasonable for solar-like objects.

6.9 Characterization of multiple system candidates

For the detailed characterization of solar twin candidates that show excess luminosity, we used their complete available photometric and spectroscopic information. The excess luminosity can only be explained by the presence of an additional stellar component or a star that is hotter or larger than the Sun. Both of those cases can be investigated and confirmed by the data and models described in Sections 6.6 and 6.8. In the scope of our comparative methodology, we constructed a broad collection of synthetic single, double, and triple stellar systems that were compared and fitted to the observations.

As the measured *Gaia* DR2 parallaxes, and therefore inferred distances, of some objects, are poorly fitted or highly uncertain, the distance results provided by Bailer-Jones *et al.* [170] yield three distinct distance estimates - the mode of an inferred distance distribution (`r_est`) and a near and distant distance (`r_lo`, and `r_hi`) estimation, between which 68 % of the distance estimations are distributed. As the actual shape of the distribution is not known and could be highly skewed, we did not draw multiple possible distances from the distribution, but only used its most probable value (mode).

6.9.1 Fitting procedure

A complete characterization and exploration fitting procedure for every stellar configuration (single, binary, and triple system) consist of four consecutive steps that are detailed in the following sections. As we are investigating solar twin spectra, the initial assumption for the iron abundance of the system is set to [Fe/H] = 0. This also includes the assumption that stars in a system are of the same age, at similar evolutionary stages, and were formed from a similarly enriched material. If that is true, we can set iron abundance to be equal for all stars in the system. This notion is supported by the simulations [370] and studies [371] of field stars showing that close stars are very likely to be co-natal if their velocity separation is small.

The observed systems must be composed of multiple main sequence stars. Otherwise, the giant companion would dominate the observables, and the system would not be a spectroscopic match to the Sun. Therefore their parameters $T_{\rm eff}$ and $\log g$ are drawn from the middle of the main sequence determined by *The Cannon* parameters. The Kiel diagram of the stars with valid parameters and model of the main sequence isochrone used in the fitting procedure are shown in Figure 6.10.

6.9.2 Photometric fitting - first step

With those initial assumptions in mind, we begin with the construction of the photometric signature of the selected stellar configuration. To find the best model that describes the observations, we employ a Bayesian Markov chain Monte Carlo (MCMC) fitting approach [372], where the varied parameter is the effective temperature of the components. The selected $T_{\rm eff}$ values, and inferred $\log g$ (Equation 6.3), are fed to the photometric model (Section 6.8.2) to predict a photometric signature of an individual component. Multiple stellar signatures are combined together into

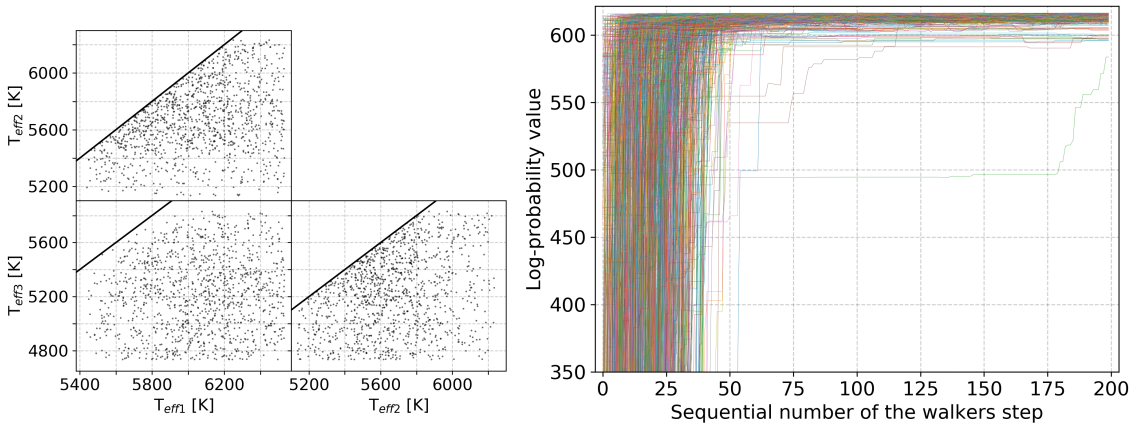

Figure 6.11: Initial distribution of walker parameters considered in the photometric fit. To ensure unique solutions with increasing $T_{\rm eff}$, combinations above the linear line were not considered in the fit. Convergence of walkers in the initial photometric fit. The plot shows a log-probability of the posteriors shown in the Figure 6.12. Walkers converge to the same value of log-probability after 50 steps. Every walker is plotted with different colour.

a single unresolved stellar source using the following equation:

$$M_{model} = -2.5 \log_{10}\Big(\sum_{i=1}^{n_s} 10^{-0.4 M_i}\Big);\ n_s = [1, 2, 3], \tag{6.4}$$

where M_i denotes absolute magnitude of a star in one of the used photometric bands, and n_s number of components in a system. The newly constructed photometric signature at selected $T_{\rm eff}$ values is compared to the observations using the photometric log-likelihood function $\ln p_P$ defined as:

$$\ln p_P(T_{\rm eff}|M,\sigma) = -\frac{1}{2}\sum_{i=1}^{n_p}\Big[\frac{(M_i - M_{model,i})^2}{\sigma_i^2} + ln(2\pi\sigma_i^2)\Big], \tag{6.5}$$

where M and σ represent extinction corrected absolute magnitudes, and their measured uncertainties that were taken for multiple published catalogues presented in Section 6.6. The constructed photometric model of a multiple system is represented by the variable M_{model} and the number of photometric bands by n_p. The maximum, and most common value for n_p is 13, but in some cases, it can drop to as low as 8. The MCMC procedure is employed to maximise this log-likelihood and find the best fitting stellar components.

To determine the best possible combination of $T_{\rm eff}$ values, we initiate the fit with 1200 uniformly distributed random combinations of initial temperature values that span the parameter space shown in Figure 6.10. The number of initial combinations is intentionally high in order to sufficiently explore the temperature space. Excessive or repeated variations of initial parameters are rearranged by a prior limitation that the temperatures of components must be decreasing, therefore $T_{\rm eff1} \geq T_{\rm eff2} \geq T_{\rm eff3}$ in the case of a triple system (example of used initial walker parameters is shown in Figure 6.11). The initial conditions are run for 200 steps. The number of steps was selected in such a way to ensure a convergence of all considered cases (example of the walkers convergence is shown in Figure 6.11). The distribution of priors for

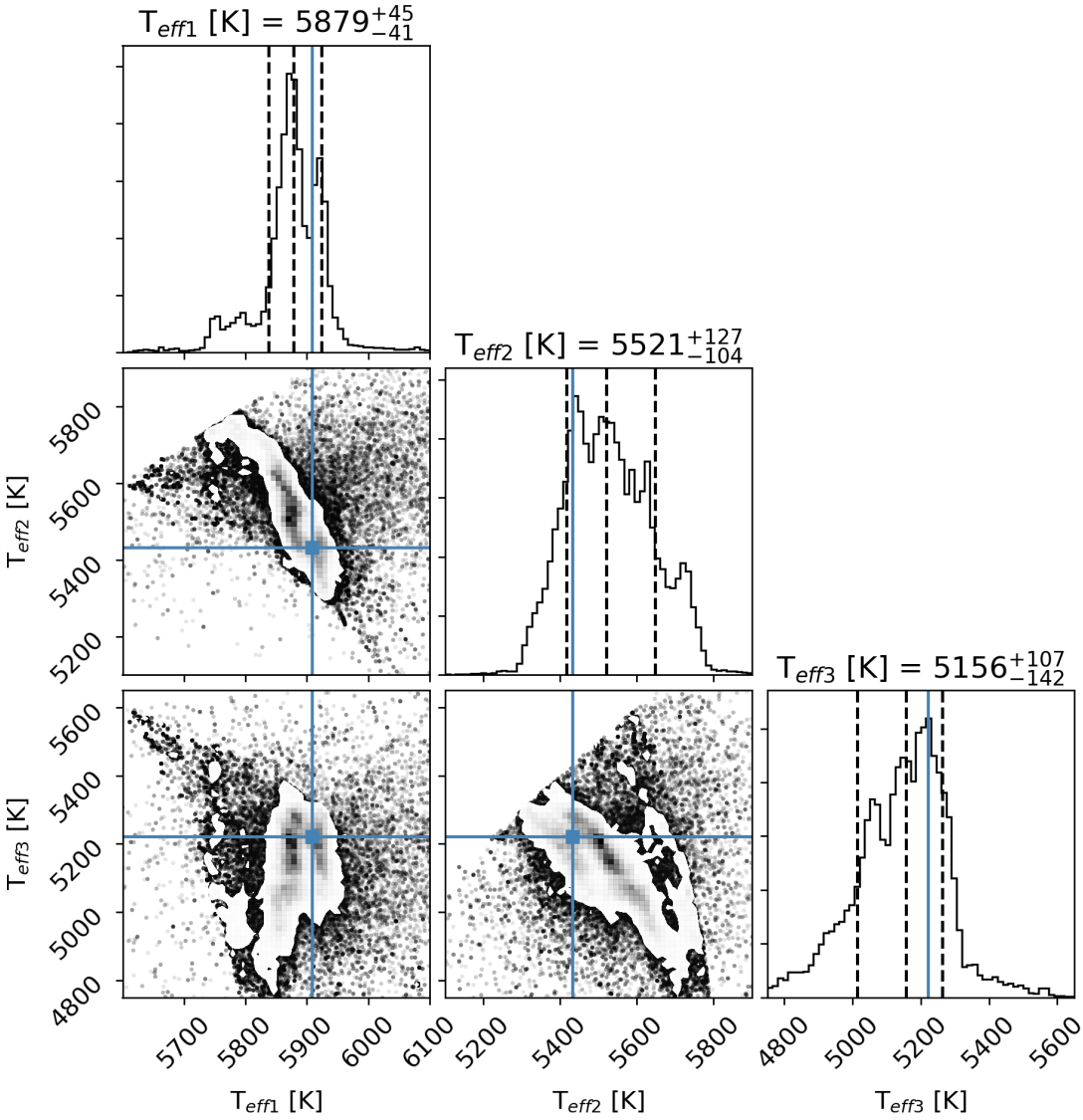

Figure 6.12: Distribution of considered posteriors during the initial MCMC photometric fit for one of the objects that was at the end classified as a triple candidate. As the plots show the first step in the fitting procedure that explores the complete parameter space, the distributions are not expected to be smooth because of possible multiple local minima. Values indicated with blue crosses on the scatter plots represent medians for the 10 % of the best fitting solutions. Value above the plots represent mean value of the histogram and its difference towards 16^{th} and 84^{th} percentile of the histogram. The same percentiles are also marked on histograms with black dashed lines.

such a run is shown as a corner plot in Figure 6.12. After that, only the best 150 walkers are kept, their values perturbed by 2 %, and run for another 200 steps to determine the posterior distribution of the parameters varied during the MCMC fit. This two-step run is needed to speed up the process and discard solutions with lower $\ln p_P$. During the initial tests, we found that the investigated parameter space can have multiple local minima which attract walkers, especially in the case of a triple system.

After the completion of all MCMC steps, the considered parameter combinations

are ordered by their log-likelihood in descending order. Of those, only 10 % of the best fitting combinations are used to compute final $T_{\mathrm{eff}[1-3]}$ values. They are computed as the median of the selected best combinations.

6.9.3 Spectroscopic fitting - second step

After an effective temperature of the components has been determined, we proceed with the evaluation of how well they reproduce the observed spectrum. As a majority of the fitted systems do not consist of multiple components with a T_{eff} equal to solar, the [Fe/H] of the system must be slightly changed to equalize absorption strength of the simulated and observed spectral lines.

To determine the [Fe/H] of the system, we first compute a simulated spectrum for every component using a spectroscopic model described in Section 6.8.1. Individual spectra are afterwards combined using equations

$$\begin{aligned} r_{12} &= \frac{L_{2,\lambda}}{L_{1,\lambda}};\ \lambda = [1,2,3,4] \\ f_{model,\lambda} &= \frac{f_{1,\lambda}}{1+r_{12}} + \frac{f_{2,\lambda}}{1+1/r_{12}} \end{aligned} \tag{6.6}$$

in the case of a binary system, or equations

$$\begin{aligned} r_{12} &= \frac{L_{2,\lambda}}{L_{1,\lambda}},\quad r_{13} = \frac{L_{3,\lambda}}{L_{1,\lambda}} \\ f_{model,\lambda} &= \frac{f_{1,\lambda}}{1+r_{12}+r_{13}} + \frac{r_{12}f_{2,\lambda}}{1+r_{12}+r_{13}} + \frac{r_{13}f_{3,\lambda}}{1+r_{12}+r_{13}} \end{aligned} \tag{6.7}$$

in the case of a triple system.

Individual normalised spectra, denoted by $f_{n,\lambda}$ in Equations 6.6 and 6.7, are weighted by the luminosity ratios between the components (r_{xy}) and then summed together.

As the HERMES spectrum covers four spectral ranges, whose distribution of spectral energy depends on stellar T_{eff}, different luminosity ratios also have to be used for every spectral arm. Of all of the used photometric systems, APASS filters B, g', r', and i' have the best spectral match with blue, green, red, and infrared HERMES arm. The modelled APASS magnitudes of the same stars are used to compute luminosity ratios between them.

The described summation of the spectra introduces an additional assumption about the analysed object. With this step, we assume that components have a negligible internal spread of projected radial velocities that could otherwise introduce asymmetries in the shape of observed spectral lines. The assumption allows us to combine individual components without any wavelength corrections.

Similarly, as in the previous case, a Bayesian MCMC fitting procedure was used to maximise the spectroscopic log-likelihood $\ln p_S$ defined as:

$$\ln p_S([\mathrm{Fe/H}]|f,\sigma) = -\frac{1}{2}\sum_{\lambda}\Big[\frac{(f_\lambda - f_{model,\lambda})^2}{\sigma_\lambda^2} + ln(2\pi\sigma_\lambda^2)\Big], \tag{6.8}$$

where f and σ represent the observed spectrum and its per-pixel uncertainty, respectively. A modelled spectrum of the system, at selected [Fe/H], is represented by the

variable f_{model} and the number of wavelength pixels in that model by λ. Combined, all four spectral bands consist of almost $16,000$ pixels. $T_{\rm eff}$ values of the components are fixed for all considered cases.

The MCMC fit is initiated with 150 randomly selected [Fe/H] values, whose uniform distribution is centred at the initial [Fe/H] value of the system and has a span of 0.4 dex. All of the initiated walkers are run for 100 steps. At every [Fe/H] level, a new simulated spectrum composite is generated and compared to the observed spectrum by computing log-likelihood $\ln p_S$ of a selected [Fe/H] value. The range of possible [Fe/H] values considered in the fit is limited by a flat prior between -0.5 and 0.4.

By the definition of [Fe/H] in the scope of GALAH *The Cannon* analysis, the parameter describes stellar iron abundance and not its metallicity as commonly used in the literature. Therefore only spectral absorption regions of un-blended Fe atomic lines are used to compute the spectral log-likelihood. Having to fit only one variable at a time, the solution is easily found and computed as a median value of all posteriors considered in the fit.

6.9.4 Final fit - third step

A changed value of [Fe/H] for the system will introduce subtle changes to its photometric signature, therefore we re-initiate the photometric fitting procedure. It is equivalent to the procedure described in Section 6.9.3, but with much narrower initial conditions. These new initial conditions are uniformly drawn from the distribution centred at $T_{\rm eff}$ values determined in the first step of the fitting procedure. The width of the uniform distribution is equal to 100 K. Drawn initial conditions are afterwards run through the same procedure as described before.

At this point, the second and third step in the fitting procedure can be repeatedly run several times to further pinpoint the best solution. We found out that further refinement was not needed in our case as it did not influence the determined number of stars in the system.

6.9.5 Number of stellar components - final classification

The fitting procedure described above was used to evaluate observations of every multiple stellar candidate to determine whether they belong to a single, binary or triple stellar system. This resulted in the following set of results for every configuration: predicted $T_{\rm eff}$ of the components, [Fe/H] of the system, simulated spectrum, and simulated photometric signature of the system.

As the photometric and spectroscopic fits do not always agree on the best configuration, the following set of steps and rules was applied to classify results in one of six classes presented in Table 6.3.

- Compute χ^2 between the simulated photometric signature of the modelled system and extinction corrected absolute photometric observations for every considered stellar configuration.
- Compute χ^2 between the simulated spectrum of the modelled system and the complete GALAH observed spectrum for every considered stellar configuration.

Table 6.3: Number of different systems discovered by the fitting procedure performed on possible multiple stars that exhibit excess luminosity.

Configuration classification	Number of systems
1 star	2
$\geq$ 1 star	14
2 stars	27
$\geq$ 2 stars	14
3 stars	6
Inconclusive	1
Total objects	64

- Independently select the best fitting configuration with the lowest χ^2 for photometric and spectroscopic fit.
- If the best photometric and spectroscopic fit point to the same configuration, then the system is classified as having a number of stars defined by both fitting procedures.
- If the best photometric and spectroscopic fit do not point to the same configuration, then the system is classified as having at least as many stars as determined by the prediction with a lower number of stars (e.g. $\geq$ 2 stars).
- If the difference between those two predictions is greater than 1 (e.g. photometric fit points to a single star and spectroscopic to a triple star), then the system is classified as inconclusive.

The classification produced using these rules is shown as colour coded *Gaia* colour-magnitude diagram in Figure 6.13.

6.9.6 Quality flags

In additional to our final classification, we also provide an additional quality checks that might help to identify cases for which our method might return questionable determination of a stellar configuration. Every of those checks, listed in Table 6.4, is represented by one bit of a parameter `flag` in the final published table (Table 6.8). The first bit gives us an indication of whether the object could have an uncertain astrometric solution, whereas the second and the third bits indicate if the final fitted solution has a worse match with the observations than the parameters produced by the *The Cannon* pipeline. To evaluate this, we used the original stellar parameters reported for the object to construct their photometric and spectroscopic synthetic model that was compared to the observations by computing their χ^2 similarity (`m_sim_p` and `m_sim_f` in Table 6.8). The resulting fitted spectrum or photometric signature is marked as deviating if its similarity towards observations is worse than for the reported one star parameters. This might not be the best

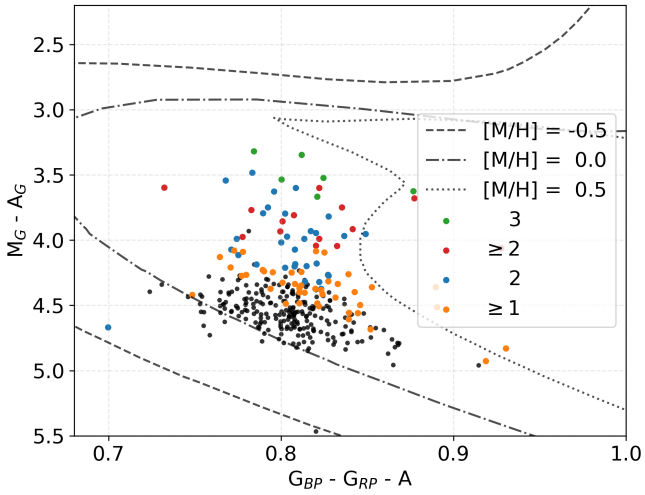

Figure 6.13: Showing the same data points as Figure 6.9 with indicated definite triple (green dots), possible triple (red dots), binary (blue dots), and possible binary stellar systems (orange dots). All other classes are shown with black dots. Overplotted are PARSEC isochrones [369] for stars with the age of 4.5 Gyr and different metallicities.

Table 6.4: Explanation of the used binary quality flags in the final classification of the stellar configuration. A raised bit could indicate possible problems or mismatches in the determined configuration. Symbol X in the last two descriptions represents the best fitting configuration, therefore X = [1,2,3].

Raised bit	Description
0	None of the flags was raised
1st bit	High astrometric uncertainty (RUWE > 1.4)
2nd bit	Deviating photometric fit (`sX_sim_p` > `m_sim_p`)
3rd bit	Deviating spectroscopic fit (`sX_sim_s` > `m_sim_s`)

indication of possible mismatch as it is common that *The Cannon* parameters of the analysed multiple candidates deviate from the main sequence in Kiel diagram (Figure 6.10) and therefore fall into less populated parameter space, where they can skew the single star models (Sections 6.8.1 and 6.8.2).

6.10 Characterization of single star candidates

Once we concluded our analysis of the set of 64 objects that showed obvious signs of excess luminosity, we then proceeded to study the remaining 265 objects that are most probably not part of a complex stellar system. To explore their composition, they were analysed with the same procedure as multiple candidates (Section 6.9). Before running the procedure, we omitted the option to fit for a triple system as they clearly do not possess enough excess luminosity for that kind of a system.

The obtained results were analysed and classified using the same set of rules

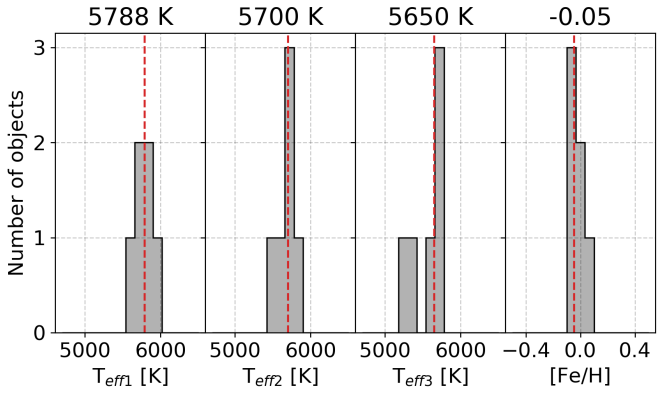

Figure 6.14: Parameters of triple stellar systems discovered and characterized by our analysis. Histograms represent distribution of all fitted results for 6 objects that were classified as triple stellar systems and observationally mimic solar spectrum. Median values of the distributions are given above individual histogram and indicated with dashed vertical line.

Table 6.5: Number of different systems discovered by the fitting procedure performed on stars that do not exhibit excess luminosity. Classes and their description is the same as used in Table 6.3. In addition, number of stars without parallactic measurements is added for completeness.

Configuration classification	Number of systems
1 star	230
$\geq$ 1 star	31
2 stars	4
$\geq$ 2 stars	0
3 stars	0
Inconclusive	0
Unknown parallax	0
Total objects	265

introduced in Section 6.9.5. Retrieved classes are summarized in Table 6.5 and in Figure 6.13. In the latter we see that the potential binaries are located on the top of the colour-magnitude diagram, which is consistent with the potential presence of an additional stellar source.

6.11 Orbital period constraints

The observational data we have gathered on possible multiple systems point to configurations that change slowly as a function of time. Using the observational constraints and models that describe the formation of the observed spectra, we try to set limiting values on the orbital parameters of the detected triple stellar system

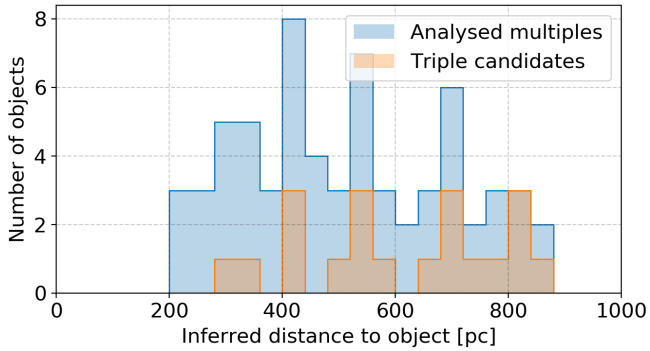

Figure 6.15: Histogram showing distribution of radial distances for the analysed set of stars. Distances were inferred by the Bayesian approach that takes into account the distribution of stars in the Galaxy [170].

candidates. In order to have a greater sample size, we use both definite (class 3) and probable (class $\geq$2) triple stars.

With the limited set of observations, we have to set assumptions about the constitution of those systems. For a hierarchical 2 + 1 system to be dynamically stable on long timescales, its ratio between the orbital period of an inner pair P_S and outer pair P_L must be above a certain limit. Eggleton [373] showed that P_L/P_S must be higher than 5. The same lower limit is noticeable when the correlation between periods of the known triple stars is plotted [374, 375].

In order to estimate the periods, we have to know the masses and distances between the stars in a system. Without complete information about the projected velocity variation in a system, masses can also be inferred from the spectral type. As we are looking at the solar twin triples, whose effective temperatures are all very similar and close to solar values (see Figure 6.14), our rough estimate is that all stars also have a solar-like mass $\sim M_\odot$. From this, we can set the inner mass ratio q_S to be close to 1 and the outer mass ratio $q_L \sim 0.5$. When we are dealing with the outermost star, the inner pair is combined into one object with twice the mass of the Sun. The likelihood of such a configuration is also supported by the observations [374] where a higher concentration of triple systems is present around those mass ratios. Contrary to our systems, twin binaries with equal masses usually have shorter orbital periods [376].

With the initial assumption about the configuration of the triple systems and masses of the stars, the periods of the inner and outer pair can be constrained to some degree as other orbital elements (inclination, ellipticity, phase ...) are not known.

6.11.1 Outer pair and *Gaia* angular resolution

Limited by its design, *Gaia* spacecraft and its on-ground data processing are in theory unable to resolve stars with the angular separation below $\sim$ 0.1 arcsec. Above that limit, their separability is governed by the flux ratio of the pair. The validation report of the latest data release [377] shows that the currently achieved angular resolution is approximately 0.4 arcsec as none of the source pairs is found closer than this separation. We used this reported angular limit to assess possible or-

bital configurations that are consistent with both the spectroscopic and astrometric observations.

Along with the final astrometric parameters, the *Gaia* data set also contains information about the goodness of the astrometric fit. Evans [60] used those parameters to confirm old and find new candidates for unresolved exoplanet hosting binaries in the data set. As we are looking at a system of multiple stars that orbit around their common centre of mass, we expect their photo-centre to slightly shift during such orbital motion. The observed wobble of the photo-centre also depends on the mass of stars in the observed system. In the case of a binary system containing two identical stars, the wobble would not be observed, but its prominence increase as the difference between their luminosities becomes more pronounced. This subtle change in the position of a photo-centre adds additional stellar movement to the astrometric fit, consequently degrading its quality. Improvements for such a motion will be added in later *Gaia* data releases. If a system has an orbital period much longer than the time-span of *Gaia* DR2 measurements (22 months), its movement has not yet affected the astrometric solution. This puts an upper limit on an orbital period as it should not be longer than few years in order to already affect the astrometric fit results.

Setting arbitrary limits to the published parameters of the astrometric fit quality (`astrometric_excess_noise` > 5, and `astrometric_gof_al` > 20 as proposed by Evans [60]), none of our 329 stars meets those requirements. This suggests that all of them are most likely well below the *Gaia* separability limit and/or have long orbital periods. Another indicator for a lower-quality astrometric fit that we can use is RUWE. Figure 6.9 shows distribution of potentially problematic large RUWE among single and multiple candidates. The latter, on average, have a much poorer fit quality that might indicate a presence of an additional parameter that needs to be considered in the astrometric fit.

Distances to triple stars, shown by the histogram in Figure 6.15 range from around 0.3 to 0.9 kpc. From this we can assume the maximal allowable distance between components of an outer pair to be in the order of $100 - 350$ AU, pointing to outer orbital periods larger than 500 years. To test if such systems would meet our detection constraints, we created 100,000 synthetic binary systems whose orbital parameters were uniformly distributed within the parameter ranges given in Table 6.6. Observable radial velocities of both components were computed using the following equations:

$$v_1 = \frac{2\pi \sin i}{P\sqrt{1-e^2}} \frac{aq}{1+q} \Big(\cos(\theta + \omega) + e \cos \omega \Big), \tag{6.9}$$

$$v_2 = \frac{2\pi a \sin i}{P\sqrt{1-e^2}} \frac{a}{1+q} \Big(\cos(\theta + \omega + \pi) + e \cos(\omega + \pi) \Big), \text{ and} \tag{6.10}$$

$$P = \sqrt{a^3 \frac{4\pi^2}{GM_1(1+q)}}. \tag{6.11}$$

The parameters in the above equations represent the following physical description of a binary orbit: i - inclination, a - length of a semi-major axis, q - mass ratio between the stars, e - eccentricity of the elliptical orbit, P - orbital period, θ - true anomaly of an orbit, and ω - argument of the closest separation in orbit.

Table 6.6: Ranges of the orbital parameters used for the prediction of observable radial velocity separation between stars in an outer binary pair. The range of the semi-major axis length a is set between the *Gaia* separability limit for the closest and farthest triple candidate. The uniform distribution of a is a good approximation of the real periodicity distribution published by Raghavan *et al.* [333] as we are sampling a narrow range of it. Use of the real observed distribution would in our case introduce insignificant changes in radial velocity separation as we are simulating wide, slowly rotating systems.

Parameter	Considered range
M_1	2 $M_\odot$
a	100 ... 350 AU
q	0.45 ... 0.55
$\sin i$	0 ... 1 (i = 0 ... 90 deg)
e	0.1 ... 0.8
phase	0 ... 1, used for calculation of θ
ω	0 ... 360 deg

The distribution of velocity separations ($\Delta v = v_2 - v_1$) for synthetic systems is given by Figure 6.16, where we can see that more than 99.7 % of generated configurations would produce a spectrum that would still be considered as a solar twin ($\Delta v < 6$ km s^{-1}, see Section 6.12.1 and Figure 6.21 for further clarification). If we set the semi-major axis a (in Table 6.6) to single value in the same simulation, we can find the closest separation that would still meet the same observational criteria in at least 68 % of the cases. In our simulation this happens at the mutual separation of 10 AU (and at 50 AU for 95 % of the cases). The orbital period of a such outer pair is about 18 years (and 200 years for 50 AU) and is most probably way too long to had significant effect on the quality of the astrometric fit. Considering observationally favorable orbital configurations with face-on orbits, the actual system could be much more compact than estimated.

6.11.2 Inner binary pair and formation of double lines in a spectrum

An upper estimate of orbital sizes for an outer pair gives us a confirmation that almost every considered random orbital configurations would satisfy the observational and selection constraints. To ensure the long-term stability of such a system, the inner pair must have a period that is at least five times shorter [373]. At such short orbital periods, the inner stars could potentially move sufficiently rapidly in their orbits as to produce noticeable absorption line splitting for edge-on orbits. To be confident that none of the analysed objects produces such an SB2 spectrum, we visually checked all considered spectra and found no noticeable line splitting in any of the acquired GALAH or Asiago spectra. A subtle hint about a possible broadening of the spectral lines comes from the determined $v \sin i$. The median of its distri-

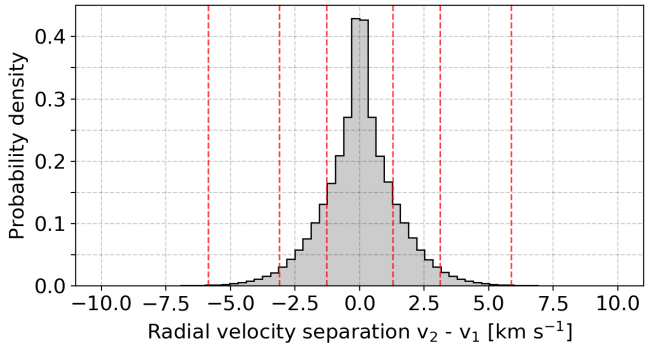

Figure 6.16: Distribution of computed radial velocity separations between a primary and secondary component of the simulated binary systems defined by the orbital parameters given in Table 6.6. Red vertical lines show 1, 2, and 3 σ probabilities of the given distribution.

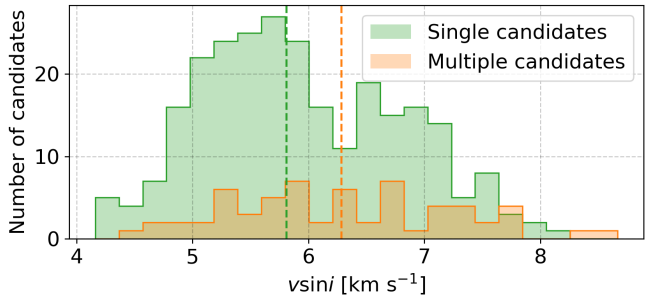

Figure 6.17: Distribution of determined $v \sin i$ for analysed single and multiple candidates. Median values of the distributions are marked with dashed vertical lines.

bution in Figure 6.17 is higher for multiple candidates with the excess projected velocity of $\sim$0.5 km s^{-1}.

Accounting for the GALAH resolving power and spectral sampling, we can estimate a minimal radial velocity separation between components of a spectrum to show clear visual signs of duplicated spectral lines. To determine a lower limit, we combined two solar spectra of different flux ratios and visually evaluated when the line splitting becomes easily noticeable. With equally bright sources, this happens at the separation of $\sim$14 km s^{-1}. When the secondary component contributes 1/3 of the total flux, minimal separation is increased to $\sim$20 km s^{-1}. A similar separation is needed when secondary contributes only 10 % of the flux. As no line splitting is observed in our analysed spectra, we can be confident that the velocity separation between the binary components was lower than that during the acquisition.

Considering the minimal ratio between outer and inner binary period, we can deduce an expected radial velocity separation of an inner pair for the widest possible orbits. As explained in previous section, we used Equations 6.9-6.11 and possible ranges of inner orbital parameters (Table 6.7) to generate a set of synthetic binary systems. The distribution of their Δv is shown in Figure 6.18 and represent inner binaries with orbital periods from 100 to 700 years. At those orbital periods, more than 92 % of the considered configurations would satisfy the condition of $\Delta v < 4$ km s^{-1},

Table 6.7: Same as Table 6.6, but for an inner binary of a hierarchical triple stellar system.

Parameter	Considered range
M_1	$M_\odot$
P_S	P_L / 5, used for calculation of a
q	0.9 ... 1.0
$\sin i$	0 ... 1 (i = 0 ... 90 deg)
e	0.1 ... 0.8
phase	0 ... 1, used for calculation of θ
ω	0 ... 360 deg

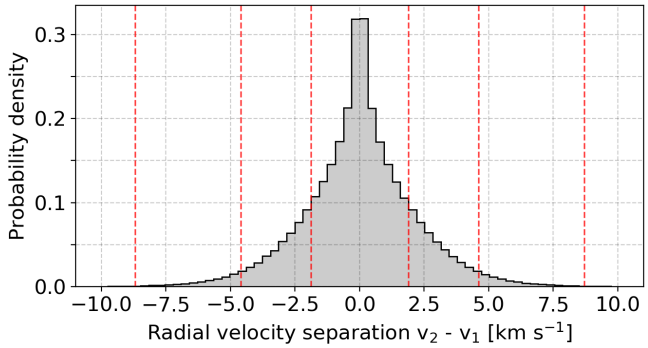

Figure 6.18: Same as Figure 6.16, but for an inner binary of a simulated triple stellar system.

ensuring that the observed composite of two equal solar spectra would still be considered as a solar twin (see Section 6.12.1 and Figure 6.21 for further clarification). If we limit our synthetic inner binaries to only one orbital period, we can estimate the most compact system that would still meet the observational criteria in at least 68 % of the cases. Those criteria are satisfied at the orbital period of 40 years with a semi-major axis of 14 AU.

6.11.3 Multi-epoch radial velocities

To support our claims about slowly changing low orbital speeds in detected triple candidates, we analysed changes in their measured radial velocities between the GALAH and other comparable all-sky surveys. Distributions of changes are presented by three histograms in Figure 6.20. There almost all velocity changes, except one, are within 5 $\mathrm{km\,s^{-1}}$. Differences between the GALAH and *Gaia* radial velocities were expected to be small as the latter reports median velocities in the time-frame that is similar to the acquisition span of the GALAH spectra. Extending the GALAH observations in the past with RAVE DR5 [96] and in the future with Asiago observations did not produce any extreme changes. For this comparison, we

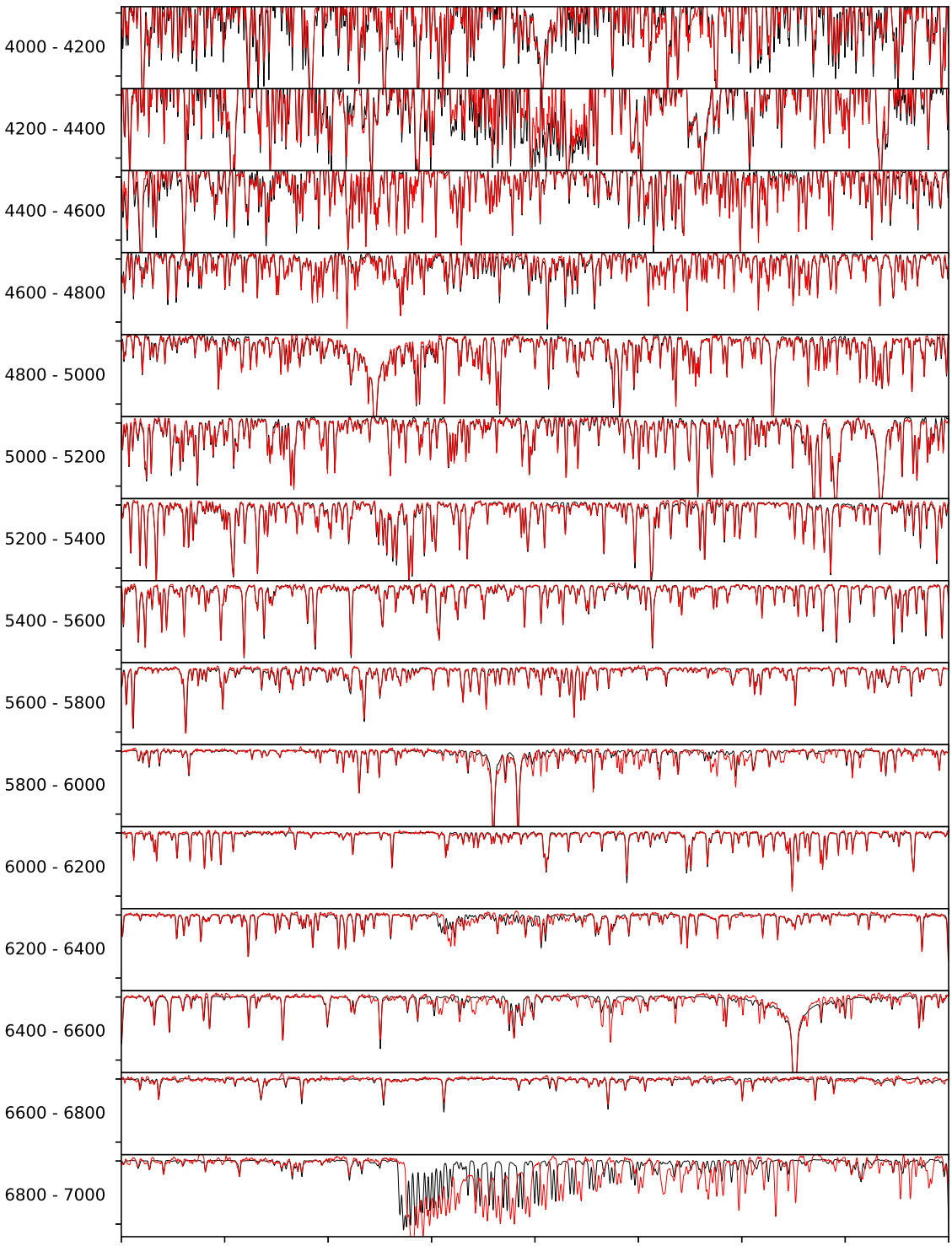

Figure 6.19: Follow up observation of the star TYC 502-985-1. Our observed and normalised spectrum is shown with the red curve and high-resolution solar spectrum convolved to the same resolving power in black. Labels on the left side give wavelength ranges of the individual subplot. To generate this plot, shown subsets of both spectra were renormalised with a linear function and the same sigma clipping levels. Combination of different barycentric and radial velocity is evident in the bottom infra-red panel, where the absorption feature of Earth atmosphere is dominant and shifted among observations.

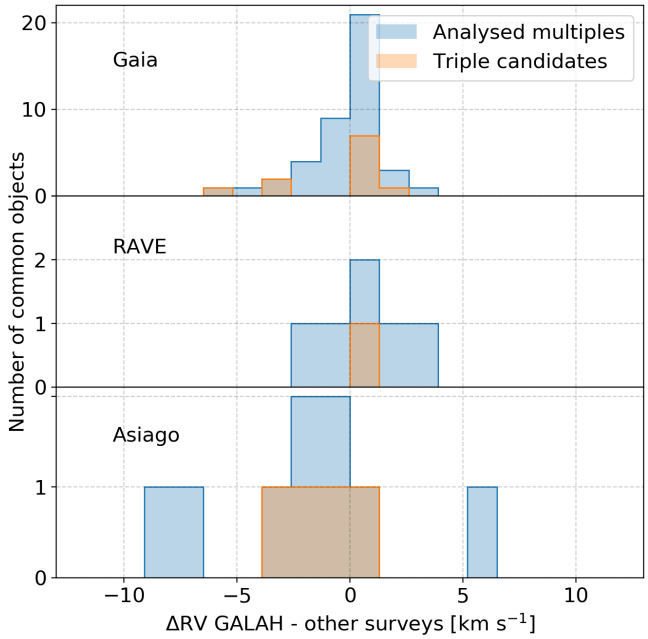

Figure 6.20: RV difference between GALAH and other large spectroscopic surveys. The name of an individual survey or observatory is given in the upper-left corner of every panel. Blue histograms represent velocity difference for all investigated multiple candidates and orange histograms for discovered triple systems only.

also have to consider the uncertainty of the measurements. They are in the order of $\sim$2 km s^{-1} for RAVE and $\sim$1 km s^{-1} for Asiago spectra. Example of the spectrum acquired for one of the best solar twins is presented and compared towards the real solar spectrum in Figure 6.11.3.

With the data synergies of those three surveys, we could produce only three observational time-series that have observations at more than two sufficiently separated times. With only three data points in each graph, not much can be said about actual orbits, especially if their temporal variability is much shorter that time between different acquisitions.

6.12 Simulations and tests

To determine the limitations of the employed algorithms, we used them to evaluate a set of synthetic photometric and spectroscopic sources. As the complete procedure depends on the criteria for the selection of solar twin candidates, we first investigated if the selection criteria allows for any broadening of the spectral lines or their multiplicity as they are both signs of a multiple system with components at different projected radial velocities.

In the second part, we generated a set of ideal synthetic systems that were analysed by the same fitting procedure as the observed data set. The results of this analysis are used to determine what kind of systems could be recognized with the fitting procedure and how the results could be used to spot suspicious combinations of fitted parameters.

In the final part, we try to evaluate selection biases that might arise from the

position of analysed stars on the *Gaia* colour-magnitude diagram and from the GALAH selection function that picks objects based on their apparent magnitude.

6.12.1 Radial velocity separation between components

To determine the minimal detectable radial velocity (RV) separation between components in a binary or a triple system, we constructed a synthetic spectrum resembling an observation of multiple Suns at a selected RV separation. The spectrum of a primary component was fixed at the rest wavelength with RV = 0 $\mathrm{km\,s^{-1}}$ and the secondary spectrum shifted to a selected velocity. After the shift, these two spectra were added together based on their assumed flux ratio.

The generated synthetic spectrum was compared to the solar spectrum with exactly the same metric as described in Section 6.2.4. Computed spectral similarities at different separations were compared to the similarities of analysed solar twin candidates. The first RV separation that produces a spectrum that is more degraded than the majority of solar twin candidates was determined to be a minimal RV at which the observed spectrum would be degraded enough that it would no longer be recognised as a solar twin. The high SNR of the generated spectrum was not taken into account for this analysis in contrast to the algorithm that was used to pinpoint solar twin candidates. Therefore we also omitted candidates with a lower similarity that in our case directly corresponds to their low SNR.

The result of this comparison is presented in the left side of Figure 6.21, where we can see that the minimal detectable separation of two equally bright stars resembling the Sun is $\sim$4 $\mathrm{km\,s^{-1}}$. In the case where a primary star contributes 2/3 of the total flux, the minimal RV increases to $\sim$6 $\mathrm{km\,s^{-1}}$ (see right side of Figure 6.21). Further increase in the ratio between their fluxes would also increase a minimal detectable separation, but only to a certain threshold from where on a secondary star would not contribute enough flux for it to be detectable, and its received flux would be comparable to the typical HERMES spectral noise. In our case, this happens when the secondary contributes less than 10 % of the total flux. These boundaries are only indicative as they also heavily depend on the quality of the acquired spectra. When a low level of noise with the Gaussian distribution ($\sigma = 0.01$) is added to the second component with a comparable luminosity, the minimal RV decreases because the similarity between spectra also decreases. In that case, the similarity for $\Delta v = 0$ is located near the mode value of similarity distribution in Figure 6.21.

6.12.2 Analysis of synthetic multiple systems

The limitations and borderline cases of the fitting procedure were tested by evaluating its performance on a set of synthetic binary and triple systems that were generated using the same models and equations as described in the fitting procedure. An ideal, virtually noiseless set comprised 95 binary and 445 triple systems whose components differ in temperature steps of 100 K. The hottest component in the system was set to values in the range between 5300 and 6200 K. The T_{eff} of the coldest component could go as low as 4800 K. The [Fe/H] of all synthetic systems was set to 0.0 to mimic solar-like conditions. Condensed results are presented in Figures 6.22.

As expected, the fitting procedure (Section 6.9) did not have any problem de-

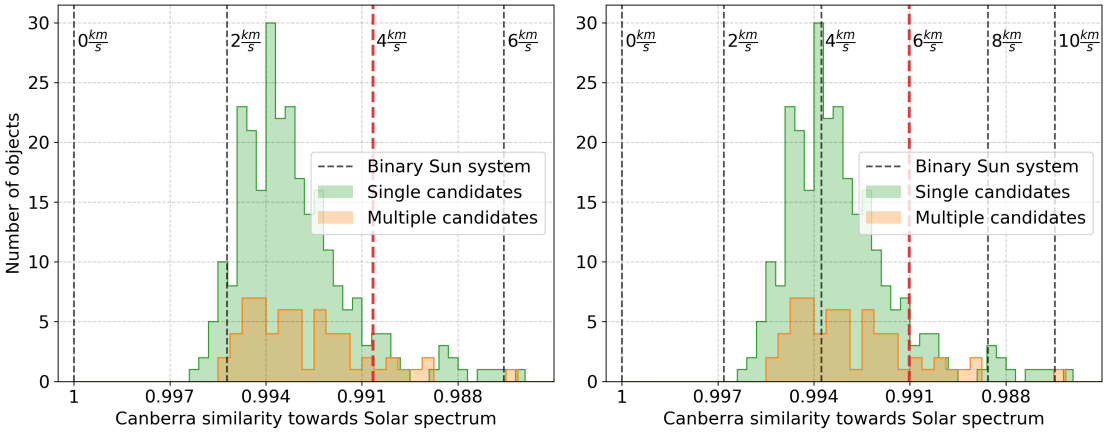

Figure 6.21: Histograms of Canberra similarities towards solar spectrum for multiple stellar candidates (orange histogram) and single star candidates (green histogram) in the red HERMES arm. Vertical dashed lines in the left plot represent the same similarity measure, but for a binary system comprising of two equally bright Suns, whose components are at different radial velocities. The separation between components is increasing in 2 $\mathrm{km\,s^{-1}}$ steps, where the leftmost vertical line represents the case where both stars are moving with the same projected velocity. The selected maximal velocity at which the composite spectrum would still be considered a solar twin is marked with the thicker red vertical line. The right plot shows the same data, but for the case of a triple system where only one component out of three has a radial velocity shift in comparison to other two. Distribution of histograms also shows that spectra of multiple system candidates are as (dis)similar as spectra of single stars.

termining the correct configuration of the synthesized system. Temperatures of the components were also correctly recovered with a median error of 0 ± 13 K for the hottest component and 0 ± 43 K for the coldest component in a triple system. A more detailed analysis with the distribution of prediction errors, where the temperature difference between components is taken into account, is presented in Figure 6.22. From that analysis, we can deduce that $T_{\rm eff}$ of the secondary component is successfully retrieved if it does not deviate by more than 1000 K from the primary. Results at such large temperature difference are inconclusive as the number of simulated systems drops rapidly. The same can be said for the tertiary component, but with the limitation that it should not be colder by more than 700 K when compared to the secondary star. Beyond that point, the uncertainty of the fitted result increases and a star is determined to be hotter as it is.

Another application of such an analysis is to identify signs that could point to a possibly faulty solution when it is comparably likely that a multiple system is comprised of two or three components. The results of such an analysis are presented in the right part of Figure 6.22, where we tried to describe a binary system with a triple system fit. From the distribution of errors, we can observe that the effective temperature of a primary star with the largest flux is recovered with the smallest fit errors whose median is 20 ± 93 K. As we are using too many components in that fit, $T_{\rm eff}$ of a secondary is reduced in order to account for the redundant tertiary component in the fit. As imposed by the limit in the fitting procedure, the tertiary $T_{\rm eff3}$ is set to as low as possible. If the same thing happens for a real observed system, this could be interpreted as a model over-fitting.

6.12.3 Triple stars across the H-R diagram

In the current stage of Galactic evolution, binary stars with a solar-like $T_{\rm eff}$ are located near a region that is also occupied by main sequence turnoff stars in the Kiel and colour-magnitude diagram (Figures 6.10 and 6.13). This, combined with the fact that older stars with a comparable initial mass and metallicity would also pass a region occupied with binaries (Figure 6.9) poses an additional challenge for detection of unresolved multiples if their spectrum does not change sufficiently during the evolution.

To analyse the possible influence of the turnoff stars and older, more evolved stars, we ran the same detection procedure as described in Section 6.2 and analysis (Section 6.9) to determine the fraction of binaries and triples at different $T_{\rm eff}$, ranging from 5100 to 6000 K, with a step of 100 K. A comparison median spectrum was computed from all spectra in the range $\Delta T_{\rm eff} \pm 60$ K, $\Delta \log g \pm 0.05$ dex and Δ[Fe/H] $\pm$ 0.05 dex, where the main sequence $\log g$ was taken from the main sequence curve shown in Figure 6.10, and [Fe/H] was set to 0.0. The limiting threshold for multiple candidates was automatically determined in the same way as for solar twin candidates (Section 6.7.1).

Condensed results of the analysis are shown in Figure 6.23, where we can see that both the fraction of stars with excessive luminosity and triple candidates starts increasing above the $T_{\rm eff} \sim 5600$ K. This probability increase might indicate that the underlying distribution of more evolved and/or hotter stars might have some effect on our selection function. On the other hand, this increasing binarity fraction coincides with other surveys that show similar trends [334].

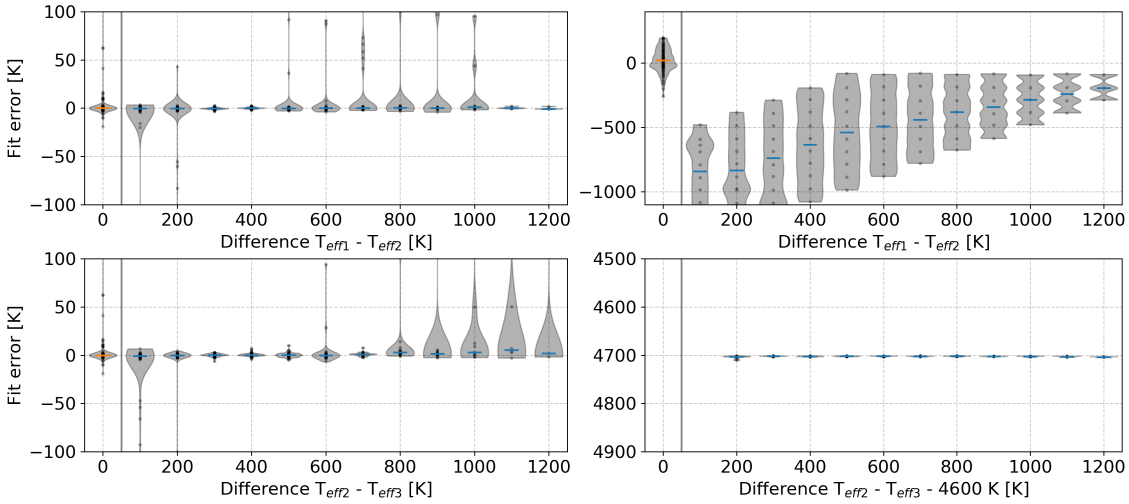

Figure 6.22: The accuracy of our analysis when the fitting procedure was applied to a synthetic triple system (left column) and synthetic binary sistem (right column). Upper panels show the distribution of $T_{\rm eff}$ prediction errors for a secondary star depending on the difference between selected temperatures of a primary and a secondary star. The lower panel shows the same relation but for a tertiary star in comparison to a secondary. As a reference, the prediction error of a primary star is shown on the left side of both panels. A strange wavy pattern in graphs is a consequence of a low number of sample points. Labels $T_{\rm eff1}$, $T_{\rm eff2}$, and $T_{\rm eff3}$ indicate decreasing effective temperatures of stars in a simulated system. In the case when we try to describe a synthetic binary system using the triple star model, reference $T_{\rm eff3}$ was set to 0 K. Therefore all results for the tertiary component try to choose a star with the temperature as low as possible, but it can not go lower as model limitation of 4700 K.

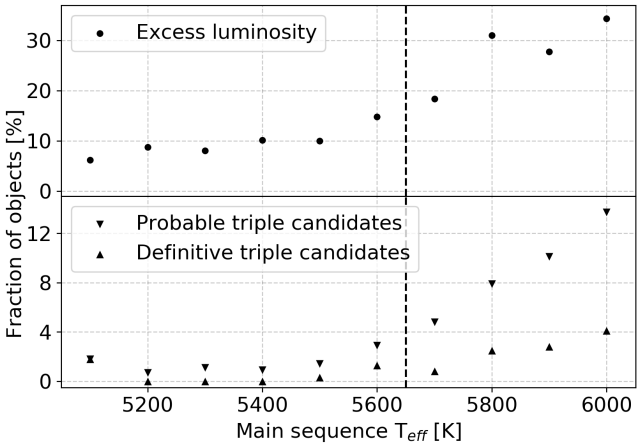

Figure 6.23: The upper panel shows the percentage of objects at different positions on the Kiel diagram shown in Figure 6.10 that show excessive luminosity and are spectroscopically similar to main sequence stars. Similarly, the lower panel shows upper and lower boundary on a percentage of triple system candidates at the same positions. For a definitive candidate, both fits must agree on a triple configuration. The strong dashed vertical line repents a point on Kiel diagram where the main sequence visually starts merging with the red-giant branch, the point where a region above the main sequence becomes polluted with more evolved stars.

As the detected triple star candidates encompass a fairly small region in a colour-magnitude space, we were interested in the degree to which our analysed spectra are similar to those of other stars in the same region. For this purpose, the GALAH objects with absolute *Gaia* corrected magnitudes in the range $3.3 < M_G < 3.6$ and $0.77 < G_{BP}\text{-}G_{RP} < 0.84$ were selected and plotted in Figure 6.24, where spectra are arranged according to their similarity. In this 2D projection (details about the construction of which are given in Traven *et al.* [84] and Buder *et al.* [104] it is obvious that spectra considered in this chapter clearly exhibit a far greater degree of mutual similarity than other spectra with similar photometric signature. As expected, many of them lie inside the region of SB2 spectra. All of our solar twin candidates were visually checked for the presence of a resolvable binary component. Nevertheless, this plot gives us additional proof of their absence as none of the analysed twins is located inside that region.

6.12.4 Observational bias - Galaxia model

Every magnitude limited survey, such as the GALAH, will introduce observational bias into the frequency of observed binary stars as their additional flux changes the volume of the Galaxy from where they are sampled. Their distances and occupied volume of space is located further from the Sun and therefore also more spacious than for comparable single stars with the same apparent magnitude and colour.

To evaluate this bias for our type of analysed stars, we created a synthetic Milky Way population of single stars using the Galaxia code [378]. First, the code was run to create the entire population of stars in the apparent V magnitude range between 10 and 16 without any colour cuts. In order for the distribution of synthe-

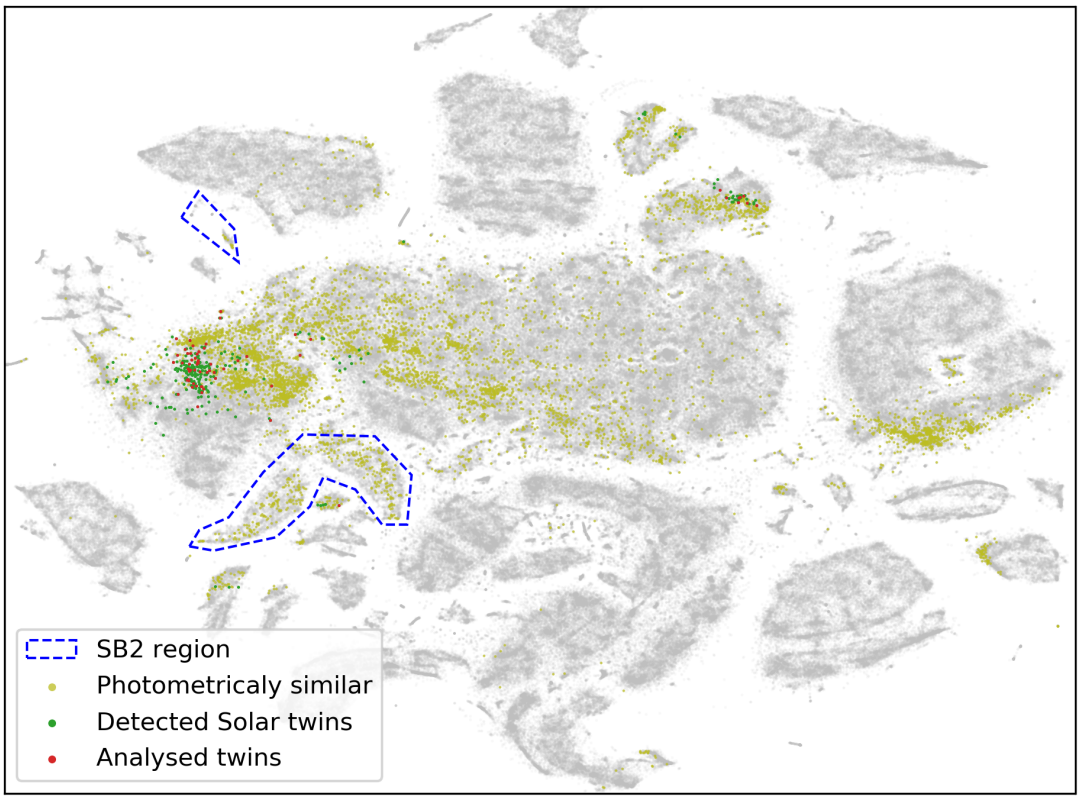

Figure 6.24: Visual representation of similarities between spectra using dimensionality reduction analysis. Clumps in this 2D projection represent morphologically similar spectra, whose features separate them from the rest of the data set. Blue-dashed polygons indicate over-densities, where SB2 spectra are located (projection and regions are taken from Figure 13 in Buder *et al.* [104]). Each grey dot represent one spectrum of GALAH survey. Green coloured dots represent objects considered in this chapter, of which objects marked with red were analysed for higher order multiplicity. Yellowish dots represent objects whose *Gaia* magnitudes fall inside the range of determined triple star candidates.

sised stars to mimic the GALAH survey as closely as possible, only stars located inside the observed GALAH 2° fields were retained. After the spatial filtering, only stars with Sun-like absolute magnitude $M_V = 4.81 \pm 0.05$ and colour index $B - V = 0.63 \pm 0.05$ were kept in the set (reference magnitudes were taken from Willmer [366]).

From the filtered synthetic data set we took three different sub-sets representing a pool of observable single solar twins ($12.0 \leq \mathrm{V} \leq 14.0$), binary twins ($12.75 \leq \mathrm{V} \leq 14.75$), and triple twins ($13.2 \leq \mathrm{V} \leq 15.2$) based on their apparent V magnitudes. Extinction and reddening were not used in this selection as their use did not significantly alter the relationship between the number of stars in sub-sets. Assuming that the frequency of multiple stars is constant and does not change in any of those sets, we can estimate the selection bias on the frequency of multiple stars. The number of stars in each sub-set was 15862, 35470, and 54567 respectively. According to those star counts, the derived frequency of binaries would be too high by a factor of 2.2 and a factor 3.4 for triple stars. Considering those factors, the fraction of unresolved triple candidates with solar-like spectra is ∼2 % and ∼11 % for binary candidates.

Accidental visual binaries that lie along the same line-of-sight and have angular separation smaller that the field-of-view of the 2dF spectroscopic fibre (2″) or smaller than the *Gaia* end-of-mission angular resolution (0.1″) were not considered in this estimation.

6.13 Conclusions

Combining multiple photometric systems with spectroscopic data from the GALAH and astrometric measurements taken by *Gaia*, we showed the possible existence of triple stellar systems with long orbital periods whose combined spectrum mimics solar spectrum. The average composition of such a system consists of three almost identical stars, where one of the stars is ∼10 K warmer than the Sun. The coldest has an effective temperature of ∼120 K below the Sun. The derived percentage of such unresolved systems would be different for nearby/close stars as they become more/less spatially resolved. In the scope of our magnitude limited survey, we sampled only a fraction of possible distances to the systems.

Without any obvious signs of the orbital periodicity in the measured radial velocities of the systems, orbital periods were loosely constrained based on the observational limits and few assumptions. By the prior assumption that the *Gaia* spacecraft sees those systems as a single light source, we showed that they could be described by orbital periods where a difference between projected velocities of components does not sufficiently degrade an observed spectrum. The spectroscopic signature and radial velocity variations were further used to put a limit on the minimum orbital period of an inner pair to be at least 20 years. Shorter periods are not completely excluded as it could happen that the spectrum was acquired in a specific orbital phase where the difference between projected orbital speeds is negligible. From the fact that analysed objects are spatially unresolvable for the *Gaia* spacecraft, the orbital size of outer binary pair can extend up to 100-350 AU and therefore have orbital periods of the order of a few hundred years.

To confirm their existence, detected systems are ideal candidates to be observed with precise interferometric measurements or high time-resolution photometers if they happen to be occulted by the Moon. Simulation of the lunar motion showed

that four of the analysed multiple candidates lie in its path if the observations would be carried out from the Asiago observatory that has suitable photon-counting detectors.

The main drawback of the analysis was found to be its separate treatment of photometric and spectroscopic information in two independent fitting procedures. In future analyses, they should be combined to acquire even more precise results as different stellar physical parameters have a different degree of impact on these two types of measurements.

6.14 Table description and summary

In the Table 6.8 we provide a list of metadata available for every object detected using the methodology described in this chapter. The complete table of detected objects and its metadata is available in electronic form at the CDS. An excerpt of the table, containing a subset of columns, for definitive and probable triple candidates is given in Table 6.9.

Table 6.8: List and description of the fields in the published catalogue of analysed objects.

Field	Unit	Description
`source_id`		*Gaia* DR2 source identifier
`sobject_id`		Unique internal per-observation star ID
`ra`	deg	Right ascension from *Gaia* DR2
`dec`	deg	Declination from *Gaia* DR2
`ruwe`		Value of re-normalized astrometric χ^2
`m_sim_p`		Photometric χ^2 for original parameters
`m_sim_f`		Spectroscopic χ^2 for original parameters
`s1_teff1`	K	$T_{\rm eff1}$ in a fitted single system
`s1_feh`		[Fe/H] of a fitted single system
`s1_sim_p`		Photometric χ^2 of a fitted single system
`s1_sim_f`		Spectroscopic χ^2 of a fitted single system
`s2_teff1`	K	$T_{\rm eff1}$ in a fitted binary system
`s2_teff2`	K	$T_{\rm eff2}$ in a fitted binary system
`s2_feh`		[Fe/H] of a fitted binary system
`s2_sim_p`		Photometric χ^2 of a fitted binary system
`s2_sim_f`		Spectroscopic χ^2 of a fitted binary system
`s3_teff1`	K	$T_{\rm eff1}$ in a fitted triple system
`s3_teff2`	K	$T_{\rm eff2}$ in a fitted triple system
`s3_teff3`	K	$T_{\rm eff3}$ in a fitted triple system
`s3_feh`		[Fe/H] of a fitted triple system
`s3_sim_p`		Photometric χ^2 of a fitted triple system
`s3_sim_f`		Spectroscopic χ^2 of a fitted triple system
`n_stars_p`		Best fitting photometric configuration
`n_stars_f`		Best fitting spectroscopic configuration
`class`		Final configuration classification
`flag`		Result quality flags

Table 6.9: Subset of results for definitive and probable solar-like triple candidates detected by our selection and fitting procedure. The complete table is given as a supplementary material to this chapter in a form of the textual CSV file. It is also available in the electronic form at the CDS portal.

source_id	ruwe	s2_teff1	s2_teff2	s2_feh	s3_teff1	s3_teff2	s3_teff3	s3_feh	class	flag
6157059919188478720	11.1	6002	6000	0.24	5825	5787	5780	0.03	3	1
6777339306532222080	8.0	5875	5819	0.10	5642	5636	5631	-0.06	3	1
6412815502155127808	1.1	5880	5316	-0.02	5922	4703	4701	0.03	>2	4
6564302331580491904	1.1	5991	4701	0.13	5945	4702	4700	0.02	>2	0
5386113598793714304	1.5	5882	5808	0.07	5878	5438	5313	-0.03	3	5
2534579880633620992	1.2	5986	5876	0.15	5750	5739	5714	-0.06	3	6
5484353352124904064	1.1	5876	5802	0.14	5844	5525	5425	-0.00	>2	0
5399712362903401984	1.3	5968	4703	0.08	5922	4703	4701	0.01	>2	4
3688523450018482432	1.1	5820	5688	-0.015	5998	4717	4702	0.09	>2	2
6220408320279116032	7.6	5923	5883	0.11	5724	5677	5669	-0.08	3	1
6198738457927949440	0.9	6052	4798	0.15	5986	4704	4701	0.10	>2	4
5822222074090352384	4.7	5651	5640	-0.04	5866	4702	4701	0.01	>2	1
6355462192511955456	1.0	5700	5692	-0.11	5471	5449	5379	-0.28	>2	4
4678229218054642176	0.9	6028	4791	0.09	5968	4702	4700	0.09	>2	4
4423111085550775680	1.6	5829	5774	-0.02	5645	5633	5398	-0.11	>2	1
6261736621608044416	4.7	5927	5890	0.07	6154	4702	4701	0.20	>2	1
5947219160131475712	1.1	6071	5843	0.22	6007	5724	5265	0.10	3	4
6661888382295654656	1.5	5725	5678	-0.07	5956	4706	4701	0.04	>2	1
6362136502970853120	1.8	5844	5826	0.013	5666	5658	5540	-0.18	>2	1
6645693508028329728	1.3	5936	4715	0.06	5833	4703	4701	-0.02	>2	4

Chapter 7

Conclusions and future prospects

The fast increase in observational data seen in the last decade with the introduction of new and improved astronomical observational facilities requires new approaches to the exploration of acquired data. The old approach of exact and painstaking analysis of individual objects have to be changed in order to grasp the full potential of large observational sets. On the other hand, large amounts of data and complex observational scenarios also lead to more complicated data reduction and analysis pipelines. As it is impossible to look through all acquired data and consider all variables, many computer algorithms have over the years been developed to help with those tasks. Of them, the most trending and occasionally misused are numerous machine learning procedures of classification, clustering, and regression.

In this thesis, we used some of the available machine learning tools to explore large astronomical datasets, mainly collected as part of the GALAH and *Gaia* large sky surveys. The data of those surveys were, when needed, supplemented with results of other spectroscopic and photometric surveys. This merger gave us additional information by observing a broader wavelength range and an increase in temporal coverage of variabilities by combing similar data sets.

The body of this thesis focuses on exploring different types of stars. Usually, the stars are generally separated into two broad groups: normal and peculiar. The first term describes the majority of the stars that are happily living through the longest periods of their lifetime. Being non-problematic, they are easily modelled and preferred for chemical analyses such as our exploration of open stellar clusters and their surrounding. During the analysis, we explored the possibility of chemical differentiation between chemical signatures of cluster members, possible past members and surrounding stars. Separation into given components for 11 open clusters was based on kinematic vectors defined by *Gaia* observations. The GALAH survey supplied the chemical signature for a subset of stars. The analysis of open clusters in our dataset showed that they are not chemically as homogeneous as theory says and as much as we would like. Not knowing the origin of this discrepancy, we shifted to the differential chemical analysis that takes out the identified trends and compares only chemical signatures of stars with the same stellar parameters, most commonly effective temperature. The results show that uncertainties of the determined abundances and their scatter limit chemical separability between an open cluster and field stars as the majority have very comparable signatures. We showed that inclusion of kinematic information helps with the delineation as it was more likely that potentially ejected stars were chemically tagged to the cluster than remaining field

stars.

The second large group of stars that are not wanted in the above-described analyses are peculiar stars. The term is inclusive and depends on the scientific question in hand and observed wavelength region. In our case, the definition includes spectroscopically detectable classes such as interacting stars, multiple stellar systems, stars in temporally short-lived evolutionary stages and spectra with unexpected chemical compositions. Our selection of all peculiar classes depended on the comparison between normal-looking spectra and observed spectra. The modelling of normal spectra was done in multiple different ways: by averaging spectra with similar parameters, running observed spectra trough a neural network autoencoder and modelling using spectrum generative approach *The Cannon*. The direct comparison gave us a difference between spectral modelling and reality. When we were looking at the specific wavelength regions, we uncovered spectra with pronounced molecular absorption bands of C_2 - stars that are carbon-rich and spectra with expressed hydrogen, [NII], and [SII] emission lines. As the emission lines are results of different ongoing processes on a star and around it, we tried to delineate them on chromospheric and nebular contributions using the derived line parameters such as position, strength and shape. Combination of normal single star spectra can also be used for the spectral disentanglement of multiple stellar systems. Our methodology was used to describe over-luminous stars whose spectra did not show signs of multiple stars in a system. Therefore we assumed that comprising stars in the system must be orbiting on wide orbits. By combining photometric and spectroscopic single star models, we were able to reproduce the observable quantities and confirm the existence of triple systems with near-identical stars. Orbits of such proposed systems were also modelled to confirm that their configuration is consistent with observational limitations.

All tabulated results presented in this thesis are also published as freely available catalogues on the astronomical online database collection VizieR and directly from the publishers' website. The compiled catalogues could serve as a starting point for diverse additional lines of research which focus on the exact physics behind identified peculiar stars and their spectra. Some of the possibilities for future studies were already given in the text and will be considered as potential observing possibilities at the available observing facilities, such as Asiago observatory where we are conducting spectroscopic observations almost every month. Most often, spectroscopic data must be complemented with photometric sets to explore or confirm additional possible physical scenarios behind interesting objects.

The hype of the big data era in astronomy is still to increase as many new telescopes and instruments are currently under development and construction. Their observations are planned to start in the following years. Until then, the *Gaia* satellite is continuously scanning our sky in order to bring us the most precise distances and movements for billions of stars that we are all waiting for. The next grander data release is still at least a year away, but everyone is already questioning how much more can it do for science in the field of galactic archaeology. In our case, the updated distances could completely change the membership, shape and structure of open clusters and reclassify exciting possible triple stars to dull single free-floating stars because of their parallactic uncertainty. Until then, we have to double-check our applied quality filters and trust in the data and parameters currently disposable at our hands. The journey into big astronomical data does not seem to be stopping

anywhere in the next years.

Bibliography

[1] C. Chiappini, F. Matteucci and D. Romano, *Abundance Gradients and the Formation of the Milky Way*, **554**, 1044 (2001).

[2] T. Naab and J. P. Ostriker, *Theoretical Challenges in Galaxy Formation*, **55**, 59 (2017).

[3] C. J. Lada and E. A. Lada, *Embedded Clusters in Molecular Clouds*, **41**, 57 (2003).

[4] V. V. Gvaramadze, A. Gualandris and S. Portegies Zwart, *On the origin of high-velocity runaway stars*, **396**, 570 (2009).

[5] E. Gaburov, J. Lombardi, James C. and S. Portegies Zwart, *On the onset of runaway stellar collisions in dense star clusters - II. Hydrodynamics of three-body interactions*, **402**, 105 (2010).

[6] T. Ryu, N. W. C. Leigh and R. Perna, *Formation of runaway stars in a star-cluster potential*, **470**, 3049 (2017).

[7] A. Gualandris and S. Portegies Zwart, *A hypervelocity star from the Large Magellanic Cloud*, **376**, L29 (2007).

[8] J. Kos, J. Bland-Hawthorn, K. Freeman, S. Buder, G. Traven, G. M. De Silva, S. Sharma, M. Asplund, L. Duong, J. Lin, K. Lind, S. Martell, J. D. Simpson, D. Stello, D. B. Zucker, T. Zwitter, B. Anguiano, G. Da Costa, V. D'Orazi, J. Horner, P. R. Kafle, G. Lewis, U. Munari, D. M. Nataf, M. Ness, W. Reid, K. Schlesinger, Y.-S. Ting and R. Wyse, *The GALAH survey: chemical tagging of star clusters and new members in the Pleiades*, **473**, 4612 (2018).

[9] A. McBride and M. Kounkel, *Runaway Young Stars near the Orion Nebula*, **884**, 6 (2019).

[10] K. Bekki, *Dynamical friction of star clusters against disc field stars in galaxies: implications on stellar nucleus formation and globular cluster luminosity functions*, **401**, 2753 (2010).

[11] S. Röser and E. Schilbach, *Praesepe (NGC 2632) and its tidal tails*, **627**, A4 (2019).

[12] F. C. Yeh, G. Carraro, M. Montalto and A. F. Seleznev, *Ruprecht 147: A Paradigm of Dissolving Star Cluster*, **157**, 115 (2019).

[13] S. Meingast and J. Alves, *Extended stellar systems in the solar neighborhood. I. The tidal tails of the Hyades*, **621**, L3 (2019).

[14] Y. Zhang, S.-Y. Tang, W. P. Chen, X. Pang and J. Z. Liu, *Diagnosing the Stellar Population and Tidal Structure of the Blanco 1 Star Cluster*, **889**, 99 (2020).

[15] G. Carraro, *Photometry of dissolving star cluster candidates. The cases of NGC 7036 and NGC 7772*, **385**, 471 (2002).

[16] C. Bonatto, E. Bica and D. B. Pavani, *NGC 2180: A disrupting open cluster?*, **427**, 485 (2004).

[17] R. Carrera, M. Pasquato, A. Vallenari and et al., *Extended halo of NGC 2682 (M 67) from Gaia DR2*, **627**, A119 (2019).

[18] S. F. Portegies Zwart, P. Hut, J. Makino and S. L. W. McMillan, *On the dissolution of evolving star clusters*, **337**, 363 (1998).

[19] G. R. I. Moyano Loyola and J. R. Hurley, *Stars on the run: escaping from stellar clusters*, **434**, 2509 (2013).

[20] R. Wielen, *On the Lifetimes of Galactic Clusters*, **13**, 300 (1971).

[21] R. Wielen, *Dissolution of star clusters in galaxies*, in *The Harlow-Shapley Symposium on Globular Cluster Systems in Galaxies*, IAU Symposium, Vol. 126, edited by J. E. Grindlay and A. G. D. Philip (1988) pp. 393–406.

[22] H. Monteiro and W. S. Dias, *Distances and ages from isochrone fits of 150 open clusters using Gaia DR2 data*, **487**, 2385 (2019).

[23] D. Bossini, A. Vallenari, A. Bragaglia and et al., *Age determination for 269 Gaia DR2 open clusters*, **623**, A108 (2019).

[24] G. Carraro, *Open cluster remnants: an observational overview*, Bulletin of the Astronomical Society of India **34**, 153 (2006).

[25] D. B. Pavani and E. Bica, *Characterization of open cluster remnants*, **468**, 139 (2007).

[26] E. Bica, B. X. Santiago, C. M. Dutra, H. Dottori, M. R. de Oliveira and D. Pavani, *Dissolving star cluster candidates*, **366**, 827 (2001).

[27] E. E. Salpeter, *The Luminosity Function and Stellar Evolution.*, **121**, 161 (1955).

[28] J. M. Scalo, *The Stellar Initial Mass Function*, **11**, 1 (1986).

[29] G. Chabrier, *Galactic Stellar and Substellar Initial Mass Function*, **115**, 763 (2003).

[30] P. Kroupa, *On the variation of the initial mass function*, **322**, 231 (2001).

[31] S. Hony, D. A. Gouliermis, F. Galliano and et al., *Star formation rates from young-star counts and the structure of the ISM across the NGC 346/N66 complex in the SMC*, **448**, 1847 (2015).

[32] D. Raboud and J. C. Mermilliod, *Evolution of mass segregation in open clusters: some observational evidences*, **333**, 897 (1998).

[33] I. A. Bonnell and M. B. Davies, *Mass segregation in young stellar clusters*, **295**, 691 (1998).

[34] R. de Grijs, G. F. Gilmore, R. A. Johnson and A. D. Mackey, *Mass segregation in young compact star clusters in the Large Magellanic Cloud - II. Mass functions*, **331**, 245 (2002).

[35] C. Bonatto and E. Bica, *Mass segregation in M 67 with 2MASS*, **405**, 525 (2003).

[36] S. Dib, S. Schmeja and R. J. Parker, *Structure and mass segregation in Galactic stellar clusters*, **473**, 849 (2018).

[37] R. de Grijs, C. Li and A. M. Geller, *The dynamical importance of binary systems in young massive star clusters*, ArXiv e-prints (2015).

[38] P. Kroupa, M. G. Petr and M. J. McCaughrean, *Binary stars in young clusters: models versus observations of the Trapezium Cluster*, **4**, 495 (1999).

[39] P. J. T. Leonard, *Star counts in the open cluster NGC 2420*, **95**, 108 (1988).

[40] S. Schmeja, N. V. Kharchenko, A. E. Piskunov, S. Röser, E. Schilbach, D. Froebrich and R.-D. Scholz, *Global survey of star clusters in the Milky Way. III. 139 new open clusters at high Galactic latitudes*, **568**, A51 (2014).

[41] Gaia Collaboration, F. van Leeuwen, A. Vallenari, C. Jordi, L. Lindegren, U. Bastian, T. Prusti, J. H. J. de Bruijne, A. G. A. Brown, C. Babusiaux and et al., *Gaia Data Release 1. Open cluster astrometry: performance, limitations, and future prospects*, **601**, A19 (2017).

[42] Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti and et al., *Gaia Data Release 2. Summary of the contents and survey properties*, **616**, A1 (2018).

[43] T. Cantat-Gaudin, C. Jordi, A. Vallenari and et al., *A Gaia DR2 view of the open cluster population in the Milky Way*, **618**, A93 (2018).

[44] A. Castro-Ginard, C. Jordi, X. Luri, T. Cantat-Gaudin and L. Balaguer-Núñez, *Hunting for open clusters in Gaia DR2: the Galactic anticentre*, **627**, A35 (2019).

[45] U. Bastian, *Gaia 8: Discovery of a star cluster containing beta Lyrae – and of a larger old (extinct) star formation complex surrounding it*, arXiv e-prints , arXiv:1909.04612 (2019).

[46] L. Liu and X. Pang, *A Catalog of Newly Identified Star Clusters in Gaia DR2*, **245**, 32 (2019).

[47] G. Sim, S. H. Lee, H. B. Ann and S. Kim, *207 New Open Star Clusters within 1 kpc from Gaia Data Release 2*, Journal of Korean Astronomical Society **52**, 145 (2019).

[48] T. Cantat-Gaudin, A. Krone-Martins, N. Sedaghat and et al., *Gaia DR2 unravels incompleteness of nearby cluster population: new open clusters in the direction of Perseus*, **624**, A126 (2019).

[49] A. Castro-Ginard, C. Jordi, X. Luri, J. Álvarez Cid-Fuentes, L. Casamiquela, F. Anders, T. Cantat-Gaudin, M. Monguió, L. Balaguer-Núñez, S. Solà and R. M. Badia, *Hunting for open clusters in {Gaia} DR2:* 582 *new OCs in the Galactic disc*, arXiv e-prints , arXiv:2001.07122 (2020).

[50] H. Baumgardt, *The nature of some doubtful open clusters as revealed by HIPPARCOS*, **340**, 402 (1998).

[51] G. Carraro, *NGC 6994: An open cluster which is not an open cluster*, **357**, 145 (2000).

[52] E. Han, J. L. Curtis and J. T. Wright, *The Putative Old, Nearby Cluster Lodén 1 Does Not Exist*, **152**, 7 (2016).

[53] J. Kos, G. de Silva, S. Buder, J. Bland -Hawthorn, S. Sharma, M. Asplund, V. D'Orazi, L. Duong, K. Freeman, G. F. Lewis, J. Lin, K. Lind, S. L. Martell, K. J. Schlesinger, J. D. Simpson, D. B. Zucker, T. Zwitter, T. R. Bedding, K. Čotar, J. Horner, T. Nordlander, D. Stello, Y.-S. Ting and G. Traven, *The GALAH survey and Gaia DR2: (non-)existence of five sparse high-latitude open clusters*, **480**, 5242 (2018).

[54] T. Cantat-Gaudin and F. Anders, *Clusters and mirages: cataloguing stellar aggregates in the Milky Way*, **633**, A99 (2020).

[55] J. Choi, C. Conroy, Y.-S. Ting, P. A. Cargile, A. Dotter and B. D. Johnson, *Star Cluster Ages in the Gaia Era*, **863**, 65 (2018).

[56] D. R. Soderblom, *The Ages of Stars*, **48**, 581 (2010).

[57] J. Kos, J. Bland-Hawthorn, M. Asplund, S. Buder, G. F. Lewis, J. Lin, S. L. Martell, M. K. Ness, S. Sharma, G. M. De Silva, J. D. Simpson, D. B. Zucker, T. Zwitter, K. Čotar and L. Spina, *Discovery of a 21 Myr old stellar population in the Orion complex⋆*, **631**, A166 (2019).

[58] E. Terlevich, *Evolution of n-body open clusters*, **224**, 193 (1987).

[59] A. F. Seleznev, *Open-cluster density profiles derived using a kernel estimator*, **456**, 3757 (2016).

[60] D. F. Evans, *Evidence for Unresolved Exoplanet-hosting Binaries in Gaia DR2*, Research Notes of the American Astronomical Society **2**, 20 (2018).

[61] D. Birko, T. Zwitter and et al., *Single-lined Spectroscopic Binary Star Candidates from a Combination of the RAVE and Gaia DR2 Surveys*, **158**, 155 (2019).

[62] A. Widmark, B. Leistedt and D. W. Hogg, *Inferring Binary and Trinary Stellar Populations in Photometric and Astrometric Surveys*, **857**, 114 (2018).

[63] Gaia Collaboration, *Gaia Data Release 2. Observational Hertzsprung-Russell diagrams*, **616**, A10 (2018).

[64] K. Čotar, T. Zwitter, G. Traven, J. Kos, M. Asplund, J. Bland-Hawthorn, S. Buder, V. D'Orazi, G. M. de Silva, J. Lin, S. L. Martell, S. Sharma, J. D. Simpson, D. B. Zucker, J. Horner, G. F. Lewis, T. Nordlander, Y.-S. Ting, R. A. Wittenmyer and Galah Collaboration, *The GALAH survey: unresolved triple Sun-like stars discovered by the Gaia mission*, **487**, 2474 (2019).

[65] V. Belokurov, Z. Penoyre, S. Oh and et al., *Unresolved stellar companions with Gaia DR2 astrometry*, arXiv e-prints , arXiv:2003.05467 (2020).

[66] K. Freeman and J. Bland-Hawthorn, *The New Galaxy: Signatures of Its Formation*, **40**, 487 (2002).

[67] J. Bland-Hawthorn, T. Karlsson, S. Sharma, M. Krumholz and J. Silk, *The Chemical Signatures of the First Star Clusters in the Universe*, **721**, 582 (2010).

[68] D. W. Hogg, A. R. Casey, M. Ness, H.-W. Rix, D. Foreman-Mackey, S. Hasselquist, A. Y. Q. Ho, J. A. Holtzman, S. R. Majewski, S. L. Martell, S. Mészáros, D. L. Nidever and M. Shetrone, *Chemical Tagging Can Work: Identification of Stellar Phase-space Structures Purely by Chemical-abundance Similarity*, **833**, 262 (2016).

[69] R. Garcia-Dias, C. Allende Prieto, J. Sánchez Almeida and P. Alonso Palicio, *Machine learning in APOGEE. Identification of stellar populations through chemical abundances*, **629**, A34 (2019).

[70] F. Anders, C. Chiappini, B. X. Santiago, G. Matijevič, A. B. Queiroz, M. Steinmetz and G. Guiglion, *Dissecting stellar chemical abundance space with t-SNE*, **619**, A125 (2018).

[71] P. Jofré, U. Heiter and C. Soubiran, *Accuracy and Precision of Industrial Stellar Abundances*, **57**, 571 (2019).

[72] J. Bovy, *The Chemical Homogeneity of Open Clusters*, **817**, 49 (2016).

[73] S. Blanco-Cuaresma, C. Soubiran, U. Heiter and et al., *Testing the chemical tagging technique with open clusters*, **577**, A47 (2015).

[74] A. Dotter, C. Conroy, P. Cargile and M. Asplund, *The Influence of Atomic Diffusion on Stellar Ages and Chemical Tagging*, **840**, 99 (2017).

[75] C. Bertelli Motta, A. Pasquali, J. Richer and et al., *The Gaia-ESO Survey: evidence of atomic diffusion in M67?*, **478**, 425 (2018).

[76] L. Casamiquela, Y. Tarricq, C. Soubiran, S. Blanco-Cuaresma, P. Jofré, U. Heiter and M. Tucci Maia, *Differential abundances of open clusters and their tidal tails: Chemical tagging and chemical homogeneity*, **635**, A8 (2020).

[77] M. Baratella, V. D'Orazi, G. Carraro and et al., *The Gaia-ESO Survey: a new approach to chemically characterising young open clusters*, arXiv e-prints , arXiv:2001.03179 (2020).

[78] A. R. Casey, J. C. Lattanzio, A. Aleti and et al., *A Data-driven Model of Nucleosynthesis with Chemical Tagging in a Lower-dimensional Latent Space*, **887**, 73 (2019).

[79] A. H. W. Küpper, P. Kroupa, H. Baumgardt and D. C. Heggie, *Peculiarities in velocity dispersion and surface density profiles of star clusters*, **407**, 2241 (2010).

[80] M. E. K. Williams, M. Steinmetz, S. Sharma and et al., *The Dawning of the Stream of Aquarius in RAVE*, **728**, 102 (2011).

[81] I. Carrillo, I. Minchev, G. Kordopatis and et al., *Is the Milky Way still breathing? RAVE-Gaia streaming motions*, ArXiv e-prints (2017).

[82] G. W. Preston, *The chemically peculiar stars of the upper main sequence*, **12**, 257 (1974).

[83] L. Tomasella, U. Munari and T. Zwitter, *A High-resolution, Multi-epoch Spectral Atlas of Peculiar Stars Including RAVE, GAIA , and HERMES Wavelength Ranges*, **140**, 1758 (2010).

[84] G. Traven, G. Matijevič, T. Zwitter, M. Žerjal, J. Kos, M. Asplund, J. Bland-Hawthorn, A. R. Casey, G. De Silva, K. Freeman, J. Lin, S. L. Martell, K. J. Schlesinger, S. Sharma, J. D. Simpson, D. B. Zucker, B. Anguiano, G. Da Costa, L. Duong, J. Horner, E. A. Hyde, P. R. Kafle, U. Munari, D. Nataf, C. A. Navin, W. Reid and Y.-S. Ting, *The Galah Survey: Classification and Diagnostics with t-SNE Reduction of Spectral Information*, **228**, 24 (2017).

[85] G. Matijevič, T. Zwitter, U. Munari and et al., *Double-lined Spectroscopic Binary Stars in the Radial Velocity Experiment Survey*, **140**, 184 (2010).

[86] K. El-Badry, H.-W. Rix, Y.-S. Ting, D. R. Weisz, M. Bergemann, P. Cargile, C. Conroy and A.-C. Eilers, *Signatures of unresolved binaries in stellar spectra: implications for spectral fitting*, **473**, 5043 (2018).

[87] K. El-Badry, Y.-S. Ting, H.-W. Rix and et al., *Discovery and characterization of 3000+ main-sequence binaries from APOGEE spectra*, **476**, 528 (2018).

[88] L. Allen, S. T. Megeath, R. Gutermuth, P. C. Myers, S. Wolk, F. C. Adams, J. Muzerolle, E. Young and J. L. Pipher, *The Structure and Evolution of Young Stellar Clusters*, in *Protostars and Planets V*, edited by B. Reipurth, D. Jewitt and K. Keil (2007) p. 361, arXiv:astro-ph/0603096 [astro-ph] .

[89] M. Nakano, K. Sugitani, M. Watanabe, N. Fukuda, D. Ishihara and M. Ueno, *Wide-field Survey of Emission-line Stars in IC 1396*, **143**, 61 (2012).

[90] G. Traven, T. Zwitter, S. Van Eck and et al., *The Gaia-ESO Survey: Catalogue of Hα emission stars*, **581**, A52 (2015).

[91] Gaia Collaboration, T. Prusti, de Bruijne and et al., *The Gaia mission*, **595**, A1 (2016).

[92] S. L. Martell, S. Sharma, S. Buder and et al., *The GALAH survey: observational overview and Gaia DR1 companion*, **465**, 3203 (2017).

[93] S. R. Majewski, R. P. Schiavon, P. M. Frinchaboy and et al., *The Apache Point Observatory Galactic Evolution Experiment (APOGEE)*, **154**, 94 (2017).

[94] X.-Q. Cui, Y.-H. Zhao, Y.-Q. Chu and et al., *The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)*, Research in Astronomy and Astrophysics **12**, 1197 (2012).

[95] B. Yanny, C. Rockosi, H. J. Newberg and et al., *SEGUE: A Spectroscopic Survey of 240,000 Stars with g = 14-20*, **137**, 4377-4399 (2009).

[96] A. Kunder, G. Kordopatis, M. Steinmetz and et al., *The Radial Velocity Experiment (RAVE): Fifth Data Release*, **153**, 75 (2017).

[97] G. Gilmore, S. Randich, M. Asplund, J. Binney, P. Bonifacio, J. Drew, S. Feltzing, A. Ferguson, R. Jeffries, G. Micela and et al., *The Gaia-ESO Public Spectroscopic Survey*, The Messenger **147**, 25 (2012).

[98] R. S. de Jong, O. Bellido-Tirado, C. Chiappini and et al., *4MOST: 4-metre multi-object spectroscopic telescope*, in , Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8446 (2012) p. 84460T.

[99] G. Dalton, S. C. Trager, D. C. Abrams and et al., *WEAVE: the next generation wide-field spectroscopy facility for the William Herschel Telescope*, in , Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8446 (2012) p. 84460P.

[100] J. Kos, J. Lin, T. Zwitter, M. Žerjal, S. Sharma, J. Bland-Hawthorn, M. Asplund, A. R. Casey, G. M. De Silva, K. C. Freeman, S. L. Martell, J. D. Simpson, K. J. Schlesinger, D. Zucker, B. Anguiano, C. Bacigalupo, T. R. Bedding, C. Betters, G. Da Costa, L. Duong, E. Hyde, M. Ireland, P. R. Kafle, S. Leon-Saval, G. F. Lewis, U. Munari, D. Nataf, D. Stello, C. G. Tinney, G. Traven, F. Watson and R. A. Wittenmyer, *The GALAH survey: the data reduction pipeline*, **464**, 1259 (2017).

[101] R. Ahumada, C. Allende Prieto, A. Almeida and et al., *The Sixteenth Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra*, arXiv e-prints , arXiv:1912.02905 (2019).

[102] M. Steinmetz, G. Matijevic, H. Enke and et al., *The Sixth Data Release of the Radial Velocity Experiment (RAVE) – I: Survey Description, Spectra and Radial Velocities*, arXiv e-prints , arXiv:2002.04377 (2020).

[103] M. Ness, D. W. Hogg, H.-W. Rix, A. Y. Q. Ho and G. Zasowski, *The Cannon: A data-driven approach to Stellar Label Determination*, **808**, 16 (2015).

[104] S. Buder, M. Asplund, L. Duong, J. Kos, K. Lind, M. K. Ness, S. Sharma, J. Bland-Hawthorn, A. R. Casey, G. M. De Silva, V. D'Orazi, K. C. Freeman, G. F. Lewis, J. Lin, S. L. Martell, K. J. Schlesinger, J. D. Simpson, D. B. Zucker, T. Zwitter, A. M. Amarsi, B. Anguiano, D. Carollo, L. Casagrande, K. Čotar, P. L. Cottrell, G. Da Costa, X. D. Gao, M. R. Hayden, J. Horner, M. J. Ireland, P. R. Kafle, U. Munari, D. M. Nataf, T. Nordlander, D. Stello, Y.-S. Ting, G. Traven, F. Watson, R. A. Wittenmyer, R. F. G. Wyse, D. Yong, J. C. Zinn and M. Žerjal, *The GALAH Survey: second data release*, **478**, 4513 (2018).

[105] Y.-S. Ting, C. Conroy, H.-W. Rix and P. Cargile, *The Payne: Self-consistent ab initio Fitting of Stellar Spectra*, **879**, 69 (2019).

[106] T. Yang and X. Li, *An autoencoder of stellar spectra and its application in automatically estimating atmospheric parameters*, **452**, 158 (2015).

[107] H. W. Leung and J. Bovy, *Deep learning of multi-element abundances from high-resolution spectroscopic data*, **483**, 3255 (2019).

[108] R. Wang, A. L. Luo, J.-J. Chen, W. Hou, S. Zhang, Y.-H. Zhao, X.-R. Li, Y.-H. Hou and LAMOST MRS Collaboration, *SPCANet: Stellar Parameters and Chemical Abundances Network for LAMOST-II Medium Resolution Survey*, **891**, 23 (2020).

[109] R. Olney, M. Kounkel, C. Schillinger, M. T. Scoggins, Y. Yin, E. Howard, K. R. Covey, B. Hutchinson and K. G. Stassun, *APOGEE Net: Improving the derived spectral parameters for young stars through deep learning*, arXiv e-prints , arXiv:2002.08390 (2020).

[110] B. Chen, E. D'Onghia, S. A. Pardy, A. Pasquali, C. Bertelli Motta, B. Hanlon and E. K. Grebel, *Chemodynamical Clustering Applied to APOGEE Data: Rediscovering Globular Clusters*, **860**, 70 (2018).

[111] N. Price-Jones and J. Bovy, *Blind chemical tagging with DBSCAN: prospects for spectroscopic surveys*, **487**, 871 (2019).

[112] P. Jofré, P. Das, J. Bertranpetit and R. Foley, *Cosmic phylogeny: reconstructing the chemical history of the solar neighbourhood with an evolutionary tree*, **467**, 1140 (2017).

[113] S. Blanco-Cuaresma and D. Fraix-Burnet, *A phylogenetic approach to chemical tagging. Reassembling open cluster stars*, **618**, A65 (2018).

[114] L. Lindegren, *"Re-normalising the astrometric chi-square in Gaia DR2", Gaia Technical Note: GAIA-C3-TN-LU-LL-124-0* (2018).

[115] I. Becker, K. Pichara, M. Catelan, P. Protopapas, C. Aguirre and F. Nikzat, *Scalable end-to-end recurrent neural network for variable star classification*, **493**, 2981 (2020).

[116] Y. Bai, J. Liu, Z. Bai, S. Wang and D. Fan, *Machine-learning Regression of Stellar Effective Temperatures in the Second Gaia Data Release*, **158**, 93 (2019).

[117] G. Marton, P. Ábrahám, E. Szegedi-Elek and et al., *Identification of Young Stellar Object candidates in the Gaia DR2 x AllWISE catalogue with machine learning methods*, **487**, 2522 (2019).

[118] A. Helmi, J. Veljanoski, M. A. Breddels, H. Tian and L. V. Sales, *A box full of chocolates: The rich structure of the nearby stellar halo revealed by Gaia and RAVE*, **598**, A58 (2017).

[119] N. W. Borsato, S. L. Martell and J. D. Simpson, *Identifying stellar streams in Gaia DR2 with data mining techniques*, **492**, 1370 (2020).

[120] B. Ostdiek, L. Necib, T. Cohen, M. Freytsis, M. Lisanti, S. Garrison-Kimmel, A. Wetzel, R. E. Sanderson and P. F. Hopkins, *Cataloging Accreted Stars within Gaia DR2 using Deep Learning*, arXiv e-prints , arXiv:1907.06652 (2019).

[121] L. Necib, B. Ostdiek, M. Lisanti, T. Cohen, M. Freytsis and S. Garrison-Kimmel, *Chasing Accreted Structures within Gaia DR2 using Deep Learning*, arXiv e-prints , arXiv:1907.07681 (2019).

[122] T. Marchetti, E. M. Rossi, G. Kordopatis, A. G. A. Brown, A. Rimoldi, E. Starkenburg, K. Youakim and R. Ashley, *An artificial neural network to discover hypervelocity stars: candidates in Gaia DR1/TGAS*, **470**, 1388 (2017).

[123] Y. Bai, J.-F. Liu and S. Wang, *Machine learning classification of Gaia Data Release 2*, Research in Astronomy and Astrophysics **18**, 118 (2018).

[124] C. A. L. Bailer-Jones, M. Fouesneau and R. Andrae, *Quasar and galaxy classification in Gaia Data Release 2*, **490**, 5615 (2019).

[125] Y. Bai, J. Liu, Y. Wang and S. Wang, *Machine-learning Regression of Extinction in the Second Gaia Data Release*, **159**, 84 (2020).

[126] A. Castro-Ginard, C. Jordi, X. Luri, J. Álvarez Cid-Fuentes, L. Casamiquela, F. Anders, T. Cantat-Gaudin, M. Monguió, L. Balaguer-Núñez, S. Solà and R. M. Badia, *Hunting for open clusters in Gaia DR2: 582 new open clusters in the Galactic disc*, **635**, A45 (2020).

[127] N. V. Kharchenko, A. E. Piskunov, E. Schilbach, S. Röser and R.-D. Scholz, *Global survey of star clusters in the Milky Way. II. The catalogue of basic parameters*, **558**, A53 (2013).

[128] R. Hoogerwerf, J. H. J. de Bruijne and P. T. de Zeeuw, *The Origin of Runaway Stars*, **544**, L133 (2000).

[129] B. Gustafsson, B. Edvardsson, K. Eriksson, U. G. Jørgensen, Å. Nordlund and B. Plez, *A grid of MARCS model atmospheres for late-type stars. I. Methods and general properties*, **486**, 951 (2008).

[130] L. van der Maaten, *Barnes-Hut-SNE*, ArXiv e-prints (2013).

[131] V. Adibekyan, E. Delgado-Mena, S. Feltzing and et al., *Sun-like stars unlike the Sun: Clues for chemical anomaliesof cool stars*, Astronomische Nachrichten **338**, 442 (2017).

[132] J. Meléndez, M. Asplund, B. Gustafsson and D. Yong, *The Peculiar Solar Composition and Its Possible Relation to Planet Formation*, **704**, L66 (2009).

[133] K. Biazzo, L. Pasquini, P. Bonifacio, S. Randich and L. R. Bedin, *True solar analogues in the open cluster M67 .*, **80**, 125 (2009).

[134] F. Liu, M. Asplund, D. Yong, J. Meléndez, I. Ramírez, A. I. Karakas, M. Carlos and A. F. Marino, *The chemical compositions of solar twins in the open cluster M67*, **463**, 696 (2016).

[135] S. B. Howell, C. Sobeck, M. Haas and et al., *The K2 Mission: Characterization and Early Results*, **126**, 398 (2014).

[136] G. R. Ricker, J. N. Winn, R. Vanderspek and et al., *Transiting Exoplanet Survey Satellite (TESS)*, in , Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9143 (2014) p. 914320, arXiv:1406.0151 [astro-ph.EP] .

[137] M. Cropper, D. Katz, P. Sartoretti, T. Prusti and et al., *Gaia Data Release 2. Gaia Radial Velocity Spectrometer*, **616**, A5 (2018).

[138] Gaia Collaboration, *Gaia Data Release 2. The celestial reference frame (Gaia-CRF2)*, **616**, A14 (2018).

[139] L. Lindegren, J. Hernández, A. Bombrun and et al., *Gaia Data Release 2. The astrometric solution*, **616**, A2 (2018).

[140] X. Luri, A. G. A. Brown, L. M. Sarro and et al., *Gaia Data Release 2. Using Gaia parallaxes*, **616**, A9 (2018).

[141] D. W. Evans, M. Riello, F. De Angeli and et al., *Gaia Data Release 2. Photometric content and validation*, **616**, A4 (2018).

[142] M. Riello, F. De Angeli, D. W. Evans and et al., *Gaia Data Release 2. Processing of the photometric data*, **616**, A3 (2018).

[143] C. Soubiran, G. Jasniewicz, L. Chemin and et al., *Gaia Data Release 2. The catalogue of radial velocity standard stars*, **616**, A7 (2018).

[144] P. Sartoretti, D. Katz, M. Cropper and et al., *Gaia Data Release 2. Processing the spectroscopic data*, **616**, A6 (2018).

[145] B. Holl, M. Audard, K. Nienartowicz and et al., *Gaia Data Release 2. Summary of the variability processing and analysis results*, **618**, A30 (2018).

[146] N. Mowlavi, I. Lecoeur-Taïbi, T. Lebzelter, L. Rimoldini, D. Lorenz, M. Audard, J. De Ridder, L. Eyer, L. P. Guy, B. Holl, G. Jevardat de Fombelle, O. Marchal, K. Nienartowicz, S. Regibo, M. Roelens and L. M. Sarro, *Gaia Data Release 2. The first Gaia catalogue of long-period variable candidates*, **618**, A58 (2018).

[147] L. Molnár, E. Plachy, Á. L. Juhász and L. Rimoldini, *Gaia Data Release 2. Validating the classification of RR Lyrae and Cepheid variables with the Kepler and K2 missions*, **620**, A127 (2018).

[148] G. Clementini, V. Ripepi, R. Molinaro and et al., *Gaia Data Release 2. Specific characterisation and validation of all-sky Cepheids and RR Lyrae stars*, **622**, A60 (2019).

[149] R. Andrae, M. Fouesneau, O. Creevey and et al., *Gaia Data Release 2. First stellar parameters from Apsis*, **616**, A8 (2018).

[150] Gaia Collaboration, *Gaia Data Release 2. Observations of solar system objects*, **616**, A13 (2018).

[151] Gaia Helpdesk, *Gaia DR2 primer: Everything you wish you had known before you started working with Gaia Data Release 2, issue 1.0* (2019), [downloaded 20. 12. 2019].

[152] G. M. De Silva, K. C. Freeman, J. Bland-Hawthorn and et al., *The GALAH survey: scientific motivation*, **449**, 2604 (2015).

[153] S. C. Barden, D. J. Jones, S. I. Barnes and et al., *HERMES: revisions in the design for a high-resolution multi-element spectrograph for the AAT*, in *Ground-based and Airborne Instrumentation for Astronomy III*, , Vol. 7735 (2010) p. 773509.

[154] A. Sheinis, B. Anguiano, M. Asplund and et al., *First light results from the High Efficiency and Resolution Multi-Element Spectrograph at the Anglo-Australian Telescope*, Journal of Astronomical Telescopes, Instruments, and Systems **1**, 035002 (2015).

[155] R. A. Wittenmyer, S. Sharma, D. Stello, S. Buder, J. Kos, M. Asplund, L. Duong, J. Lin, K. Lind, M. Ness, T. Zwitter, J. Horner, J. Clark, S. R. Kane, D. Huber, J. Bland-Hawthorn, A. R. Casey, G. M. De Silva, V. D'Orazi, K. Freeman, S. Martell, J. D. Simpson, D. B. Zucker, B. Anguiano, L. Casagrande, J. Esdaile, M. Hon, M. Ireland, P. R. Kafle, S. Khanna, J. P. Marshall, M. H. M. Saddon, G. Traven and D. Wright, *The K2-HERMES Survey. I. Planet-candidate Properties from K2 Campaigns 1-3*, **155**, 84 (2018).

[156] S. Sharma, D. Stello, S. Buder, J. Kos, J. Bland-Hawthorn, M. Asplund, L. Duong, J. Lin, K. Lind, M. Ness, D. Huber, T. Zwitter, G. Traven, M. Hon, P. R. Kafle, S. Khanna, H. Saddon, B. Anguiano, A. R. Casey, K. Freeman, S. Martell, G. M. De Silva, J. D. Simpson, R. A. Wittenmyer and D. B. Zucker, *The TESS-HERMES survey data release 1: high-resolution spectroscopy of the TESS southern continuous viewing zone*, **473**, 2004 (2018).

[157] M. F. Skrutskie, R. M. Cutri, R. Stiening and et al., *The Two Micron All Sky Survey (2MASS)*, **131**, 1163 (2006).

[158] G. R. Ricker, J. N. Winn, R. Vanderspek and et al., *Transiting Exoplanet Survey Satellite (TESS)*, Journal of Astronomical Telescopes, Instruments, and Systems **1**, 014003 (2015).

[159] P. de Laverny, A. Recio-Blanco, C. C. Worley and B. Plez, *The AMBRE project: A new synthetic grid of high-resolution FGKM stellar spectra*, **544**, A126 (2012).

[160] T. Zwitter, J. Kos, A. Chiavassa, S. Buder, G. Traven, K. Čotar, J. Lin, M. Asplund, J. Bland-Hawthorn, A. R. Casey, G. De Silva, L. Duong, K. C. Freeman, K. Lind, S. Martell, V. D'Orazi, K. J. Schlesinger, J. D. Simpson, S. Sharma, D. B. Zucker, B. Anguiano, L. Casagrande, R. Collet, J. Horner, M. J. Ireland, P. R. Kafle, G. Lewis, U. Munari, D. M. Nataf, M. Ness, T. Nordlander, D. Stello, Y.-S. Ting, C. G. Tinney, F. Watson, R. A. Wittenmyer and M. Žerjal, *The GALAH survey: accurate radial velocities and library of observed stellar template spectra*, **481**, 645 (2018).

[161] J. A. Valenti and N. Piskunov, *Spectroscopy made easy: A new tool for fitting observations with synthetic spectra.*, **118**, 595 (1996).

[162] N. Piskunov and J. A. Valenti, *Spectroscopy Made Easy: Evolution*, **597**, A16 (2017).

[163] Y. Osorio, C. Allende Prieto, I. Hubeny, S. Mészáros and M. Shetrone, *NLTE for APOGEE: simultaneous multi-element NLTE radiative transfer*, **637**, A80 (2020).

[164] A. M. Amarsi, P. E. Nissen and Á. Skúladóttir, *Carbon, oxygen, and iron abundances in disk and halo stars. Implications of 3D non-LTE spectral line formation*, **630**, A104 (2019).

[165] A. M. Amarsi, K. Lind, Y. Osorio, T. Nordlander, M. Bergemann, H. Reggiani, E. X. Wang, S. Buder, M. Asplund, P. S. Barklem, A. Wehrhahn, A. Skuladottir, C. Kobayashi, A. Karakas, X. D. Gao3, J. Bland-Hawthorn, G. M. De Silva, G. F. Lewis, S. L. Martell, S. Sharma, J. D. Simpson, D. B. Zucker, K. Čotar and J. Horner, *The GALAH Survey: Non-LTE departure coefficients for large spectroscopic surveys*, **submitted** (2020).

[166] S. Buder, M. Asplund, L. Duong, S. Sharma, J. Bland-Hawthorn and et al., *The GALAH Survey: Third data release*, **in preparation** (2020).

[167] U. Munari, K. Cotar, V. Andreoli, S. Dallaporta and L. N. Yalyalieva, *Recent developments in optical spectra of nova AT 2019qwf (Nova V2891 Cyg)*, The Astronomer's Telegram **13340**, 1 (2019).

[168] U. Munari, V. Joshi, D. P. K. Banerjee, K. Čotar, S. Y. Shugarov, Jurdana-Šepić, R. , R. Belligoli, A. Bergamini, M. Graziani, G. L. Righetti, A. Vagnozzi and P. Valisa, *The 2018 eruption and long-term evolution of the new high-mass Herbig Ae/Be object Gaia-18azl = VES 263*, **488**, 5536 (2019).

[169] A. Krone-Martins and A. Moitinho, *UPMASK: unsupervised photometric membership assignment in stellar clusters*, **561**, A57 (2014).

[170] C. A. L. Bailer-Jones, J. Rybizki, M. Fouesneau, G. Mantelet and R. Andrae, *Estimating Distance from Parallaxes. IV. Distances to 1.33 Billion Stars in Gaia Data Release 2*, **156**, 58 (2018).

[171] J. Bovy, *galpy: A python Library for Galactic Dynamics*, **216**, 29 (2015).

[172] M. Gebran, M. Vick, R. Monier and L. Fossati, *Chemical composition of A and F dwarfs members of the Hyades open cluster*, **523**, A71 (2010).

[173] A. M. Boesgaard, B. W. Roper and M. G. Lum, *The Chemical Composition of Praesepe (M44)*, **775**, 58 (2013).

[174] F. Liu, D. Yong, M. Asplund, I. Ramírez and J. Meléndez, *The Hyades open cluster is chemically inhomogeneous*, **457**, 3934 (2016).

[175] A. Bragaglia, X. Fu, A. Mucciarelli, G. Andreuzzi and P. Donati, *The chemical composition of the oldest nearby open cluster Ruprecht 147*, **619**, A176 (2018).

[176] T. Bensby, S. Feltzing and I. Lundström, *Elemental abundance trends in the Galactic thin and thick disks as traced by nearby F and G dwarf stars*, **410**, 527 (2003).

[177] L. Spina, J. Meléndez, A. I. Karakas and et al., *The temporal evolution of neutron-capture elements in the Galactic discs*, **474**, 2580 (2018).

[178] J. Lin, M. Asplund, Y.-S. Ting, L. Casagrand e, S. Buder, J. Bland-Hawthorn, A. R. Casey, G. M. De Silva, V. D'Orazi, K. C. Freeman, J. Kos, K. Lind, S. L. Martell, S. Sharma, J. D. Simpson, T. Zwitter, D. B. Zucker, I. Minchev, K. Čotar, M. Hayden, J. Horner, G. F. Lewis, T. Nordlander, R. F. G. Wyse and M. Žerjal, *The GALAH survey: temporal chemical enrichment of the galactic disc*, **491**, 2043 (2020).

[179] H. J. G. L. M. Lamers and M. Gieles, *Clusters in the solar neighbourhood: how are they destroyed?*, **455**, L17 (2006).

[180] K. Čotar, T. Zwitter, J. Kos, U. Munari, S. L. Martell, M. Asplund, J. Bland -Hawthorn, S. Buder, G. M. de Silva, K. C. Freeman, S. Sharma, B. Anguiano, D. Carollo, J. Horner, G. F. Lewis, D. M. Nataf, T. Nordlander, D. Stello, Y.-S. Ting, C. Tinney, G. Traven, R. A. Wittenmyer and Galah Collaboration, *The GALAH survey: a catalogue of carbon-enhanced stars and CEMP candidates*, **483**, 3196 (2019).

[181] A. Secchi, *Schreiben des Herrn Professors Secchi an den Herausgeber*, Astronomische Nachrichten **73**, 129 (1869).

[182] I. Iben, Jr., *Carbon star formation and neutron-rich isotope formation in low-mass asymptotic giant branch stars*, **275**, L65 (1983).

[183] Z. Han, P. P. Eggleton, P. Podsiadlowski and C. A. Tout, *The formation of barium and CH stars and related objects*, **277**, 1443 (1995).

[184] J. Yoon, T. C. Beers, V. M. Placco, K. C. Rasmussen, D. Carollo, S. He, T. T. Hansen, I. U. Roederer and J. Zeanah, *Observational Constraints on First-star Nucleosynthesis. I. Evidence for Multiple Progenitors of CEMP-No Stars*, **833**, 20 (2016).

[185] P. Battinelli, S. Demers, C. Rossi and K. S. Gigoyan, *Extension of the C Star Rotation Curve of the Milky Way to 24 kpc*, Astrophysics **56**, 68 (2013).

[186] G. Bothun, J. H. Elias, G. MacAlpine, K. Matthews, J. R. Mould, G. Neugebauer and I. N. Reid, *Carbon stars at high Galactic latitude*, **101**, 2220 (1991).

[187] B. Margon, S. F. Anderson, H. C. Harris and et al., *Faint High-Latitude Carbon Stars Discovered by the Sloan Digital Sky Survey: Methods and Initial Results*, **124**, 1651 (2002).

[188] R. A. Downes, B. Margon, S. F. Anderson and et al., *Faint High-Latitude Carbon Stars Discovered by the Sloan Digital Sky Survey: An Initial Catalog*, **127**, 2838 (2004).

[189] R. F. Griffin and R. O. Redman, *Photoelectric measurements of the λ4200 A CN band and the G band in G8-K5 spectra*, **120**, 287 (1960).

[190] R. D. McClure and S. van den Bergh, *Five-color intermediate-band photometry of stars, clusters, and galaxies.*, **73**, 313 (1968).

[191] L. Häggkvist and T. Oja, *Narrow-band photometry of late-type stars*, **1**, 199 (1970).

[192] D. Moro and U. Munari, *The Asiago Database on Photometric Systems (ADPS). I. Census parameters for 167 photometric systems*, **147**, 361 (2000).

[193] M. Fiorucci and U. Munari, *The Asiago Database on Photometric Systems (ADPS). II. Band and reddening parameters*, **401**, 781 (2003).

[194] N. Christlieb, P. J. Green, L. Wisotzki and D. Reimers, *The stellar content of the Hamburg/ESO survey II. A large, homogeneously-selected sample of high latitude carbon stars*, **375**, 366 (2001).

[195] P. Green, *Innocent Bystanders: Carbon Stars from the Sloan Digital Sky Survey*, **765**, 12 (2013).

[196] Y. S. Lee, T. C. Beers, T. Masseron, B. Plez, C. M. Rockosi, J. Sobeck, B. Yanny, S. Lucatello, T. Sivarani, V. M. Placco and D. Carollo, *Carbon-enhanced Metal-poor Stars in SDSS/SEGUE. I. Carbon Abundance Estimation and Frequency of CEMP Stars*, **146**, 132 (2013).

[197] W. Ji, W. Cui, C. Liu, A. Luo, G. Zhao and B. Zhang, *Carbon Stars from LAMOST DR2 Data*, **226**, 1 (2016).

[198] Y.-B. Li, A.-L. Luo, C.-D. Du and et al., *Carbon Stars Identified from LAMOST DR4 Using Machine Learning*, **234**, 31 (2018).

[199] J. E. Norris, S. G. Ryan and T. C. Beers, *Extremely Metal-poor Stars. IV. The Carbon-rich Objects*, **488**, 350 (1997).

[200] W. Aoki, J. E. Norris, S. G. Ryan, T. C. Beers and H. Ando, *The Chemical Composition of Carbon-rich, Very Metal Poor Stars: A New Class of Mildly Carbon Rich Objects without Excess of Neutron-Capture Elements*, **567**, 1166 (2002).

[201] R. Cayrel, E. Depagne, M. Spite, V. Hill, F. Spite, P. François, B. Plez, T. Beers, F. Primas, J. Andersen, B. Barbuy, P. Bonifacio, P. Molaro and B. Nordström, *First stars V - Abundance patterns from C to Zn and supernova yields in the early Galaxy*, **416**, 1117 (2004).

[202] P. S. Barklem, N. Christlieb and T. C. Beers, *Metal-poor star abundances from the HERES project*, in *13th Cambridge Workshop on Cool Stars, Stellar Systems and the Sun*, ESA Special Publication, Vol. 560, edited by F. Favata, G. A. J. Hussain and B. Battrick (2005) p. 433.

[203] J. G. Cohen, A. McWilliam, S. Shectman, I. Thompson, N. Christlieb, J. Melendez, S. Ramirez, A. Swensson and F.-J. Zickgraf, *Carbon Stars in the Hamburg/ESO Survey: Abundances*, **132**, 137 (2006).

[204] W. Aoki, T. C. Beers, N. Christlieb, J. E. Norris, S. G. Ryan and S. Tsangarides, *Carbon-enhanced Metal-poor Stars. I. Chemical Compositions of 26 Stars*, **655**, 492 (2007).

[205] J. E. Norris, N. Christlieb, A. J. Korn, K. Eriksson, M. S. Bessell, T. C. Beers, L. Wisotzki and D. Reimers, *HE 0557-4840: Ultra-Metal-Poor and Carbon-Rich*, **670**, 774 (2007).

[206] J. K. Hollek, A. Frebel, I. U. Roederer, C. Sneden, M. Shetrone, T. C. Beers, S. Kang and C. Thom, *The Chemical Abundances of Stars in the Halo (CASH) Project. II. A Sample of 14 Extremely Metal-poor Stars*, **742**, 54 (2011).

[207] D. Yong, J. E. Norris, M. S. Bessell, N. Christlieb, M. Asplund, T. C. Beers, P. S. Barklem, A. Frebel and S. G. Ryan, *The Most Metal-poor Stars. II. Chemical Abundances of 190 Metal-poor Stars Including 10 New Stars with [Fe/H] = -3.5*, **762**, 26 (2013).

[208] I. U. Roederer, G. W. Preston, I. B. Thompson, S. A. Shectman, C. Sneden, G. S. Burley and D. D. Kelson, *A Search for Stars of Very Low Metal Abundance. VI. Detailed Abundances of 313 Metal-poor Stars*, **147**, 136 (2014).

[209] T. Hansen, C. J. Hansen, N. Christlieb and et al., *An Elemental Assay of Very, Extremely, and Ultra-metal-poor Stars*, **807**, 173 (2015).

[210] H. R. Jacobson, S. Keller, A. Frebel and et al., *High-Resolution Spectroscopic Study of Extremely Metal-Poor Star Candidates from the SkyMapper Survey*, **807**, 171 (2015).

[211] J. G. Cohen, N. Christlieb, I. Thompson, A. McWilliam, S. Shectman, D. Reimers, L. Wisotzki and E. Kirby, *Normal and Outlying Populations of the Milky Way Stellar Halo at [Fe/H] -2*, **778**, 56 (2013).

[212] T. C. Beers, G. W. Preston and S. A. Shectman, *A search for stars of very low metal abundance. II*, **103**, 1987 (1992).

[213] S. Rossi, T. C. Beers and C. Sneden, *Carbon Abundances for Metal-Poor Stars Based on Medium-Resolution Spectra*, in *The Third Stromlo Symposium: The Galactic Halo*, Astronomical Society of the Pacific Conference Series, Vol. 165, edited by B. K. Gibson, R. S. Axelrod and M. E. Putman (1999) p. 264.

[214] J. G. Cohen, S. Shectman, I. Thompson, A. McWilliam, N. Christlieb, J. Melendez, F.-J. Zickgraf, S. Ramírez and A. Swenson, *The Frequency of Carbon Stars among Extremely Metal-poor Stars*, **633**, L109 (2005).

[215] S. Lucatello, S. Tsangarides, T. C. Beers, E. Carretta, R. G. Gratton and S. G. Ryan, *The Binary Frequency Among Carbon-enhanced, s-Process-rich, Metal-poor Stars*, **625**, 825 (2005).

[216] S. Rossi, T. C. Beers, C. Sneden, T. Sevastyanenko, J. Rhee and B. Marsteller, *Estimation of Carbon Abundances in Metal-Poor Stars. I. Application to the Strong G-Band Stars of Beers, Preston, and Shectman*, **130**, 2804 (2005).

[217] A. Frebel, N. Christlieb, J. E. Norris, T. C. Beers, M. S. Bessell, J. Rhee, C. Fechner, B. Marsteller, S. Rossi, C. Thom, L. Wisotzki and D. Reimers, *Bright Metal-poor Stars from the Hamburg/ESO Survey. I. Selection and Follow-up Observations from 329 Fields*, **652**, 1585 (2006).

[218] B. E. Marsteller, *The frequency of carbon-enhanced metal-poor stars and the origin of carbon in the universe*, Ph.D. thesis, Michigan State University (2007).

[219] D. Carollo, T. C. Beers, J. Bovy, T. Sivarani, J. E. Norris, K. C. Freeman, W. Aoki, Y. S. Lee and C. R. Kennedy, *Carbon-enhanced Metal-poor Stars in the Inner and Outer Halo Components of the Milky Way*, **744**, 195 (2012).

[220] D. Yong, J. E. Norris, M. S. Bessell, N. Christlieb, M. Asplund, T. C. Beers, P. S. Barklem, A. Frebel and S. G. Ryan, *The Most Metal-poor Stars. III. The Metallicity Distribution Function and Carbon-enhanced Metal-poor Fraction*, **762**, 27 (2013).

[221] V. M. Placco, A. Frebel, T. C. Beers and R. J. Stancliffe, *Carbon-enhanced Metal-poor Star Frequencies in the Galaxy: Corrections for the Effect of Evolutionary Status on Carbon Abundances*, **797**, 21 (2014).

[222] J. Yoon, T. C. Beers, S. Dietz, Y. S. Lee, V. M. Placco, G. Da Costa, S. Keller, C. I. Owen and M. Sharma, *Galactic Archeology with the AEGIS Survey: The Evolution of Carbon and Iron in the Galactic Halo*, **861**, 146 (2018).

[223] T. Suda, M. Aikawa, M. N. Machida, M. Y. Fujimoto and I. Iben, Jr., *Is HE 0107-5240 A Primordial Star? The Characteristics of Extremely Metal-Poor Carbon-Rich Stars*, **611**, 476 (2004).

[224] E. Starkenburg, M. D. Shetrone, A. W. McConnachie and K. A. Venn, *Binarity in carbon-enhanced metal-poor stars*, **441**, 1217 (2014).

[225] H. Umeda and K. Nomoto, *First-generation black-hole-forming supernovae and the metal abundance pattern of a very iron-poor star*, **422**, 871 (2003).

[226] H. Umeda and K. Nomoto, *Variations in the Abundance Pattern of Extremely Metal-Poor Stars and Nucleosynthesis in Population III Supernovae*, **619**, 427 (2005).

[227] N. Tominaga, N. Iwamoto and K. Nomoto, *Abundance Profiling of Extremely Metal-poor Stars and Supernova Properties in the Early Universe*, **785**, 98 (2014).

[228] P. Banerjee, Y.-Z. Qian and A. Heger, *s-Process in Massive Carbon-Enhanced Metal-Poor Stars*, 10.1093/mnras/sty2251 (2018).

[229] D. Carollo, K. Freeman, T. C. Beers, V. M. Placco, J. Tumlinson and S. L. Martell, *Carbon-enhanced Metal-poor Stars: CEMP-s and CEMP-no Subclasses in the Halo System of the Milky Way*, **788**, 180 (2014).

[230] M. A. Cruz, H. Cogo-Moreira and S. Rossi, *Searching for chemical classes among metal-poor stars using medium-resolution spectroscopy*, **475**, 4781 (2018).

[231] T. C. Beers and N. Christlieb, *The Discovery and Analysis of Very Metal-Poor Stars in the Galaxy*, **43**, 531 (2005).

[232] M. Spite, E. Caffau, P. Bonifacio, F. Spite, H.-G. Ludwig, B. Plez and N. Christlieb, *Carbon-enhanced metal-poor stars: the most pristine objects?*, **552**, A107 (2013).

[233] J. K. Hollek, A. Frebel, V. M. Placco, A. I. Karakas, M. Shetrone, C. Sneden and N. Christlieb, *The Chemical Abundances of Stars in the Halo (CASH) Project. III. A New Classification Scheme for Carbon-enhanced Metal-poor Stars with s-process Element Enhancement*, **814**, 121 (2015).

[234] R. C. Johnson, *The Structure and Origin of the Swan Band Spectrum of Carbon*, Philosophical Transactions of the Royal Society of London Series A **226**, 157 (1927).

[235] A. Alksnis, A. Balklavs, U. Dzervitis, I. Eglitis, O. Paupers and I. Pundure, *General Catalog of Galactic Carbon Stars by C. B. Stephenson. Third Edition*, Baltic Astronomy **10**, 1 (2001).

[236] L. van der Maaten and G. Hinton, *Visualizing Data using t-SNE*, Journal of Machine Learning Research **9**, 2579 (2008).

[237] M. Wattenberg, F. Viégas and I. Johnson, *How to Use t-SNE Effectively*, Distill 10.23915/distill.00002 (2016).

[238] J. Sperauskas, L. Začs, W. J. Schuster and V. Deveikis, *The Binary Nature of CH-Like Stars*, **826**, 85 (2016).

[239] J. Bergeat, A. Knapik and B. Rutily, *The pulsation modes and masses of carbon-rich long period variables*, **390**, 987 (2002).

[240] T. Lloyd Evans, *Carbon stars*, Journal of Astrophysics and Astronomy **31**, 177 (2010).

[241] P. Battinelli and S. Demers, *Variability of halo carbon stars*, **544**, A10 (2012).

[242] P. Battinelli and S. Demers, *The variability of carbon stars in the Sagittarius dwarf spheroidal galaxy*, **553**, A93 (2013).

[243] P. Battinelli and S. Demers, *Miras among C stars*, **568**, A100 (2014).

[244] B. Margon, T. Kupfer, K. Burdge, T. A. Prince, S. R. Kulkarni and D. L. Shupe, *The Binary Dwarf Carbon Star SDSS J125017.90+252427.6*, **856**, L2 (2018).

[245] D. S. P. Dearborn, J. Liebert, M. Aaronson, C. C. Dahn, R. Harrington, J. Mould and J. L. Greenstein, *On the nature of the dwarf carbon star G77-61*, **300**, 314 (1986).

[246] L. J. Whitehouse, J. Farihi, P. J. Green, T. G. Wilson and J. P. Subasavage, *Dwarf carbon stars are likely metal-poor binaries and unlikely hosts to carbon planets*, **479**, 3873 (2018).

[247] J. Farihi, A. R. Arendt, H. S. Machado and L. J. Whitehouse, *Evidence for halo kinematics among cool carbon-rich dwarfs*, **477**, 3801 (2018).

[248] T. T. Hansen, J. Andersen, B. Nordström, T. C. Beers, V. M. Placco, J. Yoon and L. A. Buchhave, *The role of binaries in the enrichment of the early Galactic halo. II. Carbon-enhanced metal-poor stars: CEMP-no stars*, **586**, A160 (2016).

[249] C. Abia, H. M. J. Boffin, J. Isern and R. Rebolo, *IY Hya - A new super Li-rich carbon star*, **245**, L1 (1991).

[250] I.-J. Sackmann, R. L. Smith and K. H. Despain, *Carbon and eruptive stars: surface enrichment of lithium, carbon, nitrogen, and ^{13}C by deep mixing.*, **187**, 555 (1974).

[251] P. C. Keenan and W. W. Morgan, *The Classification of the Red Carbon Stars.*, **94**, 501 (1941).

[252] C. Barnbaum, R. P. S. Stone and P. C. Keenan, *A Moderate-Resolution Spectral Atlas of Carbon Stars: R, J, N, CH, and Barium Stars*, **105**, 419 (1996).

[253] Y. Komiya, T. Suda, H. Minaguchi, T. Shigeyama, W. Aoki and M. Y. Fujimoto, *The Origin of Carbon Enhancement and the Initial Mass Function of Extremely Metal-poor Stars in the Galactic Halo*, **658**, 367 (2007).

[254] T. Masseron, J. A. Johnson, B. Plez, S. van Eck, F. Primas, S. Goriely and A. Jorissen, *A holistic approach to carbon-enhanced metal-poor stars*, **509**, A93 (2010).

[255] V. M. Placco, C. R. Kennedy, S. Rossi, T. C. Beers, Y. S. Lee, N. Christlieb, T. Sivarani, D. Reimers and L. Wisotzki, *A Search for Unrecognized Carbon-Enhanced Metal-Poor Stars in the Galaxy*, **139**, 1051 (2010).

[256] C. Abate, O. R. Pols, R. G. Izzard and A. I. Karakas, *Carbon-enhanced metal-poor stars: a window on AGB nucleosynthesis and binary evolution. II. Statistical analysis of a sample of 67 CEMP-s stars*, **581**, A22 (2015).

[257] J. Yoon, T. C. Beers, V. M. Placco, K. C. Rasmussen, D. Carollo, S. He, T. T. Hansen, I. U. Roederer and J. Zeanah, *VizieR Online Data Catalog: Carbon-enhanced metal-poor (CEMP) star abundances (Yoon+, 2016)*, VizieR Online Data Catalog **183** (2017).

[258] H. Koppelman, A. Helmi and J. Veljanoski, *One Large Blob and Many Streams Frosting the nearby Stellar Halo in GaiaDR2*, **860**, L11 (2018).

[259] R. D. McClure and A. W. Woodsworth, *The binary nature of the barium and CH stars. III - Orbital parameters*, **352**, 709 (1990).

[260] A. Jorissen, S. Van Eck, H. Van Winckel, T. Merle, H. M. J. Boffin, J. Andersen, B. Nordström, S. Udry, T. Masseron, L. Lenaerts and C. Waelkens, *Binary properties of CH and carbon-enhanced metal-poor stars*, **586**, A158 (2016).

[261] K. Čotar, T. Zwitter, G. Traven, J. Bland-Hawthorn, S. Buder, M. R. Hayden, J. Kos, G. F. Lewis, S. L. Martell, T. Nordland er, D. Stello, J. Horner, Y.-S. Ting and M. Žerjal, *The GALAH survey: Characterization of emission-line stars with spectral modelling using autoencoders*, arXiv e-prints , arXiv:2006.03062 (2020).

[262] R. Cayrel, C. van't Veer-Menneret, N. F. Allard and C. Stehlé, *The Hα Balmer line as an effective temperature criterion. I. Calibration using 1D model stellar atmospheres*, **531**, A83 (2011).

[263] A. M. Amarsi, T. Nordlander, P. S. Barklem, M. Asplund, R. Collet and K. Lind, *Effective temperature determinations of late-type stars based on 3D non-LTE Balmer line formation*, **615**, A139 (2018).

[264] R. E. Giribaldi, M. L. Ubaldo-Melo, G. F. Porto de Mello, L. Pasquini, H. G. Ludwig, S. Ulmer-Moll and D. Lorenzo-Oliveira, *Accurate effective temperature from Hα profiles*, **624**, A10 (2019).

[265] M. Ness, D. W. Hogg, H. W. Rix, M. Martig, M. H. Pinsonneault and A. Y. Q. Ho, *Spectroscopic Determination of Masses (and Implied Ages) for Red Giants*, **823**, 114 (2016).

[266] M. Bergemann, A. Serenelli, R. Schönrich and et al., *The Gaia-ESO Survey: Hydrogen lines in red giants directly trace stellar mass*, **594**, A120 (2016).

[267] P. S. Barklem, N. Piskunov and B. J. O'Mara, *Self-broadening in Balmer line wing formation in stellar atmospheres*, **363**, 1091 (2000).

[268] N. F. Allard, J. F. Kielkopf, R. Cayrel and C. van't Veer-Menneret, *Self-broadening of the hydrogen Balmer α line*, **480**, 581 (2008).

[269] L. Lancaster, J. Greene, Y.-S. Ting, S. E. Koposov, B. J. S. Pope and R. L. Beaton, *A Mystery in Chamaeleon: Serendipitous Discovery of a Galactic Symbiotic Nova*, arXiv e-prints , arXiv:2002.07852 (2020).

[270] B. Reipurth, A. Pedrosa and M. T. V. T. Lago, *Hα emission in pre-main sequence stars. I. an atlas of line profiles.*, **120**, 229 (1996).

[271] C. E. Jones, C. Tycner and A. D. Smith, *The Variability of Hα Equivalent Widths in Be Stars*, **141**, 150 (2011).

[272] J. Silaj, C. E. Jones, T. A. A. Sigut and C. Tycner, *The Hα Profiles of Be Shell Stars*, **795**, 82 (2014).

[273] R. Ignace, S. K. Gray, M. A. Magno, G. D. Henson and D. Massa, *A Study of Hα Line Profile Variations in β Lyr*, **156**, 97 (2018).

[274] T. Kogure and K.-C. Leung, *Astrophysics and Space Science Library*, Vol. 342 (2007).

[275] R. J. White and G. Basri, *Very Low Mass Stars and Brown Dwarfs in Taurus-Auriga*, **582**, 1109 (2003).

[276] A. Natta, L. Testi, J. Muzerolle, S. Randich, F. Comerón and P. Persi, *Accretion in brown dwarfs: An infrared view*, **424**, 603 (2004).

[277] A. R. Witham, C. Knigge, J. E. Drew, R. Greimel, D. Steeghs, B. T. Gänsicke, P. J. Groot and A. Mampaso, *The IPHAS catalogue of Hα emission-line sources in the northern Galactic plane*, **384**, 1277 (2008).

[278] B. Mathew, A. Subramaniam and B. C. r. Bhatt, *Be phenomenon in open clusters: results from a survey of emission-line stars in young open clusters*, **388**, 1879 (2008).

[279] G. Matijevič, T. Zwitter, O. Bienaymé and et al., *Exploring the Morphology of RAVE Stellar Spectra*, **200**, 14 (2012).

[280] J. E. Drew, E. Gonzalez-Solares, R. Greimel and et al., *The VST Photometric Hα Survey of the Southern Galactic Plane and Bulge (VPHAS+)*, **440**, 2036 (2014).

[281] A. Aret, M. Kraus and M. Šlechta, *Spectroscopic survey of emission-line stars - I. B[e] stars*, **456**, 1424 (2016).

[282] E. H. Nikoghosyan, A. V. Vardanyan and K. G. Khachatryan, *The Search and Study of PMS Stars with Hα Emission*, in *Astronomical Surveys and Big Data*, Astronomical Society of the Pacific Conference Series, Vol. 505, edited by A. Mickaelian, A. Lawrence and T. Magakian (2016) p. 66, arXiv:1512.02729 [astro-ph.SR] .

[283] L. Kohoutek and R. Wehmeyer, *Catalogue of H-alpha emission stars in the Northern Milky Way*, **134**, 255 (1999).

[284] W. A. Reid and Q. A. Parker, *Emission-line stars discovered in the UKST Hα survey of the Large Magellanic Cloud - I. Hot stars*, **425**, 355 (2012).

[285] W. Hou, A. L. Luo, J.-Y. Hu and et al., *A catalog of early-type emission-line stars and Hα line profiles from LAMOST DR2*, Research in Astronomy and Astrophysics **16**, 138 (2016).

[286] T. J. Bohuski, *Structure of the H II regions M8 and M20. II. Electron densities.*, **184**, 93 (1973).

[287] K. P. Raju, C. D. Prasad, J. N. Desai and L. Mishra, *A kinematic study of orion nebula in the emission line [S II] 6731 Å*, **204**, 205 (1993).

[288] V. Escalante and C. Morisset, *The NII spectrum of the Orion nebula*, **361**, 813 (2005).

[289] F. Damiani, R. Bonito, L. Magrini and et al., *Gaia-ESO Survey: Gas dynamics in the Carina nebula through optical emission lines*, **591**, A74 (2016).

[290] F. Damiani, R. Bonito, L. Prisinzano and et al., *The Gaia-ESO Survey: dynamics of ionized and neutral gas in the Lagoon nebula (M 8)*, **604**, A135 (2017).

[291] M. Žerjal, T. Zwitter, G. Matijevič and et al., *Chromospherically Active Stars in the RAdial Velocity Experiment (RAVE) Survey. I. The Catalog*, **776**, 127 (2013).

[292] R. L. Kurucz, *SYNTHE spectrum synthesis programs and line data* (1993).

[293] U. Munari, R. Sordo, F. Castelli and T. Zwitter, *An extensive library of 2500 10 500 Å synthetic spectra*, **442**, 1127 (2005).

[294] P. de Laverny, *The AMBRE grid: a new library of cool stars high-resolution synthetic spectra in the optical domain*, in *Astronomical Society of India Conference Series*, Astronomical Society of India Conference Series, Vol. 6, edited by P. Prugniel and H. P. Singh (2012) p. 53.

[295] M. Ness, D. W. Hogg, H. W. Rix, A. Y. Q. Ho and G. Zasowski, *The Cannon: A data-driven approach to Stellar Label Determination* (2015) p. 16, arXiv:1501.07604 [astro-ph.SR] .

[296] H.-r. Qin, J.-m. Lin and J.-y. Wang, *Stacked Denoising Autoencoders Applied to Star/Galaxy Classification*, **41**, 282 (2017).

[297] H. Shen, D. George, E. A. Huerta and Z. Zhao, *Denoising Gravitational Waves with Enhanced Deep Recurrent Denoising Auto-Encoders*, arXiv e-prints , arXiv:1903.03105 (2019).

[298] W. Li, H. Xu, Z. Ma and et al., *Separating the EoR signal with a convolutional denoising autoencoder: a deep-learning-based method*, **485**, 2628 (2019).

[299] X.-R. Li, R.-Y. Pan and F.-Q. Duan, *Parameterizing Stellar Spectra Using Deep Neural Networks*, Research in Astronomy and Astrophysics **17**, 036 (2017).

[300] R.-y. Pan and X.-r. Li, *Stellar Atmospheric Parameterization Based on Deep Learning*, **41**, 318 (2017).

[301] A. Karmakar, D. Mishra and A. Tej, *Stellar Cluster Detection using GMM with Deep Variational Autoencoder*, arXiv e-prints , arXiv:1809.01434 (2018).

[302] T.-Y. Cheng, N. Li, C. J. Conselice, A. Aragón-Salamanca, S. Dye and R. B. Metcalf, *Identifying Strong Lenses with Unsupervised Machine Learning using Convolutional Autoencoder*, arXiv e-prints , arXiv:1911.04320 (2019).

[303] Z. Ma, H. Xu, J. Zhu and et al., *A Machine Learning Based Morphological Classification of 14,245 Radio AGNs Selected from the Best-Heckman Sample*, **240**, 34 (2019).

[304] N. O. Ralph, R. P. Norris, G. Fang and et al., *Radio Galaxy Zoo: Unsupervised Clustering of Convolutionally Auto-encoded Radio-astronomical Images*, **131**, 108011 (2019).

[305] K. He, X. Zhang, S. Ren and J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, arXiv e-prints , arXiv:1502.01852 (2015).

[306] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, arXiv e-prints , arXiv:1412.6980 (2014).

[307] G. S. Bloom, *The Balmer decrement in the emission spectra of astronomical objects*, (1969).

[308] A. Frasca, K. Biazzo, A. C. Lanzafame and et al., *The Gaia-ESO Survey: Chromospheric emission, accretion properties, and rotation in γ Velorum and Chamaeleon I⋆⋆⋆*, **575**, A4 (2015).

[309] R. W. Hanuschik, *A flux-calibrated, high-resolution atlas of optical sky emission from UVES*, **407**, 1157 (2003).

[310] J. I. Castor and H. J. G. L. M. Lamers, *An atlas of theoretical P Cygni profiles.*, **39**, 481 (1979).

[311] G. Traven, S. Feltzing, T. Merle, M. Van der Swaelmen, K. Čotar, R. Church, T. Zwitter, Y. S. Ting, C. Sahlholdt, M. Asplund, J. Bland-Hawthorn, G. De Silva, K. Freeman, S. Martell, S. Sharma, D. Zucker, S. Buder, A. Casey, V. D'Orazi, J. Kos, G. Lewis, J. Lin, K. Lind, J. Simpson, D. Stello, U. Munari and R. A. Wittenmyer, *The GALAH survey: Multiple stars and our Galaxy. I. A comprehensive method for deriving properties of FGK binary stars*, arXiv e-prints , arXiv:2005.00014 (2020).

[312] T. Merle, S. Van Eck, A. Jorissen and et al., *The Gaia-ESO Survey: double-, triple-, and quadruple-line spectroscopic binary candidates*, **608**, A95 (2017).

[313] L. Capitanio, R. Lallement, J. L. Vergely, M. Elyajouri and A. Monreal-Ibero, *Three-dimensional mapping of the local interstellar medium with composite data*, **606**, A65 (2017).

[314] A. Davenhall and S. Leggett, *Catalogue of Constellation Boundary Data*, Royal Observatory of Edinburgh (1989).

[315] B. J. McLean, G. R. Greene, M. G. Lattanzi and B. Pirenne, *The Status of the Second Generation Digitized Sky Survey and Guide Star Catalog*, in *Astronomical Data Analysis Software and Systems IX*, Astronomical Society of the Pacific Conference Series, Vol. 216, edited by N. Manset, C. Veillet and D. Crabtree (2000) p. 145.

[316] L. Venuti, L. Prisinzano, G. G. Sacco and et al., *The Gaia-ESO Survey and CSI 2264: Substructures, disks, and sequential star formation in the young open cluster NGC 2264*, **609**, A10 (2018).

[317] J. Meléndez, W. J. Schuster, J. S. Silva, I. Ramírez, L. Casagrande and P. Coelho, *uvby-β photometry of solar twins . The solar colors, model atmospheres, and the T_{eff} and metallicity scales*, **522**, A98 (2010).

[318] J. Datson, C. Flynn and L. Portinari, *New solar twins and the metallicity and temperature scales of the Geneva-Copenhagen Survey*, **426**, 484 (2012).

[319] G. Cayrel de Strobel, N. Knowles, G. Hernandez and C. Bentolila, *In search of real solar twins.*, **94**, 1 (1981).

[320] P. Jofré, T. Mädler, G. Gilmore, A. R. Casey, C. Soubiran and C. Worley, *Climbing the cosmic ladder with stellar twins*, **453**, 1428 (2015).

[321] T. Mädler, P. Jofré, G. Gilmore, C. Clare Worley, C. Soubiran, S. Blanco-Cuaresma, K. Hawkins and A. R. Casey, *Stellar twins determine the distance of the Pleiades*, **595**, A59 (2016).

[322] P. Jofré, G. Traven, K. Hawkins and et al., *Climbing the cosmic ladder with stellar twins in RAVE with Gaia*, **472**, 2517 (2017).

[323] I. Ramírez, J. Meléndez and M. Asplund, *Accurate abundance patterns of solar twins and analogs. Does the anomalous solar chemical composition come from planet formation?*, **508**, L17 (2009).

[324] P. E. Nissen, *High-precision abundances of elements in solar twin stars. Trends with stellar age and elemental condensation temperature*, **579**, A52 (2015).

[325] I. Ramírez, J. Meléndez, J. Bean and et al., *The Solar Twin Planet Search. I. Fundamental parameters of the stellar sample*, **572**, A48 (2014).

[326] P. E. Nissen, *High-precision abundances of Sc, Mn, Cu, and Ba in solar twins. Trends of element ratios with stellar age*, **593**, A65 (2016).

[327] P. Jofré, H. Jackson and M. Tucci Maia, *Traits for chemical evolution in solar twins. Trends of neutron-capture elements with stellar age*, **633**, L9 (2020).

[328] J. Meléndez, I. Ramírez, A. I. Karakas and et al., *18 Sco: A Solar Twin Rich in Refractory and Neutron-capture Elements. Implications for Chemical Tagging*, **791**, 10.1088/0004-637X/791/1/14 (2014).

[329] R. L. Kurucz, *New atlases for solar flux, irradiance, central intensity, and limb intensity*, Memorie della Societa Astronomica Italiana Supplementi **8**, 189 (2005).

[330] P. Jofré, U. Heiter, C. Soubiran and et al., *Gaia FGK benchmark stars: Metallicity*, **564**, A133 (2014).

[331] U. Heiter, P. Jofré, B. Gustafsson, A. J. Korn, C. Soubiran and F. Thévenin, *Gaia FGK benchmark stars: Effective temperatures and surface gravities*, **582**, A49 (2015).

[332] G. N. Lance and W. T. Williams, *Mixed-Data Classificatory Programs I - Agglomerative Systems.*, Australian Computer Journal **1**, 15 (1967).

[333] D. Raghavan, H. A. McAlister, T. J. Henry, D. W. Latham, G. W. Marcy, B. D. Mason, D. R. Gies, R. J. White and T. A. ten Brummelaar, *A Survey of Stellar Families: Multiplicity of Solar-type Stars*, **190**, 1 (2010).

[334] G. Duchêne and A. Kraus, *Stellar Multiplicity*, **51**, 269 (2013).

[335] M. Moe and R. Di Stefano, *Mind Your Ps and Qs: The Interrelation between Period (P) and Mass-ratio (Q) Distributions of Binary Stars*, **230**, 15 (2017).

[336] A. Tokovinin, *From Binaries to Multiples. II. Hierarchical Multiplicity of F and G Dwarfs*, **147**, 87 (2014).

[337] C. F. Quist and L. Lindegren, *Statistics of Hipparcos binaries: probing the 1-10 AU separation range*, **361**, 770 (2000).

[338] D. Pourbaix, A. A. Tokovinin, A. H. Batten and et al., *$S_{B⁹}$: The ninth catalogue of spectroscopic binary orbits*, **424**, 727 (2004).

[339] M. A. Fernandez, K. R. Covey, N. De Lee, S. D. Chojnowski, D. Nidever, R. Ballantyne, M. Cottaar, N. Da Rio, J. B. Foster, S. R. Majewski, M. R. Meyer, A. M. Reyna, G. W. Roberts, J. Skinner, K. Stassun, J. C. Tan, N. Troup and G. Zasowski, *IN-SYNC VI. Identification and Radial Velocity Extraction for 100+ Double-Lined Spectroscopic Binaries in the APOGEE/IN-SYNC Fields*, **129**, 084201 (2017).

[340] A. Duquennoy and M. Mayor, *Multiplicity among solar-type stars in the solar neighbourhood. II - Distribution of the orbital elements in an unbiased sample*, **248**, 485 (1991).

[341] B. Nordström, M. Mayor, J. Andersen, J. Holmberg, F. Pont, B. R. Jørgensen, E. H. Olsen, S. Udry and N. Mowlavi, *The Geneva-Copenhagen survey of the Solar neighbourhood. Ages, metallicities, and kinematic properties of 14 000 F and G dwarfs*, **418**, 989 (2004).

[342] G. Matijevič, T. Zwitter, O. Bienaymé and et al., *Single-lined Spectroscopic Binary Star Candidates in the RAVE Survey*, **141**, 200 (2011).

[343] N. W. Troup, D. L. Nidever, N. De Lee and et al., *Companions to APOGEE Stars. I. A Milky Way-spanning Catalog of Stellar and Substellar Companion Candidates and Their Diverse Hosts*, **151**, 85 (2016).

[344] C. Badenes, C. Mazzola, T. A. Thompson and et al., *Stellar Multiplicity Meets Stellar Evolution and Metallicity: The APOGEE View*, **854**, 147 (2018).

[345] P. M. Garnavich, *Wide binary stars at the Galactic poles*, **335**, L47 (1988).

[346] L. M. Close, H. B. Richer and D. R. Crabtree, *A complete sample of wide binaries in the solar neighborhood*, **100**, 1968 (1990).

[347] A. Gould, J. N. Bahcall, D. Maoz and B. Yanny, *Star counts from the HST Snapshot Survey. 2: Wide binaries*, **441**, 200 (1995).

[348] A. L. Kraus and L. A. Hillenbrand, *Unusually Wide Binaries: Are They Wide or Unusual?*, **703**, 1511 (2009).

[349] E. J. Shaya and R. P. Olling, *Very Wide Binaries and Other Comoving Stellar Companions: A Bayesian Analysis of the Hipparcos Catalogue*, **192**, 2 (2011).

[350] J. J. Andrews, J. Chanamé and M. A. Agüeros, *Wide binaries in Tycho-Gaia: search method and the distribution of orbital separations*, **472**, 675 (2017).

[351] J. Coronado, M. P. Sepúlveda, A. Gould and J. Chanamé, *A distant sample of halo wide binaries from SDSS*, **480**, 4302 (2018).

[352] S. Oh, A. M. Price-Whelan, D. W. Hogg, T. D. Morton and D. N. Spergel, *Comoving Stars in Gaia DR1: An Abundance of Very Wide Separation Comoving Pairs*, **153**, 257 (2017).

[353] F. M. Jiménez-Esteban, E. Solano and C. Rodrigo, *A Catalog of Wide Binary and Multiple Systems of Bright Stars from Gaia-DR2 and the Virtual Observatory*, **157**, 78 (2019).

[354] A. Skopal, *Disentangling the composite continuum of symbiotic binaries. I. S-type systems*, **440**, 995 (2005).

[355] A. Rebassa-Mansergas, B. T. Gänsicke, P. Rodríguez-Gil, M. R. Schreiber and D. Koester, *Post-common-envelope binaries from SDSS - I. 101 white dwarf main-sequence binaries with multiple Sloan Digital Sky Survey spectroscopy*, **382**, 1377 (2007).

[356] A. Rebassa-Mansergas, A. Nebot Gómez-Morán, M. R. Schreiber, B. T. Gänsicke, A. Schwope, J. Gallardo and D. Koester, *Post-common envelope binaries from SDSS - XIV. The DR7 white dwarf-main-sequence binary catalogue*, **419**, 806 (2012).

[357] J. Ren, A. Luo, Y. Li, P. Wei, J. Zhao, Y. Zhao, Y. Song and G. Zhao, *White-dwarf-Main-sequence Binaries Identified from the LAMOST Pilot Survey*, **146**, 82 (2013).

[358] A. Rebassa-Mansergas, J. J. Ren, S. G. Parsons, B. T. Gänsicke, M. R. Schreiber, E. García-Berro, X.-W. Liu and D. Koester, *The SDSS spectroscopic catalogue of white dwarf-main-sequence binaries: new identifications from DR 9-12*, **458**, 3808 (2016).

[359] C. Bergmann, M. Endl, J. B. Hearnshaw, R. A. Wittenmyer and D. J. Wright, *Searching for Earth-mass planets around α Centauri: precise radial velocities from contaminated spectra*, International Journal of Astrobiology **14**, 173 (2015).

[360] F. R. Ferraro, E. Carretta, F. Fusi Pecci and A. Zamboni, *Binary stars in globular clusters: detection of a binary sequence in NGC 2808?*, **327**, 598 (1997).

[361] J. Hurley and C. A. Tout, *The binary second sequence in cluster colour-magnitude diagrams*, **300**, 977 (1998).

[362] A. P. Milone, A. F. Marino, L. R. Bedin, A. Dotter, H. Jerjen, D. Kim, D. Nardiello, G. Piotto and J. Cong, *The binary populations of eight globular clusters in the outer halo of the Milky Way*, **455**, 3009 (2016).

[363] M. Žerjal, M. J. Ireland, T. Nordlander, J. Lin, S. Buder, L. Casagrande, K. Čotar, G. De Silva, J. Horner, S. Martell, G. Traven and T. Zwitter, *The GALAH Survey: lithium-strong KM dwarfs*, **484**, 4591 (2019).

[364] A. A. Henden, S. Levine, D. Terrell and D. L. Welch, *APASS - The Latest Data Release*, in *American Astronomical Society Meeting Abstracts #225*, American Astronomical Society Meeting Abstracts, Vol. 225 (2015) p. 336.16.

[365] E. L. Wright, P. R. M. Eisenhardt, A. K. Mainzer and et al., *The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance*, **140**, 1868 (2010).

[366] C. N. A. Willmer, *The Absolute Magnitude of the Sun in Several Filters*, **236**, 47 (2018).

[367] L. Casagrande and D. A. VandenBerg, *On the use of Gaia magnitudes and new tables of bolometric corrections*, **479**, L102 (2018).

[368] E. F. Schlafly and D. P. Finkbeiner, *Measuring Reddening with Sloan Digital Sky Survey Stellar Spectra and Recalibrating SFD*, **737**, 103 (2011).

[369] P. Marigo, L. Girardi, A. Bressan and et al.., *A New Generation of PARSEC-COLIBRI Stellar Isochrones Including the TP-AGB Phase*, **835**, 77 (2017).

[370] H. Kamdar, C. Conroy, Y.-S. Ting, A. Bonaca, B. Johnson and P. Cargile, *A Dynamical Model for Clustered Star Formation in the Galactic Disk*, arXiv e-prints , arXiv:1902.10719 (2019).

[371] H. Kamdar, C. Conroy, Y.-S. Ting, A. Bonaca, M. Smith and A. G. A. Brown, *Stars that Move Together Were Born Together*, arXiv e-prints , arXiv:1904.02159 (2019).

[372] D. Foreman-Mackey, D. W. Hogg, D. Lang and J. Goodman, *emcee: The MCMC Hammer*, **125**, 306 (2013).

[373] P. Eggleton, *Evolutionary Processes in Binary and Multiple Stars, by Peter Eggleton, pp. . ISBN 0521855578. Cambridge, UK: Cambridge University Press, 2006.* (2006).

[374] A. Tokovinin, *Comparative statistics and origin of triple and quadruple stars*, **389**, 925 (2008).

[375] A. Tokovinin, *The Updated Multiple Star Catalog*, **235**, 6 (2018).

[376] A. A. Tokovinin, *On the origin of binaries with twin components*, **360**, 997 (2000).

[377] F. Arenou, X. Luri, C. Babusiaux, C. Fabricius and et al., *Gaia Data Release 2. Catalogue validation*, **616**, A17 (2018).

[378] S. Sharma, J. Bland-Hawthorn, K. V. Johnston and J. Binney, *Galaxia: A Code to Generate a Synthetic Survey of the Milky Way*, **730**, 3 (2011).

Razširjeni povzetek v slovenskem jeziku

7.1 Uvod

Galaksija, kot jo lahko občudujemo dandanes na podlagi opazovanj, ni nastala v enem trenutku, ampak je rezultat postopnega rojevanja zvezd in dodatnih zunanjih gravitacijskih vplivov. Njeni najmlajši razpoznavni sestavni deli so razsute zvezdne kopice, ki so nastale iz istega molekularnega oblaka snovi in zaradi tega vse zvezde v njej še vedno ohranjajo določene skupne lastnosti. Najočitnejše lastnosti so njihova zgoščena lokacija na nebu in skupna smer premikanja po Galaksiji. Poleg tega lahko opazujemo še kemično sestavo posameznih zvezd v kopici, ki nam podaja neposredno informacijo o sestavi prvotnega oblaka materiala. Kot vsaka gravitacijsko slabše povezana tvorba tudi razsute kopice sčasoma razpadejo in se pomešajo med okoliške zvezde, trajanje takega procesa pa je odvisen od njene začetne mase. Glavna mehanizma za izgubo zvezd kopice sta izmetavanje članov tekom bližnjih srečanj znotraj kopice ter počasno gravitacijsko odstranjevanje zaradi vpliva zvezd izven kopice. Gravitacijsko trenje in odstranjevanje je najučinkovitejše, ko kopica potuje v gosto naseljenem področju. Takrat zvezde v kopici vlečejo navidezno stacionarne okoliške zvezde proti sebi. Posledično izgubijo nekaj energije, zvezdam kopice se zmanjša hitrost in posledično se spremeni tudi orbita potovanja okrog centra Galaksije. Takšne interakcije lahko trajajo dokler se kopica postopoma popolno ne zlije z okoliškimi strukturami.

Čeprav so kopice zaradi svojega izvora iz skupnega homogenega oblaka snovi idealne za raziskovanje strukture in gradnikov Galaksije, je njihova pričakovana življenjska doba približno 100 milijonov let. Najgostejše strukture lahko preživijo tudi nekaj milijard let. Tako kratka življenjska doba nam skrajšuje čas v katerem so kopice na voljo za raziskovanje, vendar nam po drugi strani njihov hiter razvoj obenem ponuja možnost hkratnega opazovanja dinamično različno razvitih struktur.

7.1.1 Razsute kopice v dobi satelita *Gaia*

Najnovejši podatki s satelita *Gaia* so močno izboljšali znane podatke o zvezdah. Končno poznamo natančne informacije o svetlosti, poziciji, oddaljenosti in popolnem galaktičnem gibanju tako temnih kot svetlih zvezd po celotnem nebu. Takšen podatkovni set je pripomogel tudi k ponovnemu zanimanju za razsute kopice, kar v zadnjem času opazimo kot povečano število novih dognanj o teh strukturah.

Prvi korak v raziskovanju kopic je natančna določitev članov posamezne kopice. Prvi postopki so temeljili na preprostem izbiranju zvezd na podlagi njihove lokacije na nebu, saj kopico opazimo kot lokalno zgostitev števila zvezd na majhnem obmo-

čju. Dandanes za tako določevanje uporabljamo kompleksnejše algoritme, saj lahko uporabimo mnogo dodatnih informacij poleg pozicije zvezd. Najpogostejši način izbiranja temelji na določitvi zgostitve v kinematičnem prostoru, saj ima večina kopic izrazito drugačno lastno in radialno gibanje v primerjavi z okolico. S poznavanjem barve, navidezne svetlosti zvezde in njene oddaljenosti lahko preko teoretičnega modela razvoja zvezd na HR diagramu izločimo še preostale napačno izbrane člane in posledično določimo natančno starost kopice.

Čeprav so napake meritev paralakse za bolj oddaljene kopice primerljive z njihovo velikostjo, vseeno lahko določimo njihovo približno obliko. Zaradi merskih napak le-ta ni pričakovane sferične oblike, ampak je razpotegnjena v radialni smeri proč od opazovalca. Poznavanje oblike in kopice in razporeditve članov nam omogoča študije notranjih dinamičnih procesov in porazdelitev mase zvezd. Še bolj natančne študije dogajanja znotraj samih kopic bodo omogočene z naslednjim, tretjim izidom podatkov *Gaia*, za katerega pričakujemo še dodatno zmanjšanje merskih nedoločenosti.

7.1.2 Metoda kemičnih podpisov

Vse prej opisane metodologije potrebujejo le popolne podatke o lokaciji in gibanju zvezd. Napredek opazovalnih metod in obdelave podatkov nam v zadnjem času omogoča, da presežemo te postopke in v analize vključimo še mnogo dimenzionalne podatke o kemični sestavi zvezd - postopek poznan pod imenom metoda kemičnih podpisov [66, 67]. Ta metoda nam omogoča grupiranje zvezd na njihove izvorne oblake snovi le na podlagi njihovih kemičnih podpisov in s tem odpira vrata v področje podrobnega raziskovanja formiranja in evolucije Galaksije. Takšno popolnoma slepo kemično določanje za razsute kopice trenutno še ni bilo pokazano, razen v redkih primerih, ko je kopica zašla v drug del Galaksije in tako imela močno drugačen kemični podpis v primerjavi z okoliškimi zvezdami [68]. Večji uspeh ima metoda na razločevanju posameznih komponent Galaksije (disk, osrednja odebelitev in halo), saj so bile te formirane ob zelo različnih časih, ko je bila njihova izvorna snov različno obogatena s težjimi snovi. Te elemente, težje od helija, formirajo in oddajo v okoliški prostor razvite zvezde na koncu svojega življenjskega cikla.

7.1.3 Posebni zvezdni spektri

Metoda kemičnih podpisov ni omejena le z fizikalnim dogajanjem v Galaksiji in kvaliteto opazovalnih podatkov, ampak tudi s tipom zvezd, ki jih z njo analiziramo. Poleg množice normalnih tipov zvezd, katerih znamo opazovan spekter fizikalno rutinsko opisati, poznamo še posebne tipe zvezd, katerih spekter vsebuje nepričakovane in s tem hkrati tudi zelo zanimive spektralne posebnosti. Med posebne tipe zvezd, odvisno od fokusa raziskave, štejemo dvojne ali večkratne zvezde, zelo mlade zvezde, zvezde, ki vsebujejo nadpovprečne količine posameznega kemičnega elementa, zvezde s kompleksnimi emisijskimi profili in nenazadnje možne napake v obdelavi podatkov. Zaradi tega želimo takšne zvezde identificirati in jih izločiti iz standardnega nabora podatkov. Pri njihovi analizi lahko z avtomatskimi procedurami, ki pričakujejo normalne spektre, pridobimo potencialno napačne parametre, kar bi še dodatno oteževalo kemično določevanje skupkov sorodnih zvezd.

V vsakem obširnem pregledu neba, kot je GALAH, želimo zato pridobiti čimbolj

podroben in raznolik seznam posebnih tipov zvezd, ki smo jih opazovali pri ne selektivnem izbiranju zvezd le po njihovi navidezni svetlosti. Njihova svetlost o sami zvezdi ne pove nič drugega kot to, ali bomo tekom opazovanja pridobili ustrezno kvaliteten spekter za nadaljno obdelavo.

7.1.4 Strojno učenje in obširni pregledi neba

V zadnjem desetletju smo priča porastu zbranih podatkov v večini znanstvenih področij, vključno z astronomijo. Tako velike količine zbranih podatkov je potrebno tudi analizirati, saj nam le varno shranjene na podatkovnih medijih ne prinašajo dodatne vrednosti. Na srečo vzporedno opazujemo tudi napredek in porast števila algoritmov strojnega učenja, ki nam pomagajo in omogočajo enostavnejše ter hitrejše razumevanje zbranih podatkov. Računalnik lahko za nas postori velik del opravil, ki so jih pred tem raziskovalci morali opravljati ročno. Tu pa lahko opazimo tudi slabost takšnih metod, saj le-te ponavadi ne vsebujejo dodatnega podrobnega človeškega fizikalnega znanja, ki bi jim pomagal v nepoznanih primerih. Da se izognemo takim nevšečnostim, je potrebno pred slepo uporabo metod strojnega učenja pripraviti smiselne učne primere, ki bodo algoritmu pomagali pravilno razumeti zadani problem.

V astronomski literaturi so bile nedavno metode strojnega učenja že uspešno uporabljene za določanje zvezdnih parametrov iz spektrov [103, 104, 105], razpoznavanje posebnih tipov zvezd [84], aplikacijo metode kemičnih podpisov [8, 68, 70, 73, 112, 113] in mnogo drugih primerov, primarno uporabljenih pri razvrščanju zvezdnih spektrov. Med vsemi možnimi tipi metod strojnega učenja pri obdelavi astronomskih spektralnih podatkovnih zbirk v zadnjem času opažamo porast zanimanja za uporabo različnih arhitektur nevronskih mrež [106, 107, 108, 109].

Poleg spektroskopskih podatkov je zelo primeren vir podatkov za uporabo metod strojnega učenja tudi najaktualnejši izid *Gaia* podatkov, saj se v njem nahaja več kot milijarda zvezd. Njihova opazovanja so bila med drugim uspešno uporabljena za klasifikacijo tipov spremenljivih zvezd [115], določanje efektivne temperature zvezd [116], določanje potokov zvezd v galaktičnem haloju [118, 119], identifikacijo akrecijskih zvezd [120, 121], izsledite zelo hitrih zvezd [122], razločevanje med zvezdnimi izviri svetlobe v naši galaksiji in galaktičnimi izvori izven nje [123, 124], izračun pordečitve zaradi medzvezdnega prahu [125] in razpoznavanje še neodkritih razsutih zvezdnih kopic [126].

7.1.5 Naše raziskave

Naše raziskovanje primarno spektroskopskih podatkov razsutih zvezdnih kopic in posebnih tipov zvezd sestavljajo tri medsebojno povezane teme, ki so v nadaljnjih poglavjih tudi podrobneje predstavljene:

- Pri raziskovanju razsutih kopic smo si v Poglavju 3 pomagali s podatki satelita *Gaia*, preko katerih smo kopico in njeno okolico razdelili na več komponent. Med njimi so nas najbolje zanimale zvezde, ki bi lahko v preteklosti bile izvržene iz same kopice daleč proč od nje v območje drugih zvezd. Na identificiranih primerih smo s pomočjo metode kemičnih podpisov preizkusili, ali njene predpostavke držijo za naše podatke in ali bi uporabljeno metodo lahko aplicirali tudi za neznane strukture v Galaksiji.

- V podatkih spektroskopskega pregleda neba GALAH se poleg normalnih zvezd nahaja tudi kopica posebnih tipov zvezd, ki jih je potrebno čim učinkoviteje identificirati za natančnejše delovanje metode kemičnih podpisov. Za ta namen smo uporabili več nenadzorovanih in delno nadzorovanih metod strojnega učenja, ki so bile naučene na normaliziranih spektralnih podatkih. Tako smo med pregledom spektrov odkrili 918 zvezd z močno izraženimi molekularnimi ogljikovimi črtami, ki nakazujejo na nadpovprečno vsebnost ogljika v atmosferi zvezde. Odkrili smo tudi 10.364 spektrov z izraženimi vodikovimi emisijskimi črtami, ki so posledica različnih fizikalnih procesov v okolici zvezde.

- Zanimali so nas tudi spektralni sončevi dvojniki, saj je Sonce v astronomiji uporabno kot standardna zvezda za raznorazne kalibracije. S primerjavo spektrov smo določili 329 opazovanih GALAH spektrov, ki so najbolj podobni Soncu in pregledali njihove fizikalne parametre ter absolutni izsev. Nekatere od njih imajo mnogo močnejši izsev od Sonca, kar nakazuje na prisotnost več podobnih zvezd v opazovanem sistemu. Možnost obstoja dvojnih in trojnih zvezd, ki imitirajo spekter Sonca, smo potrdili s simulacijami in združevanjem njihovih opazovalnih podatkov.

7.2 Pregledi neba

Astronomska opazovanja so se v zadnjih desetletjih močno spremenila, saj gradimo vedno večje teleskope in kompleksnejše sisteme, ki namesto posameznih namensko izbranih objektov hkrati opazujejo več sto ali tisoč objektov. Seveda takšen način opazovanj za seboj povleče tudi zahtevnejšo obdelavo zbranih surovih podatkov in drugačen pristop k znanosti, ki jo iz obdelanih podatkov lahko pridobimo. Avtomatski, ponavadi le deloma nadzorovani sistemi za obdelavo in analizo podatkov najpogosteje niso prilagojeni za celoten možen razpon opazovanj. Zato moramo biti uporabniki še bolj pozorni na razne opozorilne zastavice, ki nakazujejo na možne nepravilnosti tekom teh procedur in tudi poznati same procedure. Le tako se lahko zavedamo vseh možnih pasti. Nekaj takšnih posebnosti pri razumevanju pridobljenih podatkov si bomo pogledali tudi v naslednjih poglavjih.

Izmed vseh možnih astronomskih podatkovnih setov, smo se v disertaciji osredotočili na sledeče tri preglede neba. Ti nam podajajo informacije o svetlosti zvezd, njihovi sestavi, oddaljenosti, premikanju po Galaksiji in mnogo drugih parametrov, ki jih lahko razberemo iz opazovanj.

7.2.1 Vesoljska misija *Gaia*

Satelit *Gaia* [91] je Evropska vesoljska agencija (Esa) izstrelila v drugo Lagranževo točko sistema Zemlja-Sonce, kjer že od julija 2014 neprestano skenira zvezdno nebo. Končni cilj misije je izdelava najbolj natančne zvezdne karte zvezd do magnitude ∼20,7 ter pridobitev točnih informacij o njihovi oddaljenosti in lastnem gibanju. Končni katalog bo tako vseboval več kot milijardo zvezd. To bo satelit zmogel le s sistematičnem pregledovanjem neba, ki je sestavljeno iz dveh neodvisno vrtečih se gibanj: vrtenja okrog lasne osi vsakih šest ur ter počasne 63-dnevne precesije lastne vrtilne osi, ki je glede na Sonce postavljena pod kotom 45°. V prvih petih letih predvidenega delovanja je *Gaia* opravila 29 precesijskih period gibanja smeri

osi, kar je vodilo v optimalno pokritost zvezdnega neba. Vsaka zvezda je bila tako opazovana približno 70-krat za določitev njene pozicije in barve ter 40-krat za izmero radialne hitrosti. V podaljšku načrtovane misije, ki se je začel julija 2019, je satelit nadaljeval svoje pregledovanje, vendar z obratno smerjo vrtenja okrog precesijske osi.

Med rotacijo satelita se opazovane zvezde počasi premikajo čez njegovo goriščno ravnino, kjer je nameščenih 106 CCD detektorjev (prikaz na Sliki 2.1). Vsaka zvezda, ki jo opazujemo, se tako zapelje čez kompleksen sistem detektorjev, ki zaporedoma določijo njeno lokacijo na nebu, spektralno porazdelitev energije in posnamejo njen ozkopasovni spekter srednje ločljivosti.

Fotometrija in astrometrija

Ob prihodu v goriščno ravnino zbrana svetloba zvezde najprej posveti na detektor za določanje izvorov (angl. Sky Mapper - SM), ki samodejno zazna zvezde. Označi zvezde, ki so svetlejše od magnitude 20,7 in temnejše od 3. magnitude saj so preostale presvetle za njegovo delovanje. Po zaznavi se zvezda premakne na površinsko največji del CCD detektorjev imenovan astrometrično polje (angl. Astrometric Field - AF). Ta zabeleži natančno pozicijo, ki bo kasneje uporabna tudi za določitev lastne hitrosti in oddaljenosti, ter svetlost že prej detektiranih zvezd. Zabeležena svetlost zvezde nakazuje njeno sešteto svetlost v širokem spektralnem področju od 3300 do 10.500 Å. Izbrano območje je poimenovano *Gaia* G fotometrični pas.

Tekom obdelave pozicij zvezd na astrometričnem polju s prilagajanjem ustrezne funkcije pridobimo še podatke o paralaksi zvezde in njeno lastno gibanje. Če je izvor svetlobe sestavljen iz več neenakomerno svetlih komponent, se bo fotometričen center počasi premikal ob njihovem kroženju. Posledično to vpliva tudi na kvaliteto prilagojene funkcije. Za zaznavo takšnih astrometričnih večkratnih zvezd bodo v prihodnosti pri analizi ustrezno prilagodili prilegajočo funkcijo in tako iz podatkov izluščili še možnost večkratnosti izvora in orbitalno periodo.

Zadnjo meritev svetlosti nato opravi še spektrofotometrični sistem [141], ki z disperzijskim detektorjem izmeri natančne svetlobne tokove v večih ozkopasovnih pod-pasovih prej omenjenega širokega *Gaia* G pasu. Modri fotometer (angl. Blue Photometer - BP) posname spekter nizke ločljivosti vseh zvezd v območju od 3300 do 6800 Å. Integrirano magnitudo označimo kot BP ali G_{BP}. Podobno rdeči fotometer (angl. Red Photometer - RP) analizira svetlobo v območju od 6300 do 10.500 Å, kjer je njena integrirana vrednost podana kot RP ali G_{RP} magnituda. Oba nizko ločljivostna spektrometra BP in RP v svojem danem območju posnameta spekter v 12 pod-območjih. Te meritve uporabnikom še niso na voljo in bodo predvidoma objavljene kot del tretjega izida *Gaia* podatkov.

Spektroskopija

Končna meritev, ki jo *Gaia* opravi, je spektroskopija celotnega opazovanega zvezdnega polja. Vgrajeni spektrometer za radialne hitrosti (angl. Radial Velocity Spectrometer - RVS, [137]) za zvezde, svetlejše od 16. magnitude, zbere spektre srednje ločljivosti z resolucijsko močjo R ∼11.700 na spektralnem območju od 8450 do 8720 Å. Razpon pokriva območja treh enkrat ioniziranih kalcijevih absorpcijskih črt, ki so primerne za določevanje radialnih hitrosti raznovrstnih zvezd na širokem

temperaturnem območju. Hitro potovanje zvezd po polju CCD detektorjev onemogoča dolgotrajno zbiranje svetlobe in s tem kvalitetnejše spektre. V končnem izidu podatkov bodo za nadaljnjo analizo vsi zbrani spektri posamezne zvezde združeni v en spekter višje kakovosti. Podobno bodo v naslednjih izdajah podatkov poleg povprečne radialne hitrosti čez celotno obdobje opazovanj objavljene še hitrosti za vsako posamezno opazovanje, kar bo omogočilo študije velikega števila dvojnih zvezd.

7.2.2 Pregled neba GALAH

Večinski del podatkov, uporabljenih za izdelavo te disertacije, je bil pridobljen iz arhiva še vedno tekočega pregleda The GALactic Archaeology with HERMES (GALAH, [152]) in njemu sorodnih pregledov, ki so bili posneti s spektrografom High Efficiency and Resolution Multi-Element Spectrograph (HERMES, [153, 154]), ki omogoča hkratno opazovanje skoraj 400 zvezd. Spektrograf je za zbiranje svetlobe nameščen na štiri-metrskem Anglo-avstralskem teleskopu (AAT) na Observatoriju Siding Springs, Avstralija. Primarni cilj pregleda GALAH je razumevanje procesov, ki so vodili do današnje strukture Galaksije, preko natančnih meritev kemičnih zastopanosti zvezd različnih komponent Galaksije. Resolucijska moč sistema je $\sim$28.000 in pokriva štiri spektralna področja (4713 – 4903 Å, 5648 – 5873 Å, 6478 – 6737 Å in 7585 – 7887 Å), ki so neodvisno zajeta in obdelana v štirih ločenih delih spektrografa (za shematičen prikaz glejte Sliko 2.3). Ta frekvenčna področja poimenujemo tudi kot modri, zeleni, rdeči in bližnje infrardeči spektralni del. Skupaj pokrivajo približno 1000 Å in vsebujejo pomembni vodikovi absorpcijski črti Hα in Hβ. Selekcijska funkcija za izbiro tarč je za svetla polja omejena na zvezde z magnitudo $10 < \mathrm{V} < 12$ in temna polja z magnitudo $12 < \mathrm{V} < 14$. Razen izogibanja delom s premajhno ali preveliko gostoto možnih zvezd okrog Galaktične ravnine ($|b| > 10°$) ni izvedene nobene dodatne vnaprejšnje izbire zvezd. S takšnim enostavnim izbiranjem tarč pridobimo čim večjo raznolikost opazovanih tipov zvezd in si obenem poenostavimo primerjave z raznoraznimi modeli Galaksije ter omogočimo populacijske študije.

Vsa hkratna opazovanja zvezd so zbrana na dvodimenzionalni sliki, ki jo je potrebno obdelati, da z nje pridobimo množico enodimenzionalnih zvezdnih spektrov. Avtomatska procedura za redukcijo opazovanj je podrobno opisana in razložena v Kos *et al.* [100]. Po obdelavi iz spektrov pridobimo radialno hitrost zvezde, njene osnovne fizikalne lastnosti in zastopanosti kemičnih elementov na površju zvezde. Tekom opazovanj se je kompleksnost algoritmov za pridobitev teh parametrov stopnjevala in v aktualni tretji objavi podatkov vključuje tudi lastnosti, ki jih je določila *Gaia*. Vključitev dodatnih informacij nam pomaga pri boljši določitvi parametra $\log g$, ki je slabo razpoznan iz izključno spektroskopskih meritev.

7.2.3 Asiago

Tekom študija sem se večkrat odpravil tudi na Observatorij Asiago v Italiji, ki ponuja zelo drugačen način opazovanja kot prej omenjena masivna pregleda neba. Uporabljen 1,82 m teleskop Copernico je v času naših obiskov okrog polne Lune prilagojen za spektroskopska opazovanja posameznih zanimivih objektov. Zaradi svojega načina delovanja je sistem primeren za podrobnejšo oziroma časovno zamaknjeno analizo spektrov zanimivih zvezd, ki smo jih prej identificirali tekom pregledovanja drugih obširnih pregledov neba. Teleskop je iz nadzorne sobe voden preko oddaljene

povezave, saj se nahaja na bližnji vzpetini Ekar na nadmorski višini 1.366 m, na katero iz observatorija pridemo peš v malce manj kot eni uri.

Nameščen spektrograf tipa Echelle omogoča zajem spektrov z resolucijsko močjo R ∼20.000 na širokem spektralnem področju med 3600 in 7400 Å. Zajet spekter je razdeljen na 30 interferenčnih redov, ki se medsebojno delno prekrivajo. To omogoča izdelavo neprekinjenega spektra zvezde na celotnem zajetem spektralnem področju. Sistem omogoča zajem spektrov z visokim razmerjem med signalom in šumom za zvezde z magnitudo V < 10.

Njegova lokacija na nasprotni Zemljini polobli kot teleskop AAT zmanjšuje presek z opazovanimi zvezdami pregleda GALAH. Kljub temu smo zbrane podatke uspešno uporabili kot dodatek k analizam v Poglavjih 4 in 6, ter kot glavni del objavljenega znanstvenega članka [168] in astronomskega telegrama [167].

7.3 Kemično in dinamično raziskovanje razsutih kopic

Med zvezdnimi polji, opazovanimi tekom pregleda GALAH, najdemo tudi člane razsutih zvezdnih kopic. Z izdajo najnovejšega kataloga *Gaia* DR2 je določevanje njihovega članstva po večini dokaj trivialno in tako so že kmalu po izidu podatkov bili objavljeni prvi rezultati. Za naše potrebe raziskovanja razsutih kopic smo pridobili in uporabili rezultate pripadnosti kopic Berkeley 32, NGC 2516, NGC 2112, NGC 6253, Blanco 1, Ruprecht 147, NGC 2632, NGC 2682, Melotte 22 in Collinder 261, ki jih je objavil Cantat-Gaudin *et al.* [43]. S podobno analizo smo sami določili še člane kopice Melotte 25, saj ta ni bila vsebovana v prej omenjenem delu.

7.3.1 Raziskovanje okolice kopic

V našem primeru nas je poleg samih znanih članov kopice zanimala še neposredna okolica kopice, saj zaradi dinamičnega dogajanja znotraj kopice le-ta počasi izgublja svoje člane, ki se pomešajo med okoliške zvezde. Izvržene članice nekaj časa še ohranijo kinematične podobnosti z matično kopico, s časom pa le-te postanejo neločljive od okoliških zvezd. V nasprotju z vektorjem potovanja, ostane površinska kemična sestava zvezde nespremenjena večino njenega življenja. Tako dolgo časa izkazuje svojo povezavo z rojstnim oblakom in ostalimi zvezdami, ki so bile ustvarjene ob istem času in na podobni lokaciji.

Za raziskovanje smo zato iz kataloga *Gaia* DR2 najprej pridobili vse dostopne informacije o zvezdah v širokem območju okrog središča kopice. Ker temnejše zvezde v tem katalogu nimajo popolne informacije o vektorju gibanja, smo jih dopolnili z radialnimi hitrostmi pridobljenimi v pregledu GALAH. S poznavanjem natančne hitrosti in galaktične lokacije zvezd kopice ter gravitacijskega potenciala Galaksije lahko z integracijo določimo pot posamezne zvezde po Galaksiji. Z uporabo programskega paketa *galpy* [171] smo tako simulirali preteklo galaktično pot vseh zvezd s kompletnim naborom informacij za 120 milijonov let v preteklost, kar je primerljivo s starostjo najmlajših kopic v našem podatkovnem setu.

S poznavanjem poti trenutnih preostalih članov kopice smo ob vsakem vmesnem integracijskem koraku na njene najbolj zunanje člane napeli odsekoma ravne ploskve, ki so definirali najmanjši volumen za zajem celotne kopice. Za določitev v

preteklosti možnih izvrženih članov smo poti vseh okoliških zvezd ob vsakem integracijskem koraku primerjali s takratnim volumnom kopice. Zaradi nedoločenosti *Gaia* parametrov smo vsako primerjavo izvedli 250-krat, kjer smo začetne parametre zvezd žrebali iz Gaussove porazdelitve, določene s srednjo vrednostjo in nedoločenostjo posameznega parametra. Iz vseh ponovitev smo določili verjetnost, da je bila okoliška zvezda nekoč izvržena iz kopice. Verjetnost določata njena pogostost prehoda volumna kopice ter čas, tekom katerega se je najdlje zadrževala znotraj nje. V končni seznam potencialnih izvrženih so na koncu prišle le zvezde, ki so volumen kopice prečkale v vsaj 68% primerih ponovitve simulacije ter znotraj volumna najdlje ostale vsaj milijon let.

7.3.2 Kemična sestava kopic in okolice

Raziskano okolje okrog kopic smo tako razdelili na tri skupine zvezd: poznani člani kopice, potencialni v preteklosti izvrženi člani, ter naključne okoliške zvezde. Po teoriji bi te komponente lahko določili tudi samo z opazovanjem njihovih kemičnih podpisov. Za preverbo izvedljivosti smo najprej narisali grafe posameznih zastopanosti v odvisnosti od efektivne temperature zvezde - za prikaze glej Slike 3.1, 3.2, 3.3 in 3.4. Prva stvar, ki nas je na njih zanimala, je seveda porazdelitev zastopanosti za znane člane kopic. Teoretični modeli pravijo, da bi zastopanosti morale biti identične za vse člane, ne glede na njihove fizikalne parametre. Tega ne vidimo v naših podatkih, saj zastopanost močno variira v odvisnosti od efektivne temperature zvezde. Efekt, ki je večji od same razpršenosti zastopanosti, je lahko vpogled v dejansko stanje ali pa odraža napake oziroma nepopolnosti tekom obdelave podatkov. Ker so trendi vse prisotni, gladki in različni med kopicami, se nam je drugi razlog zdel bolj realen. Za opis trendov smo na podatke zastopanosti v odvisnosti od efektivne temperature prilagodili poligon tretje stopnje.

Po prilagajanju smo se lotili diferencialnega postopka metode kemičnih podpisov, pri kateri smo vse zastopanosti primerjali z določenim trendom, ki naj bi opisoval stanje kemičnih elementov v opazovani kopici. Zanimalo nas je, koliko zvezd ozadja in potencialno izvrženih, ima kemični podpis podoben kopici. Podobnost smo določili tako, da smo za vsako zvezdo prešteli v kolikih elementih padejo njene vrednosti znotraj nedoločenosti okrog prilagojenega trenda. Zvezdo smo šteli kot kemično podobno, če se je s kopico ujemala vsaj v 68% elementov, ki smo jih opazovali (nekatere elemente smo pred analizo izpustili zaradi majhnega števila meritev). Rezultati za posamezne opazovane komponente so predstavljeni v Tabeli 3.3.

7.3.3 Rezultati in zaključki

Raziskovane kopice in njihove zastopanosti dajejo vtis, da raziskovanje kemičnega prostora v okviru GALAH meritev ni enostavno in se je za poenostavitev metode kemičnih podpisov potrebno omejiti vsaj na ožji temperaturni razpon zvezd in omejen volumen Galaksije. Kljub izraženim trendom zastopanosti smo s predhodnim kinematičnim razločevanjem na komponente kopice poizkusili metodo diferencialnega primerjanja zastopanosti. Pri tej metodi smo medsebojno primerjali le zvezde v ozkem temperaturnem razponu in tako efektivno izničil opažene trende. Izkazalo se je, da bi brez predhodne kinematične informacije težko razločili člane nekaterih razsutih kopic, obenem pa so bili kinematično določeni potencialno izvrženi člani

kemično podobnejši kopici kot naključne zvezde okrog nje.

7.4 Kemično posebne zvezde

Uspešnost omenjene metode kemičnih podpisov je močno odvisna tudi od tipa zvezde, saj nepričakovani posebni tipi spektrov lahko tekom obdelave spektra privedejo do napačno določenih fizikalnih parametrov zvezd in posledično tudi zastopanosti. V razsežnih neomejenih pregledih neba tako želimo čim bolj točno vedeti, ali naša opazovanja vsebujejo tudi takšne posebne zvezde. Tekom naše študije smo se zato osredotočili na nekaj zanimivih posebnih tipov. Prvi od njih so spektri, ki nakazujejo veliko vsebnost ogljika v atmosferi zvezde. Za njegovo identifikacijo smo uporabili široke molekularne SWAN pasove C_2, ki jih HERMES zajema na začetnem delu modrega spektralnega območja.

7.4.1 Nadzorovana klasifikacija

Nepričakovane stvari v spektrih je najlažje najti tako, da jih primerjamo s setom referenčnih normalnih spektrov, ki opisujejo povprečni spekter poznanih zvezd. Tak nabor smo za vsako zvezdo posebej izdelali s povprečenjem spektrov, ki so z izbranim spektrom imeli zelo podobne fizikalne parametre (ΔT_{eff}=±75 K, $\Delta \log g$=±0.125 dex in Δ[Fe/H]=±0.05 dex). Spektralne razlike smo pridobili z deljenjem opazovanega in referenčnega spektra. Naša iskana spektralna lastnost se nahaja na že vnaprej znani lokaciji pri valovni dolžini 4737 Å, kar smo izkoristili za določanje njene moči. Na predvideno območje smo prilagodili funkcijo

$$f(\lambda) = f_0 - \log \Gamma(\lambda, S, \lambda_0, A), \tag{7.1}$$

ki nam je sporočila obliko območja, njen integral pa stopnjo izraženosti ogljika v zvezdi. S pomočjo parametrov funkcije smo iz rezultatov najbolj obogatenih spektrov izločili možne napake obdelave podatkov.

7.4.2 Nenadzorovana klasifikacija

Poleg območja z najmočneje izraženim spektralnim pasom molekule C_2, ki smo ga uporabljali za nadzorovano določanje, se v njegovi okolici pojavlja še množica manj izrazitih identifikatorjev te molekule. Zanimalo nas je ali bi nenadzorovane metode strojnega učenja lahko samodejno prepoznale vse te značilnosti in naše iskane spektre razvrstile v skupno gručo. Za ta namen smo uporabili algoritem t-distributed Stochastic Neighbor Embedding (t-SNE, [236]), ki deluje v dveh korakih. Najprej je med vsemi vhodnimi spektri (obrezani na območje 4720–4890 Å) določil njihove medsebojne podobnosti z izračunom evklidskih razdalj. Na podlagi podobnosti algoritem poskuša najti optimalno preslikavo v dvo- ali tri-dimenzionalni prostor, ki je vizualno še sprejemljiv za človeka. Tekom transformacije dobimo posamezne skupine točk, ki predstavljajo podobne vhodne podatke. V našem primeru raziskovanja zvezd, bogatih z molekularnim ogljikom, je transformacija predstavljena na Sliki 4.4. Za lažje raziskovanje prikazane mape smo vanjo vrisali že prej zaznane kemično posebne spektre, kar nam je omogočilo lažje nadaljnje pregledovanje, s katerim smo odkrili še par zanimivih skupin zvezd. Te imajo poleg visoke vsebnosti ogljika še zelo nizko kovinskost, kar jih skupaj dela nadvse zanimive za nadaljnje študije.

7.4.3 Rezultati in zaključki

S kombinacijo nadzorovane in nenadzorovane metode klasifikacije smo odkrili 918 zvezd s kemično zanimivim spektrom. Z analizo njihovih fizikalnih parametrov smo odkrili, da je večina teh zvezd orjakinj, manjši del pa pripada skupini pritlikavk. Čeprav so slednje v našem primeru v manjšini, so zelo zanimive zaradi špekulacij glede njihovega točnega izvora obogatitve z ogljikom. Ena od možnosti predvideva, da je obogatitve opravila njej zelo bližnja zvezda, ki se je že dolgo nazaj razvila, napihnila in s tem kemično obogatila svojo sosedo. Znake takšnega sistema smo iskali na podlagi spreminjanja radialnih hitrosti, vendar z malim številom ponovljenih opazovanj nismo mogli potrditi obstoja kakšnega takega sistema. Za podrobnejšo analizo in potrditev bi potrebovali dodatna opazovanja, ki smo jih že pričeli izvajati na observatoriju Asiagu in v poglavju predstavili tudi enega izmed njih.

7.5 Emisijske zvezde

Podobno kot pri prejšnji določitvi posebnih tipov spektrov, smo tudi emisijske zvezde iskali z metodo direktne primerjave med normalnim oziroma referenčnim spektrom ter opazovanim spektrom. V spektru iskane značilke se nahajajo le v modrem in rdečem delu HERMES spektra, zato smo se osredotočiti le nanju.

7.5.1 Simulacija spektrov z avtoenkoderjem

Za izgradnjo referenčnih spektrov smo ponovno uporabili podatkovno usmerjeno metodo, ki lahko pokrije več dogajanja kot pa teoretični modeli. Odločili smo se za uporabo avtoenkoderja - posebne strukture nevronske mreže, ki s svojo sestavo vhodne podatke predela v dimenzijo mnogo manjšo od vhodne, nato pa jih po obratnem postopku razparkira in poskusi reproducirati vhodne podatke. Shema kodirnega dela strukture je predstavljena na Sliki 5.1. Tako reproduciran izhodni signal je ponavadi zglajen signal brez spektralnih posebnosti in posnema povprečnost spektrov, ki so podobni vhodnemu. Vse to so lastnosti, ki si jih želimo od algoritma za izdelavo primerjalnih normalnih spektrov.

Pred samo uporabo je bilo strukturo potrebno naučiti na naše podatke. Uporabili smo vse spektre, ki so imeli veljavne fizikalne parametre, ter izločili spektre že prej znanih posebnih tipov zvez [84], ki bi lahko onemogočali doseganje želenega učinka. Že majhen vnos posebnih tipov spektrov bi lahko povzročil, da se sistem nauči tudi njihovega izgleda in s tem reproducira tudi iskane značilke. Z zadostnim izločanjem nam je uspelo konstruirati sistem, ki tudi za posebne tipe spektrov vrne njim najboljši približek normalnega spektra, kar je prikazano na Slikah 5.3 in 5.4. Pridobljene vmesne skrčene značilke spektra se zelo dobro skladajo s fizikalnimi parametri zvezde (glejte Slike 5.5, 5.6 in 5.7), kar nakazuje, da se fizikalni model zvezdnega spektra in naš nenadzorovan podatkovno usmerjen model strinjata v najpomembnejših parametrih, ki pogojujejo izgled zvezdnega spektra.

7.5.2 Določanje emisijskih komponent

Po pridobitvi referenčnih spektrov za vsako zvezdo smo od njenega opazovanega spektra odšteli generiran referenčni spekter. V pridobljenem ostanku, ki poudarja

njune razlike, smo se osredotočili na kromosferske spektralne emisije črt Hα in Hβ, ter prepovedane prehode enkrat ioniziranih elementov [NII] in [SII], ki nam podajajo lastnosti razredčenega nebularnega plina v okolici zvezde. Vsak od elementov [NII] in [SII] v HERMES spektru izkazuje dve povezani emisijski črti. Za vodikovi črti smo določili naslednje parametre: ekvivalentno širino emisijskega dela, širino emisijske črte na 10% njene največje moči ter z delnim integriranjem (premik v rdeči in modri del proč od mirovne valovne dolžine vodikove črte) še asimetričnost črte, ki nakazuje na njen izvor. Manj izrazite emisije [NII] in [SII] smo analizirali s hkratnim prilagajanjem dveh povezanih Gaussovih krivulj na izražene črte posameznega elementa. Tako smo za vsak element določili njegovo število izraženih črt, skupno ekvivalentno širino ter razliko radialne hitrosti glede na opazovano zvezdo.

7.5.3 Rezultati in zaključki

Z opisano analizo smo med vsemi spektralnimi podatki odkrili 10.364 spektrov z močneje izraženimi emisijskimi črtami v Hα/Hβ območju in 4431 spektrov, ki dodatno vsebujejo še merljiv nebularni prispevek. Vse odkrite primere smo tudi narisali na zvezdno karto (glejte Sliki 5.16 in 5.17), kjer smo zaznali že vnaprej pričakovane korelacije pozicij zvezd z območji mladih zvezdnih kopic ter območji vidnega nebularnega medzvezdnega plina. Tekom analize in nadzora kvalitete rezultatov smo kot obstranski produkt izdelali še seznam dvojnih zvezd z izraženimi dvojnimi spektralnimi črtami v spektru ter označili potencialne nepravilnosti pri odštevanju ozadja tekom redukcije podatkov. Med vsemi opazovanji smo našli tudi 621 zvezd z dvema ali več opazovanji, pri katerih smo za vsaj enega potrdili prisotnost emisijskih črt. Večina ponovljenih opazovanj ne kaže spreminjanja oblike ali lokacije profila, preostali spreminjajoči objekti pa predstavljajo zanimiv podatkovni set za nadaljnje raziskave.

7.6 Soncu podobne večkratne zvezde

Izmed vseh opazovanih zvezd na nebu nam je Sonce najbolj poznano, saj njegova bližina in svetlost omogočata podrobne spektroskope, fotometrične, strukture in druge analize. Zaradi tega je Sonce uporabljeno kot referenčna zvezda za veliko fizikalnih meritev. Poznavanje večjega števila Soncu skoraj identičnih zvezd v velikih pregledih neba nam omogoča njihovo notranje umerjanje [317, 318] in medsebojno primerjavo. Zaradi tega smo se odločili, da tudi v pregledu GALAH poiščemo čim več takih zvezd, katerih spekter je čim bolj identičen Sončevemu.

7.6.1 Izbira Soncu najbolj podobnih zvezd

Izbiro zvezd smo pričeli z izdelavo Sončevega spektra, kot ga posnamemo s spektrografom HERMES. Tekom rednih opazovanj, za namen umerjanj, smo posneli tudi spektre neba ob sončevem vzhodu in zahodu, ki odražajo točen spekter Sonca. Z njihovim povprečenjem smo izdelali skoraj brezšumni referenčni spekter Sonca. Najbolj identične opazovane spektre smo izbrali tako, da smo v vsakem valovnem območju HERMES spektrografa izračunali razdaljo Canberra [332] med referenčnim in opazovanim spektrom. Tako smo dobili štiri neodvisne cenilke podobnosti za vsak

spekter. Izračunana podobnost je močno odvisna od razmerja med signalom in šumom opazovanega spektra. V primeru enoznačno izbranega pragu za izbiro bi tako pridobili le spektre z najmanj šuma, ostale pa zanemarili, čeprav bi morda dejansko bili podobnejši Soncu. V izogib problemu smo v izbiro vključili tudi moč šuma - primer postopka na Sliki 6.3. V končni seznam smo uvrstili 329 spektrov, ki so bili v vseh štirih valovnih pod-območjih razvrščeni med 7% najpodobnejših spektrov.

7.6.2 Določanje večkratnosti

Za izbran set zvezd bi pričakovali, da je njihov absoluten izsev zelo podoben oziroma identičen Sončevemu. Z združevanjem *Gaia* podatkov navidezne magnitude in oddaljenosti [170] smo na Sliki 6.7 ugotovili, da nekatere zvezde izsevajo tudi dva- in več-kratnik Sončeve svetlobe. Ker v samem spektru ni opaznih podvojenih spektralni črt, ki bi nakazovale na večkratnost objekta, smo predvidevali, da morajo biti ti sistemi sestavljeni iz skoraj identičnih zvezd na dokaj veliki medsebojni oddaljenosti z orbitalno hitrostjo, ki ne povzroči delitve spektralnih črt.

Simulacijo možnih kombinacij, ki bi nam podale opazovane vrednosti, smo razdelili na spektralni in orbitalni del. V spektralnem delu smo iz množice opazovalnih podatkov sestavili podatkovno usmerjena modela, ki sta neodvisno opisovala fotometrični in spektralni podpis enojne zvezde. S postopnim sestavljanjem naključnih kombinacij (metoda Monte Carlo markovske verige oziroma MCMC, [372]) dveh in treh različnih enojnih zvezd smo poskusili poustvariti opazovane fotometrične in spektralne podatke. Tako smo potrdili, da bi opazovane podatke s končnimi rezultati, obarvanimi na Sliki 6.13, lahko sestavili iz več zvezdnih komponent.

V orbitalnem delu analize smo želeli preveriti ali simulirani objekti tudi res lahko obstajajo znotraj opazovalnih omejitev. Pri tem smo predpostavili, da je vsak stabilen trojni hierarhičen sistem sestavljen iz dveh blizu krožečih objektov in enega oddaljenega. S simulacijo podobnosti spektrov, pri kateri je ena od komponent zamaknjena za določeno radialno hitrost, smo določili največjo možno hitrost, pri kateri se izmerjena podobnost ne spremeni preveč da sestavljen spekter ne bi bil več podoben Soncu. To hitrost smo uporabili za določitev najmanjše velikosti velike polosi orbite bližnjih objektov. Tak sistem je možen, saj je najmanjša velikost dosti manjša od mejne največje oddaljenosti zunanje zvezde. Ta je določena z mejno ločljivostjo satelita *Gaia*, pri kateri še lahko razločimo dva zvezdna izvora svetlobe.

Možnost orbitalnih dinamik smo preverili še z vključitvijo radialnih hitrosti zbranih iz pregledov *Gaia*, RAVE [96], GALAH in posameznih opazovanj zvezd, pridobljenih v Asiagu. Z združevanjem katalogov smo tako pridobili dodatno časovno dinamiko objekta, vendar za posamezno zvezdo nikoli nismo pridobili več kot treh časovnih točk. Tako majhno število onemogoča kompleksne analize, vendar nam vseeno podaja spodnjo mejo spremenljivosti hitrosti objekta. Med zvezdami nismo našli definitivnega spremenljivega izvora, saj so bile največje spremembe v rangu 5 $\mathrm{km\,s^{-1}}$in manj, kar je primerljivo z nedoločenostjo teh meritev.

7.6.3 Rezultati in zaključki

Uporabljena analiza identificiranih Sončevih dvojnikov temelji na podatkih oddaljenosti, ki se bodo s prihodnjimi podatkovnimi izidi še spreminjali in potencialno zelo spremenili rezultate. Kljub temu nam izsledki nakazujejo, da je pri uporabi

identičnih spektrov kljub temu potrebno biti zelo pozoren, saj so kljub njihovem izgledu lahko sestavljeni iz več skoraj identičnih zvezd na za opazovalca počasnih medsebojnih orbitah. V našem primeru so bili simulirani pogoji takšnih orbitalnih konfiguracij znotraj opazovalnih omejitev in posledično identificirani sistemi tudi fizikalno mogoči. Ker nas je zanimalo ali podobno velja tudi za bolj vroče in hladnejše zvezde na glavni veji, smo analizo večkratnost identičnih spektrov razširili še na njih. Rezultati na Sliki 6.23 nakazujejo že poznan trend [334] povečevanja razširjenosti večkratnih sistemov med bolj vročimi zvezdami.

7.7 Zaključki disertacije in prihodnje študije

Z napredkom opazovalnih metod in observatorijev se počasi spreminja tudi opazovalno delo astronomije. Očiten premik gre v smeri proč od dolgotrajnih analiz posameznih objektov proti masovnim pregledom neba in temu primernim obdelavam. Tega človek seveda ne more sam, zato razvija množico računalniških algoritmov, ki mu pomagajo pri tej nalogi. Med njimi trenutno najhitreje rastoči po uporabi in včasih tudi napačno uporabljeni, so algoritmi strojnega učenja, ki opravljajo naloge klasifikacije, grozdenja in regresije.

V disertaciji smo uporabili nekaj od teh orodij za raziskovanje različnih astronomskih podatkovnih setov, ki so bili primarno pridobljeni kot del pregledov neba GALAH in *Gaia*. Vsi pridobljeni rezultati so zainteresiranim brezplačno dostopni na spletnem mestu VizieR [1] in na straneh založnikov objavljenih znanstvenih člankov. S pridobitvijo dodatnih opazovanj, predvsem časovne komponente, objavljeni rezultati ponujajo še kopico možnosti za nadaljnje študije, ki bi se poglobile v fizikalna ozadja razpoznanih posebnosti. Takšne poglobljene raziskave se ponavadi izvajajo individualno za vsak objekt posebej in potrebujejo čim širši nabor opazovalnih podatkov.

Novih podatkovnih setov v bližnji prihodnosti ne bo zmanjkalo, saj smo ravno v zadnjih letih priča končnim fazam priprav in zagonu več masovnih fotometričnih in spektroskopskih pregledov neba. Do njihovega začetka pa nestrpno pričakujemo še tretji in za tem posledično končni izid podatkov *Gaia*, ki še lahko spremenijo in dopolnijo naše vedenje o sestavi in razvoju Galaksije.

[1] `http://vizier.u-strasbg.fr/`

List of publications related to this doctoral thesis

Scientific papers that were published by the author of this doctoral thesis and are related to it:

1. **K. Čotar**, T. Zwitter, G. Traven, J. Kos, M. Asplund, J. Bland-Hawthorn, S. Buder, V. D'Orazi, G. M. de Silva, J. Lin, S. L. Martell, S. Sharma, J. D. Simpson, D. B. Zucker, J. Horner, G. F. Lewis, T. Nordlander, Y.-S. Ting, R. A. Wittenmyer and Galah Collaboration, The GALAH survey: unresolved triple Sun-like stars discovered by the Gaia mission, Monthly notices of the Royal Astronomical Society, 487, 2474 (2019).

2. **K. Čotar**, T. Zwitter, J. Kos, U. Munari, S. L. Martell, M. Asplund, J. Bland-Hawthorn, S. Buder, G. M. de Silva, K. C. Freeman, S. Sharma, B. Anguiano, D. Carollo, J. Horner, G. F. Lewis, D. M. Nataf, T. Nordlander, D. Stello, Y.-S. Ting, C. Tinney, G. Traven, R. A. Wittenmyer and Galah Collaboration, The GALAH survey: a catalogue of carbon-enhanced stars and CEMP candidates, Monthly notices of the Royal Astronomical Society, 483, 3196 (2019).

Scientific papers that are currently prepared to be published by the author of this doctoral thesis and are related to it:

1. **K. Čotar**, T. Zwitter, G. Traven, J. Bland-Hawthorn, S. Buder, M. R. Hayden, J. Kos, G. F. Lewis, S. L. Martell, T. Nordland er, D. Stello, J. Horner, Y.-S. Ting and M. Žerjal, The GALAH survey: Characterization of emission-line stars with spectral modelling using autoencoders, in preparation.

Scientific papers that were co-authored by the author of this doctoral thesis and were heavily used during the production of this thesis and above listed papers:

1. S. Buder, M. Asplund, L. Duong, J. Kos, K. Lind, M. K. Ness, S. Sharma, J. Bland-Hawthorn, A. R. Casey, G. M. De Silva, V. D'Orazi, K. C. Freeman, G. F. Lewis, J. Lin, S. L. Martell, K. J. Schlesinger, J. D. Simpson, D. B. Zucker, T. Zwitter, A. M. Amarsi, B. Anguiano, D. Carollo, L. Casagrande, **K. Čotar**, P. L. Cottrell, G. Da Costa, X. D. Gao, M. R. Hayden, J. Horner, M. J. Ireland, P. R. Kafle, U. Munari, D. M. Nataf, T. Nordlander, D. Stello, Y.-S. Ting, G. Traven, F. Watson, R. A. Wittenmyer, R. F. G. Wyse, D. Yong, J. C. Zinn and M. Žerjal, The GALAH Survey: second data release, Monthly notices of the Royal Astronomical Society, 478, 4513 (2018).