Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **500,** 4849–4865 (2021) doi:10.1093/mnras/staa2524

The GALAH survey: characterization of emission-line stars with spectral modelling using autoencoders

Klemen Čotar [1]★ Tomaž Zwitter [1] Gregor Traven,[2] Joss Bland-Hawthorn [3,4] Sven Buder [3,5,6] Michael R. Hayden,[3,4] Janez Kos,[1] Geraint F. Lewis [4] Sarah L. Martell [3,7] Thomas Nordlander [3,5] Dennis Stello,[7] Jonathan Horner,[8] Yuan-Sen Ting[5,9,10,11] and Maruša Žerjal [5]
the GALAH collaboration

[1] *Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia*
[2] *Lund Observatory, Department of Astronomy and Theoretical Physics, Box 43, SE-221 00 Lund, Sweden*
[3] *ARC Centre of Excellence for All Sky Astrophysics in Three Dimensions (ASTRO-3D), Canberra, ACT 2611, Australia*
[4] *Sydney Institute for Astronomy, School of Physics, The University of Sydney, Sydney, NSW 2006, Australia*
[5] *Research School of Astronomy and Astrophysics, The Australian National University, Canberra, ACT 2611, Australia*
[6] *Max Planck Institute for Astronomy (MPIA), Koenigstuhl 17, 69117 Heidelberg, Germany*
[7] *School of Physics, University of New South Wales, Sydney, NSW 2052, Australia*
[8] *Centre for Astrophysics, University of Southern Queensland, Toowoomba, QLD 4350, Australia*
[9] *Institute for Advanced Study, Princeton, NJ 08540, USA*
[10] *Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*
[11] *Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101, USA*

Accepted 2020 August 12. Received 2020 August 7; in original form 2020 May 22

ABSTRACT

We present a neural network autoencoder structure that is able to extract essential latent spectral features from observed spectra and then reconstruct a spectrum from those features. Because of the training with a set of unpeculiar spectra, the network is able to reproduce a spectrum of high signal-to-noise ratio that does not show any spectral peculiarities, even if they are present in an observed spectrum. Spectra generated in this manner were used to identify various emission features among spectra acquired by multiple surveys using the HERMES spectrograph at the Anglo-Australian telescope. Emission features were identified by a direct comparison of the observed and generated spectra. Using the described comparison procedure, we discovered 10 364 candidate spectra with varying intensities (from partially filled-in to well above the continuum) of the Hα/Hβ emission component, produced by different physical mechanisms. A fraction of these spectra belong to the repeated observation that shows temporal variability in their emission profile. Among the emission spectra, we find objects that feature contributions from a nearby rarefied gas (identified through the emission of [N II] and [S II] lines) that was identified in 4004 spectra, which were not all identified as having Hα emission. The positions of identified emission-line objects coincide with multiple known regions that harbour young stars. Similarly, detected nebular emission spectra coincide with visually prominent nebular clouds observable in the red all-sky photographic composites.

Key words: line: profiles – methods: data analysis – catalogues – stars: activity – stars: emission line, Be – stars: peculiar.

1 INTRODUCTION

The identification of peculiar stars whose spectra contain emission lines is of interest in many areas of stellar research. The spectral complexity of such stars provides information about the physical processes occurring in and around stars. The emission features in stellar spectra might adversely impact the quality of the stellar parameters and abundances determined by automatic data analysis pipelines that are configured to produce the best results for most common stellar types. Examples of where these features might compromise spectroscopic measurements when we assume that a star is not peculiar include the determination of effective temperature (Cayrel et al. 2011; Amarsi et al. 2018; Giribaldi et al. 2019), the computation of stellar mass (Ness et al. 2016; Bergemann et al. 2016), and the effects of self-broadening on line wing formation (Barklem, Piskunov & O'Mara 2000; Allard et al. 2008). Highly accurate measurements of the hydrogen absorption profiles are needed in these cases. Any deviations in the line shapes from model predictions would produce misleading results. It is therefore important to know if the investigated line is modified by additional, unmodelled physical processes or by spectral reduction processes. Stars with evident emission lines populate a wide variety of regions on the Hertzsprung–Russell (HR) diagram. Because of possible overlaps between different stellar types, detailed photometric observations (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 example

★ E-mail: klemen.cotar@fmf.uni-lj.si

of such work is Munari et al. (2019), who performed a detailed multiband 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 semiregular asymptotic giant branch cool giant, which was its previous classification. In a similar way, Lancaster et al. (2020) 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, these authors concluded that the object was most likely a symbiotic binary star system whose emission was linked to a long-duration, low-luminosity nova phase.

Numerous physical processes that can contribute to the complex shapes of the Hα emission profile are discussed by Reipurth, Pedrosa & Lago (1996), Jones, Tycner & Smith (2011), Silaj et al. (2014) and Ignace et al. (2018), who compare observations with expected physical models. Following the classification scheme introduced by Kogure & Leung (2007), stellar emission lines 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 (White & Basri 2003; Natta et al. 2004). The procedure involves measuring simple indices (such as the equivalent width and broadening velocity) of the emission lines in the stars' spectrum.

In recent years, multiple dedicated photometric and spectroscopic surveys (e.g. Witham et al. 2008; Mathew, Subramaniam & Bhatt 2008; Matijevič et al. 2012; Nakano et al. 2012; Drew et al. 2014; Aret, Kraus & Šlechta 2016; Nikoghosyan, Vardanyan & Khachatryan 2016), and exploratory spectral classifications of large unbiased all-sky spectroscopic observational data sets (e.g. Kohoutek & Wehmeyer 1999; Reid & Parker 2012; Traven et al. 2015; Nikoghosyan et al. 2016; Hou et al. 2016; Traven et al. 2017) 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, it is possible that emission features of forbidden lines (the most commonly studied of which are the [N II] and [S II] lines) could be observed in its spectrum, providing an insight into the temperature, density, intrinsic movement and structure of the surrounding interstellar environment (Bohuski 1973; Raju et al. 1993; Escalante & Morisset 2005; Damiani et al. 2016, 2017).

Focusing on spectroscopic data, procedures for the detection of emission lines can be roughly separated into two categories: simpler procedures, searching for obvious emitters above the global continuum (Traven et al. 2015; Nikoghosyan et al. 2016; Hou et al. 2016; Nikoghosyan et al. 2016); and more complex procedures, where the observed spectrum is compared with the expected stellar spectrum of a normal star (Žerjal et al. 2013). The reference spectra in the latter category 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 latter, it is not necessary 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 there is a possibility that the strongest lines in the acquired spectrum are populated by an emission feature, which is the case for *Gaia* and RAVE spectra (Žerjal et al. 2013). In light of the future publication of *Gaia* RVS spectra as part of *Gaia* DR3 for several millions of stars, it is thus important to develop tools to identify emission-line stars. This is what we aim to do in this study via GALAH spectra.

The paper is structured as follows. We begin in Section 2 with a description of the spectroscopic data used. In Section 3 we explain our analysis pipeline, whose main components are the generation of reference spectra (Section 3.1) and the identification of multiple emission features (Sections 3.3 and 3.4). The temporal variability of detected emissions is analysed in Section 4. The results are summarized and discussed in Section 5.

2 DATA

The spectroscopic data used in this study were taken from the main GALactic Archaeology with HERMES (GALAH) survey (De Silva et al. 2015), the K2-HERMES survey (Wittenmyer et al. 2018), the TESS-HERMES survey (Sharma et al. 2018), the dedicated HERMES open clusters survey (De Silva et al. in preparation) and the HERMES Orion star-forming region survey (Kos et al. in preparation). All of the spectra were acquired by the High Efficiency and Resolution Multi-Element Spectrograph (HERMES, Barden et al. 2010; Sheinis et al. 2015), a multifibre spectrograph mounted on the 3.9-m Anglo-Australian Telescope (AAT) at the Siding Spring Observatory, Australia. The spectrograph has a resolving power of $R \sim 28\,000$ across four wavelength ranges (4713–4903, 5648–5873, 6478–6737 and 7585–7887 Å), also referred to as the blue, green, red and infrared spectral arms. Of the four, we used only the first (blue) and the third (red) in our study, as they cover the wavelength regions where interesting Balmer and forbidden emission lines can be seen and detected.

The combined data set consists of 669 845 successfully reduced stellar spectra, of which a small fraction are repeated observations. All acquired spectra were homogeneously reduced to one-dimensional spectra, continuum-normalized, shifted to the barycentric system, and finally shifted to the stellar reference frame to determine their radial velocity (a detailed description of the algorithm used can be found in Kos et al. 2017). The reduction therefore aligns the absorption lines of all spectra. All surveys combined include more spectra than the main GALAH survey alone, but at the same time break the simple rule adhered to in that main survey. It uses a simple magnitude-limited selection function (Sharma et al. in preparation), which is desirable for population studies and comparison with synthetic galactic models. The exact selection function is not important in our case, as we are not performing any population studies but are only trying to find as many emission-line objects as possible.

The stellar atmospheric parameters and individual abundances derived from our normalized spectra were analysed with the adaptation of the Spectroscopy Made Easy (SME, Valenti & Piskunov 1996; Piskunov & Valenti 2017) software that is described in depth by Buder et al. (in preparation) as part of the latest GALAH data release (DR3), which includes fully reduced spectra and derived parameters.

Our algorithm for the detection of emission-line spectra, described in detail below, uses normalized GALAH spectra that had already been corrected for telluric absorption and had sky spectral emission contributions removed. The correct sky removal (described in more detail in Section 3.5) is essential, as one of the telluric lines falls inside the range of the Hα line.

3 DETECTION AND CHARACTERIZATION

The first attempts to discover Hα/Hβ emission spectra in GALAH survey observations were made by Traven et al. (2017), who used the unsupervised dimensionality reduction technique t-SNE (van der Maaten 2013) 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. (2017) 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. During the procedure, the operator manually selects individual clump of spectra. The application returns a combined plot of all spectra in the selected clump and their median. If there are no obvious deviations from the median spectrum or all spectra look normal to the operator, the clump is classified as normal. This would hold true for weak chromospheric emissions, as they would probably not be visually distinguished in a median spectrum that is not additionally compared with some external spectral model. To broaden the range of detectability and include spectra with such marginal levels of emission, 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 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.

3.1 Spectral modelling using autoencoders

A reference or a synthetic spectrum of a normal, emission-free star can be produced by a multitude of physics-based computational stellar models (Kurucz 1993; Munari et al. 2005; de Laverny 2012) or supervised generative data-driven approaches (Ness et al. 2015; Ting et al. 2019), whose common weakness is the need for prior knowledge of at least approximate stellar parameters of the 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 various 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 the available training data set, we chose to use an autoencoder type of artificial neural network (ANN) that is rarely used to analyse astronomical data. Its current use ranges from data de-noising (Qin, Lin & Wang 2017; Shen et al. 2019; Li et al. 2019) to unsupervised feature extraction and feature-based classification (Yang & Li 2015; Li, Pan & Duan 2017; Pan & Li 2017; Karmakar, Mishra & Tej 2018; Cheng et al. 2019; Ma et al. 2019; Ralph et al. 2019).

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 per cent, 50 per cent, 25 per cent and 10 per cent 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 Fig. 1. In the middle feature layer,

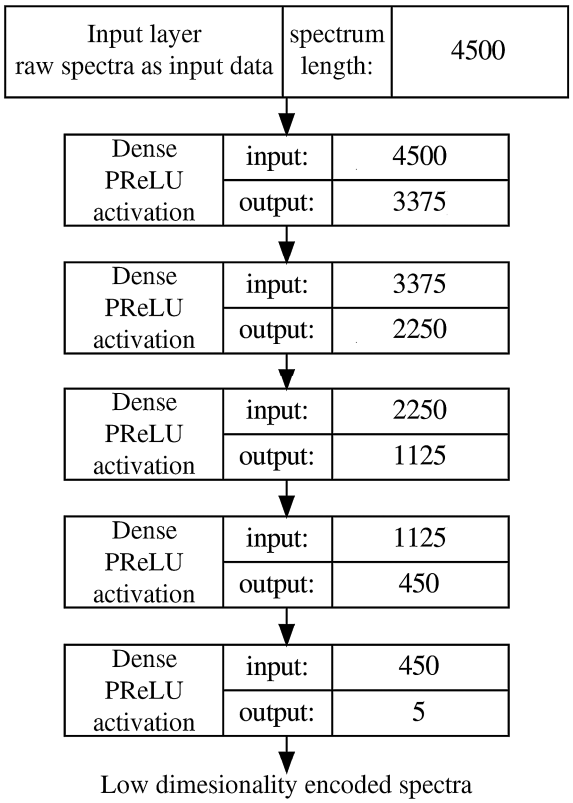

Figure 1. Visual representation of the encoder part of the autoencoder structure used for the red spectral arm. After the input spectra are encoded, they are passed through the same inverted architecture to produce modelled low-noise spectra. The value in the right-most column indicates the number of input and output connections to neighbouring 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.

the autoencoder structure reduces to only five relevant extracted features. Selecting a higher number of extracted features would 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 the extraction of a few relevant spectral features. Furthermore, because of the low number of extracted features, our decoded output spectrum has a much reduced noise compared with an input spectrum.

A visual representation of the described architecture is shown in Fig. 1. The shape of the decoding structure of the autoencoder is the same, except in reverse order. The Parametric Rectified Linear Unit (PReLU, He et al. 2015) activation function defined as

$$f(x) = \begin{cases} x, & \text{if} x > 0 \\ ax, & \text{if} x \leq 0 \end{cases} \quad (1)$$

is used for all nodes of the network, with the exception of the final output layer, which 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 (1) is optimized 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 the red arm, respectively) whose coverage is slightly wider than the range of an individual HERMES arm to account for variations in wavelength span because of the radial velocity. Observations that did not completely fill the selected range were padded with a continuum value of 1. To be classified as normal, spectra must satisfy the following selection rules: the 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. 2017) and have valid spectral parameters (`flag_sp` <16 in Buder et al. in preparation). 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 parameters. Spectra with $0<$ `flag_sp` <16 include objects with a bad astrometric solution, unreliable broadening, and low SNR, which are still useful for our training process. From Traven et al. (2017), Buder et al. (2018) and Čotar et al. (2019), we know that some GALAH spectra display a 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, which are actually treated as normal spectra in our case. Even such a rigorous filtering approach can miss some strange spectra.

After applying these quality cuts, we were left with 482 900 spectra, of which the last 10 per cent were used as an independent validation set during the training process. Before the training, normalized spectra were inverted (1 − normalized 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 optimization algorithm (Kingma & Ba 2014) for 350 epochs. At every epoch, all training spectra were divided into multiple batches of 40 000 spectra, whose content is randomized at every epoch. A batch is a subset of data that is independently used during a training process. Such splitting and randomization of training spectra into batches decreases the probability of model over-fitting. To enable the selection of the best network model, a model was saved after the end of every training epoch.

The loss score minimized by the Adam optimizer, shown in Fig. 2, was computed as a mean absolute error (MAE) between the input observed and decoded spectra, defined as

$$loss_{\mathrm{MAE}} = \frac{1}{N n_\lambda} \sum_{n=1}^{N} \sum_{i=1}^{n_\lambda} \left| f_{\mathrm{ae},n,i} - f_{\mathrm{obs},n,i} \right|, \tag{2}$$

where N represents 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 normalized observed spectrum. Such a loss function gives a 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 are less affected by any 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 (with greater prediction loss) and rejects the need for a longer model training.

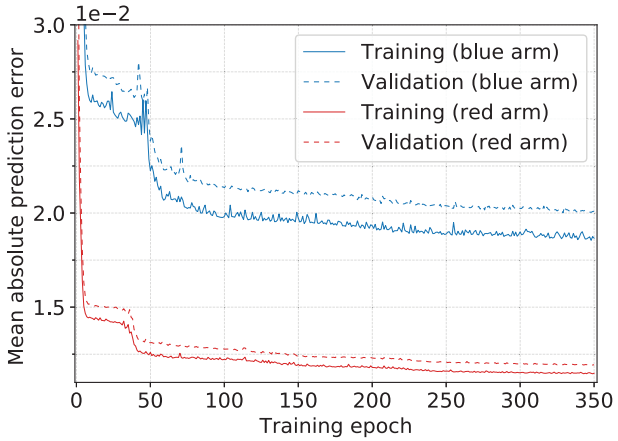

Figure 2. Prediction accuracy of the blue- and red-arm autoencoders at different training epochs. The prediction error is computed as the sum of all absolute differences between the input and output data sets (see equation 2). Shown are training (solid line) and validation (dashed line) curves, which do not show any strong model over-fitting on the training set. The curves indicate that both autoencoders learned in a similar way, because the same optimizer was used. The blue-arm model has a slightly higher loss and shows slower learning, because of the greater spectral complexity and lower signal-to-noise ratio in that wavelength region.

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 high-SNR reference spectra. Four such spectra are shown in Fig. 3.

3.2 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 according to unflagged physical parameters published in GALAH DR3.. Latent-feature scatter plots, colour-coded according todifferent combinations of stellar parameters, are presented in Fig. 4 (with T_{eff} and $\log g$ for the red arm) as well as in Figs B2 (with T_{eff} and $\log g$ for the blue arm) and B3 (with T_{eff} and [Fe/H] for the blue arm).

As expected, all plots show colour gradients induced by the changing value of the physical parameter under investigation. This confirms that the derived stellar physical parameters are spectroscopically meaningful and have the strongest influence on the appearance of acquired spectra. Approximate physical parameters of previously unanalysed or peculiar spectra can therefore be acquired by averaging the parameter values of their neighbourhood in the latent space. Similar procedures for parameter estimation have already been successfully explored by Yang & Li (2015), Pan & Li (2017) and Li et al. (2017).

3.3 Hα and Hβ emission characterization

The detection of emission components in spectra is based on the spectral difference f_{diff}, computed as

$$f_{\mathrm{diff}} = f_{\mathrm{obs}} - f_{\mathrm{ref}}, \tag{3}$$

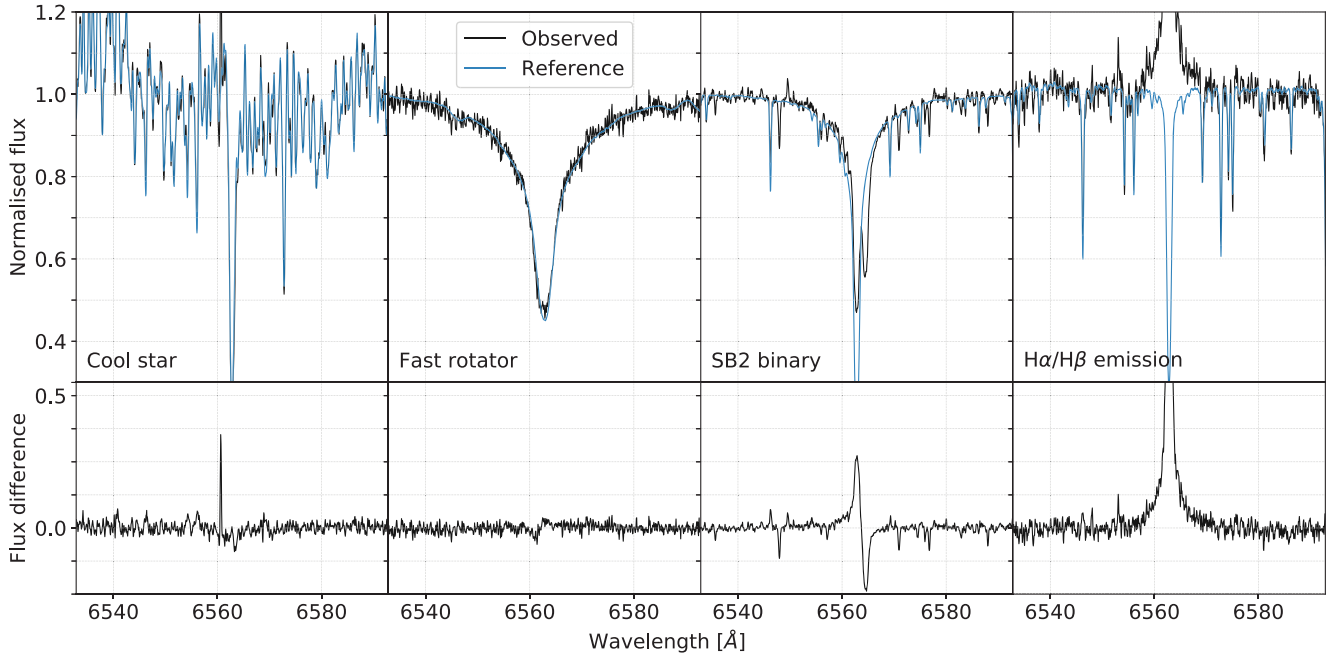

Figure 3. The diversity of spectra that must be processed by our reference spectrum generation scheme. Panels in the first row show spectra of the following normal and peculiar stars: cool, hot fast-rotating, spectroscopic binary, and Hα/Hβ emission. The examples show that the autoencoder network did reproduce the observed shapes of the normal spectra (first two) but not the peculiar spectra (last two), as desired for the reference spectrum generator. The original spectra are shown in black and reconstructed in blue. The bottom panels represent the flux difference between the observed and reference spectra.

where $f_{\rm obs}$ and $f_{\rm ref}$ are the observed spectrum and the generated reference spectrum, respectively. The results for a computed difference $f_{\rm diff}$ for an emission spectrum are shown in the top panel of Fig. 5. Ideally, this computation would enhance only the mismatch between the two spectra, with the inclusion of spectral noise, if both represent a star with the same stellar physical parameters. During the initial processing, we found that some observed spectra have slight normalization problems, and therefore we re-normalized all spectra prior to the difference computation. As the targeted reference spectrum $f_{\rm ref}$ is known and has a continuum level close to the median value of similar stars in the training set, we first compute the spectral ratio $f_{\rm div}$, defined as

$$f_{\rm div} = \frac{f_{\rm obs}}{f_{\rm ref}}. \quad (4)$$

The resulting ratio can be viewed as a proxy for a renormalization curve that would bring $f_{\rm obs}$ to the same continuum level as $f_{\rm ref}$, but would at the same time cancel out any spectral differences between them. To avoid continuum differences, we fitted $f_{\rm div}$ with a third-degree polynomial with a symmetrical 2-sigma clipping, run for five iterations. We used the polynomial fit to renormalize $f_{\rm obs}$. The renormalization brought both spectra to the same level, which decreased the probability of false detections, as their flux separation could be counted as the emission-line flux.

To obtain the first identification of an emission feature, 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 Fig. 5) is wide enough to encompass the emission profiles of all spectra, with the exception of a few that 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 3.5). The correlation between measurements of the two EWs is shown in Fig. 6, from which it is evident that the Hβ emission feature is not as strong as the Hα feature, but is comparable to it.

The EW integrator has no knowledge about the local continuum, and therefore its area represents a mismatch between observed and modeled spectra that could be the result of a partially filled-in Balmer line or a strong emission feature that stands far above the local continuum. In this paper, we make no formal distinction between partially filled-in and strong emissions, and mark all of them as emission-line spectra.

Alongside the EWs of the residual components (EW(Hα) and EW(Hβ)), we also measured two additional properties of these lines, which provide some insight into the physics of emission sources. The broadening velocity of a line is described by its width at 10 per cent of the line peak (W10 per cent(Hα) and W10 per cent(Hβ)) expressed in kilometres per second. 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 its strength drops below 10 per cent of the peak flux value. The broadening velocity is defined as the width between those two limiting cuts. As the computation is done for every object in an unsupervised way, the results are meaningful only for 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 per cent(Hα), the mass accretion could be estimated if emission is of a chromospheric origin.

The second index measured in the $f_{\rm diff}$ spectrum, which roughly describes the shape and location of an emission feature, is the asymmetry index, defined as

$$Asymmetry = \frac{|EW_{\rm red}| - |EW_{\rm blue}|}{|EW_{\rm red}| + |EW_{\rm blue}|}, \quad (5)$$

where $|EW_x|$ represents the equivalent width of the absolute difference $|f_{\rm diff}|$ on the red and blue sides of the central wavelength of the investigated Balmer line. By this definition, a line that is, as a

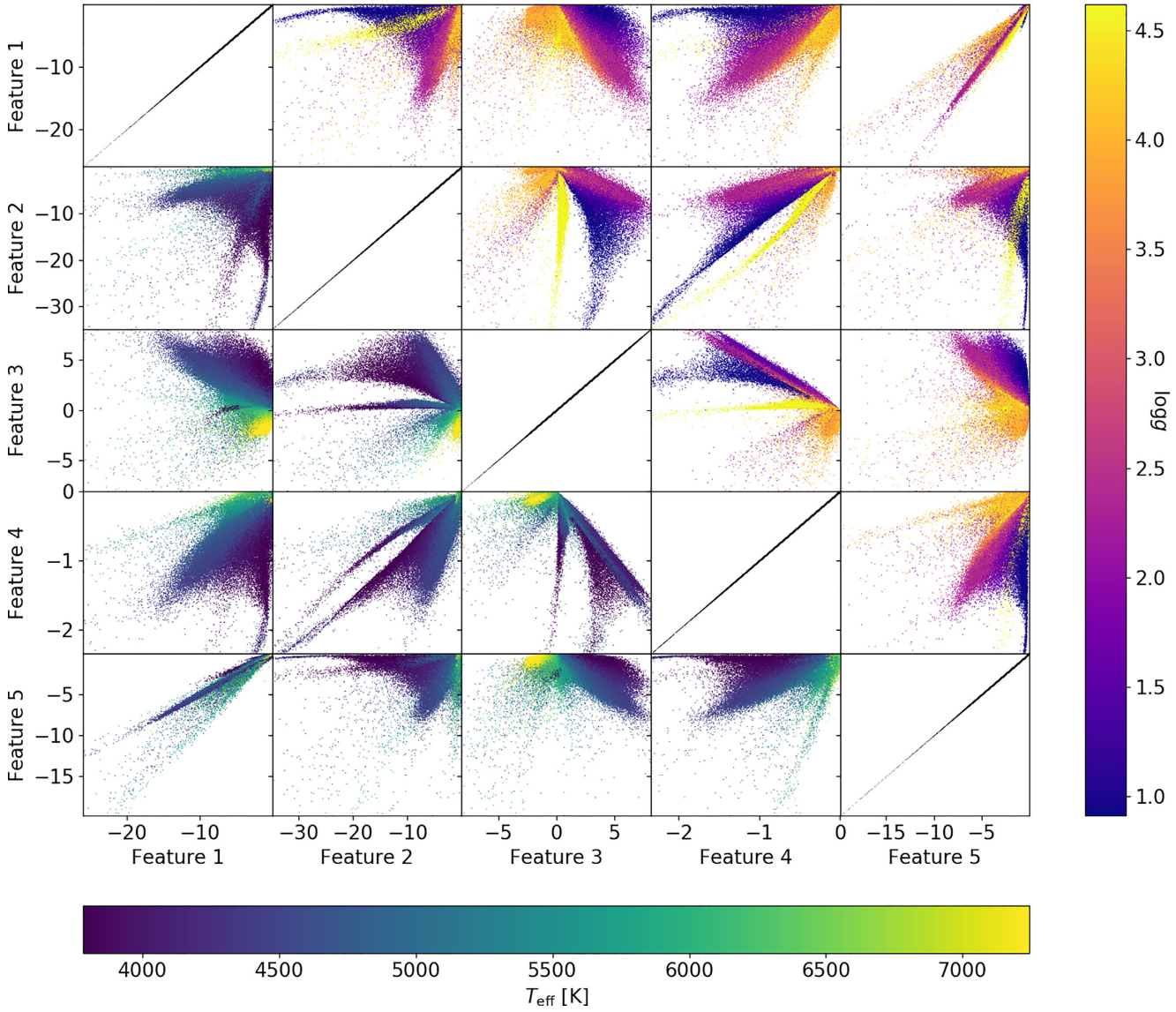

Figure 4. Correlation between extracted latent features and physical parameters. Scatter plots between different features are coloured according to the GALAH physical parameters of the original spectra. Points in the lower triangle are coloured according to their $T_{\rm eff}$ and those in the upper triangle according to their $\log g$. Associated colour mappings are given below the figure (for the lower triangle) and on its right side (for the upper triangle). The results presented are for the red-arm autoencoder.

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 distribution of the asymmetry index values for the most prominent and unflagged (see Section 3.8) emitters is shown in Fig. 7, 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 Fig. 7, where its index is scattered 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 centred component with an asymmetry index close to 0. Everything outside the central region in Fig. 7, defined by a circle with a radius of 0.25, could be thought to be of extra-stellar origin, as the lines are not perfectly aligned. The used thresholding radius value of 0.25 was defined by observing Fig. 7. Radius was selected in a such way as to encircle the main over-density of almost-symmetric emission profiles.

3.4 Detection of nebular contributions

Owing to the multiple possible origins of H emission lines (Kogure & Leung 2007), we also attempted to detect the extra-stellar nebular contributions of nearby rarefied gas. The presence of this gas 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 [N II] (6548.03 and 6583.41 Å) and [S II] (6716.47 and 6730.85 Å). Because they usually have 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 $f_{\rm diff}$). In order to automatically detect the emission strength and position of both doublets, we independently fitted two Gaussian functions with the

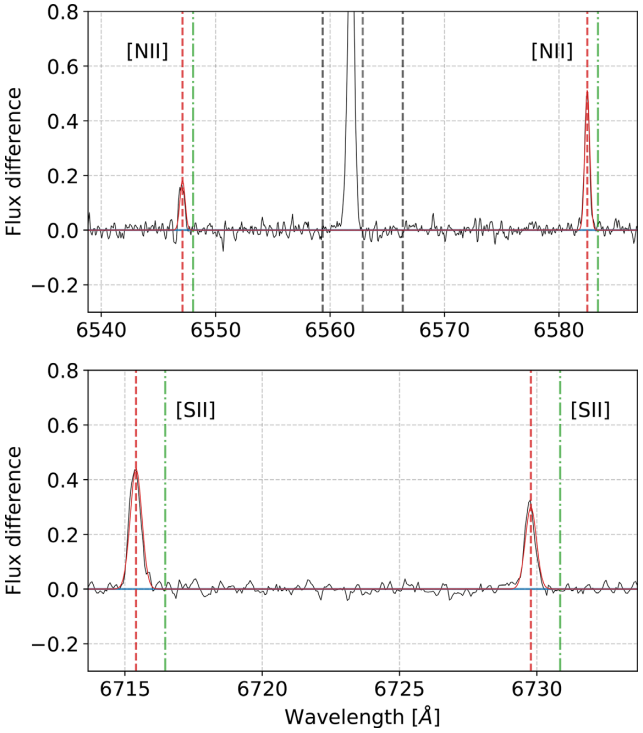

Figure 5. The two panels show two different wavelength regions of $f_{\rm diff}$ for the same star. The top panel is focused on the Hα and [N II] nebular lines, while the bottom one focuses on [S II] lines. Rest wavelengths of both nebular doublets are given by the green dash–dotted vertical lines. Their fitted locations, affected by the movement of a gas cloud, are given by the red dashed vertical lines. The EW(Hα) integration range is bounded by the central black dashed vertical lines in the top panel.

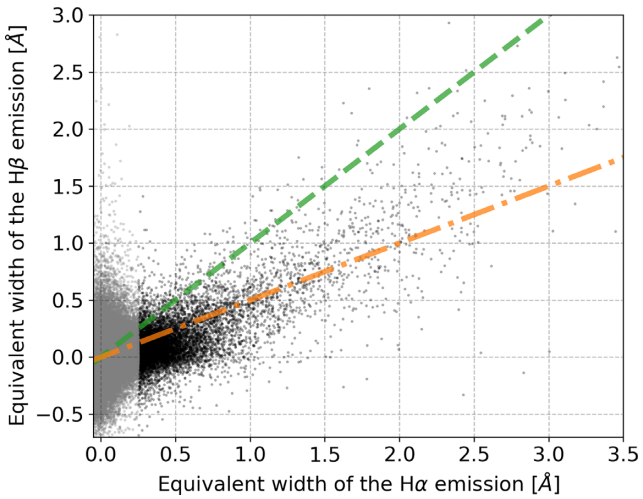

Figure 6. Correlation between EWs of the Hα and Hβ emission components for our set of detected stars (defined as having `Ha_EW` > 0.25 Å) as black dots. The remaining set of objects is shown with grey dots. All flagged objects and possible spectroscopic binaries are removed for this plot. The green dashed linear line represents the one-to-one relationship, and the orange dash–dotted line indicates cases where the EW of the Hβ line is half that of the Hα line.

same radial velocity shift for each element to $f_{\rm diff}$. Because the contributing medium is not necessarily physically related to the observed object, its radial velocity could be different, and therefore it was treated as a free parameter in our fit. Two independent

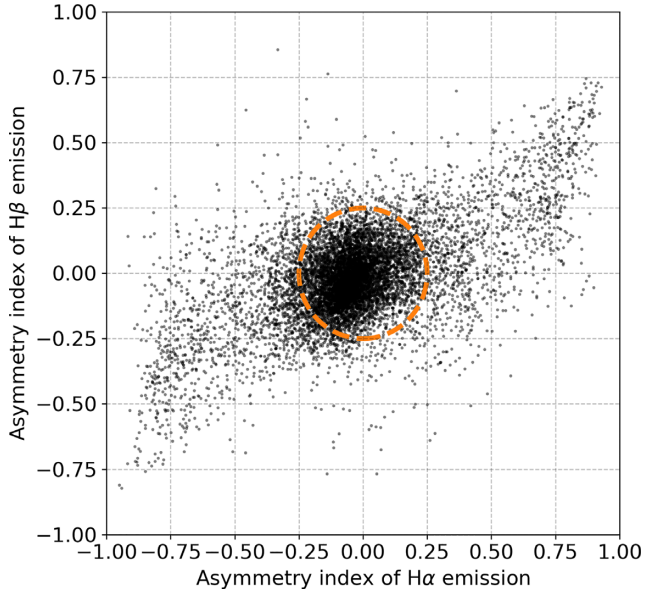

Figure 7. Asymmetry index of objects with prominent emission lines in the integration range around the investigated hydrogen Balmer lines. Objects with an index inside the orange dashed circle are considered to have a symmetric emission contribution, which can be attributed to chromospheric activity. The central circular region has a radius of asymmetric index 0.25.

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 km s^{-1} on the velocity difference. Some of the discarded outliers might be correct detections, because a few spectra show two or more peaks for each nebular line, which might point to a contribution from multiple clouds with different radial velocities. Such cases are not fully accounted for by the fitting algorithm, which identifies only the strongest emission.

In the absence of additional fitting constraints, the routine might also find two noise peaks and lock on to them. Therefore, we put an arbitrarily selected detection threshold (0.05 of relative flux) on the minimum amplitude of the fitted forbidden lines to be counted as detected. The result of this fitting and analysis procedure is a number of successfully detected peaks per element and their combined equivalent widths (EW([N II]) and EW([S II])), reported in the final published table (see Table A1). In order to filter out some possible mis-detections, we counted a spectrum as having nebular lines when at least three nebular lines above the threshold were detected. The correlations for measured radial velocities and EWs of identified objects with nebular emission are given in Figs 8 and 9 respectively.

The radial velocities of the two doublets shown in Fig. 9 give a first impression that the gas dynamics of the elements in all observed clouds are 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. (2016, 2017), who attributed it to the uncertainties in their adopted line wavelengths, which are slightly different from ours (by less than 0.05 Å), causing the velocity points to be located either above or below the identity line in Fig. 9. As this wavelength uncertainty could not explain the observed shift, we looked into the exact wavelengths of absorption lines in the redward part of the red HERMES arm. When compared with the external model spectrum, we clearly saw a slight mismatch among them. Inspection by eye revealed that the central positions of lines at ~6720 Å are shifted by approximately –0.04 Å. The combination of

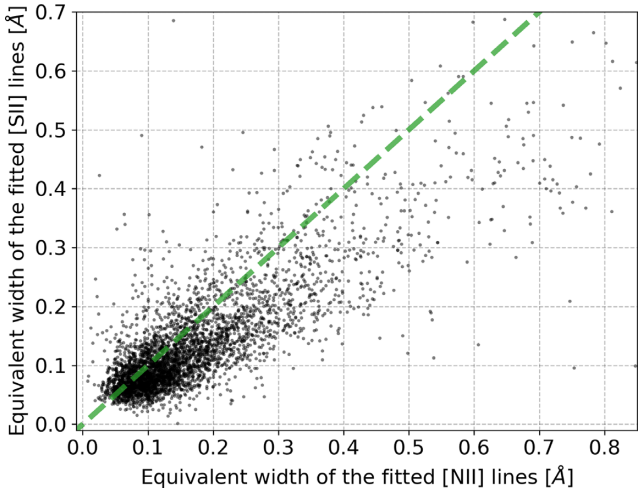

Figure 8. Correlation between the strengths of the nebular contributions from the two elements. Only cases with a small difference in the determined radial velocities, as shown in Fig. 9, are plotted.

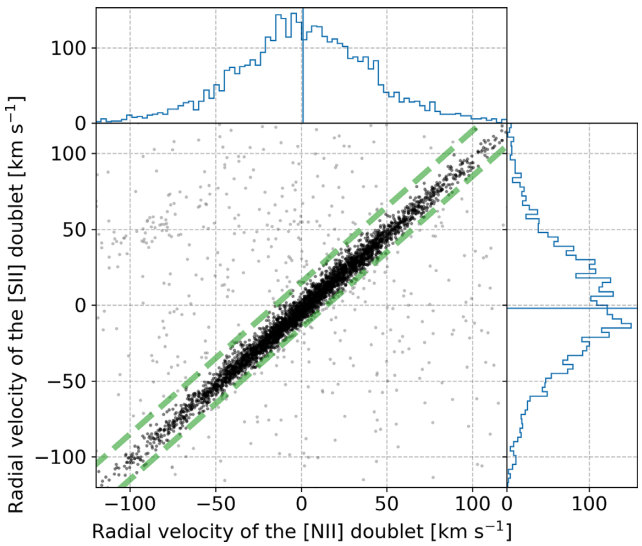

Figure 9. Correlation between the radial velocities of both assessed nebular contributions that are observable in the red arm of the HERMES spectrum. Only cases with at least three detected forbidden lines are shown. The grey dots were flagged and discarded from the final listas their absolute difference between velocities is more than 15 km s^{-1}. The limiting thresholds are depicted by the dashed straight lines. Plotted velocities are measured in the stellar rest-frame and therefore grouped towards zero velocity, meaning that they are moving together with the star. The median velocities of both doublets are shown on the histograms as solid lines. Their absolute difference is 2.98 km s^{-1}.

the two effects (0.09 Å equals 4 km s^{-1} at 6720 Å) may explain the velocity offset for the [S II] lines. The wavelength mismatch among observed and model spectra was not identified in the blueward part of the red arm.

In addition, the plot in Fig. 9 reveals that the majority of gas clouds have a different radial velocity from stars behind or inside the cloud.

As we are working with fully reduced normalized spectra, with the inclusion of sky background removal, the detection procedure would, in the case of an ideal background removal, not detect emission

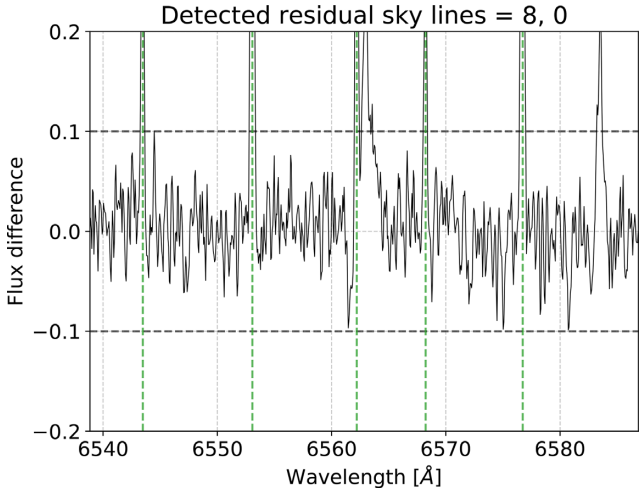

Figure 10. Sky emission lines are most evident after spectra subtraction in $f_{\rm diff}$. Green vertical dashed lines represent the expected locations of emission lines in the rest-frame of an observed star. The middle sky line in this plot falls inside the actual Hα emission feature and changes its shape from single- to double-peaked, and consequently modifies the measured equivalent width. Upper and lower thresholds for detection are given by the horizontal dashed lines. The number of detected under- and over-corrected sky lines, in this order, is given above the plot.

coming from nebular clouds. As the measured flux of the nebular contribution is very unlikely the same for physically separated object and sky fibres (see next Section 3.5 and Kos et al. 2017), 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 Fig. 8. The figure reveals a high degree of correlation, where on average [S II] lines have a lower strength than [N II] lines.

3.5 Identification of sky emission lines

Assigning a limited and relatively low number of HERMES fibres to monitor the sky in hopefully star- and galaxy-free regions imposes limitations on the quality of the sky background removal in the GALAH reduction pipeline (Kos et al. 2017). 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, the 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 the range of the analysed stellar line. For us, the most problematic sky emission line, which can alter the shape of the Hα profile, is located at 6562.7598 Å (our list of sky emission lines was taken from Hanuschik 2003). As this line can become blended with a real emission feature of the Hα or simulate its presence, we try to estimate the impact of the sky residual in the spectrum from multiple nearby emission lines. First, we select only the strongest sky lines (with parameter `Flux` ≥ 0.9 in Hanuschik 2003) and shift their reference wavelength into a stellar rest-frame. After that, we use a simple thresholding (see Fig. 10) to

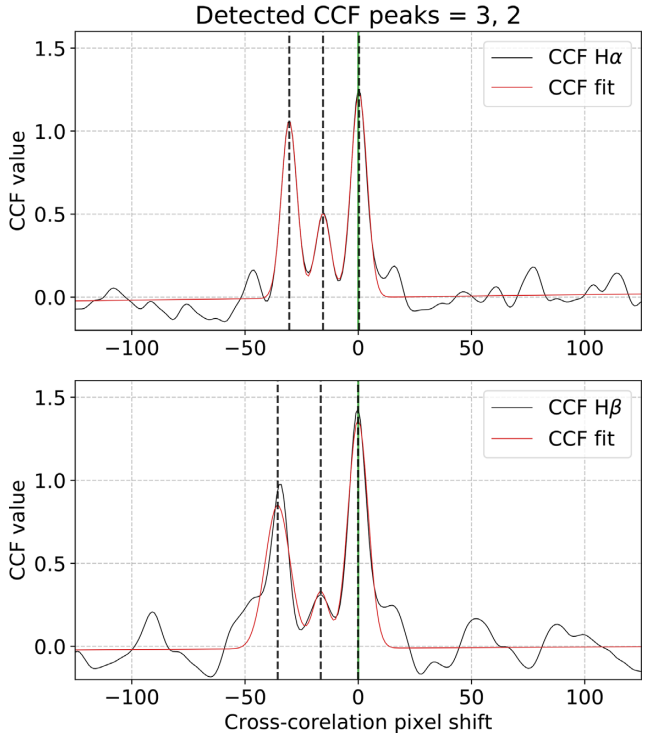

Figure 11. Detection of a spectroscopic binary candidate by cross-correlating an observed spectrum with its reference spectrum. Three Gaussian functions that are fitted to the resulting CCF (black solid curve) are depicted by their means (dashed vertical lines) and their best-fitting sum (in red). CCFs are for the red arm in the top panel and for the blue arm in the bottom panel. The number of detected peaks for both arms is given above the figure.

estimate their number. With 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 normalized flux above 10 per cent is detected, a quality flag (see Section 3.8) is raised, warning the user that the EW 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.

3.6 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 $f_{\rm diff}$ have a profile similar to the P Cygni or inverted P Cygni profile (for example, see the third spectrum in Fig. 3) that is often observed in emission-line objects (Castor & Lamers 1979). To detect SB2 candidates, we performed cross-correlation between the reference and the observed spectra, disregarding the wavelength range of $\pm$10Å around the centre of the Balmer lines in order to avoid broadening of the cross-correlation function (CCF) peak. Cross-correlation was performed independently for both (the blue and the red) HERMES spectral arms. The resulting CCF, shown as the black curve in Fig. 11, was fitted by three Gaussian functions, centred at the three strongest peaks, to describe its shape. The location, amplitude and width of these 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 a potential SB2 detection in the final results (see Table A1), where binarity indication is given independently for the two arms. However, the results for the blue arm (column `SB2_c1`) are more trustworthy. This is the consequence of the blue arm having a greater number of absorption lines, causing sharper CCF peaks. In addition, the blue arm is less polluted by fibre-crosstalk (explained in the next paragraph). For even greater completeness of detected SB2 candidates, the list of analysed binaries compiled by Traven et al. (2020) can be used. These authors combined the unsupervised spectral dimensionality reduction algorithm t-SNE and semi-supervised CCF analysis (Merle et al. 2017) 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 small subsection of it. After a thorough investigation, we discovered that this effect is caused by an improper treatment of fibre cross-talk while extracting spectra from the original 2D image (Kos et al. 2017). 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.

Furthermore, the highest peak of our CCF function is used to determine the correctness of the wavelength calibration during the reduction of the spectra (Kos et al. 2017). If the peak is shifted by more than five correlation steps (equal to about 13 km s^{-1}) from the rest wavelength of the reference spectrum, the quality flag (see Section 3.8) is raised, warning the user that the derived radial velocity, EW, and asymmetry index might be wrong in the respective arm, as the two spectra were not aligned ideally.

3.7 Resulting table

The emission indices and other computed parameters are collected in Table A1. The complete table is available in electronic form at The Strasbourg astronomical Data Center (CDS). An excerpt of the published results, containing a subset of 30 rows and the 11 most interesting columns for the strongest unflagged emitters, is given in Table B1.

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 3.8. Their use depends on the specific requirements of the user and the physics of the analysed sysytem.

3.8 Flagging, quality control and result selection

The pipeline described above runs blindly on every successfully reduced spectrum (`guess_flag` = 0, for details see Kos et al. 2017), 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, which are listed and described in detail in Table 1. A user can base their selection of results according to the desired confidence level and 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 the parameter `emiss` in Table A1 is set to one (the equivalent of true). In order to include only 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

Table 1. Quality binary flags produced during various steps of our detection and analysis pipeline. A lower value of the flag represents a lower significance to the quality of detection and classification. The final reported `flag` value in Table A1 is the 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 spectra in the red arm of a spectrum. Median squared error (MSE) between them was $\geq$0.002
16	Large difference between reference and observed spectra in the blue arm of a spectrum. MSE was $\geq$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. Four or more residual sky lines were detected. Could be a result of under- or over-correction.
2	Wavelength solution (or determined radial velocity) might be wrong in the red arm of the spectrum. Determined from the 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.

be mimicked by spectral noise or the uncertainty of the reference spectrum, or be induced by the reduction pipeline. This selection criterion discards the weakest chromospheric components, which might be of great interest for specific studies, but, at the same time, it includes emissions that only partially fill the Hα line. If the user is interested only in stronger emitters, the threshold should be raised to `Ha_EW` > 0.5 Å or above. Similarly, spectra with the strongest emissions whose emission feature stands well above the surrounding continuum can be selected by focusing on spectra with `Ha_EW` > 1.0 Å. Above that limit, no Hα absorption line is visible unless the emission feature is wavelength-shifted or has a double-peaked shape.

Table A1 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}).

4 TEMPORAL VARIABILITY

The strategy of the GALAH survey is to observe as many objects as possible, and, as a result, not many repeat observations were made. The repeated fields were mostly observed to assess the stability of the instrument. Time spans between observations are therefore of the orders of days or years. This greatly limits the possibility of finding a variable object, but still enables us to discover potentially 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 two and nine 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 Fig. 12, 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-peaked 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 per cent of the inspected spectra display a stable Hα profile. Noticeable profile shape transformation was observed in $\sim$10 per cent of the cases, and peak location change in $\sim$5 per cent of the cases. Some degree of emission intensity change was noticed in $\sim$40 per cent of the cases. Visually similar is reduction-induced variability (see the rightmost panel in Fig. 12), observed for $\sim$25 per cent 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 Fig. 12. That kind of reduction-induced variability is limited to a few observed fields.

5 DISCUSSION AND CONCLUSIONS

In this paper, we have described the development and application of a neural network autoencoder structure that is able to extract the most relevant latent features from the spectrum. Low dimensionality latent feature contains only the most basic spectral information that is 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 part of this paper, we showed that a dense autoencoder neural network structure can be reliably used for the generation of non-peculiar reference spectra if it is 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 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 over all the sky. Fig. 13 shows that we can identify a 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. The detected nebular emission in stellar spectra, shown in Fig. B4, coincides 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. 2000) such as the Antares Emission nebulae, clouds around π Sco and δ Sco, Barnard's loop, the Carina Nebula, nebulae around λ Ori, nebular veils in the constellations of Puppis, Pyxis and Antlia, and other less prominent features. The published velocities of forbidden lines in Table A1 are given in the barycentric system and can therefore be used for studies of the internal dynamics of clouds.

Our detected emission spectra have a broad range of emission components – these range from a very strong to a 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

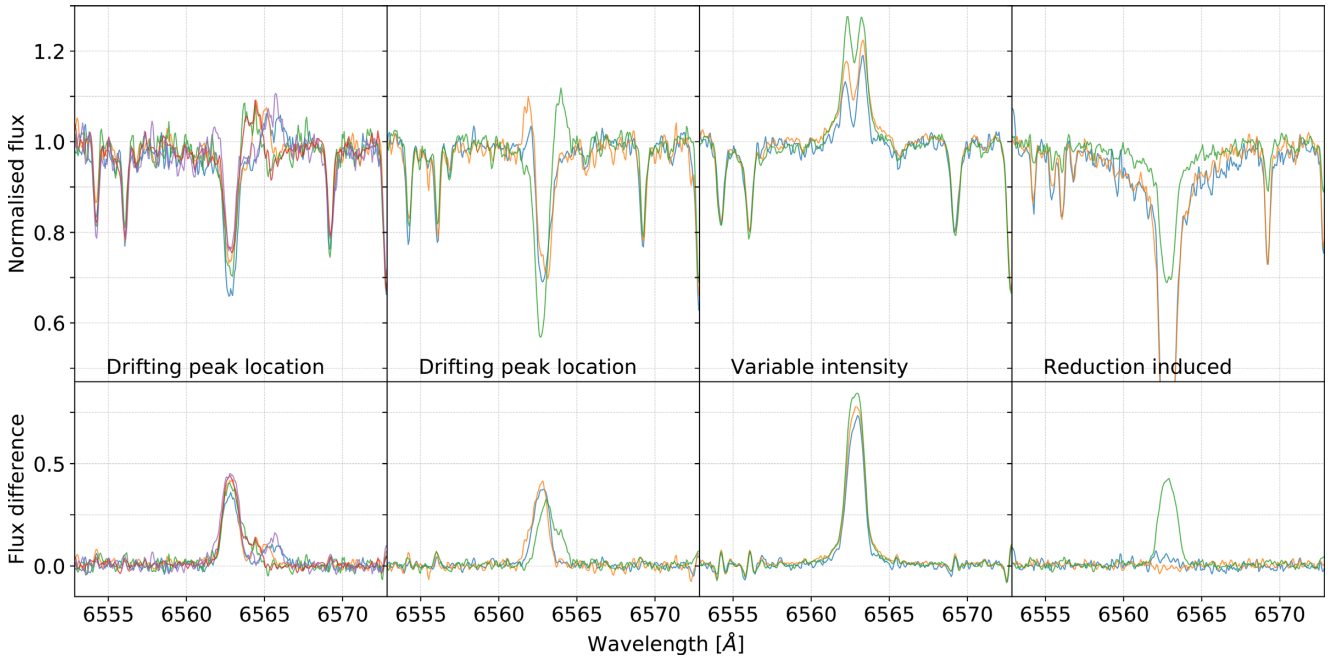

Figure 12. A sample of objects with repeated observations, where at least one of the normalized spectra (top row) contains strong Hα emission detected by comparison with the reference spectrum (bottom row). The first two objects (or columns) show the shifting location of an additional emission component peak, and the last two show variations in strength. The last example is most likely a result of a mis-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 the majority of them show the same tendency of having weaker absorption lines across the spectrum.

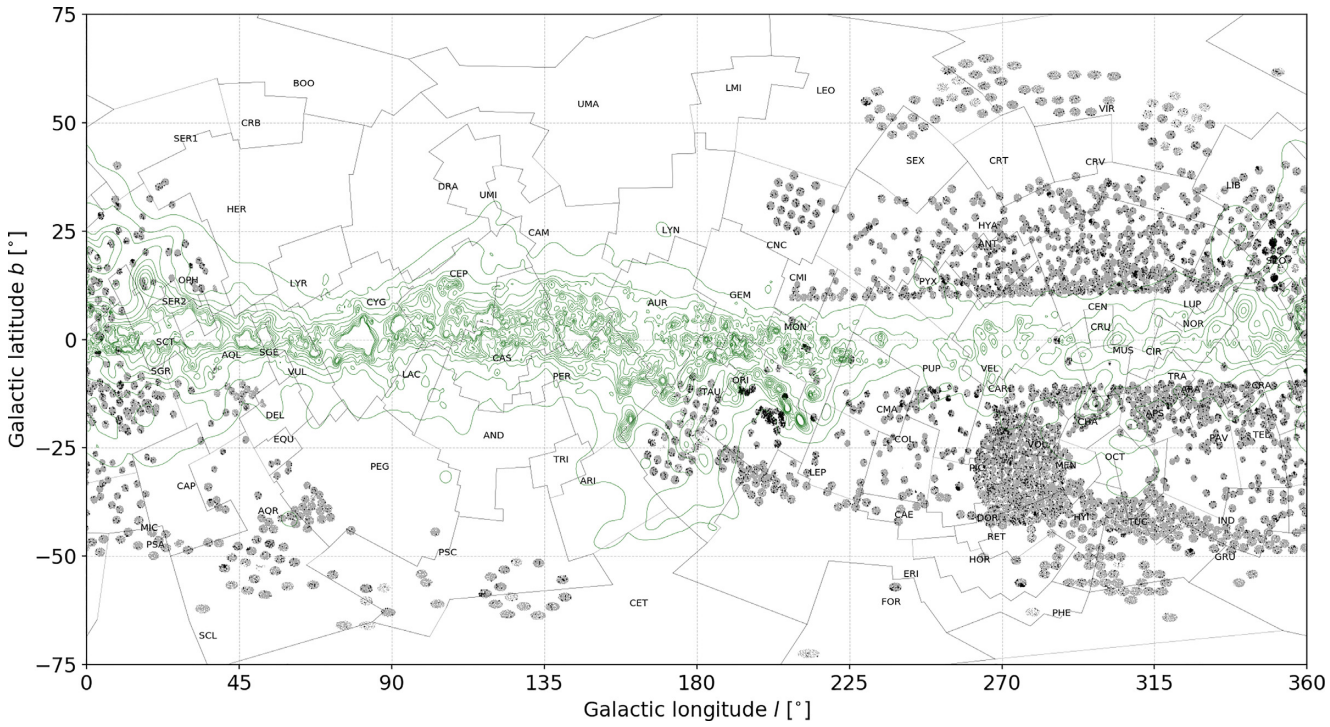

Figure 13. Spatial distribution of stars with detected Balmer emission profiles. Grey areas represent regions that were observed and analysed in this paper. The green lines represent the location of equal reddening in steps of 0.1 mag at a distance of 2 kpc. Reddening data were taken from results published by Capitanio et al. (2017). For readability, no isoline is shown above a reddening of 1 mag. Constellation boundaries were taken from Davenhall & Leggett (1989). Locations of their designations are defined by median values of constellation polygon vertices. The plot is not optimized for printing and should be explored in the higher-resolution electronic format that is a supplement to this paper.

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

To reliably detect even the weakest chromospheric emissions, the uncertainty of the 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 of improving our methodology using a variational autoencoder. Its advantage lies in the possibility of the simultaneous determination of a reference spectrum and its uncertainty, which would enable an uncertainty estimation of the measured emission-line indices.

ACKNOWLEDGEMENTS

This work is based on data acquired through the Australian Astronomical Observatory, under programmes: A/2014A/25, A/2015A/19, A2017A/18 (the GALAH survey); A/2015A/03, A/2015B/19, A/2016A/22, A/2016B/12, A/2017A/14 (the K2-HERMES K2-follow-up program); A/2016B/10 (the HERMES-TESS program); A/2015B/01 (Accurate physical parameters of Kepler K2 planet search targets); S/2015A/012 (Planets in clusters with K2). We acknowledge the traditional owners of the land on which the AAT stands, the Gamilaraay people, and pay our respects to elders past and present.

The Digitized Sky Surveys were produced at the Space Telescope Science Institute under US Government grant NAG W-2166. The images of these surveys are based on photographic data obtained using the Oschin Schmidt Telescope on Palomar Mountain, and the UK Schmidt Telescope. The plates were processed into the present compressed digital form with the permission of these institutions.

KČ, TZ and JK acknowledge financial support from the Slovenian Research Agency (research core funding no. P1-0188) and the European Space Agency (PRODEX Experiment Arrangement No. C4000127986). JBH is funded by an ARC Laureate Fellowship. SB acknowledges funds from the Australian Research Council (grants DP150100250 and DP160103747) as well as from the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research. Parts of this research were supported by the Australian Research Council (ARC) Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013.

DATA AVAILABILITY

The data underlying this article are available in the article and in its online supplementary material.

REFERENCES

Allard N. F., Kielkopf J. F., Cayrel R., van't Veer-Menneret C., 2008, A&A, 480, 581
Amarsi A. M., Nordlander T., Barklem P. S., Asplund M., Collet R., Lind K., 2018, A&A, 615, A139
Aret A., Kraus M., Šlechta M., 2016, MNRAS, 456, 1424
Barden S. C. et al., 2010, Ground-based and Airborne Instrumentation for Astronomy III, SPIE - International Society for Optics and Photonics. p. 773509
Barklem P. S., Piskunov N., O'Mara B. J., 2000, A&A, 363, 1091
Bergemann M. et al., 2016, A&A, 594, A120
Bohuski T. J., 1973, ApJ, 184, 93
Buder S. et al., 2018, MNRAS, 478, 4513
Capitanio L., Lallement R., Vergely J. L., Elyajouri M., Monreal-Ibero A., 2017, A&A, 606, A65
Castor J. I., Lamers H. J. G. L. M., 1979, ApJS, 39, 481
Cayrel R., van't Veer-Menneret C., Allard N. F., Stehlé C., 2011, A&A, 531, A83
Cheng T.-Y., Li N., Conselice C. J., Aragón-Salamanca A., Dye S., Metcalf R. B., 2020, MNRAS, 494, 3750
Čotar K. et al., 2019, MNRAS, 483, 3196
Damiani F. et al., 2016, A&A, 591, A74
Damiani F. et al., 2017, A&A, 604, A135
Davenhall A., Leggett S., 1989, Royal Observatory of Edinburgh
de Laverny P., 2012, in Prugniel P., Singh H. P., eds, Astronomical Society of India Conference Series Vol. 6, Astronomical Society of India. p. 53
De Silva G. M. et al., 2015, MNRAS, 449, 2604
Drew J. E. et al., 2014, MNRAS, 440, 2036
Escalante V., Morisset C., 2005, MNRAS, 361, 813
Giribaldi R. E., Ubaldo-Melo M. L., Porto de Mello G. F., Pasquini L., Ludwig H. G., Ulmer-Moll S., Lorenzo-Oliveira D., 2019, A&A, 624, A10
Hanuschik R. W., 2003, A&A, 407, 1157
He K., Zhang X., Ren S., Sun J., 2015, Proceedings of the IEEE international conference on computer vision, p. 1026
Hou W. et al., 2016, Res. Astron. Astrophys., 16, 138
Ignace R., Gray S. K., Magno M. A., Henson G. D., Massa D., 2018, AJ, 156, 97
Jones C. E., Tycner C., Smith A. D., 2011, AJ, 141, 150
Karmakar A., Mishra D., Tej A., 2018, IEEE Recent Advances in Intelligent Computational Systems,p. 122
Kingma D. P., Ba J., 2015, 3rd International Conference on Learning Representations
Kogure T., Leung K.-C., 2007, Astrophysics of Emission-Line Stars, Vol. 342, Springer, New York
Kohoutek L., Wehmeyer R., 1999, A&AS, 134, 255
Kos J. et al., 2017, MNRAS, 464, 1259
Kurucz R. L., 1993, SYNTHE Spectrum Synthesis Programs and Line Data, Smithsonian Astrophysical Observatory
Lancaster L., Greene J., Ting Y.-S., Koposov S. E., Pope B. J. S., Beaton R. L., 2020, AJ, 160, 125
Li W. et al., 2019, MNRAS, 485, 2628
Li X.-R., Pan R.-Y., Duan F.-Q., 2017, Res. Astron. Astrophys., 17, 036
Ma Z. et al., 2019, ApJS, 240, 34
Mathew B., Subramaniam A., Bhatt B. C. r., 2008, MNRAS, 388, 1879
Matijevič G. et al., 2012, ApJS, 200, 14
MacLean B. J., Greene G. R., Lattanzi M. G., Pirenne B., 2000, The Status of the Second Generation Digitized Sky Survey and Guide Star Catalog, Astron. Soc. Pac.. p. 145
Merle T. et al., 2017, A&A, 608, A95
Munari U., Sordo R., Castelli F., Zwitter T., 2005, A&A, 442, 1127
Munari U. et al., 2019, MNRAS, 488, 5536
Nakano M., Sugitani K., Watanabe M., Fukuda N., Ishihara D., Ueno M., 2012, AJ, 143, 61
Natta A., Testi L., Muzerolle J., Randich S., Comerón F., Persi P., 2004, A&A, 424, 603
Ness M., Hogg D. W., Rix H. W., Ho A. Y. Q., Zasowski G., 2015, ApJ, 808, 16
Ness M., Hogg D. W., Rix H. W., Martig M., Pinsonneault M. H., Ho A. Y. Q., 2016, ApJ, 823, 114
Nikoghosyan E. H., Vardanyan A. V., Khachatryan K. G., 2016, ASPC, 505, 66
Pan R.-y., Li X.-r., 2017, Chinese Astron. Astrophys., 41, 318
Piskunov N., Valenti J. A., 2017, A&A, 597, A16
Qin H.-r., Lin J.-m., Wang J.-y., 2017, Chinese Astron. Astrophys., 41, 282

Raju K. P., Prasad C. D., Desai J. N., Mishra L., 1993, Ap&SS, 204, 205
Ralph N. O. et al., 2019, PASP, 131, 108011
Reid W. A., Parker Q. A., 2012, MNRAS, 425, 355
Reipurth B., Pedrosa A., Lago M. T. V. T., 1996, A&AS, 120, 229
Sharma S. et al., 2018, MNRAS, 473, 2004
Sheinis A. et al., 2015, J. Astron. Telesc. Instrum. Syst., 1, 035002
Shen H., George D., Huerta E. A., Zhao Z., 2019, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, p. 3237
Silaj J., Jones C. E., Sigut T. A. A., Tycner C., 2014, ApJ, 795, 82
Ting Y.-S., Conroy C., Rix H.-W., Cargile P., 2019, ApJ, 879, 69
Traven G. et al., 2015, A&A, 581, A52
Traven G. et al., 2017, ApJS, 228, 24
Traven G. et al., 2020, A&A, 638, A145
Valenti J. A., Piskunov N., 1996, A&AS, 118, 595
van der Maaten L., 2013, preprint (arXiv:1301.3342)
White R. J., Basri G., 2003, ApJ, 582, 1109
Witham A. R., Knigge C., Drew J. E., Greimel R., Steeghs D., Gänsicke B. T., Groot P. J., Mampaso A., 2008, MNRAS, 384, 1277
Wittenmyer R. A. et al., 2018, AJ, 155, 84
Yang T., Li X., 2015, MNRAS, 452, 158
Žerjal M. et al., 2013, ApJ, 776, 127

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

results_emission_lines_final_mnras.fits

sky_emissions.png

sky_nebular.png

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: TABLE DESCRIPTION

In Table A1 we provide a list of metadata available for every object analysed using the methodology described in this paper. The complete table of detected objects and their metadata is available only in electronic form at the CDS and at the publisher's website.

Table A1. List and description of the fields in the published catalogue of analysed objects.

Column	Unit	Description
source_id		*Gaia* DR2 star identifier
sobject_id		GALAH internal per-spectrum unique i.d.
ra	°	Right ascension coordinate from 2MASS
dec	°	Declination coordinate from 2MASS
Ha_EW	Å	Equivalent width of a difference between observed and template spectra 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 spectra 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	km s^{-1}	Width (in km s^{-1}) of the Hα emission feature at 10 per cent 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		Indication if binarity was detected in the red arm
SB2_c1		Was binarity detected in the blue arm
NII		Number of detected [N II] peaks in the doublet
SII		Number of detected [S II] peaks in the doublet
NII_EW	Å	Equivalent width of Gaussian profiles fitted to the [N II] emission features
SII_EW	Å	Same as the NII_EW, but for the [S II] doublet
rv_NII	km s^{-1}	Intrinsic radial velocity of the [N II] doublet in the barycentric frame. To compute this velocity, we subtracted the radial velocity of a star from the doublet velocity in Fig. 9.
rv_SII	km s^{-1}	Same as rv_NII, but for the [S II] doublet
nebular		Spectrum is considered to have an additional nebular component
emiss		Spectrum is considered to have an additional Hα emission component
flag		Sum of all bitwise flags raised for a spectrum

APPENDIX B: ADDITIONAL FIGURES

In order to increase the readability and transparency of the text, additional and repeated plots are supplied as appendices to the main text.

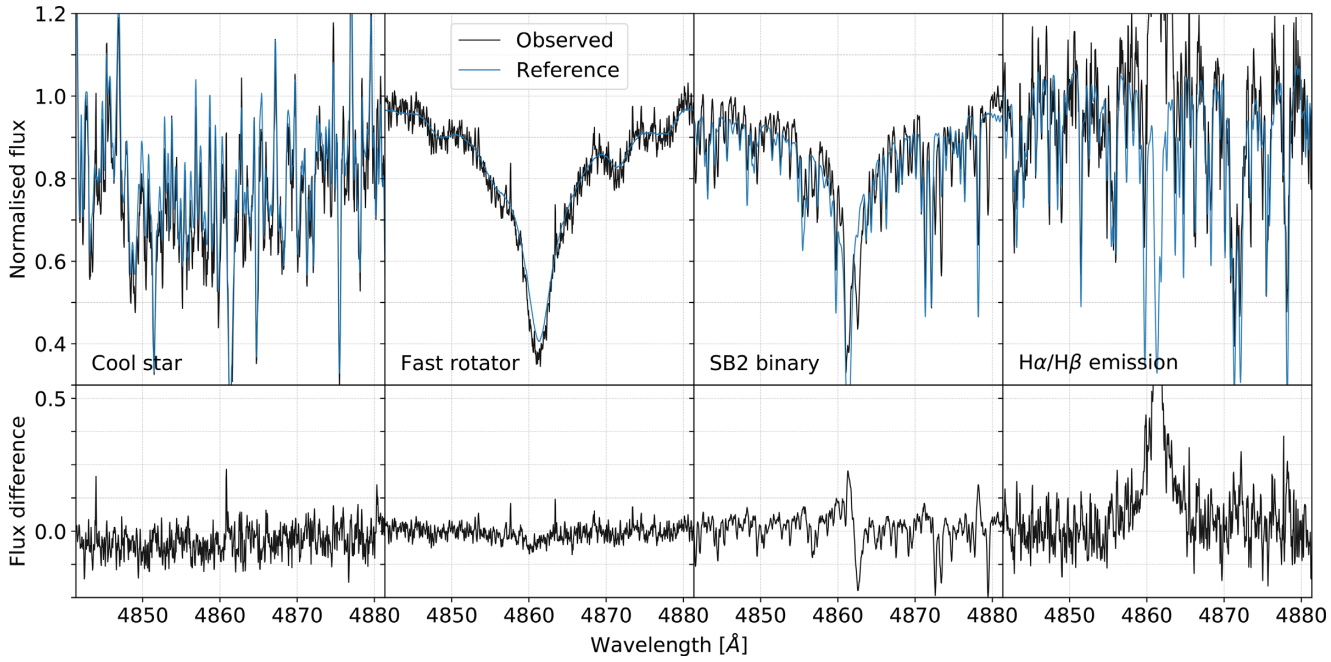

Figure B1. Same plots and objects as in Fig. 3, but for the blue spectral arm.

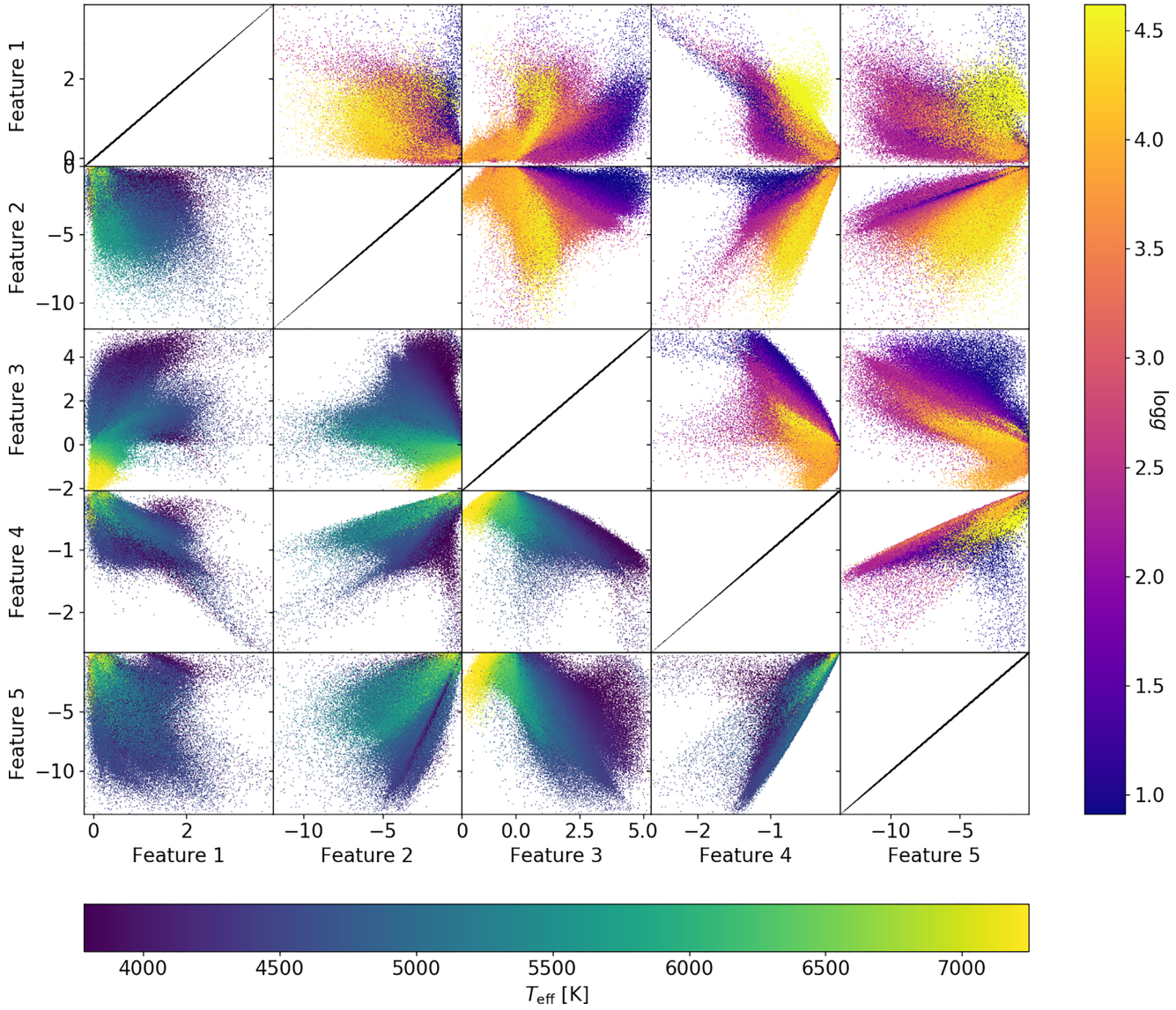

Figure B2. Same plots as shown in Fig. 4, but for the latent features of the blue HERMES band, coloured according to the parameter $T_{\rm eff}$ in lower triangle and to log g in the upper triangle.

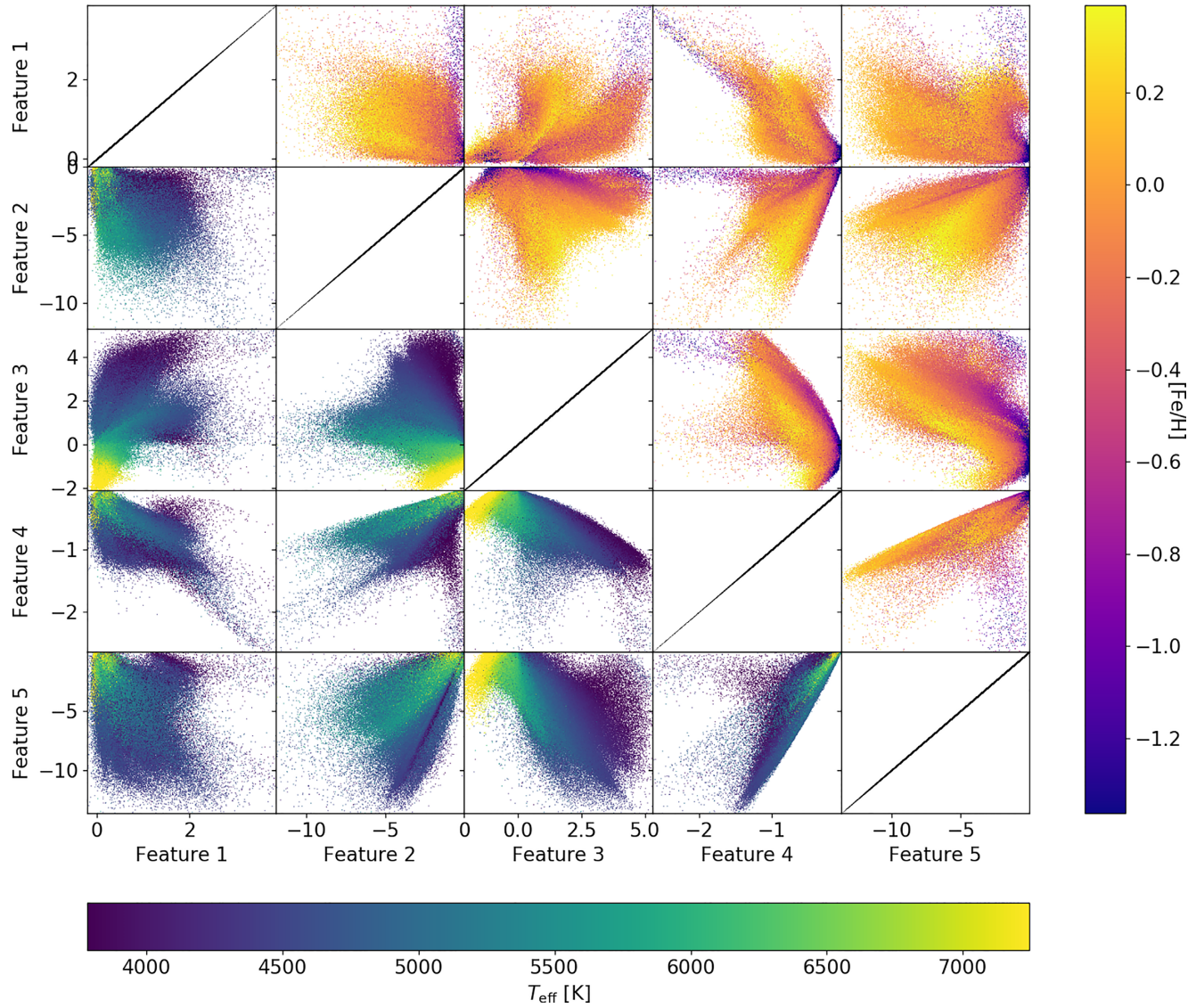

Figure B3. Same plots as shown in Fig. 4, but for the latent features of the blue HERMES band, coloured according to the parameter $T_{\rm eff}$ in the lower triangle and to [Fe/H] in the upper triangle.

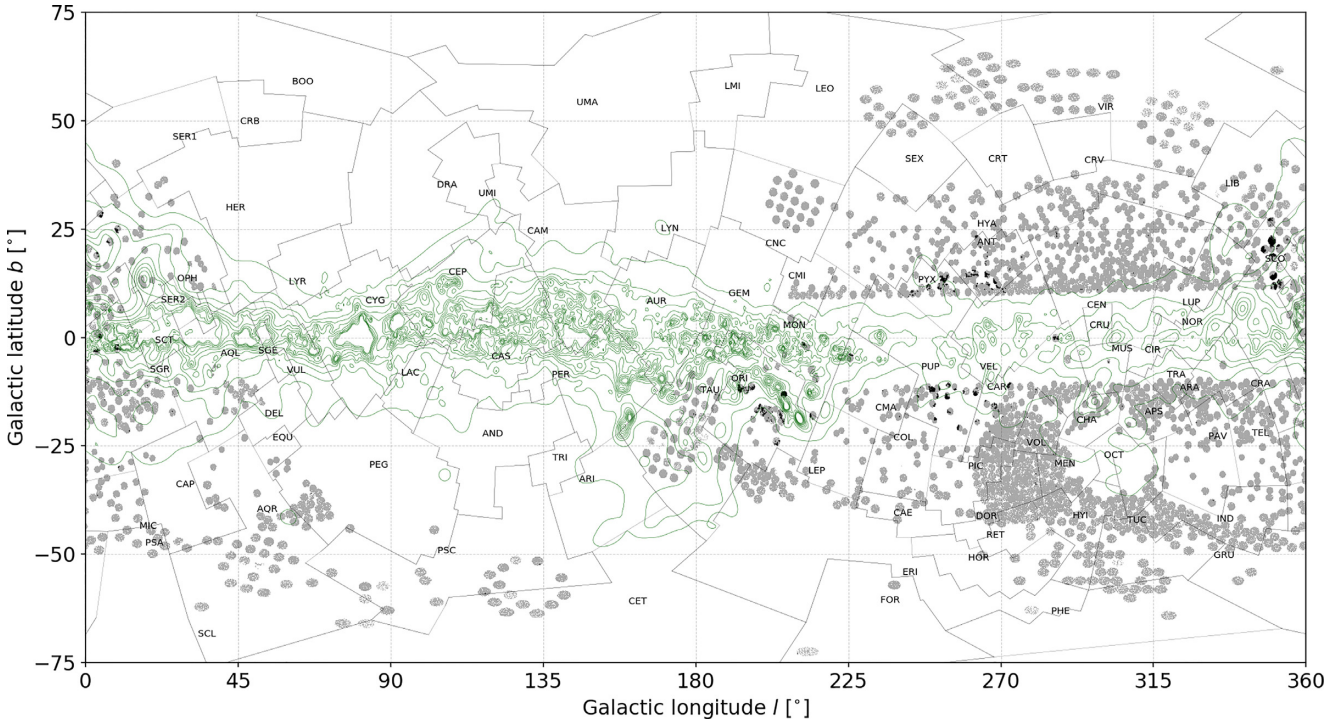

Figure B4. Same as Fig. B4, but showing stars with at least three detected nebular emission lines, shown with black dots.

Table B1. Excerpt of the 30 strongest unflagged emitters from the published table presented in detail in Table A1. The rest of the table can be downloaded in electronic form from CDS service and the publisher's 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
3220688120051681792	3.27	3.27	131.65	− 0.11	0	0	−0.01	− 11.03	− 126.68	0
2920301135326368256	3.17	3.17	204.04	− 0.16	1	0	0.07	− 102.90	84.00	0
6086395513764172800	3.11	3.11	130.44	− 0.16	0	0	0.04	− 55.69	27.51	0
6014331906769737728	3.11	3.11	133.08	− 0.07	0	1	−0.02	80.91	44.02	0
3221921742035204608	3.09	3.10	114.30	− 0.02	0	0	0.02	− 46.60	11.19	0
3181196759054802432	3.07	3.07	111.78	− 0.05	0	0	−0.02	32.48	20.05	0
6084509542084855808	3.03	3.03	191.97	− 0.03	0	0	0.00	25.25	92.65	0
3217773074208416768	3.00	3.30	343.16	− 0.76	0	0	−0.01	− 128.82	3.13	0
6244153510833938432	3.00	3.00	117.06	− 0.08	0	0	0.01	41.17	35.66	0
3220786839876234752	2.99	2.99	152.48	− 0.03	1	0	0.02	− 79.85	− 69.71	0

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