Astronomy & Astrophysics manuscript no. output
October 20, 2021

Metallicities in M dwarfs: Investigating different determination techniques

V. M. Passegger[1,2], A. Bello-García[3], J. Ordieres-Meré[4], A. Antoniadis-Karnavas[5,6], E. Marfil[7,8], C. Duque-Arribas[7], P. J. Amado[9], E. Delgado-Mena[5], D. Montes[7], B. Rojas-Ayala[10], A. Schweitzer[1], H. M. Tabernero[11], V. J. S. Béjar[12,13], J. A. Caballero[8], A. P. Hatzes[14], Th. Henning[15], S. Pedraz[16], A. Quirrenbach[17], A. Reiners[18], and I. Ribas[19,20]

[1] Hamburger Sternwarte, Gojenbergsweg 112, D-21029 Hamburg, Germany
e-mail: vpassegger@hs.uni-hamburg.de
[2] Homer L. Dodge Department of Physics and Astronomy, University of Oklahoma, 440 West Brooks Street, Norman, OK 73019, United States of America
[3] Departamento de Construcción e Ingeniería de Fabricación, Universidad de Oviedo, c/ Pedro Puig Adam, Sede Departamental Oeste, Módulo 7, 1^a planta, E-33203 Gijón, Spain
[4] Departamento de Ingeniería de Organización, Administración de Empresas y Estadística, Universidad Politécnica de Madrid, c/ José Gutiérrez Abascal 2, E-28006 Madrid, Spain
[5] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762 Porto, Portugal
[6] Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal
[7] Departamento de Física de la Tierra y Astrofísica and IPARCOS-UCM (Instituto de Física de Partículas y del Cosmos de la UCM), Facultad de Ciencias Físicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain
[8] Departamento de Astrofísica, Centro de Astrobiología (CSIC-INTA), ESAC campus, Camino Bajo del Castillo s/n, E-28691, Villanueva de la Cañada, Madrid, Spain
[9] Instituto de Astrofísica de Andalucía (IAA-CSIC), Glorieta de la Astronomía s/n, E-18008 Granada, Spain
[10] Instituto de Alta Investigación, Universidad de Tarapacá, Casilla 7D, Arica, Chile
[11] Centro de Astrobiología (CSIC-INTA), Carretera de Ajalvir km 4, Torrejón de Ardoz, 28850, Madrid, Spain
[12] Instituto de Astrofísica de Canarias, c/ Vía Láctea s/n, E-38205 La Laguna, Tenerife, Spain
[13] Departamento de Astrofísica, Universidad de La Laguna, E-38206 La Laguna, Tenerife, Spain
[14] Thüringer Landessternwarte Tautenburg, Sternwarte 5, D-07778 Tautenburg, Germany
[15] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
[16] Centro Astronómico Hispano-Alemán (CSIC-MPG), Observatorio Astronómico de Calar Alto, Sierra de los Filabres, E-04550 Gérgal, Almería, Spain
[17] Landessternwarte, Zentrum für Astronomie der Universität Heidelberg, Königstuhl 12, D-69117 Heidelberg, Germany
[18] Institut für Astrophysik, Georg-August-Universität, Friedrich-Hund-Platz 1, D-37077 Göttingen, Germany
[19] Institut de Ciències de l'Espai (CSIC-IEEC), Campus UAB, c/ de Can Magrans s/n, E-08193 Bellaterra, Barcelona, Spain
[20] Institut d'Estudis Espacials de Catalunya (IEEC), E-08034 Barcelona, Spain

Received ..; accepted ..

ABSTRACT

Deriving metallicities for sun-like stars follows well-established methods, but for cooler stars such as M dwarfs the determination is much more complicated due to the forests of molecular lines. Several methods have been developed over the last years to determine accurate and precise stellar parameters for these cool stars. However, sometimes significant differences can be found when comparing metallicities for the same star derived from different methods. In this work, we determine effective temperatures, surface gravities, and metallicities of 18 well-studied M dwarfs observed with the CARMENES high-resolution spectrograph following different approaches, including synthetic spectral fitting, the analysis of pseudo-equivalent widths, and machine learning. We analyze the discrepancies in the derived stellar parameters, including metallicity, in several analysis runs. Our goal is to minimize these discrepancies and find stellar parameters more consistent with literature. We attempt to achieve this by standardizing the most commonly used components, such as wavelength ranges, synthetic model spectra, continuum normalization methods, and stellar parameters. We see that, although such modifications work quite well in hotter main-sequence stars, they do not improve the consistency in stellar parameters in M dwarfs, leaving mean deviations of around 50-200 K in temperature, and 0.1-0.3 dex in metallicity. M dwarfs are much more complex and a standardization of the aforementioned components cannot be considered as a straightforward recipe for bringing consistency to the derived parameters. Further in-depth investigations of the employed methods would be necessary to identify and correct for the still existing discrepancies.

Key words. methods: data analysis - techniques: spectroscopic – stars: fundamental parameters – stars: late-type – stars: low-mass

1. Introduction

Precise stellar metallicity determination is an essential step to a full understanding of the dynamical and chemical evolution of the Galaxy. Several methods have been developed to study element abundances of all kinds of stars. Among those, M dwarfs are the most numerous type in our Galaxy, and therefore an accurate determination of their abundances is of utmost interest. In the fast growing field of exoplanet detection and characterization, abundance determination of the host star is also important to better understand the formation and evolution of planetary systems (e.g., Burn et al. 2021).

A popular method for deriving metallicities of M dwarfs is based on the measurement of pseudo-equivalent widths (pEWs) of spectral lines. This method was used by Neves et al. (2013, 2014), Mann et al. (2013a, 2014), Newton et al. (2014), Maldonado et al. (2015), and Khata et al. (2020), among others. Another widely used approach is spectral synthesis. There, the stellar spectrum is synthesized using stellar atmosphere models along with radiative transfer codes and atomic and molecular line lists. The PHOENIX stellar atmosphere code (Hauschildt 1992, 1993) is the basis for stellar model grids such as the BT-Settl model atmospheres (Allard et al. 2012, 2013) and the PHOENIX-ACES synthetic model grid (Husser et al. 2013).

Marfil et al. (2021) used the BT-Settl model atmospheres and the radiative transfer code `turbospectrum` (Plez 2012) to generate synthetic spectra around 75 Fe I and Ti I lines along with the γ- and ϵ-TiO bands to determine $T_{\rm eff}$, $\log g$, and [Fe/H] for 342 M dwarfs from the CARMENES survey by means of the SteParSyn code (Tabernero et al. 2018, 2021). The `turbospectrum` code was also employed by Souto et al. (2017, 2020), and Sarmento et al. (2021) together with 1-D MARCS stellar atmospheres (Gustafsson et al. 2008) to derive stellar parameters and abundances for several M dwarfs observed with the Apache Point Observatory Galactic Evolution Experiment (APOGEE, Majewski et al. 2017). Operating in a wavelength range from 1500–1700 nm and with high-resolution ($R \approx$ 22,500; Wilson et al. 2010), APOGEE is dedicated to observe red giants; additionally, APOGEE has observed around 2000 M dwarfs. Önehag et al. (2012) and Lindgren et al. (2016) fitted synthetic spectra to high-resolution CRIRES J-band spectra of M dwarfs using the Spectroscopy Made Easy package (SME, Valenti & Piskunov 1996; Valenti & Fischer 2005) with MARCS atmospheres. SME computes synthetic spectra on the fly and determines the best fit stellar parameters by χ^2-minimization with the observed spectra. Passegger et al. (2018) fitted the PHOENIX-ACES model spectra grid to high-resolution CARMENES spectra of 300 M dwarfs and derived $T_{\rm eff}$, $\log g$, and [Fe/H].

Within the last several years, machine learning has emerged as a valuable tool to predict stellar parameters for large sets of stars. Several applications of neural networks in stellar parameter determination can be found in Fabbro et al. (2018), Birky et al. (2020), Antoniadis-Karnavas et al. (2020), and Passegger et al. (2020), among others.

For a more detailed overview on previous work regarding stellar parameter determination in M dwarfs, we refer to the literature summaries in Passegger et al. (2020) and Marfil et al. (2021).

It is known from previous parameter studies that different determination methods sometimes provide significantly different results for stellar parameters for the same stars. This is shown in the comparison plots of several parameter determination studies, e.g., Figs. 5-7 in Passegger et al. (2019), Fig. 7 in Passegger et al. (2020), Figs. 1 and 5 in Lindgren et al. (2016), Figs. 13-14 in Neves et al. (2014), Fig. 13 in Rojas-Ayala et al. (2012), Figs. 10-12 in Marfil et al. (2020), Figs. 9, 11, Figs. A1-A6 in Marfil et al. (2021), and Figs. 12 and 13 in Sarmento et al. (2021). These inconsistencies challenge the reliability of the determined stellar parameters for the lowest mass stars.

However, one has to keep in mind that there **are** different kinds of inconsistencies. The most relevant cases in this context are inconsistencies between different methods and between different observations of the same star with different instruments. Since there is no way yet to measure the absolute correct physical and atmospheric properties of a given star, we have to rely on the parameters that different methods and observations provide. Deriving consistent values for the same star with different methods (and/or different instruments) can therefore be considered as a proxy for the reliability of the value of the stellar parameters and of the methods themselves.

Several studies conducted such a consistency analysis for FGK-type stars and analyzed the differences introduced when deriving abundances with different methods. Hinkel et al. (2016) investigated 4 G-type stars with high-resolution MIKE spectra (R $\approx$ 50,000) from the Magellan Planet Search Program, with an average S/N of 200, and covering the wavelength range 5050–7100 Å. Six different groups participated in the analysis and determined abundances for 10 elements (C, O, Na, Mg, Al, Si, Fe, Ni, Ba, and Eu), in 4 different runs. In Run 1 each group used their individual techniques, in Run 2 standard stellar parameters for $T_{\rm eff}$, $\log g$, and microturbulent velocity ξ were provided. Run 3 included a standard line list, whereas Run 4 was a combination between Runs 2 and 3. They found that Run 2 gave consistently better results between the elements, followed by Run 4, which suggests that stellar parameters other than abundances and/or line lists should be standardized in order to produce similar results.

A larger sample of 34 *Gaia* benchmark FGK-type stars was used by Jofré et al. (2014). The spectra were collected from HARPS ($R \approx 115,000$), NARVAL ($R \approx 80,000$), and UVES ($R \geq 70,000$), covering a spectral range from 4760–6840 Å. Seven different groups participated in this study and derived Fe abundances in three runs. Their main aim was to analyze the effects of instrumental resolution on the determination of metallicity when fixing $T_{\rm eff}$ and $\log g$ to independently derived values. Furthermore, all groups used a common line list, atomic data (see Heiter et al. 2021) and the same atmospheric models (MARCS). In the different runs, they used spectra with their original resolution and with resolution downgraded to $R = 70,000$ to study instrumental effects. They found that different resolutions result in a metallicity difference of less than 0.05 dex, and that metallicities agree when using different instruments. A comparison of the different methods showed larger standard deviations in metallicity for cooler stars (0.1 dex, $T_{\rm eff}$ < 5000 K) than for hotter stars (0.07 dex, $T_{\rm eff}$ > 5000 K). A follow-up study by Jofré et al. (2015) analyzed 10 different element abundances with eight methods taking into account NLTE corrections for Fe and errors of the fixed stellar parameters. They performed a detail analysis of systematic errors for differential and absolute abundances. For an extensive discussion on each element and NLTE effects we refer to Jofré et al. (2015).

Jofré et al. (2017) provided a detailed study of 4 *Gaia* benchmark stars, the Sun, Arcturus, 61 Cyg A, and HD 22879. Their high-resolution spectra from NARVAL and HARPS were convolved to a common resolution of 70,000. Also here, the stellar parameters $T_{\rm eff}$, $\log g$, ξ, and $v \sin i$ were fixed for each star. The analysis was performed by six different teams in eight different runs, including tests regarding continuum normalization, common line lists, hyperfine structure, alpha enhancement, and radiative transfer code. They concluded that the most important point for consistent metallicity values is a common continuum flux.

Focusing on cooler stars, Slumstrup et al. (2019) conducted a similar study for red giant stars in three open clusters. They compared several combinations of line lists and methods to derive EWs and analyzed the systematic uncertainties from a line-by-line spectroscopic analysis. As a result, they find scatter of around 170 K in $T_{\rm eff}$, 0.4 dex in $\log g$, and 0.25 dex in metallicity, showing that even for high-precision spectroscopic analysis external constraints are necessary to obtain consistent results between different methods.

Up to now, no such analysis has been performed for M dwarfs. In this work, we aim to follow the approach by Hinkel et al. (2016) to study the deviations in metallicity as well as $T_{\rm eff}$ and $\log g$ coming from different determination methods and identify ways to derive more consistent results for stars at the cool end of the main-sequence. This paper is structured as follows. Section 2 gives an overview on the methods we use in our analyses. Section 3 explains our sample of benchmark stars and the different analysis runs we perform. In Sect. 4 we present the results of the investigation, followed by a discussion in Sect. 5. A short summary is given in Sect. 6.

2. Methods

In the following we describe the four different methods we use for deriving fundamental the stellar parameters $T_{\rm eff}$, $\log g$, and [Fe/H].

2.1. Synthetic spectra fitting

2.1.1. Pass19-code

This method is fully described in Passegger et al. (2018, 2019), hereafter referred to as Pass19-code. We use a downhill simplex method with a χ^2 minimization to find the synthetic model spectrum that best fits the observed spectrum by fitting several wavelength ranges in the VIS and NIR simultaneously (see Table 2 in Passegger et al. 2019).

The PHOENIX-ACES model spectra grid (Husser et al. 2013) incorporated here is based on the PHOENIX code developed by Hauschildt (1992, 1993). Improvements to the code are described in, e.g., Hauschildt et al. (1997), Hauschildt & Baron (1999), Claret et al. (2012), and Husser et al. (2013). The one-dimensional (1D) mode of the PHOENIX code computes spherically symmetric model atmospheres, which can be used to simulate main sequence stars and brown dwarfs, including L and T spectral types, as well as white dwarfs and giants; it also includes models for expanding envelopes of novae and supernovae, and accretion disks. PHOENIX can calculate synthetic spectra in 1D or 3D and can be executed in LTE or non-LTE radiative transfer mode. Several model atmosphere grids for late-type stars are based on the PHOENIX code, for instance the NextGen models (Hauschildt et al. 1999), the AMES models (Allard et al. 2001), and the BT-Settl models (Allard et al. 2011). For the calculation of the aforementioned PHOENIX-ACES model spectra grid a new equation of state was used, which was especially designed for the formation of molecules in very cool stellar atmospheres. The grid takes into account solar chemical compositions from Asplund et al. (2009), updated with meteoritic values from Lodders et al. (2009). Since the PHOENIX-ACES grid we use has $[\alpha/\text{Fe}] = 0$, our metallicity results of [M/H] directly translate into identical [Fe/H] values. However, for certain parameter ranges an α-enhanced PHOENIX-ACES grid is available (see Husser et al. 2013).

To match the instrumental resolution and wavelength grid of observed spectra the PHOENIX-ACES model spectra are convolved with a Gaussian and linearly interpolated in wavelength. The synthetic spectra are broadened to account for the rotational velocity $v \sin i$ of the star (Reiners et al. 2018). Therefore, a separate function estimates the effect on the line spread function and the synthetic spectrum is convolved with the resulting line spread function. The pseudo-continuum of both, observed and synthetic spectra, is normalized with a linear fit within each small wavelength region that is analyzed.

The surface gravity ($\log g$) is determined from evolutionary models as in Passegger et al. (2019). This is done to break degeneracies between the parameters. The evolutionary models used in this work were taken from the PARSEC v1.2S library (Bressan et al. 2012; Chen et al. 2014, 2015; Tang et al. 2014), which provides $T_{\rm eff}$ and $\log g$ for metallicities in the range $-2.2 <$ [M/H] $< +0.7$ and different stellar ages, among other parameters. To select the appropriate isochrone we took stellar ages from Passegger et al. (2019). The $\log g$ is then calculated from this isochrone's $T_{\rm eff}$-$\log g$ relation depending on $T_{\rm eff}$ and [Fe/H] chosen by our algorithm. To get finer values we linearly interpolate for metallicities between -1.0 and $+0.7$. The PHOENIX-ACES model spectra grid is then interpolated according to these three parameters and the χ^2 is calculated between the observed and synthetic spectrum. A downhill simplex finds the best fitting synthetic spectrum with the smallest χ^2 by exploring the 2-D $T_{\rm eff}$-[Fe/H] parameter space and adjusting those parameters accordingly.

2.1.2. SteParSyn

The SteParSyn code is described in detail in Tabernero et al. (2021). It is a Bayesian implementation of the spectral synthesis technique that determines the probability distributions of the stellar atmospheric parameters ($T_{\rm eff}$, $\log g$, [Fe/H], $v \sin i$, and ζ) from a Markov Chain Monte Carlo (MCMC) approach. In general terms, the code compares a grid of synthetic spectra precomputed around certain spectral features of interest. Therefore, we use a selection of **75 magnetically insensitive** Ti i and Fe i lines, as well as the $\gamma-$ and $\epsilon-$TiO bands in a range between 5850–15800 Å. The assessment of any point in the parameter space is done in a computationally inexpensive way employing principal component analysis (PCA). The code finally returns the posterior probability distributions in the stellar atmospheric parameters along with the best synthetic fit for the input spectral features.

With the aim of avoiding any potential degeneracy in the M-dwarf parameter space, especially between $\log g$ and [Fe/H], we **assume** Gaussian prior probability distributions in $T_{\rm eff}$ and $\log g$ for all individual targets, with standard deviations of 200 K and 0.2 dex, respectively. The prior distributions are centered following Cifuentes et al. (2020), who determined $T_{\rm eff}$ from a multi-band photometric analysis by means of the Virtual Observatory Spectral energy distribution Analyser (VOSA, Bayo et al. 2008), and derived stellar radii and masses from the Stefan-Boltzmann law and the mass-radius relation presented in Schweitzer et al. (2019).

Even though any model atmosphere grid can be used along with SteParSyn, in this work we employ BT-Settl model atmospheres (Allard et al. 2012). Since the grid is alpha-enhanced, metallicities derived using this method are corrected using a simple interpolation scheme between the mass fraction Z and [Fe/H] following the standard composition in the MARCS models (Gustafsson et al. 2008), as explained in Marfil et al. (2021).

SteParSyn was also used in Tabernero et al. (2018) for the study of cool supergiants in the Magellanic clouds, as well as in Tabernero et al. (2021) for the analysis of the AGB-star candidate VX Sgr. Marfil et al. (2021) applied SteParSyn to the CARMENES GTO sample. The exoplanet host WASP-121 was also analyzed with SteParSyn using ESPRESSO spectra (Borsa et al. 2021).

2.2. Machine Learning

2.2.1. Deep Learning (DL)

This method is described in detail in Passegger et al. (2020). Artificial neural networks are machine learning methods that are constructed from a collection of artificial neurons organized in different layers, in order to learn structures from data in a similar way as the human brain. In deep learning, the neural network models consist of multiple processing layers that can learn relevant features by themselves without user interaction.

For each stellar parameter we build a convolutional deep neural network with several hidden layers. In order to learn features from the input spectrum, the networks are trained with PHOENIX-ACES synthetic models. We linearly interpolate the existing grid using `pyterpol` (Nemravová et al. 2016) to increase the number of training samples. We apply additional restrictions to our grid similar to the Pass19-code. Based on the PARSEC v1.2S evolutionary models we exclude combinations of $T_{\rm eff}$, $\log g$, and [Fe/H] that are physically unrealistic for M dwarfs (i.e., they represent stellar objects far away from the main sequence). In the end we create 449 806 synthetic model spectra for the reference set in training process.

We convolve the synthetic spectra with a Voigt profile to account for instrumental broadening using a function based on `libcerf` (Johnson S.G. 2019). The Gaussian and Lorentzian components of the Voigt function for CARMENES were determined by Nagel et al. (2021, submitted). We also take into account $v \sin i$ by broadening the synthetic spectra with a Fortran translation of the `rotational_convolution` function of `Eniric`, assuming a default limb darkening coefficient of 0.6 (see Figueira et al. 2016). For continuum normalization of the synthetic as well as the observed CARMENES spectra we employ the Gaussian Inflection Spline Interpolation Continuum (`GISIC`) routine[1]. The routine smoothens the spectrum with a Gaussian before identifying molecular bands with a numerical gradient. Then continuum points are selected and a cubic spline interpolation normalizes the continuum within the desired spectral range. The observed CARMENES spectra are corrected for the spatial motion of the stars by using a cross-correlation between the observed spectrum and a PHOENIX-ACES model spectrum. Because this results in shifts of the wavelength grid of the observations, we linearly interpolate this grid to match the wavelength grid of the synthetic spectra.

In the training, the reference set is divided into a training set (95 %) and a validation set (5 %). After running the training set through the deep neural network the training error is estimated from the difference between the output and the known input stellar parameters. Based on this error the hyper-parameters of the DL model are adjusted through backward propagation. The validation set is used to determine the validation error (mean square error, MSE) after each training epoch to verify that the adjustment of the model hyper-parameters goes in the right direction to improve the DL model and the error keeps decreasing. It also helps to avoid over-fitting of the training set, which happens when the DL model learns to describe random variations and is unable to generalize to new data. The training is finished once the minimum validation error is reached. At this point a test set of 100 randomly generated synthetic spectra is sent through the DL model to measure the test error. This presents a final test to the DL model before it is applied to observed spectra. The model performs well, if the average test error is below a certain threshold, which we define to be between $5 \cdot 10^{-4}$ and 10^{-5} depending on the stellar parameter under investigation.

As explained in Passegger et al. (2020), the range 8800–8835 Å and an individual neural network model for each stellar parameter separately give the smallest validation errors. Therefore, we also follow that approach in this work.

2.2.2. Pseudo-EW approach (ODUSSEAS)

A detailed description of the machine learning tool ODUSSEAS can be found in Antoniadis-Karnavas et al. (2020). ODUSSEAS receives 1D spectra and their resolutions as input. The method is based on measuring the pEWs of absorption lines and blended lines in the range between 5300 and 6900Å. Spectral sections that include the activity-sensitive Na doublet, Hα line, and strong telluric lines, have been excluded from the line list. The line list consists of 4104 absorption features, the same as used by Neves et al. (2014).

ODUSSEAS contains a supervised machine learning algorithm based on the "scikit learn" package of Python, in order to determine the $T_{\rm eff}$ and [Fe/H] of the stars. In the training, it is provided with both input and expected output, in order to create the machine learning models using ridge regression. The pEWs in 65 HARPS spectra are used together with their $T_{\rm eff}$ and [Fe/H] from Casagrande et al. (2008) and Neves et al. (2012), respectively, as reference for training and testing its models.

Applied to new spectra, ODUSSEAS measures the pEWs of the lines and compares them to the model generated from the HARPS spectra, convolved to the respective resolution of the new spectra. In this case, the HARPS reference spectra are convolved from their resolution of 115 000 to the CARMENES resolution of 94 600. For each new star, the resulting parameters are calculated from the mean values of 100 determinations obtained from randomly shuffling and splitting each time the training (70 % of the sample, i.e. 45 stars) and testing groups (30 % of the population, i.e. 20 stars). This iterative process of multiple runs minimizes the possible dependence of the resulting parameters on how the stars from the HARPS dataset are split for training and testing in a single measurement.

We report parameter uncertainties derived by quadratically adding the dispersion of the resulting stellar parameters and the uncertainties of the machine learning models at this resolution after having taken into consideration the intrinsic uncertainties of the reference dataset parameters during the machine learning process. Since ODUSSEAS only relies on pEWs from HARPS spectra, this method is independent of synthetic spectra.

[1] `https://pypi.org/project/GISIC/`, developed by D. D. Whitten

3. Analysis

3.1. Stellar sample

Our stellar sample of benchmark stars consists of 18 M dwarfs, listed in Table 1. All stars are part of the CARMENES GTO sample and were observed with the CARMENES[2] instrument. They were chosen such that they have a high-S/N CARMENES spectrum with a S/N of at least 75 in the optical (VIS) and near-infrared (NIR) as stated in Passegger et al. (2018). **Their spectral types cover the range between M0.0 V to M5.5 V, following the typical CARMENES GTO distribution (see Marfil et al. 2021)**. The mean S/N over all spectrograph orders in the VIS and NIR for each spectrum is listed in Table 1. There is one exception to the S/N > 75 limit in the NIR, which is J13005+056 (GJ 493.1) due to its high rotational velocity. **All sample stars, except for the two high-rotation stars, show only minimal to no stellar activity.** Each star has literature photospheric parameters determined from at least three, and up to 12 other studies.

CARMENES operates with two highly stable fiber-fed spectrographs covering 5200–9600 Å in the VIS and 9600–17100 Å in the NIR wavelength ranges. The spectral resolutions are $R \approx$ 94 600 and 80 500, respectively (Quirrenbach et al. 2018; Reiners et al. 2018). The spectrographs are mounted on the Zeiss 3.5 m telescope at the Calar Alto Observatory in Spain. The prime goal of CARMENES is the search for Earth-sized planets in the habitable zones of M dwarfs.

Zechmeister et al. (2014), Caballero et al. (2016), and Passegger et al. (2019) presented a detailed description of the data reduction. After spectral extraction, each single spectrum is corrected for telluric lines by modeling a telluric absorption spectrum with the tool `Molecfit` (Kausch et al. 2014; Smette et al. 2015). The process is described in Nagel et al. (2021, submitted). The absorption telluric spectrum is subtracted from the observed spectrum resulting in a telluric-free spectrum that is then fed into the CARMENES radial velocity pipeline `serval` (SpEctrum Radial Velocity AnaLyser; Zechmeister et al. 2018). There, a high-S/N template spectrum is constructed for each star having at least five single spectra. This is a byproduct of the radial velocity calculation, where the radial velocities of the single spectra are derived from a least-square fit against the template. In this work, we apply our methods to these high-S/N templates of our 18 benchmark stars.

The stellar photospheric parameters we collected from literature for the benchmark stars are summarized in Table A.1. Although most benchmark stars have $\upsilon \sin i < 2$ km s^{-1} (Reiners et al. 2018), there are two stars with larger values: J07558+833 (12.1 km s^{-1}) and J13005+056 (16.4 km s^{-1}). These stars are useful to investigate the performance of the algorithms when dealing with higher rotational velocities. The literature values were derived with different methods. These methods include: interferometry to estimate the stellar radius and $T_{\rm eff}$ (Boyajian et al. 2012; Ségransan et al. 2003; von Braun et al. 2014; Berger et al. 2006; Newton et al. 2015), synthetic model fitting using BT-Settl models to determine $T_{\rm eff}$ (Gaidos et al. 2014; Lépine et al. 2013; Gaidos & Mann 2014; Mann et al. 2015) and $\log g$ (Lépine et al. 2013), empirical relations to derive stellar mass in the form of mass-luminosity relations (Mann et al. 2015; Khata et al. 2020; Boyajian et al. 2012; Berger et al. 2006; Ségransan et al. 2003), mass-magnitude relations (Maldonado et al. 2015), mass-radius relations (von Braun et al. 2014), mass-$T_{\rm eff}$ relations (Gaidos & Mann 2014; Gaidos et al. 2014), empirical relations to derive the stellar radius in the form of mass-radius relations (Maldonado et al. 2015) and $T_{\rm eff}$-radius relations (Gaidos & Mann 2014; Gaidos et al. 2014; Houdebine et al. 2019), pEW measurements to determine $T_{\rm eff}$ (Maldonado et al. 2015; Neves et al. 2014; Newton et al. 2015) and [Fe/H] (Maldonado et al. 2015; Neves et al. 2014; Gaidos et al. 2014; Mann et al. 2015), the definition of spectral indices such as the H2O-K2 index to estimate $T_{\rm eff}$ (Rojas-Ayala et al. 2012), as well as the combination of the H2O-K2 index with pEWs to derive [Fe/H] (Rojas-Ayala et al. 2012; Khata et al. 2020), the stellar radius and $T_{\rm eff}$ (Khata et al. 2020), and spectral curvature indices for the determination of $T_{\rm eff}$ (Gaidos & Mann 2014). Additionally, [Fe/H] was derived by using color-magnitude metallicity relations (Dittmann et al. 2016), atomic line strength relations (Gaidos & Mann 2014), and spectral feature relations (Terrien et al. 2015). Terrien et al. (2015) used K-band magnitudes and the Dartmouth Stellar Evolution Program (Dotter et al. 2008) to derive the stellar radius, whereas Mann et al. (2015) employed the Boltzmann equation with $T_{\rm eff}$ determined from synthetic model fits. Last, but not least, Houdebine et al. (2019) derived $T_{\rm eff}$ from photometric colors. For more details on the individual methods we refer to the descriptions in the corresponding works.

In this work, it is not our aim to analyze the variations from different techniques, data sets, and observations in the literature, however, we can compare the results of our methods to the literature as a whole. Therefore, we calculated the **median** over all literature values to reduce possible biases introduced by different data sets and methods. Thus, we presume the **median** to be to some degree more accurate than single literature values and consider the closeness of our values to the literature **median** as our quality measurement. The errors for the literature **median** come from the **root-mean-squared-errors (RMSE)** of the single measurements. Further, **the median** can be effective in smoothing extreme outliers, in case of contradicting literature values.

3.2. Different runs

We analyze our stellar sample with each method in three different runs. Each run is described thoroughly in the following.

3.2.1. Run A

For the first run, Run A, each team derived the stellar parameters with their methods as described in Sect. 2 without any particular restrictions. In this way, we directly compare the algorithms themselves and see how they perform compared to literature references.

3.2.2. Run B

In this run, all teams fixed the parameters $T_{\rm eff}$ and $\log g$ to common agreed values. They were calculated for each star as **median** values from the literature and the results from all teams from Run A (see Table A.1), hereafter referred to as overall **median**. We do this in order to increase the amount of individual measurements for each star, especially when there are not many literature values available. This leaves metallicity as the only free parameter to be determined. With this setting, we are able to gain insight into how the algorithms perform if they focus on only one parameter, and if this run gives any improvements compared to the previous run.

[2] Calar Alto high-Resolution search for M dwarfs with Exo-earths with Near-infrared and optical Échelle Spectrographs, `http://carmenes.caha.es`

Table 1. Selected sample of benchmark stars.

Karmn	Name	GJ	α (J2000) [hh:nmm:ss]	δ (J2000) [hh:nmm:ss]	Spectral type [(a)]	$v \sin i$ [(b)] [km s^{-1}]	mean S/N VIS	NIR
J00067−075	GJ 1002	1002	00:06:42.35	−07:32:46.4	M5.5 V	≤ 2.0	226	318
J00183+440	GX And	15A	00:18:27.04	+44:01:29.0	M1.0 V	≤ 2.0	993	1419
J04429+189	HD 285968	176	04:42:56.49	+18:57:12.1	M2.0 V	≤ 2.0	230	171
J05314−036	HD 36395	205	05:31:28.18	−03:41:10.5	M1.5 V	≤ 2.0	518	581
J07558+833	GJ 1101	1101	07:55:51.31	+83:22:55.7	M4.5 V	12.1	92	94
J09143+526	HD 79210	338A	09:14:20.14	+52:41:03.0	M0.0 V	≤ 2.0	459	597
J09144+526	HD 79211	338B	09:14:22.00	+52:41:00.7	M0.0 V	2.3	770	796
J10508+068	EE Leo	402	10:50:51.14	+06:48:16.6	M4.0 V	≤ 2.0	319	415
J11033+359	Lalande 21185	411	11:03:19.44	+35:56:52.8	M1.5 V	≤ 2.0	1112	1553
J11054+435	BD+44 2051A	412A	11:05:22.32	+43:31:51.6	M1.0 V	≤ 2.0	633	697
J11421+267	Ross 905	436	11:42:12.13	+26:42:11.0	M2.5 V	≤ 2.0	506	1080
J13005+056	FN Vir	493.1	13:00:32.55	+05:41:11.5	M4.5 V	16.4	89	62
J13457+148	HD 119850	526	13:45:45.67	+14:53:06.9	M1.5 V	≤ 2.0	941	1053
J15194−077	HO Lib	581	15:19:25.55	−07:43:21.7	M3.0 V	≤ 2.0	341	409
J16581+257	BD+25 3173	649	16:58:08.72	+25:44:31.1	M1.0 V	≤ 2.0	384	407
J17578+046	Barnard's star	699	17:57:47.67	+04:44:16.7	M3.5 V	≤ 2.0	976	1600
J22565+165	HD 216899	880	22:56:33.69	+16:33:08.0	M1.5 V	≤ 2.0	1140	1338
J23419+441	HH And	905	23:41:55.20	+44:10:14.1	M5.0 V	≤ 2.0	309	637

Notes. [(a)]Spectral types from Alonso-Floriano et al. (2015). [(b)]Projected rotational velocities from Reiners et al. (2018).

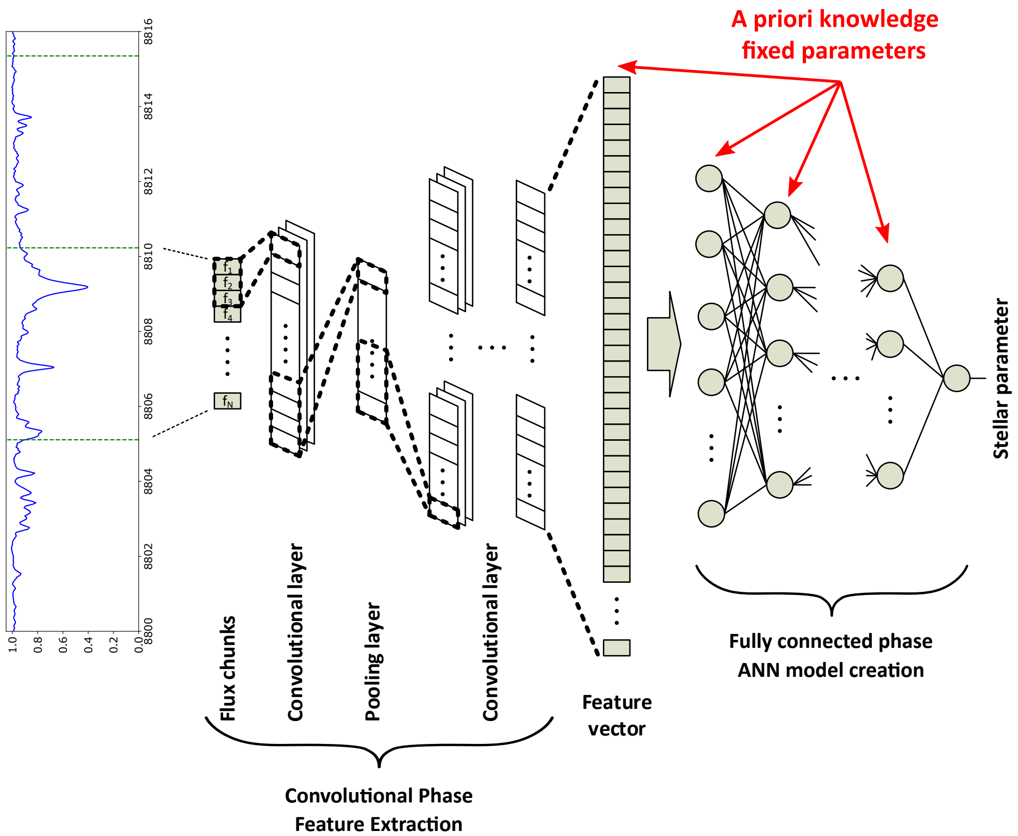

Fig. 1. Generic architecture for DL models in Passegger et al. (2020) and the different positions where we concatenated the values of the stellar parameters so that they could be fixed as needed in Run B.

The implementation is straightforward in both the Pass19-code and SteParSyn, since $T_{\rm eff}$ and $\log g$ can be kept fixed so that the downhill simplex and the MCMC chains, respectively, only explore a 1-D parameter space for metallicity. Therefore, there is only one minimum and one best-fit metallicity value. To assess the uncertainty in metallicity for the Pass19-code in the case of two fixed parameters, we follow the approach described in Passegger et al. (2016). Thus, we produce a set of

1400 synthetic spectra with uniformly distributed random parameters for $T_{\rm eff}$, $\log g$, and [Fe/H], broadened to the resolution of the CARMENES spectrographs. To simulate a S/N of ≈ 100, we add Poisson noise. The set of synthetic spectra is then sent through the Pass19-code keeping $T_{\rm eff}$ and $\log g$ fixed. This is done for the three different $v \sin i$ values of our stellar sample. The standard deviation of the mean deviation between input and derived output stellar parameter serves as an estimation for the uncertainty of the parameter.

The DL approach presented in Passegger et al. (2020) was not conceived to fix the stellar parameters as required in this run. First, we constructed models by restricting the training sample to synthetic spectra with fixed $T_{\rm eff}$ and $\log g$. This significantly reduced the training sample size and it also meant training new models for every single star. Although it is always possible to apply the DL learning process to small datasets, the results were not as good or trustworthy as the predictions we obtained with a more extensive grid. Instead, we try different architectures to take into account this prior knowledge about $T_{\rm eff}$ and $\log g$ in our DL models (see Fig. 1). In this way we are able to inject these conditions in the creation of DL models for predicting metallicity. The parameters that we fix are added at the end of the convolutional feature vector. We also consider the uncertainties of $T_{\rm eff}$ and $\log g$ from the overall **median**. For that purpose, we create two different sets of predictions. First, $T_{\rm eff}$ and $\log g$ are fixed without taking into account their uncertainties. Second, we generate 50 copies of the original flux, but with $T_{\rm eff}$ and $\log g$ extracted from a binomial distribution with the overall **median** of $T_{\rm eff}$ and $\log g$ as the center and the uncertainties of each parameter as the corresponding standard deviations. Finally, we aggregate the different predictions and create a probability density function using the Kernel Density Estimate (KDE; Rosenblatt 1956; Parzen 1962) for each benchmark star. The final result for metallicity is drawn from the maximum of the KDE, with its uncertainty being derived from the 1σ threshold.

For ODUSSEAS fixing any parameters is not possible for technical reasons, therefore this team cannot provide any metallicities for this run. Since the parameter determination process of ODUSSEAS correlates the pEWs of new spectra with the pEWs and reference stellar parameters of the HARPS dataset of same resolution, fixing $T_{\rm eff}$ of the new spectra, or even leaving out the $T_{\rm eff}$ prediction completely from the whole process, makes no difference to the derived [Fe/H] of new spectra.

3.2.3. Run C

In the last run, we standardize our methods by using the same wavelength regions, the same synthetic model spectra, and the same continuum normalization method. The analyzed wavelength regions are provided by the SteParSyn team and are summarized in Table 2. Because some of these regions fit well for hotter M dwarfs, but perform rather poorly for cooler spectral types, we manually select 35 of them that yield good fits over the whole spectral type range and use those in an additional Run C2. For the synthetic spectra, we use the PHOENIX-ACES model spectra grid as described in Sect. 2.1.1. We incorporate the same normalization method as the DL team, the `GISIC` routine (see Sect. 2.2.1). In the end, all teams are provided with normalized CARMENES and PHOENIX-ACES synthetic spectra for all wavelength regions from Table 2 to then run their individual algorithms to derive the stellar parameters $T_{\rm eff}$, $\log g$, and [Fe/H]. The line list employed by ODUSSEAS has a specific format of lower and upper wavelength boundaries for each absorption feature, which covers the range from 5300 to 6900 Å. Thus, ODUSSEAS can only use those normalized CARMENES spectral regions inside this range, in order to measure the respective absorption lines and determine the stellar parameters based on them. Their modified Run C will be designated as Run C* in the following.

Table 2. Analyzed wavelength regions for Runs C and C2. Wavelengths are given in vacuum.

Region		Run		Region		Run	
$\lambda_{\rm start}$ [Å]	$\lambda_{\rm end}$ [Å]	C	C2	$\lambda_{\rm start}$ [Å]	$\lambda_{\rm end}$ [Å]	C	C2
5867.60	5868.55	•		8437.44	8438.51	•	•
5923.32	5924.16	•		8452.76	8453.66	•	•
5979.80	5980.62	•		8469.03	8469.97	•	
6065.88	6066.77	•		8470.22	8471.29	•	
6066.73	6067.62	•		8515.87	8516.95	•	•
6086.46	6087.38	•		8516.99	8517.90	•	•
6127.49	6128.36	•		8549.98	8550.92	•	
6137.88	6138.78	•		8584.16	8585.10	•	
6138.94	6139.86	•		8613.68	8614.66	•	
6394.95	6395.81	•		8676.66	8677.61	•	•
6432.17	6433.08	•		8677.29	8678.23	•	•
6477.00	6477.84	•	•	8684.87	8685.86	•	•
6483.23	6484.09	•	•	8690.44	8691.63	•	•
6557.43	6558.32	•	•	8694.27	8695.18	•	•
6594.31	6595.20	•		8759.11	8760.09	•	
6600.51	6601.35	•		8826.06	8827.22	•	•
7050.91	7061.52	•	•	8840.36	8841.35	•	•
7084.05	7094.66	•	•	9012.65	9013.50	•	
7121.66	7132.27	•	•	9721.13	9722.16	•	•
7390.98	7391.91	•		9730.48	9731.70	•	•
7412.74	7413.67	•		9834.35	9835.38	•	
7491.21	7492.08	•		10343.25	10344.22	•	•
7497.72	7498.62	•		10381.35	10382.32	•	
7585.42	7586.36	•	•	10398.17	10399.14	•	•
7914.61	7915.50	•	•	10586.99	10588.10	•	•
8000.61	8001.72	•		10663.98	10665.14	•	
8076.92	8077.84	•	•	10777.31	10778.38	•	
8206.78	8207.65	•		11799.85	11800.97	•	
8398.70	8399.73	•	•	11886.71	11887.97	•	•
8403.31	8404.14	•	•	11952.26	11953.39	•	•
8414.15	8415.22	•	•	12814.42	12815.54	•	•
8418.80	8419.74	•	•	12922.87	12923.99	•	
8428.29	8429.39	•	•	15606.45	15607.80	•	
8436.73	8437.84	•	•	15719.19	15720.61	•	
8437.02	8447.62	•	•				

4. Results

All results for each star, run, and method are listed in Table C.1 and C.2. In the following we will discuss the results for each run. As discussed in Sect. 3.1, we compare our results to the literature **median**, assuming that the literature **median** represents accurate parameter values for each star, to investigate the consistency of our results over the different runs.

4.1. Run A

In Run A, all teams determined the stellar parameters with their methods without any restrictions. Figure 2 shows the comparison of our results with the literature **median** for each star. This gives a direct comparison of the performance of each method.

Effective temperature

It can be seen that all methods are mostly consistent with the literature **median** (purple dot) within the errors with only a few outliers. Overall, StePar Syn **reproduces the literature values best.** Compared to the literature **median**, the mean difference $\overline{\Delta T_{\rm eff}} = mean(T_{\rm eff}^{\rm our} - T_{\rm eff}^{\rm lit})$ is +20 K, meaning that, on average, SteParSyn derived $T_{\rm eff}$ 20 K hotter. Their results fall only three times outside of the error range, which is defined from the combined error bars of the literature **median** and the respective method for each star. SteParSyn is followed by the Pass19-code, which is on average 63 K hotter than the literature **median** and falls three times outside the error range. DL is on the hotter side as well, showing an average of 87 K larger than the literature **median**, and also falling three times outside the error. In contrast to the previous methods, ODUSSEAS consistently determines $T_{\rm eff}$ cooler than the other methods and on average 74 K cooler compared to the literature **median**. There are only two exceptions, where ODUSSEAS derives hotter $T_{\rm eff}$, GJ 1101 and GJ 493.1. Both stars have large $v \sin i$, which is the most likely reason for the larger $T_{\rm eff}$ values. Additionally, ODUSSEAS falls outside the error ranges five times.

Regarding large $v \sin i$, the other methods could determine values in good agreement with the literature **median** for these stars. An outlier is GJ 402, where all methods derive a $T_{\rm eff}$ of about 170 K higher than the literature **median**. Only ODUSSEAS is consistent with literature, but also here $T_{\rm eff}$ is lower compared to the other methods. In preparation of Run B, we calculate the **median** values between the literature and the results of Run A for each star. This overall **median** (red dot) is also plotted in Fig. 2 for comparison. As can be seen from the plot, the overall **median** is consistent with the literature **median** for all stars and sometimes differs only by a few K. Therefore, we consider the overall **median** as a benchmark value for each star and take it as a fixed value for our Run B.

Surface gravity

As described in Sect. 2.2.2, ODUSSEAS does not provide $\log g$. From the remaining methods, the Pass19-code performs best. The differences between the results and the literature **median** are almost always within 0.1 dex, and only once does their value fall outside the error range. On average, $\log g$ from the Pass19-code are 0.03 dex higher that the literature **median**, this difference is nearly negligible. The reason for this is likely the use of evolutionary models to constrain $\log g$. The DL method derives on average 0.04 dex lower $\log g$ than the literature **median** and lies five times outside of the error. For SteParSyn the $\log g$ has values 11 times outside the error range, on average being 0.12 dex higher than the literature **median**. In several cases, $\log g$ is significantly higher than the literature **median** and the other methods. The biggest outlier is GJ 338B, where SteParSyn derives a value of 0.53 dex larger than the literature **median**. A possible explanation for these high values could be either a still remaining degeneracy in the stellar parameter space or the synthetic gap (difference in feature distribution between synthetic and observed spectra). As shown in Marfil et al. (2021), SteParSyn retrieves tentatively higher $\log g$ values for the whole CARMENES GTO sample.

In the case of GJ 1002 the overall **median** for $\log g$ represents the **median** of all our Run A results, because there are no literature values for this star. In total, the literature and overall **median** differ less than 0.1 dex in all cases, except for GJ 493.1, which only has one literature value of 4.5 dex. Therefore, we excluded this star from the analysis of $\log g$ in this section.

Table 3. Analysis of Run A for $T_{\rm eff}$ (*top*), $\log g$ (*middle*), and [Fe/H] (*bottom*). The results for each team are provided in different columns, showing the mean difference to the literature **median**, the number of stars for which the results fall outside the error range, and the number of stars for which the results lie within 100 K and 0.1 dex, and outside 200 K and 0.2 dex of the literature **median**, for $T_{\rm eff}$, and $\log g$ and [Fe/H], respectively.

	Pass19-code	SteParSyn	DL	ODUSSEAS
$\overline{\Delta T_{\rm eff}}$ [K]	+63	+20	+87	–74
# o/s error	3	3	3	5
# <100 K	11	15	9	8
# >200 K	0	0	2	5
$\overline{\Delta \log g}$ [cgs]	+0.03	+0.12	–0.04	…
# o/s error	1	11	5	…
# <0.1 dex	15	6	10	…
# >0.2 dex	0	4	0	…
$\overline{\Delta [Fe/H]}$ [dex]	–0.02	–0.08	+0.23	–0.14
# o/s error	0	8	9	3
# <0.1 dex	7	7	3	6
# >0.2 dex	3	5	10	7

Metallicity

The bottom panel of Fig. 2 presents the results of all methods for [Fe/H]. Although this parameter will not be fixed in Run B, we calculate and plot an overall **median** for comparison. The Pass19-code performs best compared to the literature **median** and the other methods. On average, the metallicities are 0.02 dex lower than the literature **median**. All values agree with each other within their errors, for seven stars the results are within 0.1 dex difference to the literature **median**, and only for three stars the difference is larger than 0.2 dex. An explanation for this good performance can be the careful line selection of magnetically insensitive lines in the VIS and NIR. The use of multiple lines simultaneously also cancels out most of the effect coming from the synthetic gap, which especially affects DL. On this note, DL is doing worst when it comes to metallicity determination. The results for nine stars, which is half of our benchmark sample, lie outside the error range. For 10 stars the values differ by more than 0.2 dex from the literature **median**, while only three stars have differences less than 0.1 dex. On average, DL provides metallicities 0.23 dex higher than the literature **median**, tentatively deriving higher values for all but one star (GJ 205).

ODUSSEAS and SteParSyn determine tentatively lower values for metallicity, with $\overline{\Delta [Fe/H]}$ of –0.14 dex and –0.08 dex, respectively. For ODUSSEAS, three values fall outside the error range, while for SteParSyn it is eight. Six stars show differences of less than 0.1 dex with ODUSSEAS, and seven stars differ by more than 0.2 dex compared to the literature **median**. For SteParSyn, seven stars fall within 0.1 dex of the literature **median** and five stars outside of 0.2 dex.

All these numbers are summarized in Table 3 for better readability. Overall, the Pass19-code performs best in $\log g$ and [Fe/H] compared to the literature **median**. For $T_{\rm eff}$, SteParSyn would be the best choice.

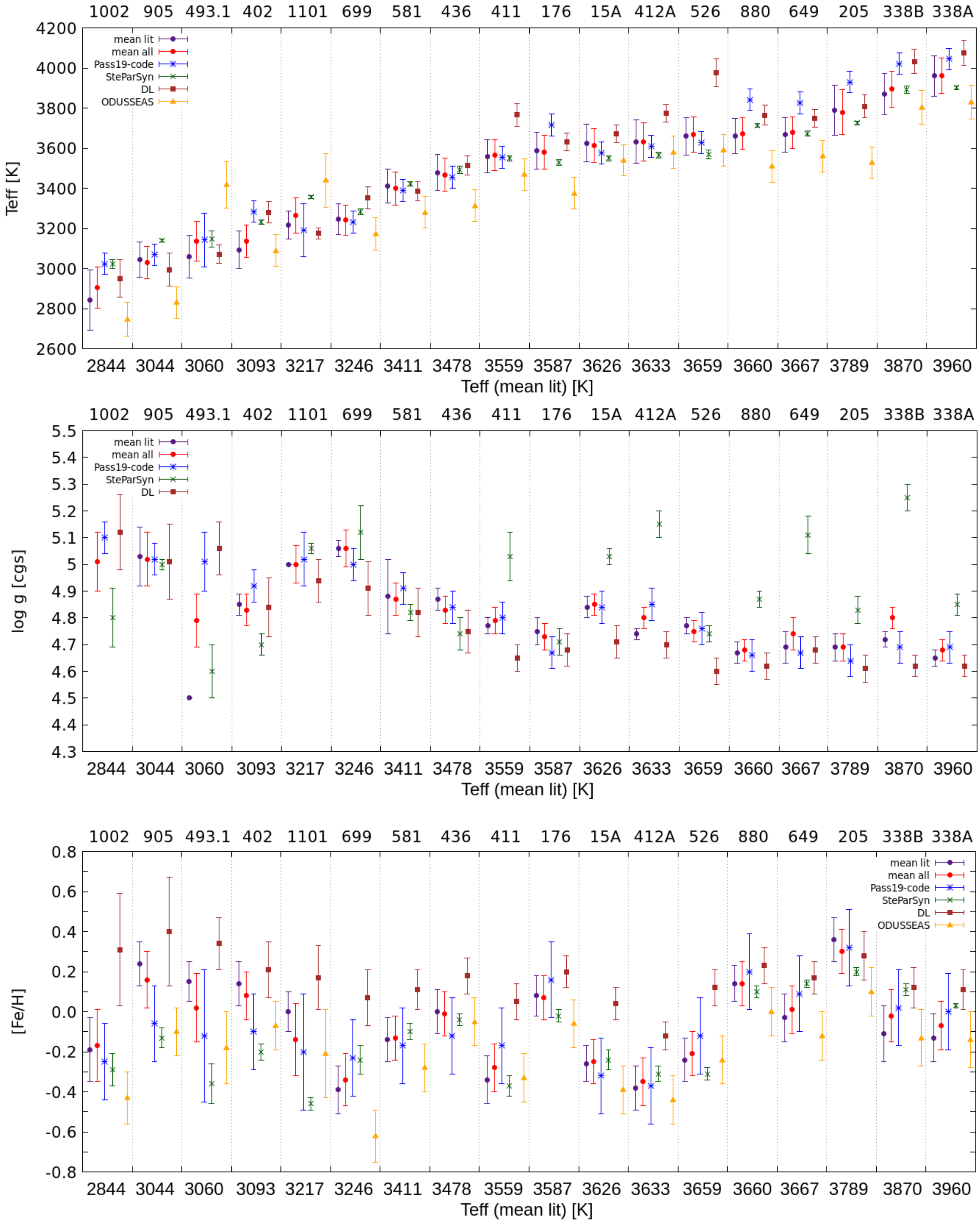

Fig. 2. Comparison of $T_{\rm eff}$ (*top*), log g (*middle*), and [Fe/H] (*bottom*) for the different methods in Run A. Each method is indicated with a different symbol and color. The **median** of all literature values and the **median** of literature+Run A are shown as purple and red dots, respectively. The x-axis indicates the $T_{\rm eff}$ **from the literature median, the top axis shows the** Gliese-Jahreiss (GJ) numbers for all sample stars. The stars are sorted by $T_{\rm eff}$ from the literature **median** to visualize possible trends.

4.2. Run B

For Run B, all teams (except ODUSSEAS, see Sect. 3.2.2) derived only [Fe/H], with $T_{\rm eff}$ and log g fixed to the **median** values determined from all teams in Run A and literature values. In Fig. 3 we show a comparison between our results and the litera-

Table 4. Same as Table 3, but for Run B, showing [Fe/H].

	Pass19-code	SteParSyn	DL
$\overline{\Delta[Fe/H]}$ [dex]	–0.06	–0.09	+0.27
# o/s error	11	4	7
# <0.1 dex	5	11	1
# >0.2 dex	4	4	12

ture **median**. Results from Run A are plotted in gray to illustrate the changes between the runs. We can see from this plot that fixing $T_{\rm eff}$ and $\log g$ does not improve the metallicities derived with the Pass19-code and DL. For both methods the differences to the literature **median** increased. This can be explained by looking at the results for $T_{\rm eff}$ from Run A. If the temperatures were further away from the overall **median**, which was used to fix this parameter in Run B, then the deviation in metallicity in Run B is larger than in Run A. Some correlation with $\log g$ can also be found in some cases. Since in Run B the parameter determination is reduced to a 1-D problem, there are no local minima anymore and only one best value for metallicity. If the fixed values $T_{\rm eff}$ and $\log g$ deviate from the best fitting values found in Run A, the deviation in metallicity will consequently increase as well. Therefore, there will be no improvement regarding metallicity, unless the other parameters $T_{\rm eff}$ and $\log g$ can be chosen freely too. We performed the same analysis of the results as for Run A, the numbers are summarized in Table 4.

On the other hand, fixing parameters slightly improved the metallicities derived by SteParSyn. The number of stars outside the error range decreased, whereas the number within 0.1 dex increased. A good example here is GJ 338B. From Run A, $T_{\rm eff}$ is close to the literature **median**, but $\log g$ is far off. By fixing $\log g$ the metallicity improves and moves closer to the literature **median**. This run suggests that there is a dependency on the stellar synthetic spectra used in the analysis. DL and the Pass19-code both rely on the PHOENIX-ACES model spectra, but do not show any improvements towards literature values, whereas SteParSyn incorporated BT-Settl model atmospheres. Therefore, the next step is for all methods to use the same synthetic models.

4.3. Run C and C2

In Run C all teams incorporate the same normalized PHOENIX-ACES model spectra and derive parameters from the same wavelength regions. As mentioned in Section 3.2.3, we carry out an additional Run C2, using a subset of 35 wavelength regions from Run C, but otherwise identical to Run C. Figure 4 presents a comparison of stellar parameters between Run A and Run C2. A comparison between Run C and C2 is shown in Fig. B.1. Table 5 summarizes the statistics of Run C and C2.

Effective temperature

We compared our results from Run C2 to those derived in Run A. Figure 4 shows that the stellar parameters do not improve from Run A to Run C2. This is most evident for $T_{\rm eff}$, where the 16 stars for SteParSyn, 13 stars for DL, and 13 stars for the Pass19-code show larger deviations to the literature **median** than in Run A. Analyzing Runs C and C2 shows that stellar parameters derived with DL agree better with the literature **median** in Run C2 than in Run C. This means an improvement towards Run C2, with all stars being closer to the literature **median** in $T_{\rm eff}$ compared to Run C. For SteParSyn, the results from Run C and C2 are a bit more ambiguous. Half of the sample stars are closer to the literature **median** in Run C for $T_{\rm eff}$, the other half in Run C2. Run C exhibits a larger $\overline{\Delta T_{\rm eff}}$ of +58 K compared to Run C2 ($\overline{\Delta T_{\rm eff}}$ = +13 K). Concerning the Pass19-code, Run C is clearly better than Run C2, giving values closer to the literature **median**. The mean difference, $\overline{\Delta T_{\rm eff}}$, amounts to +73 K, with all values being within the error range for all stars.

Table 5. Same as Table 3, but for Run C/C2.

	Pass19-code	SteParSyn	DL
$\overline{\Delta T_{\rm eff}}$ [K]	+73/+113	+58/+13	+221/+152
# o/s error	0/3	10/9	6/4
# <100 K	12/7	8/7	2/6
# >200 K	0/3	4/4	7/3
$\overline{\Delta \log g}$ [cgs]	–0.06/+0.03	–0.01/–0.01	+0.01/+0.07
# o/s error	3/4	9/11	3/2
# <0.1 dex	11/11	5/1	9/13
# >0.2 dex	4/1	6/9	3/1
$\overline{\Delta[Fe/H]}$ [dex]	+0.31/–0.17	–0.07/–0.16	–0.09/–0.07
# o/s error	5/4	7/11	2/1
# <0.1 dex	4/2	4/2	9/8
# >0.2 dex	9/9	6/10	7/5

Surface gravity

Compared to Run A, the Pass19-code and SteParSyn show similar results as in $T_{\rm eff}$, with no improvement from Run A to Run C2. On the other hand, 11 stars with $\log g$ derived from DL are closer to the literature **median** in Run C2 than in Run A. For Run C, the results from DL show an improvement towards Run C2 for eleven stars. Only six stars from SteParSyn lie closer to the literature **median** in Run C2 than in Run C, which clearly favours the results from Run C in this case. For the Pass19-code, the results derived in Run C are tentatively lower than for Run C2. Especially for hotter stars, this means that Run C is closer to the literature **median**, as can be seen from Fig. B.1.

Metallicity

Similar to $T_{\rm eff}$ and $\log g$ the results in metallicity for the Pass19-code and SteParSyn are closer to literature in Run A than in Run C2. However, DL presents a slight improvement, deriving values which are closer to the literature **median** for 13 stars in metallicity. Using multiple wavelength ranges instead of only one range, as it was done in Runs A and B, appears to trigger this improvement. Looking at Run C, eight of 18 stars have better values in Run C2 for DL. As for $\log g$, only a small number of five stars shows better results in Run C2 than in Run C for SteParSyn. This could be explained by the fact that SteParSyn is optimized for the wavelength ranges in Run C, that are originally used by the method. Analyzing a subset of these ranges, as it was done in Run C2, does not improve the results. In Run C, the Pass19-code derives consistently too high metallicities, especially for cooler M-dwarfs (see Fig. B.1, bottom panel). The selected wavelength

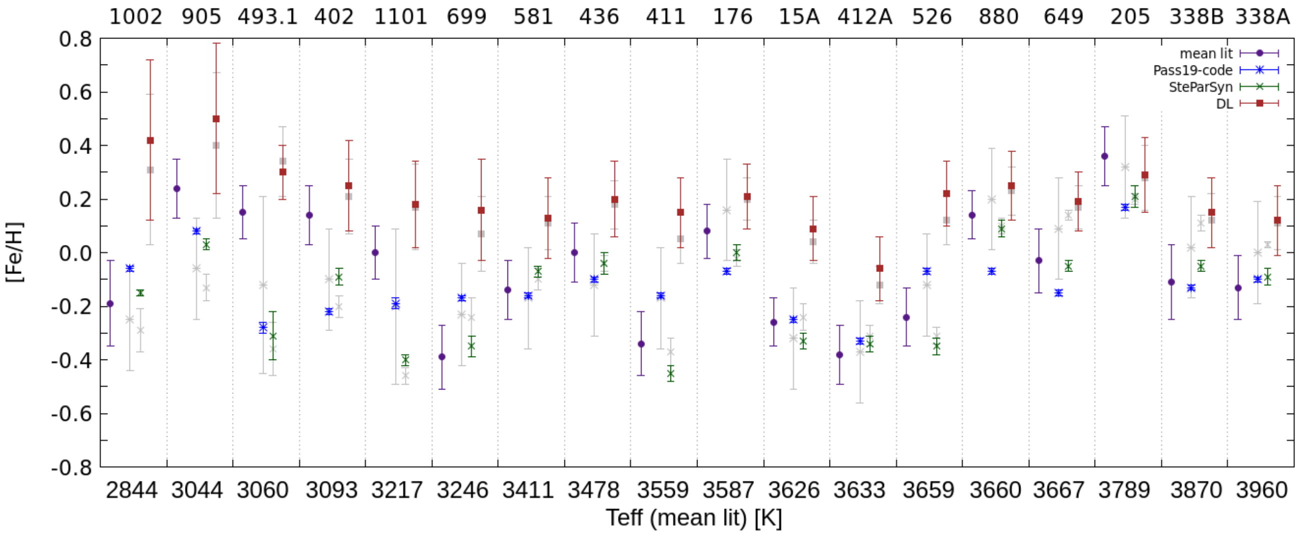

Fig. 3. Comparison of [Fe/H] for the different methods in Run B. Each method is indicated with a different symbol and color. The gray symbols indicate the results from Run A for comparison. The **median** of all literature values is shown as purple dots. The x-axis indicates T_{eff} **from the median literature, the top axis shows the** Gliese-Jahreiss (GJ) numbers for all sample stars.

ranges of Run C2, however, clearly improve the metallicity determination.

Comparing the numbers Table 5 to those of Runs A and B also illustrates the differences and the better performance of Run A. It can be seen, for example, that the number of stars with differences larger than 200 K and 0.2 dex from the literature **median** significantly increased in Run C2. An exception here is DL, which is able to decrease those numbers in Run C2.

5. Discussion

We compare the results from all runs for each method in order to analyze which run gives the best results, i.e. in which run can we find the most stars with values closest to the literature **median**. For example in case of the Pass19-code, each star has four determined [Fe/H] values, derived in Run A, B, C, and C2. For each value, we calculate the difference from the literature **median** and find the minimum difference, for example for Run A. This particular star then counts toward Run A. The procedure is repeated for all stars, and for all three stellar parameters. As a consequence, the sum over all runs for each stellar parameter is always 18. This way, we can assess which run performed best for each stellar parameter.

Figure 5 shows this number of stars for all parameters and methods. From that, we can see that for T_{eff}, all methods but ODUSSEAS perform best in Run A. Generally, the Pass19-code works better in Run A than in the other runs. Run C and C2 show a similar performance for $\log g$ and [Fe/H].

DL results get closer to the literature **median** for [Fe/H] and $\log g$ in Runs C and C2, respectively. Since the continuum normalization and synthetic spectra are the same as the ones used in Runs A and B, the only explanation for the improvement is the different wavelength regions. DL can determine [Fe/H] and $\log g$ significantly better by taking into account more wavelength regions than just the one between 8800–8835 Å, although this region seems to work well for T_{eff} alone.

Similar to the Pass19-code, SteParSyn is in general performing best in Run A, with good results declining towards Run C and C2. An exception is metallicity, which is best in Run B, directly followed by Run A. This indicates that the metallicity determinations with SteParSyn could be improved by taking independent estimates and fixing T_{eff} and $\log g$. The stellar parameters derived in Run C and C2 show tentatively larger deviations from the literature **median** than Run A, which is probably due to the different synthetic spectra used. This implies that SteParSyn is optimized for the analysis of their selected wavelength ranges with BT-Settl models.

5.1. ODUSSEAS' Run C*

As described in Sect. 3.2.3, ODUSSEAS could use only the bluest wavelength ranges provided for Run C/C2. Therefore, the results cannot be directly compared to the other methods. However, it is possible to assess the performance of the algorithm itself. Since ODUSSEAS does not rely on synthetic model spectra, and a different continuum normalization does not affect the measurement of pEWs, the only difference between Run A and C* is the choice of the wavelength ranges. In the bottom right panel of Fig. 5, it can be seen that ODUSSEAS derives the best metallicities for all 18 stars in Run A. For T_{eff}, Run A shows tentatively lower values compared to other methods and the literature. However, their modified Run C* gives significantly better results. From this, we can conclude that the wavelength ranges used in Run A are very good for deriving metallicity, but seem to be less sensitive to T_{eff}. On the other hand, the ranges used in Run C* appear to be more appropriate when it comes to temperature determination.

5.2. Comparison with interferometry

For 11 stars we can compare our results for T_{eff} and $\log g$ to independent measurements coming from interferometry (Boyajian et al. 2012; von Braun et al. 2014; Newton et al. 2015; Rabus et al. 2019). Boyajian et al. (2012) and von Braun et al. (2014) used Hipparchos parallaxes (van Leeuwen 2007) to convert the angular stellar diameter, Θ_{LD}, to stellar radius via $\Theta_{LD} = 2 \cdot R/d$, whereas Rabus et al. (2019) used Gaia DR2 data (Gaia Collabo-

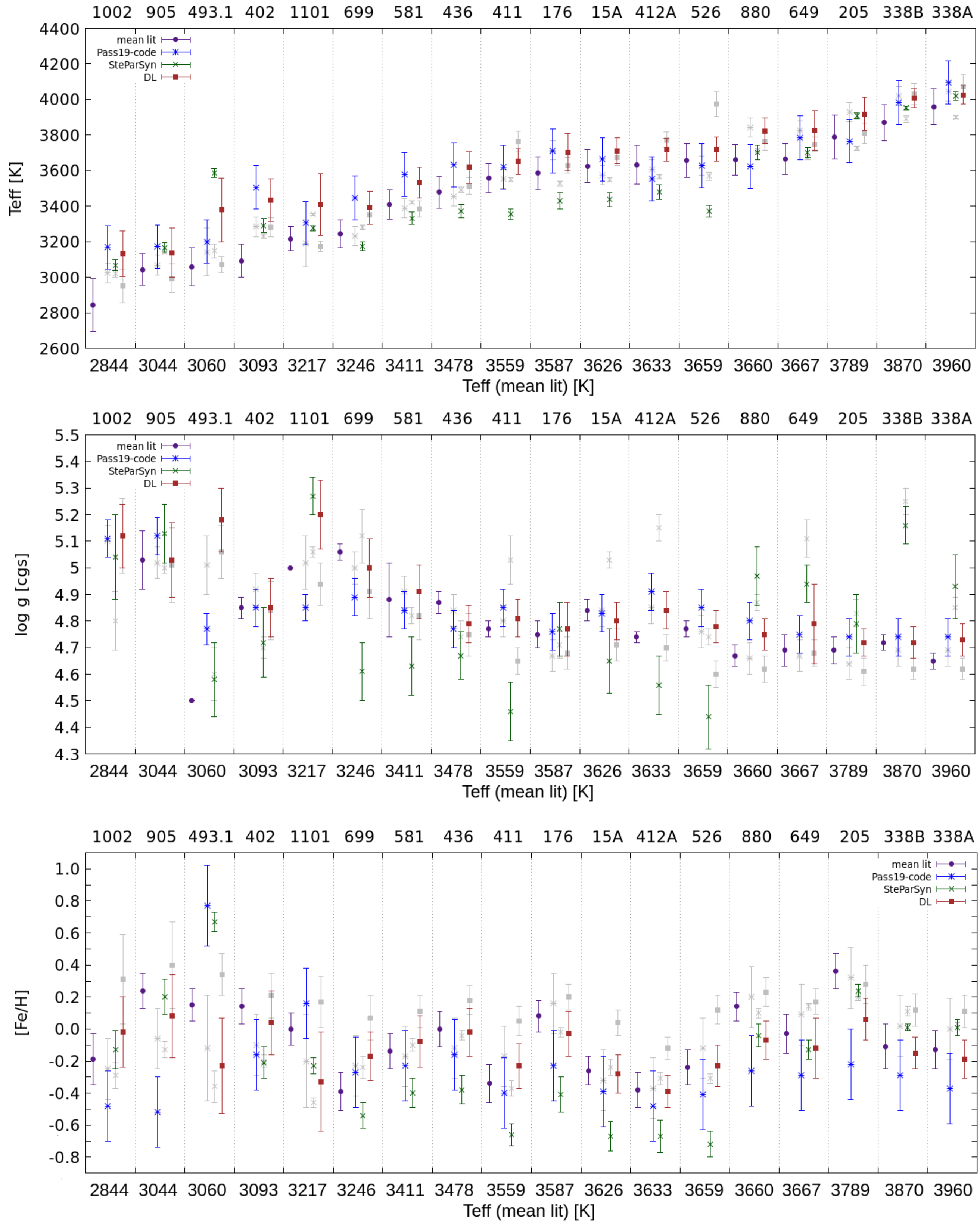

Fig. 4. Comparison of $T_{\rm eff}$ (*top*), log g (*middle*), and [Fe/H] (*bottom*) for the different methods in Run C2. Each method is indicated with a different symbol and color. The gray symbols indicate the results from Run A for comparison. The **median** of all literature values is shown as purple dots. The x-axis indicates $T_{\rm eff}$ **from the literature median, the top axis shows the** Gliese-Jahreiss (GJ) numbers for all sample stars.

ration et al. 2018). Newton et al. (2015) collected interferometric radii from the literature. When there was more than one measurement per star, they calculated the weighted mean. $T_{\rm eff}$ can be derived from the Stefan-Boltzmann law, $T_{\rm eff} = 2341 \cdot (F_{bol}/\Theta_{LD})^{1/4}$,

Table 6. Mean difference between our results and interferometric literature values, $\overline{\Delta T_{\rm eff}}$ / $\overline{\Delta \log g}$, standard deviation of the mean difference, std. dev., and Pearson correlation coefficients, r_P, for $T_{\rm eff}$ / $\log g$.

	Pass19-code	SteParSyn	DL	ODUSSEAS
Run A				
$\overline{\Delta T_{\rm eff}}$ / $\overline{\Delta \log g}$	+60 / −0.04	−18 / +0.13	+111 / −0.11	−123 / –
std. dev.	80 / 0.04	75 / 0.18	123 / 0.06	113 / –
r_P	0.950 / 0.928	0.923 / 0.300	0.817 / 0.881	0.815 / –
Run C				
$\overline{\Delta T_{\rm eff}}$ / $\overline{\Delta \log g}$	+38 / −0.08	−5 / −0.09	+169 / −0.06	−67 / –
std. dev	90 / 0.09	106 / 0.25	66 / 0.12	112 / –
r_P	0.901 / 0.795	0.922 / −0.480	0.939 / −0.021	0.815 / –
Run C2				
$\overline{\Delta T_{\rm eff}}$ / $\overline{\Delta \log g}$	+72 / −0.01	−68 / −0.07	+120 / +0.001	– / –
std. dev.	105 / 0.09	130 / 0.29	62 / 0.05	– / –
r_P	0.839 / 0.741	0.875 / −0.536	0.948 / 0.929	– / –

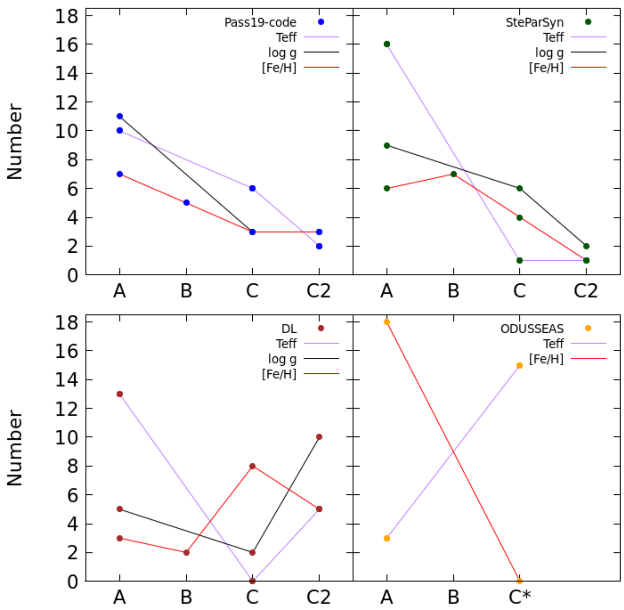

Fig. 5. Number of stars for each run where the stellar parameter lies closest to the literature **median**. Each method is shown in a separate panel. Note that ODUSSEAS does not provide $\log g$, and their Run C* differs from Run C and C2 due to restriction in the method itself.

when the bolometric flux F_{bol} is known. Boyajian et al. (2012) and von Braun et al. (2014) produced spectral energy distributions (SEDs) using flux-calibrated photometry from the literature. Rabus et al. (2019) estimated F_{bol} by integrating the flux from synthetic photometric flux points using PHOENIX-ACES synthetic spectra. Newton et al. (2015) presented interferometric $T_{\rm eff}$ from Mann et al. (2013b) and updated $T_{\rm eff}$ for three stars following their approach. Mann et al. (2013b) determined F_{bol} by comparing the measured fluxes from observed visual and NIR spectra, incorporating BT-Settl synthetic models to cover wavelength gaps in the spectra, to photometric fluxes using a correction factor to adjust the overall flux level. From the stellar radius and mass, $\log g$ can be calculated via $g = GM/R^2$. This requires the stellar mass, which cannot be measured from interferometry. Therefore, Boyajian et al. (2012) and Rabus et al. (2019) used the K-band mass-luminosity relation from Henry & McCarthy (1993), and from Benedict et al. (2016) and Mann et al. (2019), respectively. von Braun et al. (2014) determined stellar mass by deriving a mass-radius relation from the results from Boyajian et al. (2012). Although $T_{\rm eff}$ can be derived independently from interferometry, $\log g$ can be seen as semi-independent, since it involves interferometric radii, but also empirical mass-radius or mass-luminosity relations. Therefore, such quasi-interferometric $\log g$ possesses tentatively higher accuracy than $\log g$ derived from, e.g., synthetic model fits alone, and, thus, can be used as a reliable comparison.

A comparison plot is shown in Fig. 6. We calculated the mean difference, standard deviation, and Pearson correlation coefficient (r_P) between our results and the literature, presented in Table 6. A good consistency between the samples would result in a low mean difference and standard deviation, as well as a Pearson correlation coefficient close to 1.

For Run A, SteParSyn agrees quite well with interferometry in $T_{\rm eff}$ with $\overline{\Delta T_{\rm eff}}$ = −18 K, followed by the Pass19-code, which gives slightly hotter values with $\overline{\Delta T_{\rm eff}}$ = +60 K compared to interferometry. Also DL is on the hotter side ($\overline{\Delta T_{\rm eff}}$ = +111 K), whereas ODUSSEAS, as mentioned before, derived tentatively cooler temperatures ($\overline{\Delta T_{\rm eff}}$ = −123 K). In Run C (which corresponds to Run C* for ODUSSEAS), temperatures from DL and ODUSSEAS are shifted more towards the hotter side, bringing ODUSSEAS closer to the interferometric values ($\overline{\Delta T_{\rm eff}}$ = −67). This is the same behavior as in Fig. 5. In contrast, the Pass19-code provides cooler temperatures, but still mostly consistent with those from interferometry ($\overline{\Delta T_{\rm eff}}$ = +38). SteParSyn performs similar to Run A, however with some larger spread at low and high temperatures, which is represented in a higher standard deviation compared to Run A. This is similar for Run C2, where SteParSyn again yields some cooler temperatures compared to Run C. Overall, SteParSyn does best in Run A, where it shows the lowest standard deviation and highest r_P, similarly to the Pass19-code. On the other hand, DL shows a better 1:1 relation in Run C2, represented by the larger r_P (see Table 6). This indicates that the selected wavelength ranges in Run C2 lead to an improvement in the results, although there seems to

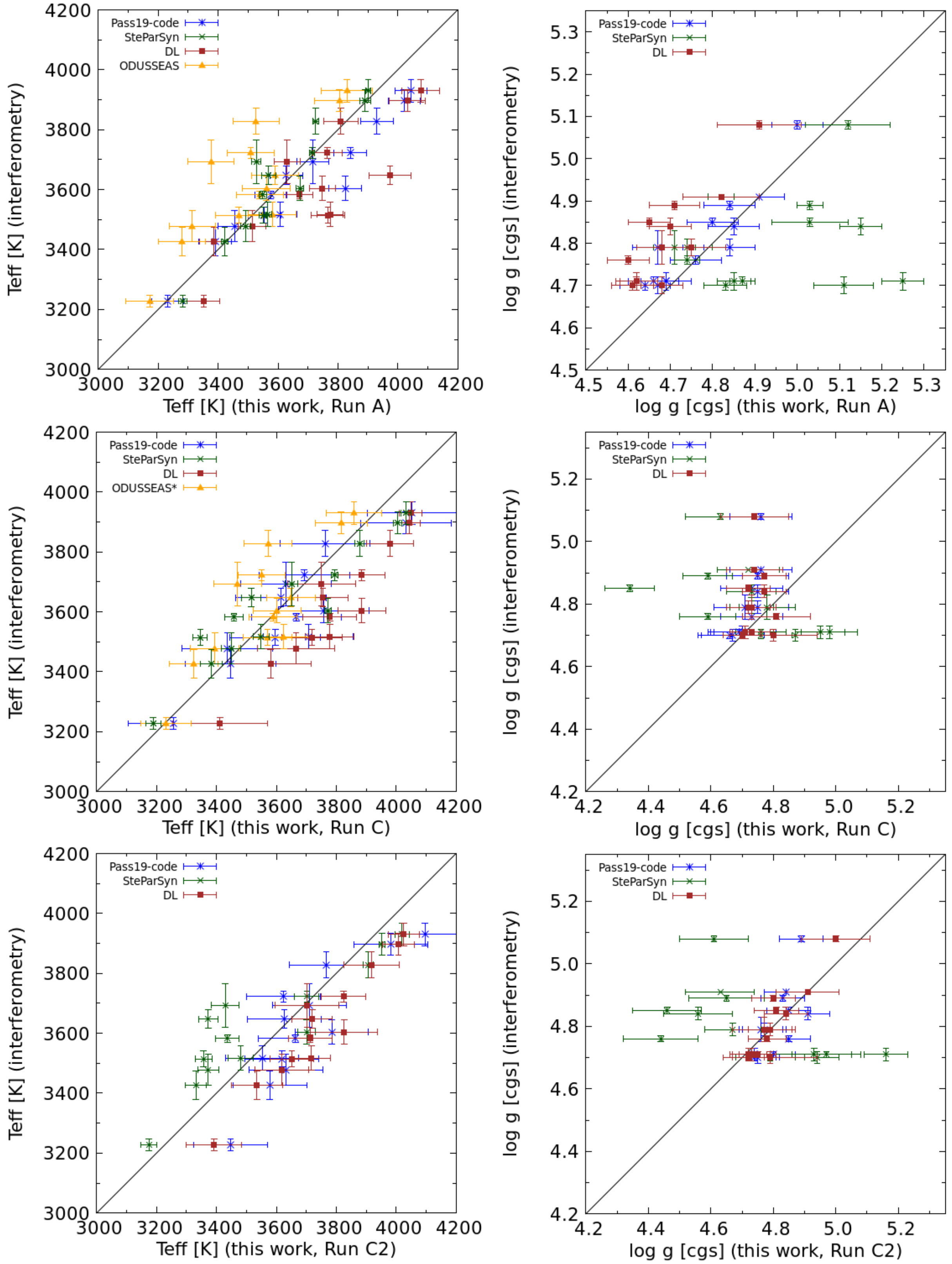

Fig. 6. Comparison of our results for T_{eff} (*left column*) and $\log g$ (*right column*) from Run A (*top*), Run C (*middle*), and Run C2 (*bottom*) with interferometric values from Boyajian et al. (2012), von Braun et al. (2014), Newton et al. (2015), and Rabus et al. (2019). If more than one value exists for a star in literature, we plot the **median** with the **RMSE** for better readability. Note that results from ODUSSEAS in Run C correspond to their Run C*. The black line indicates the 1:1 relationship.

be a general offset towards hotter temperatures compared to interferometry. Again, this is also illustrated in Fig. 5.

For $\log g$, the Pass19-code is closest to the quasi-interferometric $\log g$ for all runs, which is most likely due to the use of evolutionary models in the method. However, the values are a little lower than the literature on average. The smallest standard deviation and highest r_P is presented by Run A, as for $T_{\rm eff}$. The results given by the Pass19-code in Run C are systematically lower than interferometric ones, but they improve in Run C2. DL follows the relation in general, but those results are lower as well. A great improvement is shown for DL in Run C and even more in Run C2, where $\overline{\Delta \log g}$ decreases to as low as +0.001 dex, which can be clearly attributed to the use of multiple wavelength ranges. Overall, the values from SteParSyn show a large spread with a high standard deviation and low r_P. The spread is persistent in Run C and C2, although the mean difference of all SteParSyn values moves closer to the 1:1 relation.

Overall, this comparison is very similar to the comparison of the literature **median** and yet another indication that, for most stars, Run A gives very good results compared to the literature **median**, with the exception of DL, where the analysis of multiple wavelength ranges results in better measurements of $\log g$ and a higher correlation in $T_{\rm eff}$.

A similar analysis could be done for metallicity, should there be independent measurements from a hotter FGK-type binary companion.

5.3. Consistency between methods

Finally, we analyze the consistency between the methods we employed in this experiment. A statistical analysis similar to Table 6 is presented in Table 7. We compare each method to each of the other methods to reveal trends. We plotted all these combination for further visualization in Figs. B.3-B.5 for $T_{\rm eff}$, $\log g$, and [Fe/H] for Run A; [Fe/H] for Run B is presented in Fig. B.6, $T_{\rm eff}$, $\log g$, and [Fe/H] for Run C in Figs. B.7-B.9, and $T_{\rm eff}$, $\log g$, and [Fe/H] for Run C2 in Figs. B.10-B.12.

Run A

For $T_{\rm eff}$, SteParSyn and the Pass19-code show the best correlation with r_P of 0.974, also being the only methods with a spread, i.e., standard deviation of less than 100 K between them. Both methods compare well with DL, although DL shows higher deviations at higher $T_{\rm eff}$. As illustrated in previous comparisons (see Figs. 2 and 6) ODUSSEAS derives much lower $T_{\rm eff}$ values, on average 130 K cooler compared to the other methods.

The Pass19-code and DL correlate quite well in $\log g$, whereas SteParSyn exhibits a large spread compared to both other methods (see also Figs. 2 and 6).

SteParSyn, the Pass19-code, and ODUSSEAS are in good agreement in [Fe/H], having little mean differences and a large r_P. The direct comparison between SteParSyn and ODUSSEAS displays a slightly larger spread and therefore a smaller r_P. DL derives much higher [Fe/H] values, which are on average 0.3 dex more metal-rich compared to the other methods. This behaviour can also be seen in Fig. 2.

Run B

As in Run A, SteParSyn and the Pass19-code are most consistent with a small mean difference of only 0.03 dex, and a similar standard deviation. However, the values are not so well correlated, exhibiting a smaller r_P of 0.691 (compared to $r_P = 0.804$ in Run A). DL performs even worse than in Run A, with $\overline{\Delta[\mathrm{Fe/H}]}$ being +0.33 dex and +0.36 dex compared to the Pass19-code and SteParSyn, respectively.

Run C and C2

As described in Sect. 3.2.3 and 5.2, ODUSSEAS could only use wavelength ranges between 5300 and 6900 Å. Therefore, a direct comparison of results from this Run C* with the results from Runs C and C2 from the other methods is not meaningful. However, we included ODUSSEAS in our analysis here for completeness and to visualize relative changes between Runs C and C2 for the other methods.

A comparison of Run C with C2 reveals only minor differences for $T_{\rm eff}$. It can be seen from the numbers in Table 7 and the plot in Fig. B.7 that DL performs a bit better in Run C2, where it derives slightly lower $T_{\rm eff}$ and therefore exhibits a smaller $\overline{\Delta T_{\rm eff}}$ compared to the other methods. SteParSyn and the Pass19-code show a somewhat smaller $\overline{\Delta T_{\rm eff}}$ and spread in Run C. This is also clearly shown by the comparison of the Pass19-code and ODUSSEAS in Run C and C2. Since for both runs the Pass19-code is compared to Run C* of ODUSSEAS, relative improvements between the runs are revealed. Overall, it can be said that the correlation coefficients for $T_{\rm eff}$ are a little larger in Run C compared to Run C2.

On the other hand, there is almost no correlation in $\log g$ for any of the methods. The only notable improvement toward Run C2 is given between the Pass19-code and DL, which present a little higher correlation and smaller $\overline{\Delta \log g}$ in Run C2. This can be attributed to an improvement of DL in Run C2, as already described in in Sect. 4 and 5.2.

A clear difference can be seen for [Fe/H] between Run C and C2 (see Figs. B.9 and B.12). In Run C all methods appear more separated, also having higher mean differences. They determine, in general, higher [Fe/H] values, especially the Pass19-code, which is depicted in Fig. B.1 as well. For Run C2 the derived values are more metal-poor, causing the results to move closer together for all methods. This reduces the mean differences, although the spread and correlation coefficient are not improved necessarily for all methods.

Overall, the largest correlation between the methods for all stellar parameters is found in Run A, however we can see some trends. The determination of $\log g$ with SteParSyn is in general not very well constrained: it has a large mean difference and spread compared to the other methods. The correlation increases toward Run C and C2, however the reason for this is not clear. ODUSSEAS shows the best consistency in $T_{\rm eff}$ in Run C*, when comparing to Run C of the other methods, with the smallest mean difference and a slightly better r_P than in Run A. For [Fe/H], Run A as well as Run C2 show small $\overline{\Delta[\mathrm{Fe/H}]}$ in general, however in Run C2 the spread is larger and r_P is smaller, therefore, the consistency in [Fe/H] is better in Run A. Only DL is able to improve the consistency toward Run C/C2, with negligible differences between C and C2. Consequently, r_P increases and $\overline{\Delta[\mathrm{Fe/H}]}$ decreases.

Possible improvements to increase the consistency are very specific to the method, there is no general recipe. For DL, the values in [Fe/H] are tentatively too high, using more wavelength ranges can improve [Fe/H] in Run C and C2. However, for the determination of $T_{\rm eff}$ one wavelength range serves well. ODUSSEAS derives consistently lower $T_{\rm eff}$, the use of different wavelength ranges, as it was done in Run C*, would increase the consistency with our other methods. The analysis in this section confirms our findings in Sect. 4 and 5.2.

Table 7. Statistical analysis between our methods for Runs A, B, C, and C2. Mean difference, $\overline{\Delta}$, standard deviation of the mean difference, std. dev., and Pearson correlation coefficients, r_P, for $T_{\rm eff}$, $\log g$, and [Fe/H] for all combinations of our methods. Note that the results of Run C* from ODUSSEAS technically cannot be compared to the other methods, but are shown here for completeness. For more details, see Sect. 5.3.

Run A	$T_{\rm eff}$			$\log g$			[Fe/H]		
	$\overline{\Delta}$	std. dev.	r_P	$\overline{\Delta}$	std. dev.	r_P	$\overline{\Delta}$	std. dev.	r_P
Pass-19 – SteParSyn	+43	96	0.974	−0.07	0.24	−0.074	+0.06	0.12	0.804
Pass-19 – DL	−24	120	0.937	+0.06	0.06	0.930	−0.25	0.16	0.468
Pass-19 – ODUSSEAS	+136	178	0.829	...	...	...	+0.12	0.11	0.821
SteParSyn – DL	−67	137	0.947	+0.14	0.26	−0.189	−0.30	0.21	0.195
SteParSyn – ODUSSEAS	+93	132	0.882	...	...	...	+0.07	0.16	0.639
DL – ODUSSEAS	+161	176	0.860	...	...	...	+0.37	0.15	0.530
Run B									
Pass-19 – SteParSyn	...	...	...	...	...	...	+0.03	0.13	0.691
Pass-19 – DL	...	...	...	...	...	...	−0.33	0.10	0.643
SteParSyn – DL	...	...	...	...	...	...	−0.36	0.17	0.411
Run C									
Pass-19 – SteParSyn	+15	123	0.900	−0.06	0.22	0.177	+0.38	0.38	0.346
Pass-19 – DL	−148	86	0.954	−0.06	0.13	0.339	+0.40	0.30	0.359
Pass-19 – ODUSSEAS*	+74	152	0.844	...	...	...	+0.57	0.45	−0.222
SteParSyn – DL	−163	101	0.931	+0.00	0.22	0.285	+0.02	0.29	0.451
SteParSyn – ODUSSEAS*	+59	180	0.777	...	...	...	+0.19	0.36	0.235
DL – ODUSSEAS*	+222	114	0.901	...	...	...	+0.17	0.17	0.757
Run C2									
Pass-19 – SteParSyn	+100	172	0.782	+0.05	0.24	0.187	−0.01	0.31	0.555
Pass-19 – DL	−39	87	0.941	−0.03	0.13	0.547	−0.10	0.33	−0.162
Pass-19 – ODUSSEAS*	+114	194	0.715	...	...	...	+0.09	0.46	−0.495
SteParSyn – DL	−140	135	0.869	−0.08	0.25	0.227	−0.09	0.33	0.392
SteParSyn – ODUSSEAS*	+14	181	0.772	...	...	...	+0.10	0.42	0.087
DL – ODUSSEAS*	+153	144	0.848	...	...	...	+0.19	0.18	0.704

Notes. (*) corresponding to Run C*.

6. Summary and conclusions

We applied four different methods, including synthetic spectral fitting, pEW measurements, and machine learning, to derive the stellar parameters $T_{\rm eff}$, $\log g$, and [Fe/H] for a sample of 18 M dwarfs from high-resolution and high-S/N CARMENES spectra. Our analysis consisted of four different Runs: Run A allowed each team to use their method without restrictions, Run B fixed $T_{\rm eff}$ and $\log g$ to derive only [Fe/H], and in Run C and C2 all teams incorporated the same synthetic model spectra, continuum normalization method, and wavelength ranges.

Although we provide several new stellar parameters for our sample, it was not our goal to measure more precise or accurate parameters of these stars in the context of a catalog, but to identify and understand discrepancies in the parameters between our groups, i.e. the different parameter determination methods, with the aim to minimize these discrepancies in order to make a step forward to more consistent parameter determinations. At the beginning of this experiment we expected that a standardization of underlying synthetic models, wavelength ranges, and reducing the dimension of the parameter space by fixing stellar parameters would account for the inconsistencies between our results and literature **medians**. Therefore, we assumed to find the best agreement between our methods, and compared to the literature, in Run C and C2. However, we find that this is not necessarily the case as it is for FGK-stars (e.g. Jofré et al. 2014, 2017), and that our methods generally show the largest consistency with literature when used in their original setting without any standardizations. In general, the mean differences to the literature **median** are below 100 K in $T_{\rm eff}$ for all methods, and also below 0.1 dex in [Fe/H] for the Pass19-code and SteParSyn. In Run C and C2 these differences increase significantly, up to over 200 K and 0.3 dex for some of our methods.

This is an indication that each team had calibrated their methods and optimized them to the use of certain wavelength ranges and synthetic spectra. It also implies that there might be other components responsible for the remaining differences we see in stellar parameters, which requires a more thorough in-depth investigation of the methods themselves and underlying concepts. One example are stellar atmosphere models and their corresponding spectra. Although constant improvements are made, they still suffer from some deficiencies. Various sets of synthetic spectra also show differences when comparing the same spectral lines, due to the use of different equations of state, line lists, and other hyper-parameters. It cannot be excluded that these deviations contribute to the disagreements in derived stellar parameters found in this work. However, we were able to shed some light on deficiencies of some methods, e.g., that the DL method would benefit from the use of multiple wavelength ranges, and that ODUSSEAS could improve $T_{\rm eff}$ determination by using different sets of lines **or by changing the $T_{\rm eff}$ reference scale from Casagrande et al. (2008) to the values derived based on calibration using interferometry (e.g., Khata et al. 2021).**

Acknowledgements. CARMENES is an instrument for the Centro Astronómico Hispano-Alemán de Calar Alto (CAHA, Almería, Spain). CARMENES is funded by the German Max-Planck-Gesellschaft (MPG), the Spanish Consejo Superior de Investigaciones Científicas (CSIC), European Regional Development Fund (ERDF) through projects FICTS-2011-02, ICTS-2017-07-CAHA-4, and CAHA16-CE-3978, and the members of the CARMENES Consortium (Max-Planck-Institut für Astronomie, Instituto de Astrofísica de Andalucía, Landessternwarte Königstuhl, Institut de Ciències de l'Espai, Insitut für Astrophysik Göttingen, Universidad Complutense de Madrid, Thüringer Landessternwarte Tautenburg, Instituto de Astrofísica de Canarias, Hamburger Sternwarte, Centro de Astrobiología and Centro Astronómico Hispano-Alemán), with additional contributions by the Spanish Ministry of Economy, the German Science Foundation through the Major Research Instrumentation Programme and DFG Research Unit FOR2544 "Blue Planets around Red Stars", the Klaus Tschira Stiftung, the states of Baden-Württemberg and Niedersachsen, and by the Junta de Andalucía. E.D.M. and A.A.K. acknowledge the support by the Investigador FCT contract IF/00849/2015/CP1273/CT0003. We acknowledge financial support from NASA through grant NNX17AG24G, the Agencia Estatal de Investigación of the Ministerio de Ciencia through fellowship FPU15/01476, Innovación y Universidades and the ERDF through projects PID2019-109522GB-C51/2/3/4, AYA2016-79425-C3-1/2/3-P and AYA2018-84089, the Fundação para a Ciência e a Tecnologia through and ERDF through grants UID/FIS/04434/2019, UIDB/04434/2020 and UIDP/04434/2020, PTDC/FIS-AST/28953/2017, and PTDC/FIS-AST/32113/2017, and COMPETE2020 - Programa Operacional Competitividade e Internacionalização POCI-01-0145-FEDER-028953, and POCI-01-0145-FEDER-032113. This research has been funded by the Spanish State Research Agency (AEI) Project No. MDM-2017-0737 Unidad de Excelencia "María de Maeztu"- Centro de Astrobiología (CSIC/INTA).

References

Allard, F., Hauschildt, P. H., Alexander, D. R., Tamanai, A., & Schweitzer, A. 2001, ApJ, 556, 357
Allard, F., Homeier, D., & Freytag, B. 2011, in Astronomical Society of the Pacific Conference Series, Vol. 448, 16th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun, ed. C. Johns-Krull, M. K. Browning, & A. A. West, 91
Allard, F., Homeier, D., & Freytag, B. 2012, Philosophical Transactions of the Royal Society of London Series A, 370, 2765
Allard, F., Homeier, D., Freytag, B., et al. 2013, Memorie della Societa Astronomica Italiana Supplementi, 24, 128
Alonso-Floriano, F. J., Morales, J. C., Caballero, J. A., et al. 2015, A&A, 577, A128
Antoniadis-Karnavas, A., Sousa, S. G., Delgado-Mena, E., et al. 2020, A&A, 636, A9
Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, ARA&A, 47, 481
Bayo, A., Rodrigo, C., Barrado Y Navascués, D., et al. 2008, A&A, 492, 277
Benedict, G. F., Henry, T. J., Franz, O. G., et al. 2016, AJ, 152, 141
Berger, D. H., Gies, D. R., McAlister, H. A., et al. 2006, ApJ, 644, 475
Birky, J., Hogg, D. W., Mann, A. W., & Burgasser, A. 2020, ApJ, 892, 31
Borsa, F., Allart, R., Casasayas-Barris, N., et al. 2021, A&A, 645, A24
Boyajian, T. S., von Braun, K., van Belle, G., et al. 2012, ApJ, 757, 112
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Burn, R., Schlecker, M., Mordasini, C., et al. 2021, arXiv e-prints, arXiv:2105.04596
Caballero, J. A., Guàrdia, J., López del Fresno, M., et al. 2016, in Proc. SPIE, Vol. 9910, Observatory Operations: Strategies, Processes, and Systems VI, 99100E
Casagrande, L., Flynn, C., & Bessell, M. 2008, MNRAS, 389, 585
Chen, Y., Bressan, A., Girardi, L., et al. 2015, MNRAS, 452, 1068
Chen, Y., Girardi, L., Bressan, A., et al. 2014, MNRAS, 444, 2525
Cifuentes, C., Caballero, J. A., Cortés-Contreras, M., et al. 2020, A&A, 642, A115
Claret, A., Hauschildt, P. H., & Witte, S. 2012, A&A, 546, A14
Dittmann, J. A., Irwin, J. M., Charbonneau, D., & Newton, E. R. 2016, ApJ, 818, 153
Dotter, A., Chaboyer, B., Jevremović, D., et al. 2008, ApJS, 178, 89
Fabbro, S., Venn, K. A., O'Briain, T., et al. 2018, MNRAS, 475, 2978
Figueira, P., Adibekyan, V. Z., Oshagh, M., et al. 2016, A&A, 586, A101
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
Gaidos, E. & Mann, A. W. 2014, ApJ, 791, 54
Gaidos, E., Mann, A. W., Lépine, S., et al. 2014, MNRAS, 443, 2561
Gustafsson, B., Edvardsson, B., Eriksson, K., et al. 2008, A&A, 486, 951
Hauschildt, P. H. 1992, J. Quant. Spectr. Rad. Transf., 47, 433
Hauschildt, P. H. 1993, J. Quant. Spectr. Rad. Transf., 50, 301
Hauschildt, P. H., Allard, F., & Baron, E. 1999, ApJ, 512, 377
Hauschildt, P. H. & Baron, E. 1999, Journal of Computational and Applied Mathematics, 109, 41
Hauschildt, P. H., Baron, E., & Allard, F. 1997, ApJ, 483, 390
Heiter, U., Lind, K., Bergemann, M., et al. 2021, A&A, 645, A106
Henry, T. J. & McCarthy, Donald W., J. 1993, AJ, 106, 773
Hinkel, N. R., Young, P. A., Pagano, M. D., et al. 2016, ApJS, 226, 4
Houdebine, É. R., Mullan, D. J., Doyle, J. G., et al. 2019, AJ, 158, 56
Husser, T.-O., Wende-von Berg, S., Dreizler, S., et al. 2013, A&A, 553, A6
Jofré, P., Heiter, U., Soubiran, C., et al. 2015, A&A, 582, A81
Jofré, P., Heiter, U., Soubiran, C., et al. 2014, A&A, 564, A133
Jofré, P., Heiter, U., Worley, C. C., et al. 2017, A&A, 601, A38
Johnson S.G., Cervellino A., W. J. 2019, libcerf, numeric library for complex error functions, version 1.13, `https://jugit.fz-juelich.de/mlz/libcerf`
Kausch, W., Noll, S., Smette, A., et al. 2014, in Astronomical Society of the Pacific Conference Series, Vol. 485, Astronomical Data Analysis Software and Systems XXIII, ed. N. Manset & P. Forshay, 403
Khata, D., Mondal, S., Das, R., & Baug, T. 2021, MNRAS, 507, 1869
Khata, D., Mondal, S., Das, R., Ghosh, S., & Ghosh, S. 2020, MNRAS, 493, 4533
Lépine, S., Hilton, E. J., Mann, A. W., et al. 2013, AJ, 145, 102
Lindgren, S., Heiter, U., & Seifahrt, A. 2016, A&A, 586, A100
Lodders, K., Palme, H., & Gail, H. P. 2009, Landolt Börnstein, 4B, 712
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, The Astronomical Journal, 154, 94
Maldonado, J., Affer, L., Micela, G., et al. 2015, A&A, 577, A132
Mann, A. W., Brewer, J. M., Gaidos, E., Lépine, S., & Hilton, E. J. 2013a, AJ, 145, 52
Mann, A. W., Deacon, N. R., Gaidos, E., et al. 2014, AJ, 147, 160
Mann, A. W., Dupuy, T., Kraus, A. L., et al. 2019, ApJ, 871, 63
Mann, A. W., Feiden, G. A., Gaidos, E., Boyajian, T., & von Braun, K. 2015, ApJ, 804, 64
Mann, A. W., Gaidos, E., & Ansdell, M. 2013b, ApJ, 779, 188
Marfil, E., Tabernero, H. M., Montes, D., et al. 2021, arXiv e-prints, arXiv:2110.07329
Marfil, E., Tabernero, H. M., Montes, D., et al. 2020, MNRAS, 492, 5470
Nagel, E., Czesla, S., Kaminski, A., et al. 2021, submitted, A&A
Nemravová, J. A., Harmanec, P., Brož, M., et al. 2016, A&A, 594, A55
Neves, V., Bonfils, X., Santos, N. C., et al. 2012, A&A, 538, A25
Neves, V., Bonfils, X., Santos, N. C., et al. 2013, A&A, 551, A36
Neves, V., Bonfils, X., Santos, N. C., et al. 2014, A&A, 568, A121
Newton, E. R., Charbonneau, D., Irwin, J., et al. 2014, AJ, 147, 20
Newton, E. R., Charbonneau, D., Irwin, J., & Mann, A. W. 2015, ApJ, 800, 85
Önehag, A., Heiter, U., Gustafsson, B., et al. 2012, A&A, 542, A33
Parzen, E. 1962, The Annals of Mathematical Statistics, 33, 1065
Passegger, V. M., Bello-García, A., Ordieres-Meré, J., et al. 2020, A&A, 642, A22
Passegger, V. M., Reiners, A., Jeffers, S. V., et al. 2018, A&A, 615, A6
Passegger, V. M., Schweitzer, A., Shulyak, D., et al. 2019, A&A, 627, A161
Passegger, V. M., Wende-von Berg, S., & Reiners, A. 2016, A&A, 587, A19
Plez, B. 2012, Turbospectrum: Code for spectral synthesis
Quirrenbach, A., Amado, P. J., Ribas, I., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10702, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 107020W
Rabus, M., Lachaume, R., Jordán, A., et al. 2019, MNRAS, 484, 2674
Reiners, A., Zechmeister, M., Caballero, J. A., et al. 2018, A&A, 612, A49
Rojas-Ayala, B., Covey, K. R., Muirhead, P. S., & Lloyd, J. P. 2012, ApJ, 748, 93
Rosenblatt, M. 1956, The Annals of Mathematical Statistics, 27, 832
Sarmento, P., Rojas-Ayala, B., Delgado Mena, E., & Blanco-Cuaresma, S. 2021, A&A, 649, A147
Schweitzer, A., Passegger, V. M., Cifuentes, C., et al. 2019, A&A, 625, A68
Ségransan, D., Kervella, P., Forveille, T., & Queloz, D. 2003, A&A, 397, L5
Slumstrup, D., Grundahl, F., Silva Aguirre, V., & Brogaard, K. 2019, A&A, 622, A111
Smette, A., Sana, H., Noll, S., et al. 2015, A&A, 576, A77

Souto, D., Cunha, K., García-Hernández, D. A., et al. 2017, ApJ, 835, 239
Souto, D., Cunha, K., Smith, V. V., et al. 2020, ApJ, 890, 133
Tabernero, H. M., Dorda, R., Negueruela, I., & González-Fernández, C. 2018, MNRAS, 476, 3106
Tabernero, H. M., Dorda, R., Negueruela, I., & Marfil, E. 2021, A&A, 646, A98
Tabernero, H. M., Marfil, E., Montes, D., & González Hernández, J. I. 2021, arXiv e-prints, arXiv:2110.00444
Tang, J., Bressan, A., Rosenfield, P., et al. 2014, MNRAS, 445, 4287
Terrien, R. C., Mahadevan, S., Bender, C. F., Deshpande, R., & Robertson, P. 2015, ApJ, 802, L10
Valenti, J. A. & Fischer, D. A. 2005, ApJS, 159, 141
Valenti, J. A. & Piskunov, N. 1996, A&AS, 118, 595
van Leeuwen, F. 2007, Hipparcos, the New Reduction of the Raw Data, Vol. 350
von Braun, K., Boyajian, T. S., van Belle, G. T., et al. 2014, MNRAS, 438, 2413
Wilson, J. C., Hearty, F., Skrutskie, M. F., et al. 2010, in Ground-based and Airborne Instrumentation for Astronomy III, ed. I. S. McLean, S. K. Ramsay, & H. Takami, Vol. 7735, International Society for Optics and Photonics (SPIE), 554 – 567
Zechmeister, M., Anglada-Escudé, G., & Reiners, A. 2014, A&A, 561, A59
Zechmeister, M., Reiners, A., Amado, P. J., et al. 2018, A&A, 609, A12

Appendix A: Literature summary of sample stars

Table A.1. Collection of stellar parameters from the literature.

Karmn	Author	$T_{\rm eff}$ [K]	$\log g$ [dex]	[Fe/H] [dex]	R [$R_\odot$]	M [$M_\odot$]
J00067−075	Neves et al. (2014)	2718 ± 150	...	−0.27 ± 0.20	...	...
	Terrien et al. (2015)	...	...	−0.11 ± 0.10	...	...
	Houdebine et al. (2019)	2970 ± 149	...	...	0.127 ± 0.011	...
	Literature **median**	2844 ± 149	...	−0.19 ± 0.16	...	...
	Literature & Run A **median**	2906 ± 103	5.01 ± 0.11	−0.17 ± 0.18	...	...
J00183+440	Berger et al. (2006)	3747 ± 112	4.89 ± 0.07	...	0.379 ± 0.006	0.404 ± 0.040
	Boyajian et al. (2012)	3563 ± 11	4.89 ± 0.04^{c}	...	0.387 ± 0.002	0.423 ± 0.042
	Gaidos et al. (2014)	3669 ± 67	4.79 ± 0.02^{c}	−0.29 ± 0.11	0.470 ± 0.040	0.500 ± 0.060
	Gaidos & Mann (2014)	3693 ± 91	4.77 ± 0.03^{c}	−0.26 ± 0.08	0.490 ± 0.050	0.520 ± 0.070
	Houdebine et al. (2019)	3656 ± 183	...	...	0.365 ± 0.014	...
	Khata et al. (2020)	3493 ± 103	4.82 ± 0.04^{c}	−0.19 ± 0.08	0.385 ± 0.027	0.357 ± 0.017
	Mann et al. (2015)	3603 ± 60	4.86 ± 0.01^{c}	−0.30 ± 0.08	0.388 ± 0.013	0.398 ± 0.040
	Newton et al. (2015)	3534 ± 79	...	...	0.388 ± 0.028	...
		3602 ± 13^{int}	...	...	0.386 ± 0.002^{int}	...
	Ségransan et al. (2003)	3698 ± 95	4.89 ± 0.02^{c}	...	0.383 ± 0.020	0.414 ± 0.021
	Terrien et al. (2015)	...	...	−0.26 ± 0.10	0.395 ± 0.004	...
	Literature **median**	3626 ± 94	4.84 ± 0.04	−0.26 ± 0.09	...	...
	Literature & Run A **median**	3614 ± 84	4.85 ± 0.04	−0.25 ± 0.11	...	...
J04429+189	Gaidos et al. (2014)	3680 ± 99	4.78 ± 0.03^{c}	+0.04 ± 0.11	0.480 ± 0.050	0.510 ± 0.070
	Gaidos & Mann (2014)	3721 ± 82	4.76 ± 0.03^{c}	+0.14 ± 0.08	0.500 ± 0.050	0.530 ± 0.070
	Houdebine et al. (2019)	3542 ± 177	...	...	0.453 ± 0.027	...
	Khata et al. (2020)	3377 ± 110	4.87 ± 0.06^{c}	−0.01 ± 0.09	0.338 ± 0.032	0.312 ± 0.015
	Lépine et al. (2013)	3550 ± 57	4.50	...	...	...
	Maldonado et al. (2015)	3603 ± 68	4.75 ± 0.04	+0.03 ± 0.09	0.510 ± 0.047	0.520 ± 0.052
	Mann et al. (2015)	3680 ± 60	4.82 ± 0.01^{c}	+0.14 ± 0.08	0.452 ± 0.019	0.492 ± 0.049
	Neves et al. (2014)	3355 ± 110	...	−0.01 ± 0.09	...	0.500 ± 0.030
	Newton et al. (2015)	3574 ± 78	...	...	0.514 ± 0.029	...
		3701 ± 90^{int}	...	...	0.453 ± 0.022^{int}	...
	Rojas-Ayala et al. (2012)	3581 ± 20	...	+0.15 ± 0.17	...	...
	Terrien et al. (2015)	...	...	+0.12 ± 0.10	0.478 ± 0.010	...
	von Braun et al. (2014)	3679 ± 77	4.78 ± 0.09^{c}	...	0.453 ± 0.022	0.450 ± 0.135
	Literature **median**	3587 ± 93	4.75 ± 0.05	+0.08 ± 0.10	...	...
	Literature & Run A **median**	3581 ± 85	4.73 ± 0.05	+0.07 ± 0.11	...	...
J05314−036	Boyajian et al. (2012)	3801 ± 9	4.71 ± 0.04^{c}	...	0.574 ± 0.004	0.615 ± 0.062
	Gaidos et al. (2014)	3701 ± 61	4.77 ± 0.02^{c}	...	0.490 ± 0.040	0.520 ± 0.060
	Gaidos & Mann (2014)	3895 ± 84	4.72 ± 0.01^{c}	+0.43 ± 0.08	0.560 ± 0.040	0.600 ± 0.070
	Houdebine et al. (2019)	3696 ± 185	...	...	0.588 ± 0.019	...
	Khata et al. (2020)	3849 ± 275	4.69 ± 0.08^{c}	...	0.553 ± 0.149	0.554 ± 0.193
	Maldonado et al. (2015)	3800 ± 68	4.68 ± 0.05	+0.00 ± 0.09	0.580 ± 0.052	0.600 ± 0.056
	Mann et al. (2015)	3801 ± 60	4.71 ± 0.01^{c}	+0.49 ± 0.08	0.581 ± 0.019	0.633 ± 0.063
	Neves et al. (2014)	3670 ± 110	...	+0.19 ± 0.09	...	0.600 ± 0.070
	Newton et al. (2015)	3872 ± 75	...	...	0.597 ± 0.027	...
		3850 ± 22^{int}	...	...	0.574 ± 0.004^{int}	...
	Rojas-Ayala et al. (2012)	4012 ± 106	...	+0.35 ± 0.17	...	...
	Ségransan et al. (2003)	3520 ± 170	4.54 ± 0.06^{c}	...	0.702 ± 0.063	0.631 ± 0.031
	Terrien et al. (2015)	...	...	+0.69 ± 0.10	0.587 ± 0.040	...
	Literature **median**	3789 ± 125	4.69 ± 0.05	+0.36 ± 0.11	...	...
	Literature & Run A **median**	3779 ± 112	4.69 ± 0.05	+0.30 ± 0.11	...	...
J07558+833	Dittmann et al. (2016)	...	...	+0.00 ± 0.10	...	...
	Gaidos et al. (2014)	3183 ± 60	...	...	<0.19	<0.14
	Lépine et al. (2013)	3250 ± 76	5.00	...	...	...
	Literature **median**	3217 ± 68	5.00	+0.00 ± 0.10	...	...
	Literature & Run A **median**	3265 ± 87	5.00 ± 0.07	−0.14 ± 0.18	...	...
J09143+526	Boyajian et al. (2012)	3907 ± 35	4.71 ± 0.02^{c}	...	0.577 ± 0.013	0.622 ± 0.062
	Gaidos et al. (2014)	3991 ± 66	4.69 ± 0.01^{c}	−0.26 ± 0.11	0.590 ± 0.040	0.630 ± 0.070
	Houdebine et al. (2019)	3921 ± 196	...	...	0.602 ± 0.020	...

Table A.1. Collection of stellar parameters from the literature (cont.)

Karmn	Author	$T_{\rm eff}$ [K]	$\log g$ [dex]	[Fe/H] [dex]	R [$R_{\odot}$]	M [$M_{\odot}$]
	Khata et al. (2020)	4002 ± 125	4.48 ± 0.05^{c}	−0.08 ± 0.09	0.617 ± 0.051	0.424 ± 0.024
	Mann et al. (2015)	3920 ± 60	4.74 ± 0.00^{c}	−0.01 ± 0.08	0.550 ± 0.026	0.607 ± 0.061
	Newton et al. (2015)	3955 ± 106	...	...	0.571 ± 0.029	...
		3953 ± 41^{int}	...	...	0.577 ± 0.013^{int}	...
	Rojas-Ayala et al. (2012)	4031 ± 56	...	−0.18 ± 0.17	...	...
	Literature **median**	3960 ± 100	4.65 ± 0.03	−0.13 ± 0.12	...	...
	Literature & Run A **median**	3961 ± 88	4.68 ± 0.04	−0.07 ± 0.12	...	...
J09144+526	Boyajian et al. (2012)	3867 ± 37	4.71 ± 0.02^{c}	...	0.567 ± 0.014	0.600 ± 0.060
	Gaidos et al. (2014)	3770 ± 87	4.74 ± 0.03^{c}	...	0.520 ± 0.050	0.550 ± 0.07
	Houdebine et al. (2019)	3921 ± 196	...	...	0.600 ± 0.040	...
	Khata et al. (2020)	3844 ± 127	...	−0.07 ± 0.09	0.582 ± 0.047	...
	Newton et al. (2015)	3892 ± 92	...	...	0.562 ± 0.028	...
		3926 ± 37^{int}	...	...	0.567 ± 0.014^{int}	...
	Rojas-Ayala et al. (2012)	3869 ± 15	...	−0.15 ± 0.17	...	...
	Literature **median**	3870 ± 103	4.72 ± 0.03	−0.11 ± 0.14	...	...
	Literature & Run A **median**	3894 ± 89	4.80 ± 0.04	−0.02 ± 0.13	...	...
J10508+068	Gaidos et al. (2014)	3238 ± 81	5.02 ± 0.06^{c}	+0.13 ± 0.11	0.190 ± 0.080	0.140 ± 0.100
	Gaidos & Mann (2014)	3400 ± 63	4.92 ± 0.05^{c}	+0.20 ± 0.08	0.320 ± 0.050	0.310 ± 0.060
	Houdebine et al. (2019)	3099 ± 155	...	...	0.334 ± 0.031	...
	Khata et al. (2020)	2388 ± 113	...	+0.18 ± 0.10	...	0.155 ± 0.007
	Lépine et al. (2013)	3100 ± 76	4.50	...	...	...
	Mann et al. (2015)	3238 ± 60	4.94 ± 0.01^{c}	+0.16 ± 0.08	0.276 ± 0.012	0.246 ± 0.025
	Neves et al. (2014)	2943 ± 110	...	+0.03 ± 0.09	...	...
	Rojas-Ayala et al. (2012)	3334 ± 23	...	+0.20 ± 0.17	...	...
	Terrien et al. (2015)	...	...	+0.11 ± 0.10	...	...
	Literature **median**	3093 ± 93	4.85 ± 0.04	+0.14 ± 0.11	...	...
	Literature & Run A **median**	3136 ± 82	4.83 ± 0.06	+0.08 ± 0.12	...	...
J11033+359	Boyajian et al. (2012)	3465 ± 17	4.85 ± 0.04^{c}	...	0.392 ± 0.004	0.403 ± 0.040
	Gaidos et al. (2014)	3593 ± 66	4.81 ± 0.02^{c}	...	0.440 ± 0.040	0.460 ± 0.060
	Gaidos & Mann (2014)	3679 ± 110	4.78 ± 0.04^{c}	−0.30 ± 0.08	0.480 ± 0.060	0.510 ± 0.080
	Houdebine et al. (2019)	3602 ± 180	...	...	0.362 ± 0.013	...
	Khata et al. (2020)	3560 ± 104	4.74 ± 0.04^{c}	−0.18 ± 0.13	0.419 ± 0.028	0.355 ± 0.016
	Lépine et al. (2013)	3530 ± 39	4.50	...	...	...
	Mann et al. (2015)	3563 ± 60	4.84 ± 0.01^{c}	−0.38 ± 0.08	0.389 ± 0.013	0.386 ± 0.039
	Newton et al. (2015)	3532 ± 85	...	...	0.401 ± 0.029	...
		3532 ± 17^{int}	...	...	0.392 ± 0.003^{int}	...
	Rojas-Ayala et al. (2012)	3526 ± 18	...	−0.41 ± 0.17	...	...
	Ségransan et al. (2003)	3570 ± 42	4.85 ± 0.00^{c}	...	0.393 ± 0.008	0.403 ± 0.020
	Terrien et al. (2015)	...	...	−0.41 ± 0.10	0.378 ± 0.042	...
	Literature **median**	3559 ± 82	4.77 ± 0.03	−0.34 ± 0.12	...	...
	Literature & Run A **median**	3566 ± 76	4.79 ± 0.05	−0.28 ± 0.12	...	...
J11054+435	Boyajian et al. (2012)	3497 ± 39	4.84 ± 0.02^{c}	...	0.398 ± 0.009	0.403 ± 0.040
	Gaidos et al. (2014)	3702 ± 65	4.77 ± 0.02^{c}	−0.41 ± 0.11	0.490 ± 0.040	0.520 ± 0.060
	Gaidos & Mann (2014)	3743 ± 84	4.75 ± 0.03^{c}	−0.32 ± 0.08	0.510 ± 0.050	0.540 ± 0.070
	Houdebine et al. (2019)	3692 ± 185	...	...	0.370 ± 0.030	...
	Lépine et al. (2013)	3560 ± 44	4.50	...	...	...
	Mann et al. (2015)	3619 ± 60	4.86 ± 0.01^{c}	−0.37 ± 0.08	0.383 ± 0.013	0.390 ± 0.039
	Newton et al. (2015)	3664 ± 227	...	...	0.425 ± 0.041	...
		3537 ± 41^{int}	...	...	0.398 ± 0.009^{int}	...
	Rojas-Ayala et al. (2012)	3684 ± 20	...	−0.40 ± 0.17	...	...
	Terrien et al. (2015)	...	...	−0.38 ± 0.10	0.378 ± 0.004	...
	Literature **median**	3633 ± 109	4.74 ± 0.02	−0.38 ± 0.11	...	...
	Literature & Run A **median**	3633 ± 95	4.80 ± 0.04	−0.35 ± 0.12	...	...
J11421+267	Gaidos et al. (2014)	3479 ± 61	4.88 ± 0.05^{c}	+0.07 ± 0.11	0.370 ± 0.050	0.380 ± 0.060
	Gaidos & Mann (2014)	3606 ± 72	4.82 ± 0.04^{c}	+0.00 ± 0.08	0.440 ± 0.050	0.470 ± 0.060
	Houdebine et al. (2019)	3464 ± 173	...	...	0.403 ± 0.012	...
	Khata et al. (2020)	3534 ± 106	4.86 ± 0.04^{c}	−0.06 ± 0.08	0.418 ± 0.029	0.460 ± 0.021

Table A.1. Collection of stellar parameters from the literature (cont.)

Karmn	Author	$T_{\rm eff}$ [K]	$\log g$ [dex]	[Fe/H] [dex]	R [$R_\odot$]	M [$M_\odot$]
	Lépine et al. (2013)	3400 ± 62	5.00	…	…	…
	Mann et al. (2015)	3479 ± 60	4.78 ± 0.01^c	+0.01 ± 0.08	0.449 ± 0.019	0.445 ± 0.044
	Neves et al. (2014)	3354 ± 110	…	−0.03 ± 0.09	…	…
	Newton et al. (2015)	3477 ± 81	…	…	0.400 ± 0.028	…
		3520 ± 66^{int}	…	…	0.455 ± 0.018^{int}	…
	Rojas-Ayala et al. (2012)	3469 ± 17	…	+0.04 ± 0.17	…	…
	Terrien et al. (2015)	…	…	−0.06 ± 0.10	0.431 ± 0.008	…
	Literature **median**	3478 ± 90	4.87 ± 0.04	−0.00 ± 0.11	…	…
	Literature & Run A **median**	3468 ± 81	4.83 ± 0.05	−0.01 ± 0.11	…	…
J13005+056	Dittmann et al. (2016)	…	…	+0.07 ± 0.10	…	…
	Gaidos et al. (2014)	3090 ± 77	…	+0.28 ± 0.11	<0.19	<0.14
	Houdebine et al. (2019)	3140 ± 157	…	…	0.191 ± 0.013	…
	Lépine et al. (2013)	2950 ± 61	4.50	…	…	…
	Newton et al. (2015)	…	…	…	0.170 ± 0.043	…
	Terrien et al. (2015)	…	…	+0.09 ± 0.10	…	…
	Literature **median**	3060 ± 107	4.50	+0.15 ± 0.10	…	…
	Literature & Run A **median**	3137 ± 100	4.79 ± 0.10	+0.02 ± 0.17	…	…
J13457+148	Berger et al. (2006)	3662 ± 110	4.75 ± 0.07	…	0.493 ± 0.033	0.502 ± 0.050
	Boyajian et al. (2012)	3618 ± 31	4.78 ± 0.03^c	…	0.484 ± 0.008	0.520 ± 0.052
	Gaidos et al. (2014)	3703 ± 73	4.77 ± 0.02^c	−0.45 ± 0.11	0.490 ± 0.040	0.520 ± 0.060
	Gaidos & Mann (2014)	3792 ± 92	4.75 ± 0.03^c	−0.18 ± 0.09	0.520 ± 0.050	0.560 ± 0.070
	Houdebine et al. (2019)	3650 ± 183	…	…	0.494 ± 0.033	…
	Khata et al. (2020)	3707 ± 103	4.78 ± 0.01^c	−0.15 ± 0.10	0.502 ± 0.027	0.554 ± 0.044
	Maldonado et al. (2015)	3609 ± 68	4.79 ± 0.04	−0.10 ± 0.09	0.470 ± 0.047	0.470 ± 0.052
	Mann et al. (2015)	3649 ± 60	4.74 ± 0.01^c	−0.31 ± 0.08	0.478 ± 0.016	0.465 ± 0.046
	Neves et al. (2014)	3515 ± 110	…	−0.22 ± 0.09	…	0.500 ± 0.030
	Newton et al. (2015)	3716 ± 125	…	…	0.450 ± 0.033	…
		3646 ± 34^{int}	…	…	0.484 ± 0.008^{int}	…
	Rojas-Ayala et al. (2012)	3642 ± 17	…	−0.30 ± 0.17	…	…
	Literature **median**	3659 ± 95	4.77 ± 0.03	−0.24 ± 0.11	…	…
	Literature & Run A **median**	3667 ± 88	4.75 ± 0.04	−0.21 ± 0.11	…	…
J15194−077	Gaidos et al. (2014)	3413 ± 61	4.90 ± 0.05^c	−0.21 ± 0.11	0.330 ± 0.050	0.320 ± 0.060
	Gaidos & Mann (2014)	3357 ± 73	4.94 ± 0.05^c	−0.10 ± 0.08	0.290 ± 0.060	0.270 ± 0.080
	Houdebine et al. (2019)	3423 ± 171	…	…	0.285 ± 0.008	…
	Khata et al. (2020)	3475 ± 119	4.71 ± 0.08^c	−0.11 ± 0.12	0.364 ± 0.045	0.251 ± 0.015
	Maldonado et al. (2015)	3419 ± 68	4.95 ± 0.08	−0.20 ± 0.09	0.300 ± 0.078	0.290 ± 0.086
	Mann et al. (2015)	3395 ± 60	4.92 ± 0.01^c	−0.15 ± 0.08	0.311 ± 0.012	0.292 ± 0.029
	Neves et al. (2014)	3248 ± 110	…	−0.20 ± 0.09	…	0.300 ± 0.020
	Newton et al. (2015)	3354 ± 74	…	…	0.329 ± 0.027	…
		3487 ± 62^{int}	…	…	0.299 ± 0.010^{int}	…
	Rojas-Ayala et al. (2012)	3534 ± 18	…	−0.10 ± 0.17	…	…
	Terrien et al. (2015)	…	…	−0.06 ± 0.10	0.322 ± 0.050	…
	Literature **median**	3411 ± 83	4.88 ± 0.14	−0.14 ± 0.11	…	…
	Literature & Run A **median**	3399 ± 82	4.87 ± 0.06	−0.13 ± 0.11	…	…
J16581+257	Gaidos et al. (2014)	3744 ± 65	4.75 ± 0.02^c	−0.08 ± 0.11	0.510 ± 0.040	0.540 ± 0.060
	Houdebine et al. (2019)	3705 ± 185	…	…	0.497 ± 0.020	…
	Khata et al. (2020)	3654 ± 117	4.74 ± 0.05^c	−0.03 ± 0.12	0.466 ± 0.036	0.438 ± 0.019
	Lépine et al. (2013)	3590 ± 39	4.50	…	…	…
	Mann et al. (2015)	3700 ± 60	4.75 ± 0.01^c	+0.03 ± 0.08	0.507 ± 0.018	0.534 ± 0.053
	Newton et al. (2015)	3683 ± 79	…	…	0.497 ± 0.028	…
		3604 ± 46^{int}	…	…	0.539 ± 0.016^{int}	…
	Rojas-Ayala et al. (2012)	3733 ± 20	…	−0.04 ± 0.17	…	…
	Terrien et al. (2015)	…	…	−0.04 ± 0.10	0.505 ± 0.006	…
	von Braun et al. (2014)	3590 ± 45	4.70 ± 0.10^c	…	0.539 ± 0.016	0.540 ± 0.162
	Literature **median**	3667 ± 87	4.69 ± 0.06	−0.03 ± 0.12	…	…
	Literature & Run A **median**	3678 ± 78	4.74 ± 0.06	+0.01 ± 0.12	…	…
J17578+046	Boyajian et al. (2012)	3224 ± 10	5.06 ± 0.04^c	…	0.187 ± 0.001	0.146 ± 0.015

Table A.1. Collection of stellar parameters from the literature (cont.)

Karmn	Author	$T_{\rm eff}$ [K]	$\log g$ [dex]	[Fe/H] [dex]	R [$R_\odot$]	M [$M_\odot$]
	Dittmann et al. (2016)	...	...	−0.44 ± 0.10	...	...
	Gaidos et al. (2014)	3237 ± 60	...	−0.32 ± 0.11	<0.19	<0.14
	Gaidos & Mann (2014)	3247 ± 61	5.05 ± 0.04^c	−0.32 ± 0.08	0.190 ± 0.060	0.150 ± 0.080
	Houdebine et al. (2019)	3266 ± 163	...	...	0.186 ± 0.010	...
	Mann et al. (2015)	3228 ± 60	5.09 ± 0.01^c	−0.40 ± 0.08	0.186 ± 0.007	0.155 ± 0.015
	Neves et al. (2014)	3338 ± 110	...	−0.51 ± 0.09	...	0.160 ± 0.010
	Newton et al. (2015)	3248 ± 81	...	...	0.188 ± 0.029	...
		3238 ± 11^{int}	...	...	0.187 ± 0.001^{int}	...
	Rojas-Ayala et al. (2012)	3266 ± 29	...	−0.39 ± 0.17	...	...
	Ségransan et al. (2003)	3163 ± 65	5.05 ± 0.01^c	...	0.196 ± 0.008	0.158 ± 0.008
	Terrien et al. (2015)	...	...	−0.34 ± 0.10	0.183 ± 0.002	...
	Literature **median**	3246 ± 78	5.06 ± 0.03	−0.39 ± 0.12	...	...
	Literature & Run A **median**	3242 ± 75	5.06 ± 0.07	−0.34 ± 0.13	...	...
J22565+165	Berger et al. (2006)	3373 ± 101	4.53 ± 0.07	...	0.689 ± 0.044	0.586 ± 0.059
	Boyajian et al. (2012)	3713 ± 11	4.71 ± 0.04^c	...	0.548 ± 0.005	0.569 ± 0.057
	Gaidos et al. (2014)	3673 ± 60	4.78 ± 0.02^c	+0.18 ± 0.11	0.480 ± 0.040	0.510 ± 0.060
	Gaidos & Mann (2014)	3786 ± 87	4.75 ± 0.03^c	+0.17 ± 0.08	0.520 ± 0.050	0.560 ± 0.070
	Houdebine et al. (2019)	3661 ± 183	...	...	0.567 ± 0.019	...
	Lépine et al. (2013)	3520 ± 39	4.50	...	...	...
	Maldonado et al. (2015)	3736 ± 68	4.71 ± 0.04	−0.01 ± 0.09	0.550 ± 0.047	0.570 ± 0.052
	Mann et al. (2015)	3720 ± 60	4.71 ± 0.01^c	+0.21 ± 0.08	0.549 ± 0.018	0.574 ± 0.057
	Neves et al. (2014)	3602 ± 110	...	+0.03 ± 0.09	...	0.580 ± 0.030
	Newton et al. (2015)	3749 ± 76	...	...	0.555 ± 0.028	...
		3731 ± 16^{int}	...	...	0.548 ± 0.005^{int}	...
	Terrien et al. (2015)	...	...	+0.26 ± 0.10	0.545 ± 0.003	...
	Literature **median**	3660 ± 87	4.67 ± 0.04	+0.14 ± 0.09	...	...
	Literature & Run A **median**	3673 ± 79	4.68 ± 0.04	+0.14 ± 0.11	...	...
J23419+441	Dittmann et al. (2016)	...	...	+0.17 ± 0.10	...	...
	Gaidos et al. (2014)	3005 ± 62	...	...	<0.19	<0.14
	Gaidos & Mann (2014)	3067 ± 60	...	+0.29 ± 0.08	<0.19	<0.14
	Houdebine et al. (2019)	3032 ± 152	...	...	0.098 ± 0.003	...
	Khata et al. (2020)	3104 ± 117	5.05 ± 0.15^c	+0.26 ± 0.11	0.197 ± 0.038	0.161 ± 0.007
	Lépine et al. (2013)	3110 ± 43	5.00	...	...	...
	Mann et al. (2015)	2930 ± 60	5.04 ± 0.01^c	+0.23 ± 0.08	0.189 ± 0.008	0.145 ± 0.015
	Rojas-Ayala et al. (2012)	3058 ± 65	...	+0.19 ± 0.17	...	...
	Terrien et al. (2015)	...	...	+0.32 ± 0.10	...	...
	Literature **median**	3044 ± 88	5.03 ± 0.11	+0.24 ± 0.11	...	...
	Literature & Run A **median**	3031 ± 80	5.02 ± 0.10	+0.16 ± 0.14	...	...

Notes. $^{(c)}$ $\log g$ calculated from M and R, $^{(int)}$ interferometric measurement.

Appendix B: Additional plots

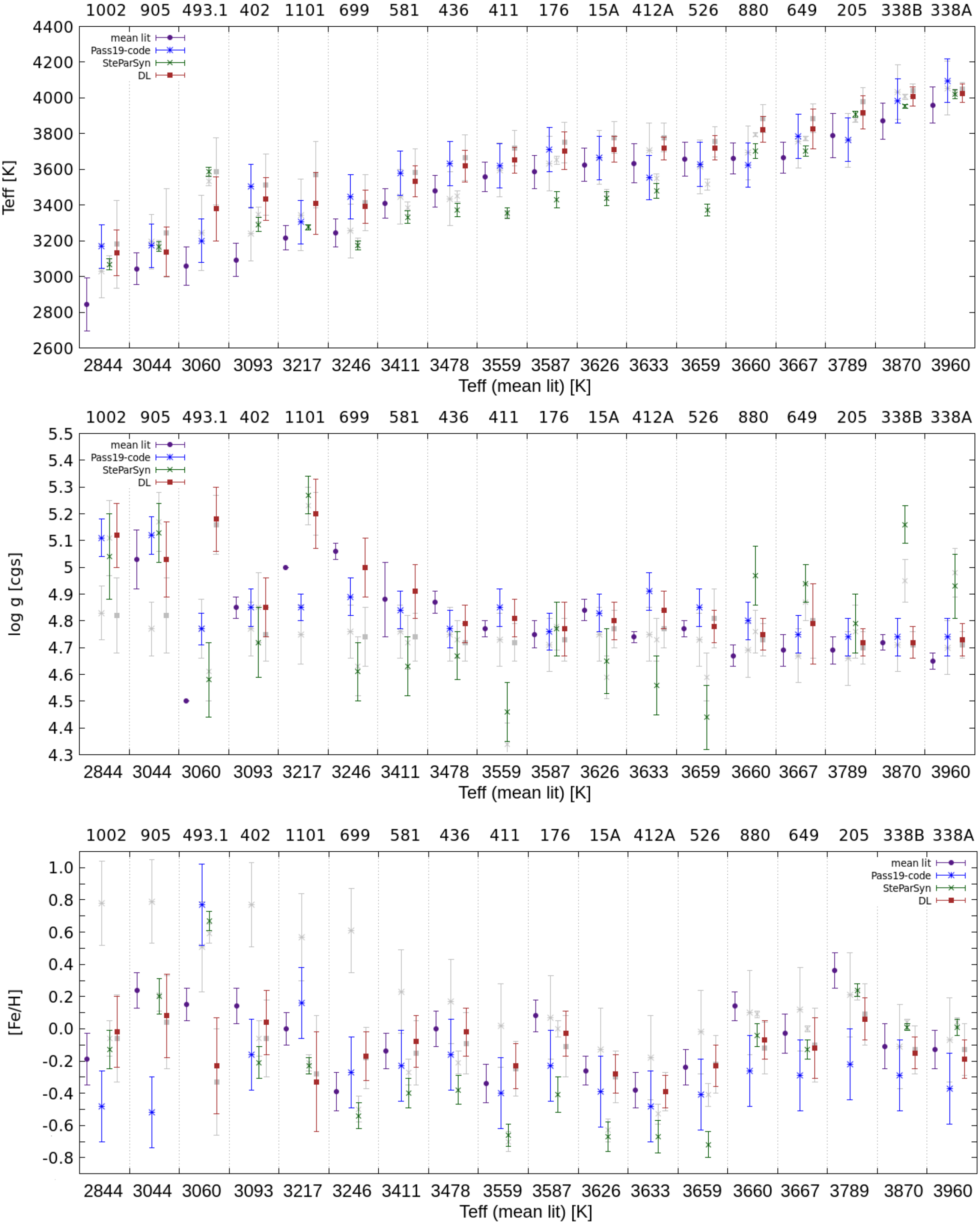

Fig. B.1. Same as Fig. 4, but the gray symbols represent results from Run C.

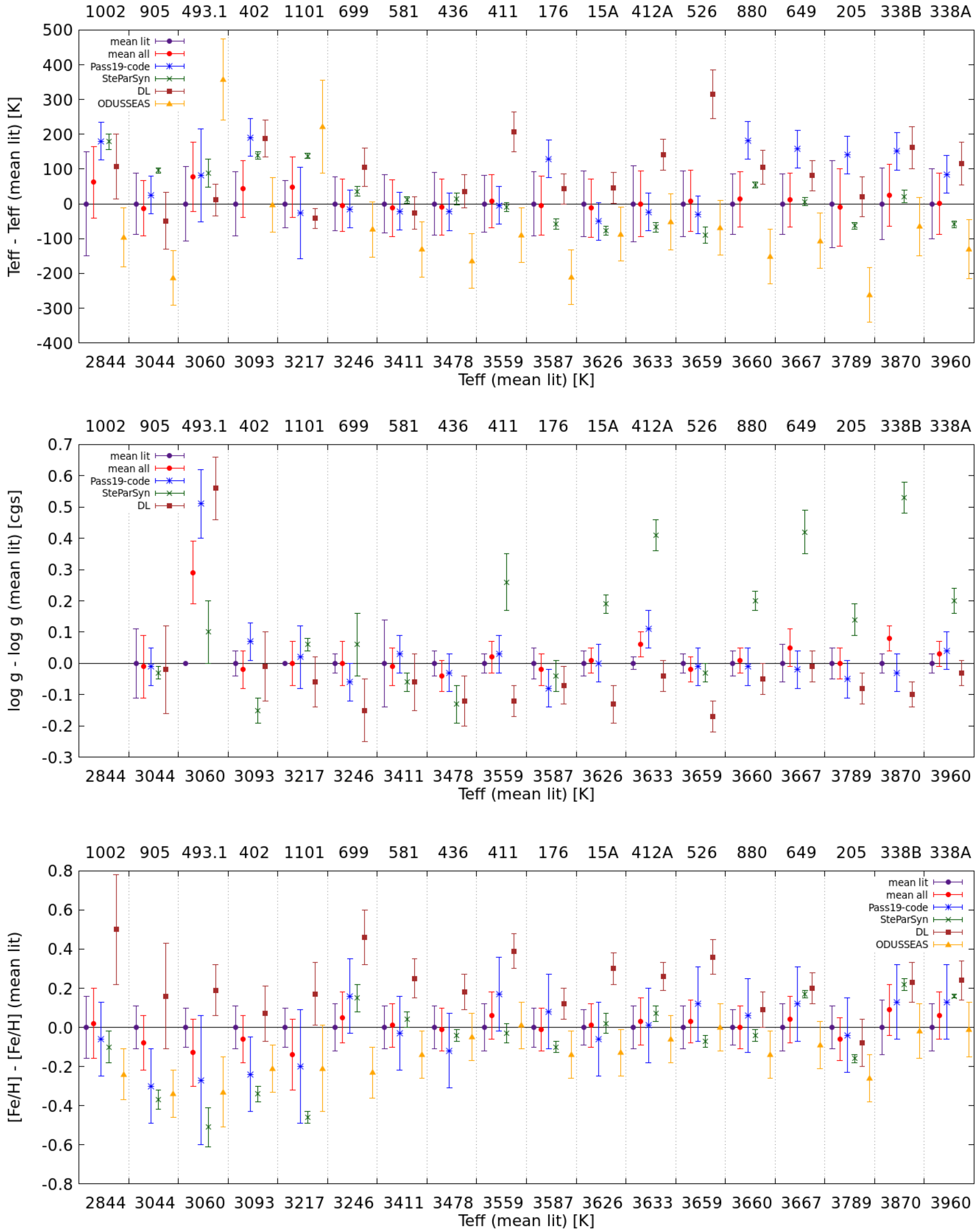

Fig. B.2. Same as Fig. 2, but showing the difference between the stellar parameter and the literature **median** on the y-axis.

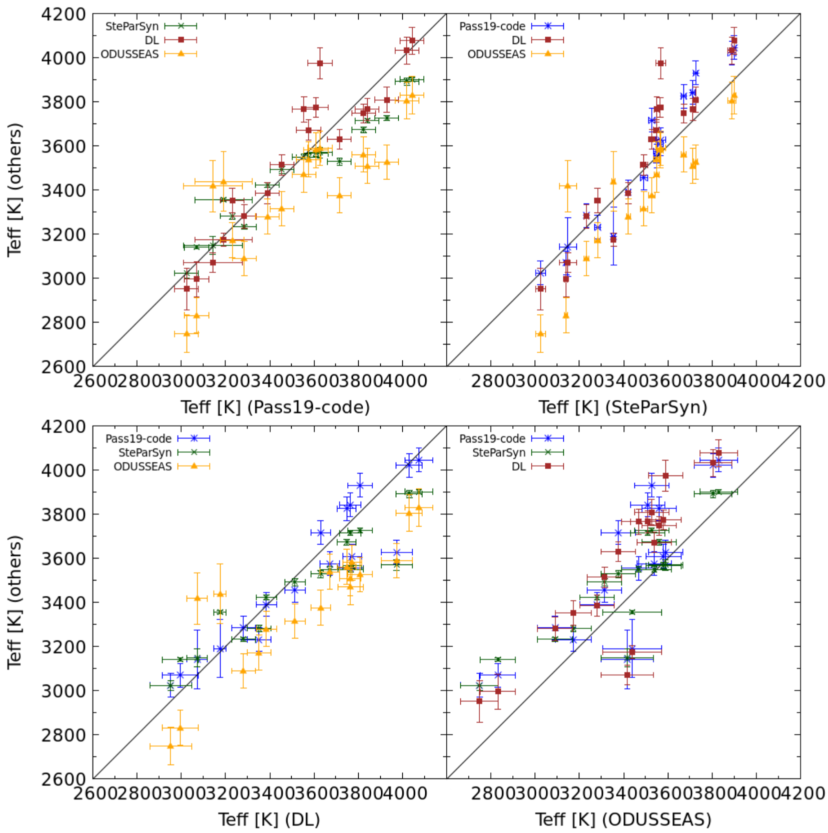

Fig. B.3. Comparison between our methods, showing the derived T_{eff} in Run A. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the x-axis label) to all other methods.

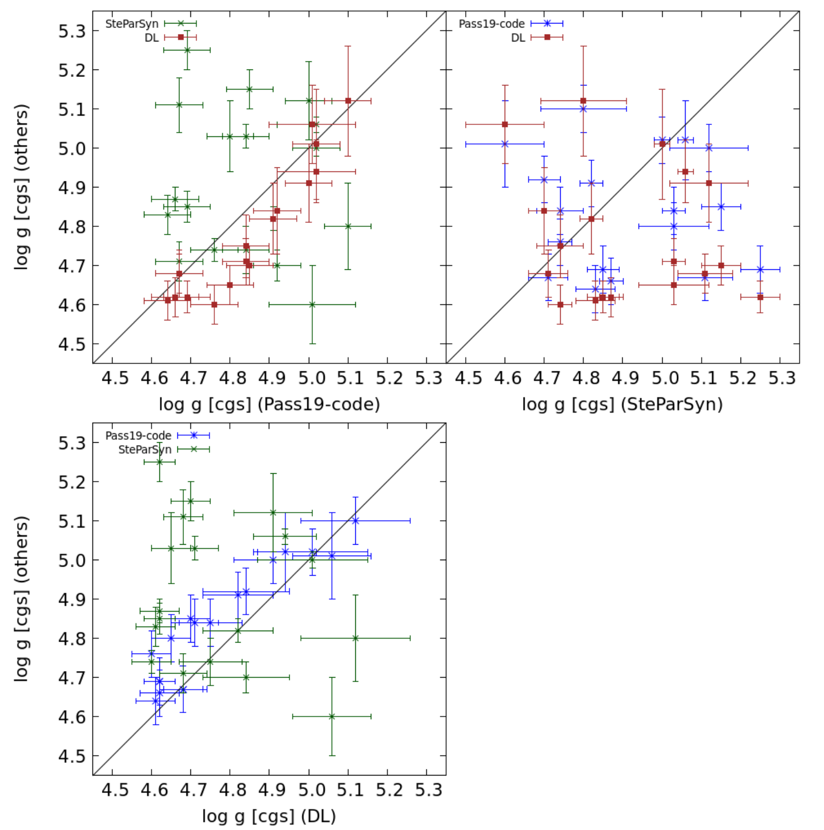

Fig. B.4. Comparison between our methods, showing the derived $\log g$ in Run A. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the x-axis label) to all other methods. Note that ODUSSEAS did not derive $\log g$.

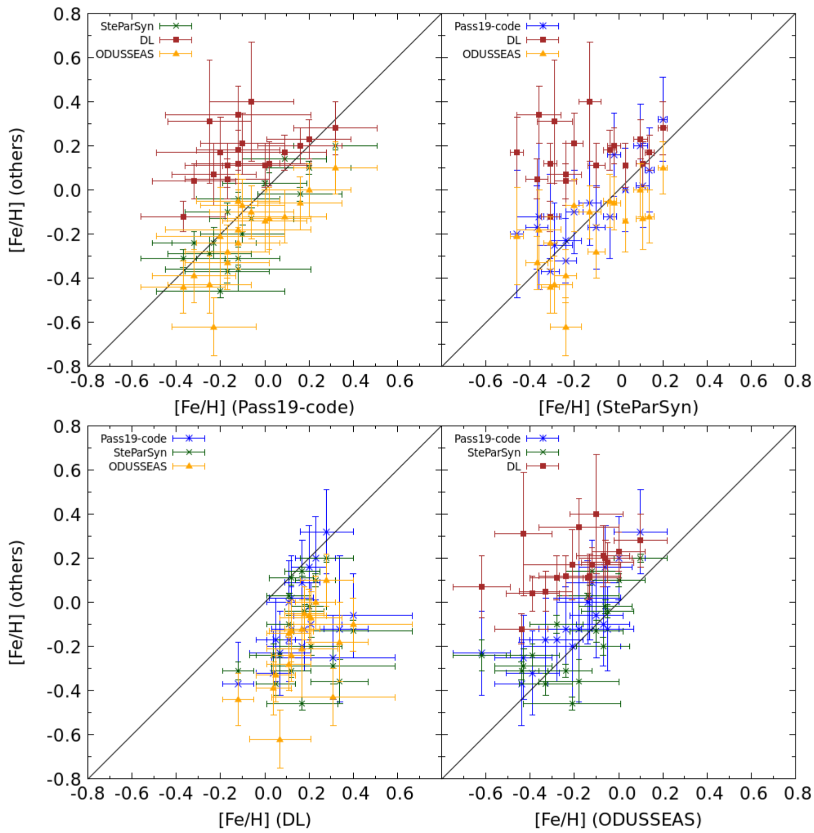

Fig. B.5. Comparison between our methods, showing the derived [Fe/H] in Run A. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the *x*-axis label) to all other methods.

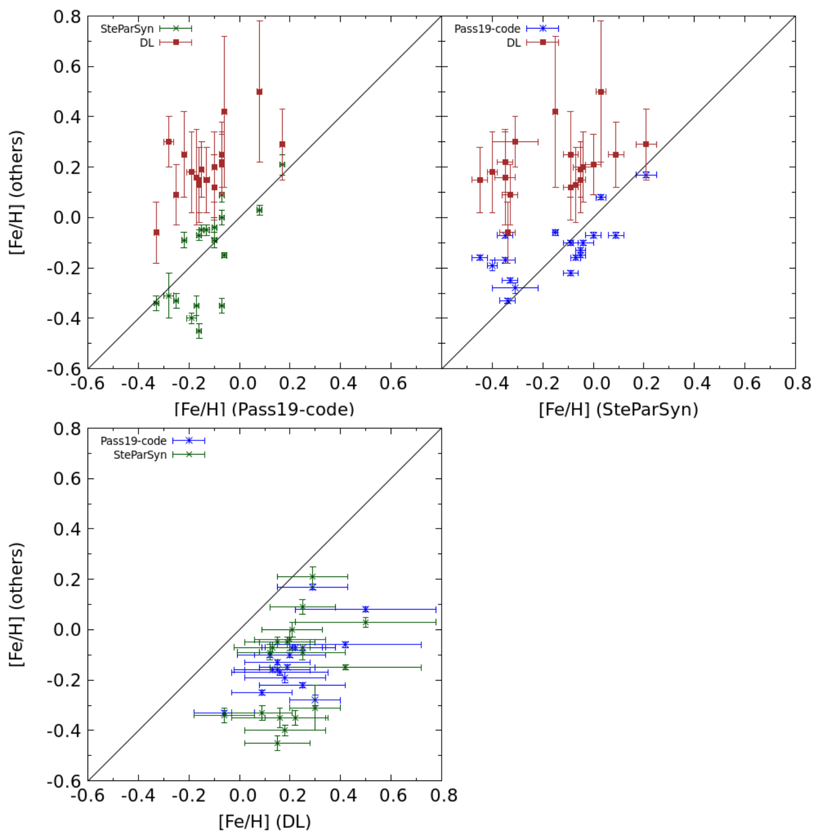

Fig. B.6. Comparison between our methods, showing the derived [Fe/H] in Run B. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the *x*-axis label) to all other methods. Note that ODUSSEAS did not participate in Run B.

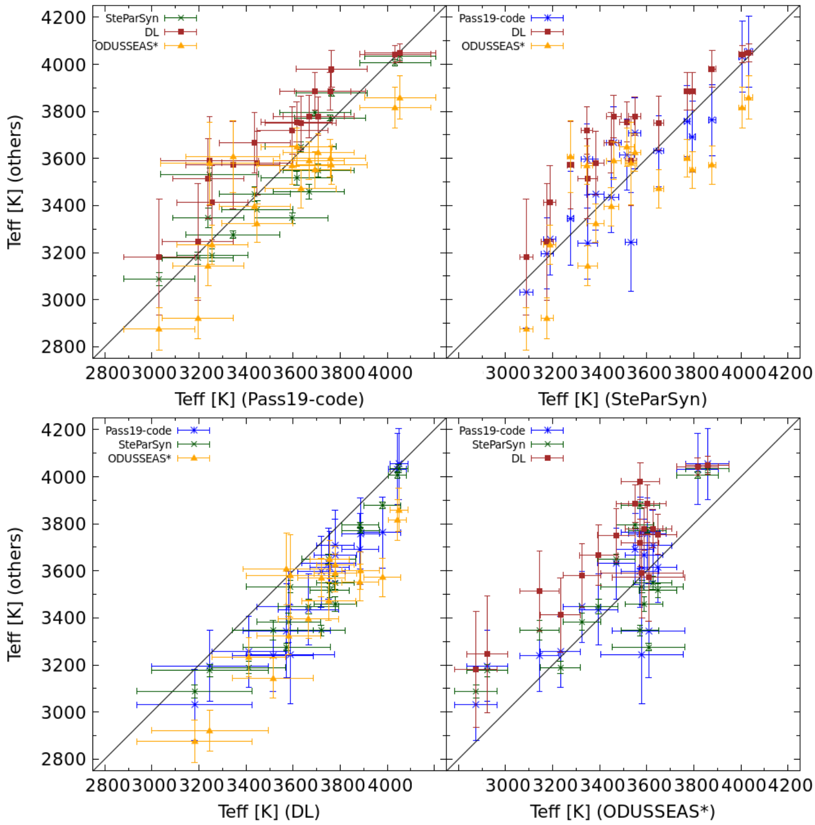

Fig. B.7. Comparison between our methods, showing the derived $T_{\rm eff}$ in Run C. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the x-axis label) to all other methods. Note that the values from ODUSSEAS correspond to Run C*.

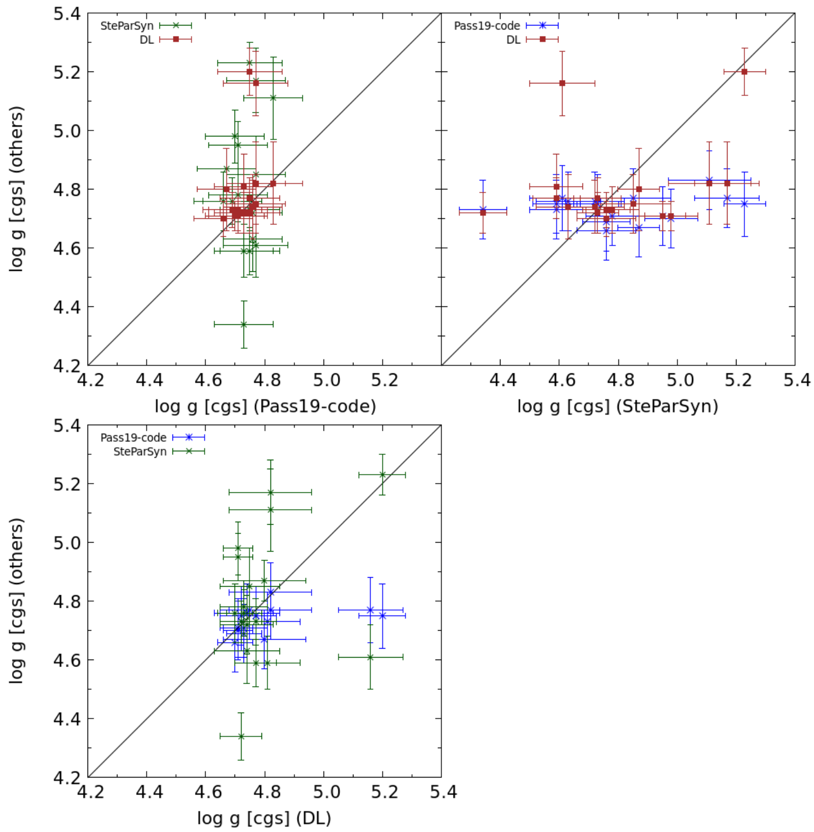

Fig. B.8. Comparison between our methods, showing the derived $\log g$ in Run C. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the x-axis label) to all other methods. Note that ODUSSEAS did not derive $\log g$.

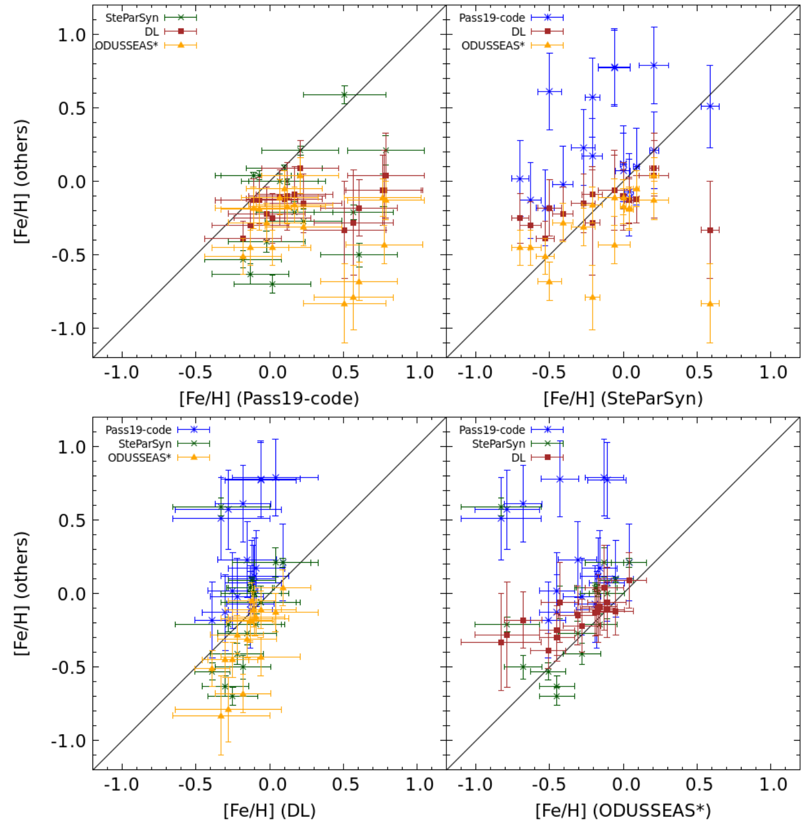

Fig. B.9. Comparison between our methods, showing the derived [Fe/H] in Run C. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the x-axis label) to all other methods. Note that the values from ODUSSEAS correspond to Run C*.

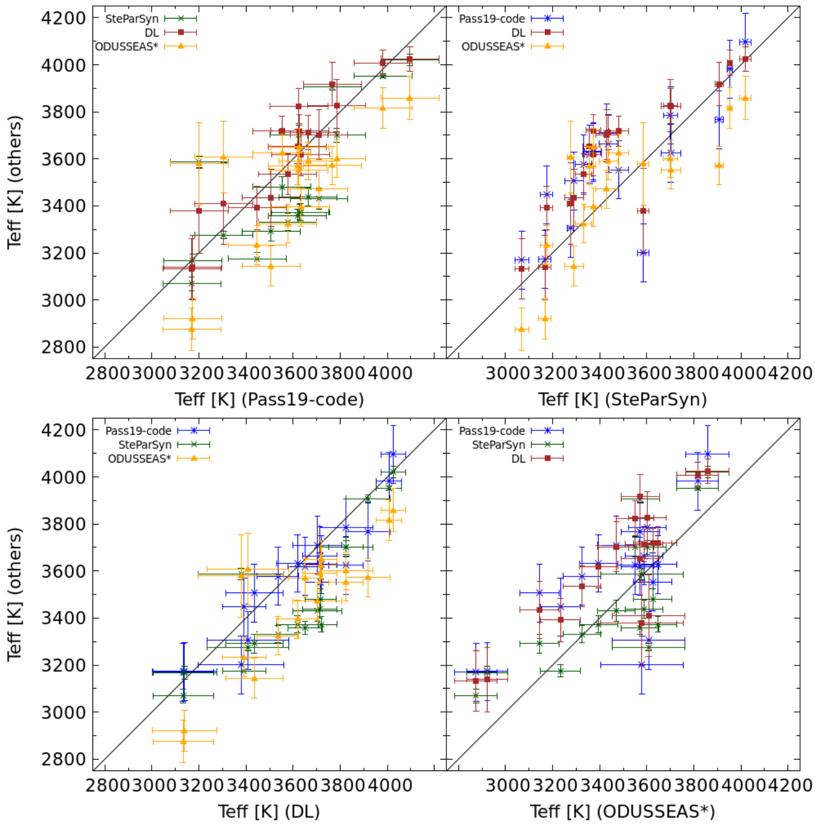

Fig. B.10. Comparison between our methods, showing the derived T_{eff} in Run C2. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the x-axis label) to all other methods. Note that the values from ODUSSEAS correspond to Run C*.

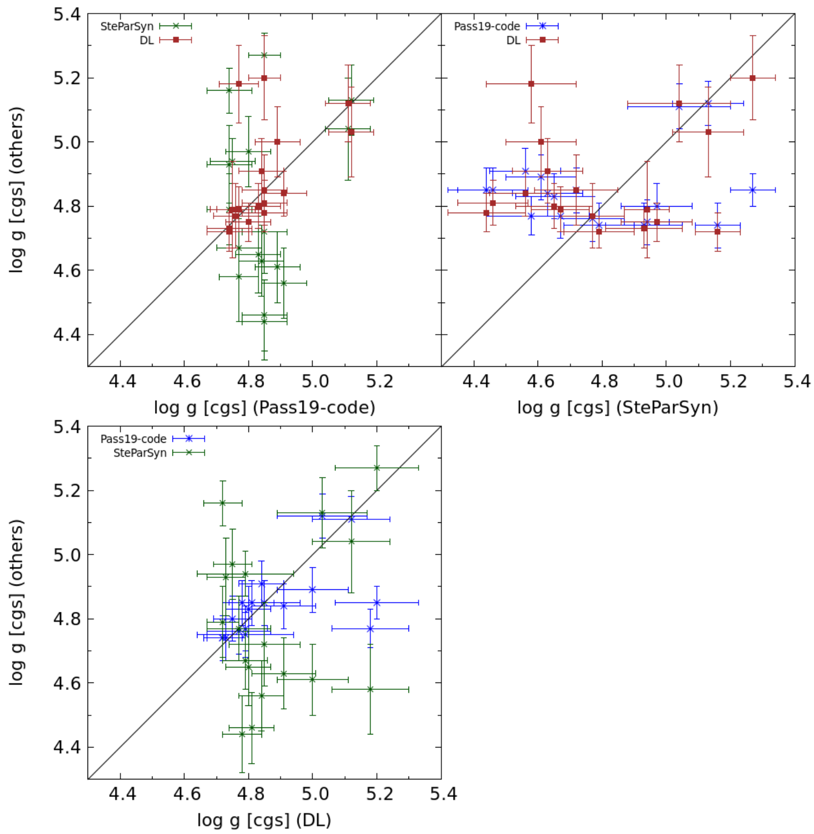

Fig. B.11. Comparison between our methods, showing the derived log g in Run C2. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the x-axis label) to all other methods. Note that ODUSSEAS did not derive log g.

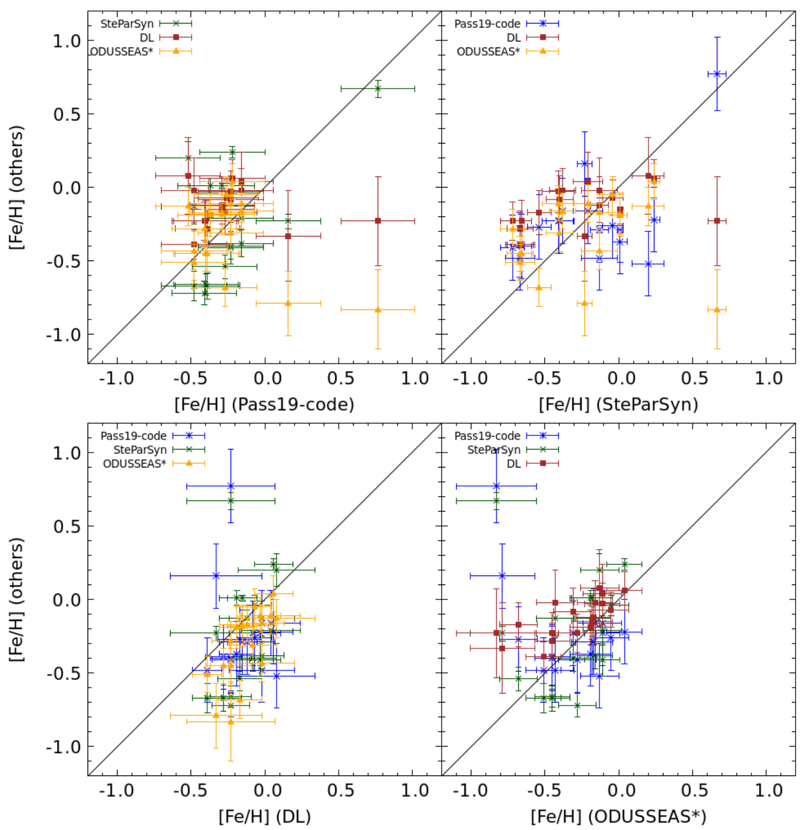

Fig. B.12. Comparison between our methods, showing the derived [Fe/H] in Run C2. Each method is indicated by a different color and symbol. Each panel compares one method (denoted by the x-axis label) to all other methods. Note that the values from ODUSSEAS correspond to Run C*.

Appendix C: Results

Table C.1. Stellar parameters for each method from Runs A and B.

Karmn	Method	Run A			Run B		
		$T_{\rm eff}$ [K]	$\log g$ [dex]	[Fe/H] [dex]	$T_{\rm eff}$ [K]	$\log g$ [dex]	[Fe/H] [dex]
J00067−075	Pass19-code	3024 ± 54	5.10 ± 0.06	−0.25 ± 0.19	2906 ± 103	5.01 ± 0.11	−0.06 ± 0.01
	SteParSyn	3023 ± 22	4.80 ± 0.11	−0.29 ± 0.08	2906 ± 103	5.01 ± 0.11	−0.15 ± 0.01
	DL	2951 ± 94	5.12 ± 0.14	+0.31 ± 0.28	2906 ± 103	5.01 ± 0.11	+0.42 ± 0.30
	ODUSSEAS	2748 ± 85	...	−0.43 ± 0.13	...	...	...
J00183+440	Pass19-code	3576 ± 54	4.84 ± 0.06	−0.32 ± 0.19	3614 ± 84	4.85 ± 0.04	−0.25 ± 0.01
	SteParSyn	3549 ± 13	5.03 ± 0.03	−0.24 ± 0.05	3614 ± 84	4.85 ± 0.04	−0.33 ± 0.03
	DL	3672 ± 45	4.71 ± 0.06	+0.04 ± 0.08	3614 ± 84	4.85 ± 0.04	+0.09 ± 0.12
	ODUSSEAS	3539 ± 78	...	−0.39 ± 0.12	...	...	...
J04429+189	Pass19-code	3716 ± 54	4.67 ± 0.06	+0.16 ± 0.19	3581 ± 85	4.73 ± 0.05	−0.07 ± 0.01
	SteParSyn	3528 ± 15	4.71 ± 0.05	−0.02 ± 0.03	3581 ± 85	4.73 ± 0.05	+0.00 ± 0.03
	DL	3630 ± 44	4.68 ± 0.06	+0.20 ± 0.08	3581 ± 85	4.73 ± 0.05	+0.21 ± 0.12
	ODUSSEAS	3376 ± 78	...	−0.06 ± 0.12	...	...	...
J05314−036	Pass19-code	3930 ± 54	4.64 ± 0.06	+0.32 ± 0.19	3779 ± 112	4.69 ± 0.05	+0.17 ± 0.01
	SteParSyn	3726 ± 10	4.83 ± 0.05	+0.20 ± 0.02	3779 ± 112	4.69 ± 0.05	+0.21 ± 0.04
	DL	3809 ± 57	4.61 ± 0.05	+0.28 ± 0.12	3779 ± 112	4.69 ± 0.05	+0.29 ± 0.14
	ODUSSEAS	3527 ± 78	...	+0.10 ± 0.12	...	...	...
J07558+833	Pass19-code	3191 ± 131	5.02 ± 0.10	−0.20 ± 0.29	3265 ± 87	5.00 ± 0.07	−0.19 ± 0.02
	SteParSyn	3355 ± 7	5.06 ± 0.02	−0.46 ± 0.03	3265 ± 87	5.00 ± 0.07	−0.40 ± 0.02
	DL	3175 ± 29	4.94 ± 0.08	+0.17 ± 0.16	3265 ± 87	5.00 ± 0.07	+0.18 ± 0.16
	ODUSSEAS	3439 ± 134	...	−0.21 ± 0.22	...	...	...
J09143+526	Pass19-code	4045 ± 54	4.69 ± 0.06	+0.00 ± 0.19	3961 ± 88	4.68 ± 0.04	−0.10 ± 0.01
	SteParSyn	3901 ± 9	4.85 ± 0.04	+0.03 ± 0.01	3961 ± 88	4.68 ± 0.04	−0.09 ± 0.03
	DL	4076 ± 62	4.62 ± 0.04	+0.11 ± 0.10	3961 ± 88	4.68 ± 0.04	+0.12 ± 0.13
	ODUSSEAS	3830 ± 85	...	−0.14 ± 0.14	...	...	...
J09144+526	Pass19-code	4021 ± 54	4.69 ± 0.06	+0.02 ± 0.19	3894 ± 89	4.80 ± 0.04	−0.13 ± 0.01
	SteParSyn	3891 ± 18	5.25 ± 0.05	+0.11 ± 0.03	3894 ± 89	4.80 ± 0.04	−0.05 ± 0.02
	DL	4032 ± 60	4.62 ± 0.04	+0.12 ± 0.10	3894 ± 89	4.80 ± 0.04	+0.15 ± 0.13
	ODUSSEAS	3805 ± 84	...	−0.13 ± 0.14	...	...	...
J10508+068	Pass19-code	3284 ± 54	4.92 ± 0.06	−0.10 ± 0.19	3136 ± 82	4.83 ± 0.06	−0.22 ± 0.01
	SteParSyn	3232 ± 11	4.70 ± 0.04	−0.20 ± 0.04	3136 ± 82	4.83 ± 0.06	−0.09 ± 0.03
	DL	3281 ± 54	4.84 ± 0.11	+0.21 ± 0.14	3136 ± 82	4.83 ± 0.06	+0.25 ± 0.17
	ODUSSEAS	3090 ± 79	...	−0.07 ± 0.12	...	...	...
J11033+359	Pass19-code	3555 ± 54	4.80 ± 0.06	−0.17 ± 0.19	3566 ± 76	4.79 ± 0.05	−0.16 ± 0.01
	SteParSyn	3550 ± 12	5.03 ± 0.09	−0.37 ± 0.05	3566 ± 76	4.79 ± 0.05	−0.45 ± 0.03
	DL	3766 ± 57	4.65 ± 0.05	+0.05 ± 0.09	3566 ± 76	4.79 ± 0.05	+0.15 ± 0.13
	ODUSSEAS	3469 ± 78	...	−0.33 ± 0.12	...	...	...
J11054+435	Pass19-code	3609 ± 54	4.85 ± 0.06	−0.37 ± 0.19	3633 ± 95	4.80 ± 0.04	−0.33 ± 0.01
	SteParSyn	3566 ± 14	5.15 ± 0.05	−0.31 ± 0.04	3633 ± 95	4.80 ± 0.04	−0.34 ± 0.03
	DL	3774 ± 44	4.70 ± 0.05	−0.12 ± 0.07	3633 ± 95	4.80 ± 0.04	−0.06 ± 0.12
	ODUSSEAS	3581 ± 80	...	−0.44 ± 0.12	...	...	...
J11421+267	Pass19-code	3455 ± 54	4.84 ± 0.06	−0.12 ± 0.19	3468 ± 81	4.83 ± 0.05	−0.10 ± 0.01
	SteParSyn	3492 ± 17	4.74 ± 0.06	−0.04 ± 0.03	3468 ± 81	4.83 ± 0.05	−0.04 ± 0.04
	DL	3514 ± 47	4.75 ± 0.08	+0.18 ± 0.09	3468 ± 81	4.83 ± 0.05	+0.20 ± 0.14
	ODUSSEAS	3314 ± 78	...	−0.05 ± 0.12	...	...	...
J13005+056	Pass19-code	3142 ± 134	5.01 ± 0.11	−0.12 ± 0.33	3137 ± 100	4.79 ± 0.10	−0.28 ± 0.02
	SteParSyn	3148 ± 40	4.60 ± 0.10	−0.36 ± 0.10	3137 ± 100	4.79 ± 0.10	−0.31 ± 0.09
	DL	3071 ± 46	5.06 ± 0.10	+0.34 ± 0.13	3137 ± 100	4.79 ± 0.10	+0.30 ± 0.10
	ODUSSEAS	3417 ± 117	...	−0.18 ± 0.18	...	...	...
J13457+148	Pass19-code	3628 ± 54	4.76 ± 0.06	−0.12 ± 0.19	3667 ± 88	4.75 ± 0.04	−0.07 ± 0.01
	SteParSyn	3569 ± 23	4.74 ± 0.03	−0.31 ± 0.03	3667 ± 88	4.75 ± 0.04	−0.35 ± 0.03
	DL	3975 ± 70	4.60 ± 0.05	+0.12 ± 0.09	3667 ± 88	4.75 ± 0.04	+0.22 ± 0.12
	ODUSSEAS	3590 ± 78	...	−0.24 ± 0.12	...	...	...
J15194−077	Pass19-code	3390 ± 54	4.91 ± 0.06	−0.17 ± 0.19	3399 ± 82	4.87 ± 0.06	−0.16 ± 0.01
	SteParSyn	3422 ± 10	4.82 ± 0.03	−0.10 ± 0.04	3399 ± 82	4.87 ± 0.06	−0.07 ± 0.02

Table C.1. continued.

Karmn	Method	Run A $T_{\rm eff}$ [K]	Run A $\log g$ [dex]	Run A [Fe/H] [dex]	Run B $T_{\rm eff}$ [K]	Run B $\log g$ [dex]	Run B [Fe/H] [dex]
	DL	3385 ± 47	4.82 ± 0.09	+0.11 ± 0.10	3399 ± 82	4.87 ± 0.06	+0.13 ± 0.15
	ODUSSEAS	3280 ± 79	...	−0.28 ± 0.12	...	...	...
J16581+257	Pass19-code	3825 ± 54	4.67 ± 0.06	+0.09 ± 0.19	3678 ± 78	4.74 ± 0.06	−0.15 ± 0.01
	SteParSyn	3673 ± 12	5.11 ± 0.07	+0.14 ± 0.02	3678 ± 78	4.74 ± 0.06	−0.05 ± 0.02
	DL	3748 ± 43	4.68 ± 0.05	+0.17 ± 0.08	3678 ± 78	4.74 ± 0.06	+0.19 ± 0.11
	ODUSSEAS	3561 ± 79	...	−0.12 ± 0.12	...	...	...
J17578+046	Pass19-code	3231 ± 54	5.00 ± 0.06	−0.23 ± 0.19	3242 ± 75	5.06 ± 0.07	−0.17 ± 0.01
	SteParSyn	3282 ± 14	5.12 ± 0.10	−0.24 ± 0.07	3242 ± 75	5.06 ± 0.07	−0.35 ± 0.04
	DL	3352 ± 55	4.91 ± 0.10	+0.07 ± 0.14	3242 ± 75	5.06 ± 0.07	+0.16 ± 0.19
	ODUSSEAS	3172 ± 80	...	−0.62 ± 0.13	...	...	...
J22565+165	Pass19-code	3842 ± 54	4.66 ± 0.06	+0.20 ± 0.19	3673 ± 79	4.68 ± 0.04	−0.07 ± 0.01
	SteParSyn	3714 ± 9	4.87 ± 0.03	+0.10 ± 0.03	3673 ± 79	4.68 ± 0.04	+0.09 ± 0.03
	DL	3765 ± 49	4.62 ± 0.05	+0.23 ± 0.09	3673 ± 79	4.68 ± 0.04	+0.25 ± 0.13
	ODUSSEAS	3509 ± 79	...	+0.00 ± 0.12	...	...	...
J23419+441	Pass19-code	3069 ± 54	5.02 ± 0.06	−0.06 ± 0.19	3031 ± 80	5.02 ± 0.10	+0.08 ± 0.01
	SteParSyn	3140 ± 7	5.00 ± 0.02	−0.13 ± 0.05	3031 ± 80	5.02 ± 0.10	+0.03 ± 0.02
	DL	2995 ± 81	5.01 ± 0.14	+0.40 ± 0.27	3031 ± 80	5.02 ± 0.10	+0.50 ± 0.28
	ODUSSEAS	2831 ± 79	...	−0.10 ± 0.12	...	...	...

Table C.2. Stellar parameters for each method from Runs C and C2.

Karmn	Method	Run C $T_{\rm eff}$ [K]	Run C $\log g$ [dex]	Run C [Fe/H] [dex]	Run C2 $T_{\rm eff}$ [K]	Run C2 $\log g$ [dex]	Run C2 [Fe/H] [dex]
J00067−075	Pass19-code	3031 ± 151	4.83 ± 0.10	+0.78 ± 0.26	3169 ± 123	5.11 ± 0.07	−0.48 ± 0.22
	SteParSyn	3088 ± 28	5.11 ± 0.14	−0.06 ± 0.11	3069 ± 29	5.04 ± 0.16	−0.13 ± 0.12
	DL	3181 ± 246	4.82 ± 0.14	−0.06 ± 0.27	3133 ± 129	5.12 ± 0.12	−0.02 ± 0.22
	ODUSSEAS	2875 ± 90*	...	−0.43 ± 0.13*	...	...	...
J00183+440	Pass19-code	3667 ± 151	4.75 ± 0.10	−0.13 ± 0.26	3664 ± 123	4.83 ± 0.07	−0.39 ± 0.22
	SteParSyn	3459 ± 31	4.59 ± 0.08	−0.63 ± 0.07	3437 ± 39	4.65 ± 0.12	−0.67 ± 0.09
	DL	3779 ± 90	4.77 ± 0.07	−0.30 ± 0.16	3713 ± 73	4.80 ± 0.07	−0.28 ± 0.12
	ODUSSEAS	3589 ± 80*	...	−0.45 ± 0.12*	...	...	...
J04429+189	Pass19-code	3632 ± 151	4.71 ± 0.10	+0.07 ± 0.26	3710 ± 123	4.76 ± 0.07	−0.23 ± 0.22
	SteParSyn	3651 ± 21	4.78 ± 0.09	+0.00 ± 0.05	3430 ± 46	4.77 ± 0.10	−0.41 ± 0.11
	DL	3751 ± 114	4.73 ± 0.08	−0.11 ± 0.19	3703 ± 105	4.77 ± 0.10	−0.03 ± 0.14
	ODUSSEAS	3471 ± 81*	...	−0.11 ± 0.12*	...	...	...
J05314−036	Pass19-code	3763±151	4.66±0.10	+0.21 ± 0.26	3766 ± 123	4.74 ± 0.07	−0.22 ± 0.22
	SteParSyn	3878 ± 15	4.76 ± 0.10	+0.21 ± 0.03	3908 ± 17	4.79 ± 0.11	+0.24 ± 0.04
	DL	3980 ± 78	4.70 ± 0.06	+0.09 ± 0.19	3918 ± 93	4.72 ± 0.05	+0.06 ± 0.13
	ODUSSEAS	3572 ± 81*	...	+0.04 ± 0.12*	...	...	...
J07558+833	Pass19-code	3345 ± 199	4.75 ± 0.11	+0.57 ± 0.27	3305 ± 123	4.85 ± 0.05	+0.16 ± 0.22
	SteParSyn	3276 ± 15	5.23 ± 0.07	−0.21 ± 0.05	3276 ± 15	5.27 ± 0.07	−0.23 ± 0.05
	DL	3572 ± 186	5.20 ± 0.08	−0.28 ± 0.36	3409 ± 173	5.20 ± 0.13	−0.33 ± 0.31
	ODUSSEAS	3608 ± 154*	...	−0.79 ± 0.22*	...	...	...
J09143+526	Pass19-code	4054 ± 151	4.70 ± 0.10	−0.07 ± 0.26	4096 ± 151	4.74 ± 0.07	−0.37 ± 0.22
	SteParSyn	4034 ± 17	4.98 ± 0.09	+0.04 ± 0.03	4020 ± 24	4.93 ± 0.12	+0.01 ± 0.05
	DL	4049 ± 38	4.71 ± 0.05	−0.13 ± 0.16	4026 ± 52	4.73 ± 0.06	−0.19 ± 0.12
	ODUSSEAS	3859 ± 92*	...	−0.19 ± 0.14*	...	...	...
J09144+526	Pass19-code	4033 ± 151	4.71 ± 0.10	−0.11 ± 0.26	3982 ± 123	4.74 ± 0.07	−0.29 ± 0.22
	SteParSyn	4006 ± 14	4.95 ± 0.08	+0.04 ± 0.02	3953 ± 9	5.16 ± 0.07	+0.01 ± 0.02
	DL	4043 ± 37	4.71 ± 0.05	−0.13 ± 0.15	4008 ± 54	4.72 ± 0.06	−0.15 ± 0.10
	ODUSSEAS	3816 ± 87*	...	−0.18 ± 0.14*	...	...	...

Table C.2. continued.

Karmn	Method	Run C $T_{\rm eff}$ [K]	Run C $\log g$ [dex]	Run C [Fe/H] [dex]	Run C2 $T_{\rm eff}$ [K]	Run C2 $\log g$ [dex]	Run C2 [Fe/H] [dex]
J10508+068	Pass19-code	3239 ± 151	4.77 ± 0.10	+0.77 ± 0.26	3506 ± 123	4.85 ± 0.07	−0.16 ± 0.22
	SteParSyn	3348 ± 42	4.85 ± 0.13	−0.06 ± 0.11	3291 ± 40	4.72 ± 0.13	−0.21 ± 0.10
	DL	3514 ± 171	4.75 ± 0.10	−0.06 ± 0.24	3435 ± 121	4.85 ± 0.11	+0.04 ± 0.20
	ODUSSEAS	3144 ± 84*	...	−0.11 ± 0.13*	...	...	...
J11033+359	Pass19-code	3597 ± 151	4.73 ± 0.10	+0.02 ± 0.26	3619 ± 123	4.85 ± 0.07	−0.40 ± 0.22
	SteParSyn	3346 ± 24	4.34 ± 0.08	−0.70 ± 0.06	3357 ± 28	4.46 ± 0.11	−0.66 ± 0.07
	DL	3719 ± 101	4.72 ± 0.07	−0.25 ± 0.17	3652 ± 73	4.81 ± 0.07	−0.23 ± 0.14
	ODUSSEAS	3571 ± 81*	...	−0.45 ± 0.12*	...	...	...
J11054+435	Pass19-code	3707 ± 151	4.75 ± 0.10	−0.18 ± 0.26	3553 ± 123	4.91 ± 0.07	−0.48 ± 0.22
	SteParSyn	3549 ± 27	4.73 ± 0.08	−0.53 ± 0.06	3481 ± 42	4.56 ± 0.11	−0.67 ± 0.10
	DL	3779 ± 81	4.77 ± 0.07	−0.39 ± 0.12	3717 ± 65	4.84 ± 0.07	−0.39 ± 0.10
	ODUSSEAS	3625 ± 81*	...	−0.51 ± 0.12*	...	...	...
J11421+267	Pass19-code	3436 ± 151	4.75 ± 0.10	+0.17 ± 0.26	3632 ± 123	4.77 ± 0.07	−0.16 ± 0.22
	SteParSyn	3449 ± 32	4.73 ± 0.07	−0.21 ± 0.07	3373 ± 36	4.67 ± 0.09	−0.38 ± 0.09
	DL	3666 ± 128	4.72 ± 0.07	−0.09 ± 0.19	3618 ± 89	4.79 ± 0.07	−0.02 ± 0.15
	ODUSSEAS	3395 ± 81*	...	−0.16 ± 0.12*	...	...	...
J13005+056	Pass19-code	3245 ± 210	4.77 ± 0.11	+0.51 ± 0.28	3201 ± 123	4.77 ± 0.06	+0.77 ± 0.25
	SteParSyn	3533 ± 5	4.61 ± 0.11	+0.59 ± 0.06	3586 ± 24	4.58 ± 0.14	+0.67 ± 0.06
	DL	3589 ± 188	5.16 ± 0.11	−0.33 ± 0.33	3379 ± 181	5.18 ± 0.12	−0.23 ± 0.30
	ODUSSEAS	3579 ± 175*	...	−0.83 ± 0.27*	...	...	...
J13457+148	Pass19-code	3615 ± 151	4.73 ± 0.10	−0.02 ± 0.26	3627 ± 123	4.85 ± 0.07	−0.41 ± 0.22
	SteParSyn	3516 ± 31	4.59 ± 0.09	−0.41 ± 0.07	3373 ± 32	4.44 ± 0.12	−0.72 ± 0.08
	DL	3755 ± 84	4.81 ± 0.11	−0.22 ± 0.18	3720 ± 68	4.78 ± 0.06	−0.23 ± 0.13
	ODUSSEAS	3648 ± 81*	...	−0.28 ± 0.13*	...	...	...
J15194−077	Pass19-code	3447 ± 151	4.76 ± 0.10	+0.23 ± 0.26	3578 ± 123	4.84 ± 0.07	−0.23 ± 0.22
	SteParSyn	3383 ± 36	4.72 ± 0.10	−0.27 ± 0.08	3332 ± 35	4.63 ± 0.11	−0.40 ± 0.09
	DL	3581 ± 135	4.74 ± 0.09	−0.15 ± 0.20	3535 ± 87	4.91 ± 0.10	−0.08 ± 0.16
	ODUSSEAS	3325 ± 81*	...	−0.31 ± 0.13*	...	...	...
J16581+257	Pass19-code	3758 ± 151	4.67 ± 0.10	+0.12 ± 0.26	3785 ± 123	4.75 ± 0.07	−0.29 ± 0.22
	SteParSyn	3772 ± 12	4.87 ± 0.07	+0.00 ± 0.02	3701 ± 29	4.94 ± 0.07	−0.13 ± 0.06
	DL	3886 ± 80	4.80 ± 0.14	−0.10 ± 0.23	3826 ± 112	4.79 ± 0.15	−0.12 ± 0.19
	ODUSSEAS	3602 ± 80*	...	−0.17 ± 0.12*	...	...	...
J17578+046	Pass19-code	3256 ± 151	4.76 ± 0.10	+0.61 ± 0.26	3448 ± 123	4.89 ± 0.07	−0.27 ± 0.22
	SteParSyn	3189 ± 26	4.63 ± 0.11	−0.50 ± 0.08	3175 ± 26	4.61 ± 0.11	−0.54 ± 0.08
	DL	3412 ± 157	4.74 ± 0.11	−0.18 ± 0.19	3392 ± 92	5.00 ± 0.11	−0.17 ± 0.15
	ODUSSEAS	3233 ± 84*	...	−0.68 ± 0.13*	...	...	...
J22565+165	Pass19-code	3693 ± 151	4.69 ± 0.10	+0.10 ± 0.26	3624 ± 123	4.80 ± 0.07	−0.26 ± 0.22
	SteParSyn	3795 ± 12	4.76 ± 0.08	+0.09 ± 0.02	3702 ± 41	4.97 ± 0.11	−0.04 ± 0.07
	DL	3885 ± 79	4.73 ± 0.06	−0.12 ± 0.16	3824 ± 74	4.75 ± 0.06	−0.07 ± 0.12
	ODUSSEAS	3551 ± 79*	...	−0.05 ± 0.12*	...	...	...
J23419+441	Pass19-code	3195 ± 151	4.77 ± 0.10	+0.79 ± 0.26	3173 ± 123	5.12 ± 0.07	−0.52 ± 0.22
	SteParSyn	3176 ± 26	5.17 ± 0.11	+0.21 ± 0.10	3167 ± 27	5.13 ± 0.11	+0.20 ± 0.11
	DL	3246 ± 248	4.82 ± 0.14	+0.04 ± 0.29	3139 ± 137	5.03 ± 0.14	+0.08 ± 0.26
	ODUSSEAS	2922 ± 87*	...	−0.83 ± 0.27*	...	...	...

Notes. (*) corresponding to Run C*.