A&A 671, A61 (2023)
https://doi.org/10.1051/0004-6361/202244765


Astronomy & Astrophysics

# The *Gaia*-ESO Survey: Preparing the ground for 4MOST and WEAVE galactic surveys

## Chemical evolution of lithium with machine learning⋆,⋆⋆,⋆⋆⋆

S. Nepal[1,2], G. Guiglion[3,1], R. S. de Jong[1], M. Valentini[1], C. Chiappini[1], M. Steinmetz[1], M. Ambrosch[4], E. Pancino[5], R. D. Jeffries[6], T. Bensby[7], D. Romano[8], R. Smiljanic[9], M. L. L. Dantas[9], G. Gilmore[10], S. Randich[5], A. Bayo[11], M. Bergemann[12,3], E. Franciosini[5], F. Jiménez-Esteban[13], P. Jofré[14], L. Morbidelli[5], G. G. Sacco[5], G. Tautvaišienė[4], and S. Zaggia[15]

[1] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
e-mail: snepal@aip.de
[2] Institut für Physik und Astronomie, Universität Potsdam, Karl-Liebknecht-Str. 24/25, 14476 Potsdam, Germany
[3] Max Planck Institute for Astronomy, Könnigstuhl 17, 69117 Heidelberg, Germany
e-mail: guiglion@mpia.de
[4] Institute of Theoretical Physics and Astronomy, Vilnius University, Sauletekio av. 3, 10257 Vilnius, Lithuania
[5] INAF, Osservatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, 50125 Firenze, Italy
[6] Astrophysics Group, Keele University, Keele, Staffordshire ST5 5BG, UK
[7] Lund Observatory, Department of Astronomy and Theoretical Physics, Box 43, 22100 Lund, Sweden
[8] INAF, Osservatorio di Astrofisica e Scienza dello Spazio, Via Gobetti 93/3, 40129 Bologna, Italy
[9] Nicolaus Copernicus Astronomical Center, Polish Academy of Sciences, ul. Bartycka 18, 00-716 Warsaw, Poland
[10] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[11] European Southern Observatory, Karl Schwarzschild-Straße 2, 85748 Garching bei München, Germany
[12] Niels Bohr International Academy, Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark
[13] Departamento de Astrofísica, Centro de Astrobiología (CSIC-INTA), ESAC Campus, Camino Bajo del Castillo s/n, 28692 Villanueva de la Cañada, Madrid, Spain
[14] Núcleo de Astronomía, Facultad de Ingeniería y Ciencias, Universidad Diego Portales (UDP), Av. Ejército Libertador 441, Santiago, de Chile
[15] INAF, Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio, 5, 35122 Padova, Italy

Received 18 August 2022 / Accepted 5 December 2022

**ABSTRACT**

*Context.* With its origin coming from several sources (Big Bang, stars, cosmic rays) and given its strong depletion during its stellar lifetime, the lithium element is of great interest as its chemical evolution in the Milky Way is not well understood at present. To help constrain stellar and galactic chemical evolution models, numerous and precise lithium abundances are necessary for a large range of evolutionary stages, metallicities, and Galactic volume.

*Aims.* In the age of stellar parametrization on industrial scales, spectroscopic surveys such as APOGEE, GALAH, RAVE, and LAMOST have used data-driven methods to rapidly and precisely infer stellar labels (atmospheric parameters and abundances). To prepare the ground for future spectroscopic surveys such as 4MOST and WEAVE, we aim to apply machine learning techniques to lithium measurements and analyses.

*Methods.* We trained a convolution neural network (CNN), coupling *Gaia*-ESO Survey iDR6 stellar labels ($T_{\rm eff}$, log($g$), [Fe/H], and A(Li)) and GIRAFFE HR15N spectra, to infer the atmospheric parameters and lithium abundances for ∼40 000 stars. The CNN architecture and accompanying notebooks are available online via GitHub.

*Results.* We show that the CNN properly learns the physics of the stellar labels, from relevant spectral features through a broad range of evolutionary stages and stellar parameters. The lithium feature at 6707.8 Å is successfully singled out by our CNN, among the thousands of lines in the GIRAFFE HR15N setup. Rare objects such as lithium-rich giants are found in our sample. This level of performance is achieved thanks to a meticulously built, high-quality, and homogeneous training sample.

*Conclusions.* The CNN approach is very well adapted for the next generations of spectroscopic surveys aimed at studying (among other elements) lithium, such as the 4MIDABLE-LR/HR (4MOST Milky Way disk and bulge low- and high-resolution) surveys. In this context, the caveats of machine-learning applications should be appropriately investigated, along with the realistic label uncertainties and upper limits for abundances.

**Key words.** techniques: spectroscopic – methods: data analysis – surveys – stars: fundamental parameters – stars: abundances – Galaxy: stellar content

⋆ Full Table 1 is only available at the CDS via anonymous ftp to cdsarc.cds.unistra.fr (130.79.128.5) or via https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/671/A61
⋆⋆ https://github.com/SamirNepal/Li_CNN_2022
⋆⋆⋆ Based on observations collected with ESO telescopes at the La Silla Paranal Observatory in Chile, for the *Gaia*-ESO Large Public Spectroscopic Survey (188.B-3002, 193.B-0936, 197.B-1074).

## 1. Introduction

The element lithium[1] (Li) is of particular interest in astrophysics given its complex origin and evolution. Lithium was produced during the big bang (BB), and its primordial abundance can be used to constrain the standard model of cosmology. The standard BB nucleosynthesis (SBBN) model predicts the primordial lithium abundance to be A(Li)[2] ∼2.75 dex (Pitrou et al. 2018). Attempts to obtain an astrophysical measurement of this primordial Li using old, warm ($T_{\rm eff}$ > 5600 K), metal-poor ([Fe/H] < −1.5 dex) halo dwarf stars has resulted in observation of a thin spread of lithium abundance that is independent of metallicity and effective temperature – referred to as the "Spite plateau," with A(Li) ∼ 2.2 dex (Spite & Spite 1982; Bonifacio & Molaro 1997). This difference of a factor of three between the theoretical prediction and observation brings on the famous cosmological lithium problem (e.g., Fields 2011).

At later times, Li is produced at two distinct sources; in the interstellar medium (ISM) via a spallative interaction of galactic cosmic rays and the ISM through the $p$+C,N,O or $\alpha$+C,N,O reaction channels (Reeves et al. 1970) as well as in stellar sources such as asymptotic giant branch (AGB) stars (McKellar 1940), and red giants (Sackmann & Boothroyd 1999), as well as core-collapse supernovae and novae (D'Antona & Matteucci 1991; Izzo et al. 2015). However, the stellar yields for the different sources are not well constrained and present large uncertainties (Matteucci et al. 1995; Romano et al. 1999, 2001; Prantzos et al. 2017; Randich & Magrini 2021).

One production channel for Li in the stars is known as the Cameron–Fowler mechanism (Cameron & Fowler 1971) whereby $^7$Be is first formed in temperatures hotter than $4 \times 10^7$ K via the reaction $^3\mathrm{He} + \alpha \rightarrow {}^7\mathrm{Be} + \gamma$. The fresh $^7$Be must then be quickly moved to cooler layers by convection, where it decays to $^7$Li and is conserved and eventually released to the ISM. This mechanism explains the existence of Li-rich giants (Brown et al. 1989; Charbonnel & Balachandran 2000; Hong-liang & Jian-rong 2022). Lithium could also be produced via the $\nu$-process taking place in the external shells of collapsing massive stars (Woosley & Weaver 1995; Kusakabe et al. 2019).

Additionally, Li can already be easily destroyed in stars by the proton capture reaction $^7\mathrm{Li}(p,\alpha)^4\mathrm{He}$ at temperatures as low as $2.5 \times 10^6$ K as early as the pre-main sequence (PMS) and in later stages, whenever that temperature is reached (Pinsonneault 1997). For example, the meteoritic A(Li) is ∼3.26 dex (Lodders & Palme 2009), which represents the initial ISM Li for the Sun; whereas the Solar photospheric abundance of only A(Li) ∼ 1.05 dex (Grevesse et al. 2007) suggests an internal destruction by a factor >150.

In order to investigate the stellar and galactic evolution of lithium, we need a statistically robust and homogeneous sample, such that a large metallicity domain and different evolutionary stages are covered. In recent years, due to the availability of larger samples of stars (typically several hundred), it has become possible to study lithium abundance in the context of chemical evolution of the thick and thin disks, internal destruction in stars, galactic chemical evolution, and exoplanet connection (Lambert & Reddy 2004; Ramírez et al. 2012; Delgado Mena et al. 2015; Bensby & Lind 2018). For example, Guiglion et al. (2016) used high-resolution spectra from ESO to homogeneously build a Li catalog composed of 7300 stars, while studying the lithium evolution in the Milky Way. Most recently, the number of stars with available Li abundances has rapidly increased thanks to large-scale Milky Way spectroscopic surveys such as *Gaia*-ESO (Fu et al. 2018; Randich et al. 2020; Magrini et al. 2021b; Romano et al. 2021), LAMOST (Gao et al. 2019), and GALAH (Gao et al. 2020), contributing significantly to our understanding of the evolution of Li.

One way to precisely measure atmospheric parameters and chemical abundances in stellar atmosphere is to use stellar spectroscopy. Lithium abundance is usually derived from the Li doublet at 6707.8 Å, shown in Fig. 1, which is the strongest Li feature in the optical wavelength regime. Other neutral Li lines at 6103 Å and 8126 Å have also been used for Li abundance analysis (Gratton & D'Antona 1989), but these lines are very weak and they are only detectable and measurable in high-resolution and/or at high-Li abundances. The 6707.8 Å Li line strength has a strong dependence on the star's effective temperature and Li abundance. The Li doublet blends with the Fe I line, thus making it challenging for classical spectroscopic pipelines to provide precise Li abundances at intermediate and low resolution or in the presence of noise.

Over the last three decades, the community has generally measured Li abundances using classical spectroscopic pipelines[3] (SME, Valenti & Piskunov 1996; MOOG, Sneden et al. 2012). In the era of future large spectroscopic surveys such as 4MOST (de Jong et al. 2019), and WEAVE (Dalton 2016), a number of $10^7$ spectra will be gathered and supplemented by the wealth of astrometric and photometric data provided by the *Gaia* satellite (Gaia Collaboration 2016, 2021; Lindegren et al. 2021). The community will have to adapt their methods and machine learning is believed to be the way forward.

Machine learning (ML) tools are becoming popular for all research fields where it is necessary to quickly process large amount of data and/or automatically learn the complex correlations from high-dimensional data. One family of extremely versatile ML algorithms are neural networks (NN), which have become very popular and successfully applied in many other astronomy fields, such as gravitational lensing (Petrillo et al. 2017), the search for open clusters in *Gaia* data (Castro-Ginard et al. 2020), detecting outliers in astronomical imaging data sets (Margalef-Bentabol et al. 2020) detecting gravitational waves (Lin & Wu 2021), photometric redshift predictions (Lima et al. 2022), and many more. Neural networks have actually been used in astrophysical applications for a long time, even though their architecture was relatively simple compared to the modern networks. For example: Bailer-Jones et al. (1997) used NN to parametrize $T_{\rm eff}$, log($g$), and [M/H] from stellar spectra and Bailer-Jones et al. (1998) used NN and principal component analysis (PCA) to classify spectral types.

Such machine learning approaches have also started to play an important role in the derivation of stellar labels. Such methods transfer the knowledge from a reference set of data, a so-called "training sample," to a larger set of data to derive the stellar labels. The reference set of data can be constructed from either empirical data or by employing spectral synthesis models. The Cannon (Ness et al. 2015) is one of the pioneering data-driven spectroscopic analysis tools, while the Payne (Ting et al. 2019) has demonstrated that we can combine physical stellar models using neural networks as a function to generate spectra, instead

[1] Unless differently indicated, by lithium (Li) we refer to the main isotope of lithium, $^7$Li.

[2] $\mathrm{A(Li)} = \log\,(N_{\rm Li}/N_{\rm H}) + 12$.

[3] Classical pipelines refer to the tools that typically compare the observed spectrum to a model spectrum based on a line-list, a model atmosphere, and a prediction on the line shape as well as intensity (curve of growth) based on a model. These pipelines provide the stellar labels for training in the context of machine learning methods.

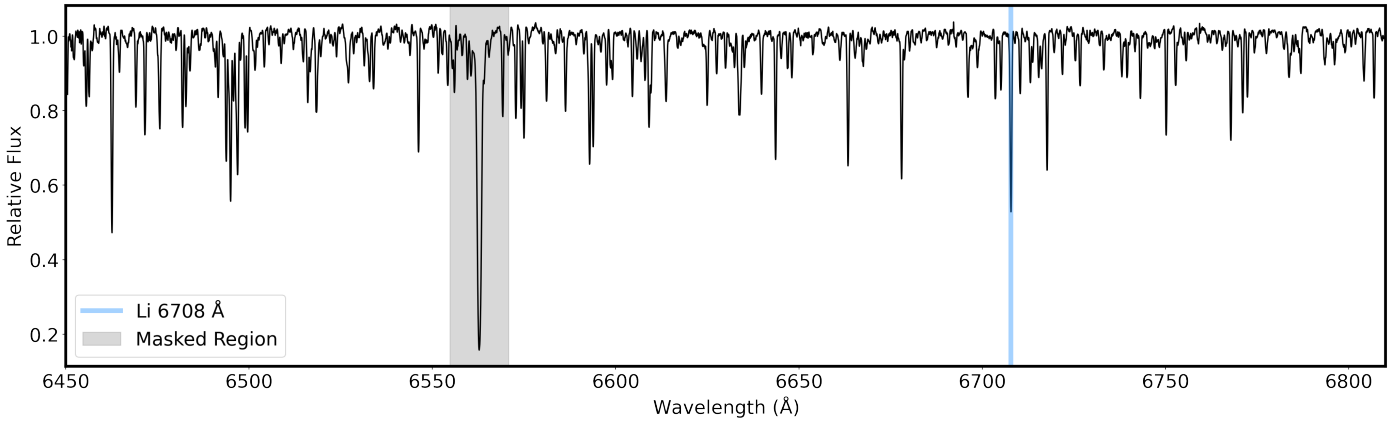


**Fig. 1.** Example GIRAFFE HR15N spectrum. This spectrum is of a star with labels: $T_{\rm eff}$ = 4897 K, $\log(g) = 2.55$ dex, [Fe/H] = −0.11 dex, and A(Li) = 2.63 dex. Lithium spectral feature is shaded with blue, while the gray shaded region centred at H$\alpha$ is masked and not used in the spectral analysis using CNN.

of a quadratic polynomial function (as in the case of Cannon). It is important to note that the Payne uses noiseless synthetic spectra as the training set. A modification of the Payne tool, named data driven-Payne (Xiang et al. 2019), has also been applied to the LAMOST low-resolution spectra.

A few recent studies used a class of neural networks called convolutional neural networks (CNN; LeCun et al. 1989; LeCun & Bengio 1995) to derive atmospheric parameters and chemical abundances from both high- and low-resolution stellar spectra. Such CNNs are very efficient at feature extraction, hence, they can be used to learn about the spectral features in stellar spectra and relate it to the atmospheric parameters and chemical abundances. Fabbro et al. (2018) developed the StarNet pipeline based on a CNN and a synthetic training set. Bialek et al. (2020) applied StarNet to *Gaia*-ESO Survey UVES instrument spectra by training the CNN with various synthetic spectral grids while mitigating the "synthetic gap". Leung & Bovy (2019) developed the astroNN tool (capable of handling missing labels) trained on observational data to derive 22 stellar parameters and chemical abundances based on APOGEE DR14 spectra and labels. Zhang et al. (2019) used StarNet to estimate the atmospheric parameters and chemical abundances of LAMOST low-resolution spectra, based on the high resolution APOGEE labels. Guiglion et al. (2020) performed similar label transfer from APOGEE DR16 to the intermediate-resolution RAVE survey, in addition to combining astrometry and photometry as additional inputs. Guiglion et al. (2020) showed that it is possible to improve the quality of predicted effective temperature and surface gravity by lifting the degeneracy in $\log(g)$ using the absolute magnitudes. Very recently, novel methods such as auto-encoders and generative domain adaptation have also been implemented for stellar spectroscopy (e.g., in O'Briain et al. 2021; Čotar et al. 2021). These research efforts and the developments in future spectroscopic surveys, computational power, and improved ML techniques are the motivation for preparing the ML ground for future spectroscopic surveys.

The main aim of this work is to provide reliable atmospheric parameters and Li abundances for a large sample of spectra and use it to study lithium evolution in the Milky Way. We adopted a CNN as a supervised ML method and our training labels are as follows: effective temperature, $T_{\rm eff}$, surface gravity, $\log(g)$, iron abundance, [Fe/H], and lithium abundance, A(Li). Any supervised ML method demands a very careful choice of training labels, as the trends and biases present in the training data are also learned and, hence, easily transferred to the predicted labels. This paper goes together with the work of Ambrosch et al. (2023), which focuses on the chemical evolution of Al and Mg abundances with CNN from GES GIRAFFE HR10 and HR21 spectra.

The paper is organized as follows. In Sect. 2, we present the spectral data set adopted in this study. In Sect. 3, we detail the CNN procedure. The catalog of lithium abundances is presented in Sect. 4, while its validation is done in Sect. 5. We present two scientific application of our catalog in Sect. 6 and we summarize our work and draw some future prospects in Sect. 7.

## 2. Observation and data

Our preliminary goal is to prepare the ground for 4MOST and WEAVE Li analyses. We looked for public spectra similar to the red arm of these two surveys, with associated high-quality lithium and atmospheric parameters. We adopted the *Gaia*-ESO Survey (GES, Gilmore et al. 2012; Randich & Gilmore 2013) data. Spectra was gathered by GES for all major Galactic components (halo, bulge, and thin and thick disks), including a large number of open and globular clusters, as well as calibration observations such as benchmark stars, radial velocity ($V_{\rm rad}$) standards, and asteroseismic CoRoT/K2 fields (see Bragaglia et al. 2022; Pancino et al. 2017; Stonkutė et al. 2016; Valentini et al. 2016). For this study, we use the spectra and parameters and abundances from the internal Data Release 6 (iDR6)[4].

The spectra were obtained using the GIRAFFE instrument of the Fibre Large Array Multi Element Spectrograph (FLAMES; Pasquini et al. 2002) located at Very Large Telescope (VLT) Observatory at Cerro Paranal (ESO) in Chile. We used the H665.0/HR15N setup that includes the Li doublet at 6708 Å. The HR15N setup is centred at 6650 Å, and covers the domain [6470–6790] Å with a resolving power $R$ = 19 200, very similar to the WEAVE and 4MOST HR red arm. The GES-iDR6 also comprises Li abundances for ∼6400 UVES spectra, which, however, we do not use in this work.

The spectroscopic analysis within GES was performed by multiple data analysis nodes which use different spectroscopic tools, but adopting the same line list and model atmospheres

[4] http://ges.roe.ac.uk,http://casu.ast.cam.ac.uk/gaiaeso

(Smiljanic et al. 2014; Lanzafame et al. 2015; Heiter et al. 2021; Gilmore et al. 2022; Randich et al. 2022; Worley et al., in prep.). The atmospheric parameters from each of the nodes are homogenized to provide a single measurement and associated uncertainty as the node-to-node dispersion. The different methods can be summarized into three categories: (i) equivalent width (EW) analysis where the atmospheric parameter determination is based on the excitation and ionization balance of the Fe lines; (ii) spectral synthesis method that estimates atmospheric parameters from a $\chi^2$ fit to the observed spectra; and (iii) multilinear regression method that derives atmospheric parameters and abundances by projecting the observed spectrum into vector functions that are constructed as the best linear combination of synthetic spectra from a grid. Here, we adopted the GES-iDR6 atmospheric parameters, $T_{\rm eff}$, and $\log(g)$, as well as the [Fe/H] abundance ratio.

GES-iDR6 provides one-dimensional local thermodynamical equilibrium (1D LTE) abundances for $^7$Li, measured using the EW measurement of the spectral feature at 6707.8 Å. The measured EWs are converted to lithium abundances using curves of growth (only one GES node contributed to Li determinations; see Sect. 2.1 of Romano et al. 2021, and Franciosini et al. 2022). For the GIRAFFE spectra, the Li line is blended with a nearby FeI line at 6707.4 Å, hence, a correction was applied. When the Li spectral line is very weak or not visible, an upper limit to the abundance is provided. GES also provides a flag for Li abundances (UPPER_COMBINED_LI1, 0 = detection, 1 = upper limit); an upper limit is provided when the 6707.8 Å Li line is undetected, as a result of too low values for the signal-to-noise ratio (S/N) or too little lithium (see Franciosini et al. 2022 for details).

### 2.1. Training and observed sample

To build the training sample[5], we applied several selection criteria. Starting with the total of 41 710 HR15N spectra, we selected objects with S/N > 40 pix$^{-1}$ (see Sect. 4.2 below) and applied the following cuts for labels: $4000 < T_{\rm eff} < 7000$ K, $1.0 < \log(g) < 5.0$ dex, $-2.0 <$ [Fe/H] $< 0.5$ dex and $0 <$ A(Li) $< 4.0$ dex. We further cleaned the training sample by applying uncertainty cuts of e$T_{\rm eff} < 100$ K, e$\log(g) < 0.3$ dex, e[Fe/H] $< 0.2$ dex, and eA(Li) $< 0.5$ dex. We rejected stars with Li upper limits. We also applied an uncertainty cut on the radial velocity E_VRAD $< 0.5$ km s$^{-1}$ (see Sect. 3.2.3). Spectra with GES flags for data reduction and analysis problems (TECH) and for peculiarities affecting the spectra (PECULI) were also rejected (see Gilmore et al. 2022 for more details). During the training, some variable and high proper motion stars were identified with significant variability in flux seen in their multiple observations. As GES provides the same homogenized labels for these multiple observations, these objects were subsequently removed from the training. The training sample is then composed of 7031 spectra and respective labels. The remaining 33 119 spectra, not included in the training sample, comprise the observed sample. We do not provide labels for 1560 spectra due to missing $V_{\rm rad}$ or very high $V_{\rm rad}$ values, shifting the spectrum out of the desired wavelength range after correction.

Next, we applied radial velocity correction to the GES continuum-normalized spectra and removed the random cosmic features. Any pixel value exceeding median of the continuum by over five sigma is replaced by a median of the continuum. Negative pixel values are replaced by a median of the continuum+lines. The spectra were then re-sampled to a common wavelength coverage $\lambda \in [6450–6810]$ Å, while keeping the original pixel separation of 0.05 Å.

The HR15N sample consists of many young objects that have strong H$\alpha$ emission lines. Since dealing with this is out of the scope of the current work, we masked the region of 16 Å around H$\alpha$. The only requirement for the observed sample was that the radial velocity should be present in the recommended radial velocity catalog provided with the *Gaia*-ESO survey iDR6. Spectra with S/N values as low as 2 are present in the observed sample. The implication of such a low S/N on the CNN predictions are discussed later (see Appendix A.3). As GES provides repeated observations, some stars have multiple spectra available with varying S/N values. These repeated spectra are present in both training and observed samples and provide a good test for the consistency of the CNN.

### 2.2. Pre-processing training and observed sample

We used Scikit-learn (Pedregosa et al. 2011) for pre-processing. Using the train_test_split function, we adopted 25% of the total training sample data as test set (leading to 1758 spectra and associated labels). The test set is not directly used for training of the CNN model, but it is only used to monitor the performance of the trained models at the end of each epoch (see Appendix A.2). The train set is then composed of 5273 spectra (75% of the training sample). Train and test sets are uniformly distributed across the label range, as homogeneity is crucial to help the CNN generalizing instead of over- or underfitting. We refer to Sect. 2.3 for a further discussion on homogeneity.

We normalized the stellar labels to values between 0 and 1, using the MinMax normalization function. Normalizing all the stellar labels within same value range helps train the CNN with easier and faster convergence to the loss function global minimum.

### 2.3. The *t*-SNE method for homogeneity check and outlier detections

To check the homogeneity of our train and test sets, we apply the *t*-distributed stochastic neighbour embedding (*t*-SNE; Van der Maaten & Hinton 2008), an unsupervised ML method. It works by assigning similar objects in the high-dimensional space with a higher probability distribution and, hence, modeling them closer together in the lower dimensional map, while dissimilar objects are mapped further apart. Overall, *t*-SNE has been widely used in astrophysical applications (Matijevič et al. 2017; Anders et al. 2018). For example, Anders et al. (2018) successfully applied *t*-SNE to their study of the stellar abundance space and identifying substructures as well as chemically peculiar stars.

We plotted the *t*-SNE maps (perplexity = 50)[6] for the whole training data set (7031 spectra with ~7000 pixels each) in Fig. 2. The axes value themselves have no physical meaning, while the nearby points represent similar spectra. The right-most plot shows how well the train and test sets follow each other in the *t*-SNE. This is only possible if they are homogeneously distributed across the range of labels. The figure shows a few outliers identified by the *t*-SNE; we checked these spectra and found them

[5] Throughout the paper, "training sample" refers to the whole data use for training and cross-validation purpose; "train set" and "test set" refer to 75% and 25% of the "training sample," respectively.

[6] Perplexity is a parameter that sets the number of effective nearest neighbours; a higher value is usually recommended for larger samples.

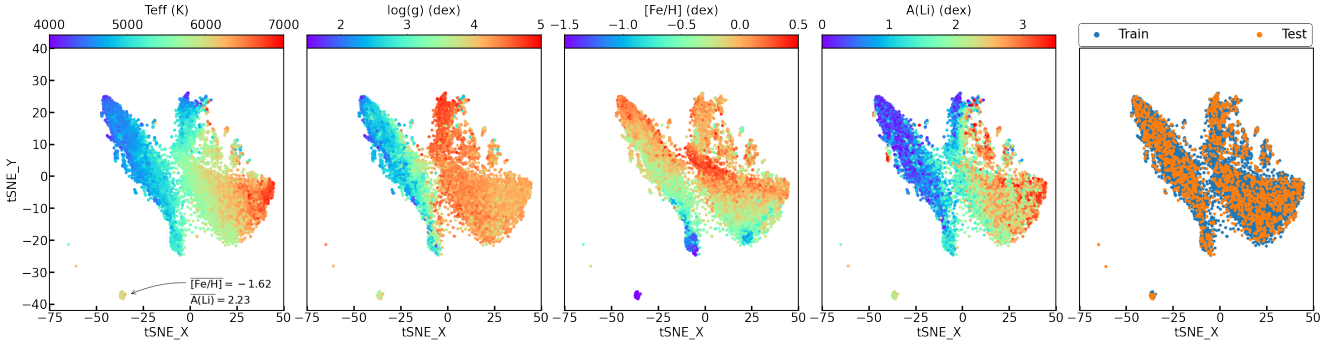


**Fig. 2.** 2D projection of $t$-SNE output for the 7031 spectra of the training sample, colored by the labels $T_{\rm eff}$, log($g$), [Fe/H] and A(Li) respectively. The right-most plot shows the $t$-SNE as the train and test sets to highlight their similar distribution across the label range. In the left subplot, we show the mean [Fe/H] and A(Li) for the highlighted island that consists of Spite plateau-like stars in the globular cluster NGC 6752.

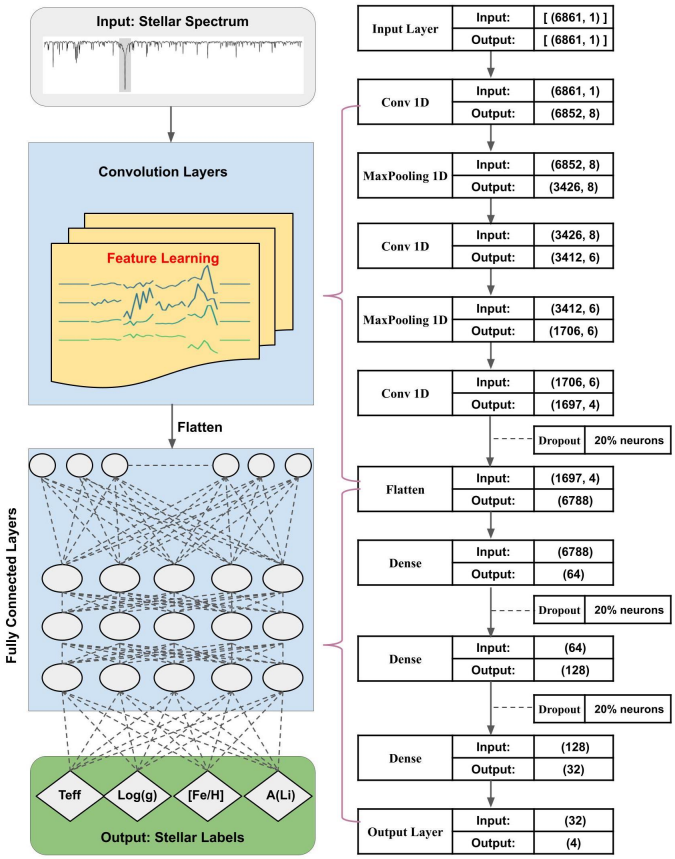


**Fig. 3.** Architecture of the CNN adopted for this study is shown as a block diagram on the left and its detailed structure with layers is shown on the right panel. The model can be divided into four distinct sections: input layer, convolution layers, fully connected layers, and output layer, with a total of 448 134 trainable parameters. The numbers, for example, (6861, 1) and (6852, 8), represent the shape of input and output of first Conv1D layer.

to have low S/N and we see they are affected by bad cosmic ray removal. The island at tSNE_$X$ = −25 and tSNE_$Y$ = −45, consists of Spite plateau-like stars ($\overline{\rm [Fe/H]} = -1.62$ dex, $\overline{\rm A(Li)} = 2.23$ dex) in the globular cluster NGC 6752, which represents the most metal-poor group in the training sample. The figure also shows how spectra and atmospheric parameters are correlated. This reveals that they are intrinsically linked by a high-complexity mapping, which the CNN will have to learn during its training.

## 3. Convolutional neural network for stellar parametrization

### 3.1. Architecture of the CNN

We built our CNN model with the open source deep learning library Keras (Chollet 2015), using the TENSORFLOW backend (Abadi et al. 2015). Keras provides a Python interface in a compact and easy manner to develop high-level artificial neural networks. Then, TENSORFLOW developed by the Google Brain Team, is an open-source software library for ML. We trained the CNN with the gradient-based Adam optimizer (Kingma & Ba 2014).

In deep learning methods, the final choice of the architecture is usually an outcome of a lot of experimentation with various setups and tuning of hyperparameters. The architecture of the CNN makes a significant impact on the training and prediction performances. The implementation of various architectures for stellar spectra parametrization can be found in the literature, we refer to the work referenced in Sect. 1 for further details. For this project, we built on the work of Guiglion et al. (2020) and optimized the architecture.

Figure 3 shows the architecture of our CNN. The preprocessed spectrum is provided as input and as output the CNN predicts $T_{\rm eff}$, log($g$), [Fe/H] and A(Li). The model has three convolution layers and four (3 + 1) dense layers, including the output layer (discussed in Appendix A). Studies such as Leung & Bovy (2019), Fabbro et al. (2018) have also adopted a similar architecture as a good trade-off between desired precision and computation time.

Further details on the CNN architecture, the choice of hyperparameters, and model generalization (avoiding over- or underfitting) of the CNN can be found in Appendix A.

### 3.2. Training the CNN

Our CNN model architecture, as illustrated in Fig. 3, has a total of 448 134 trainable parameters. These parameters include all the weights and biases for the different layers present in the model. The training process optimizes the values for the parameters by minimizing the value of a loss function and judges the performance of the training by calculating a metric on the test data. We use the mean squared error (MSE) as the loss function as well as the metric. The EarlyStopping callback, defined in

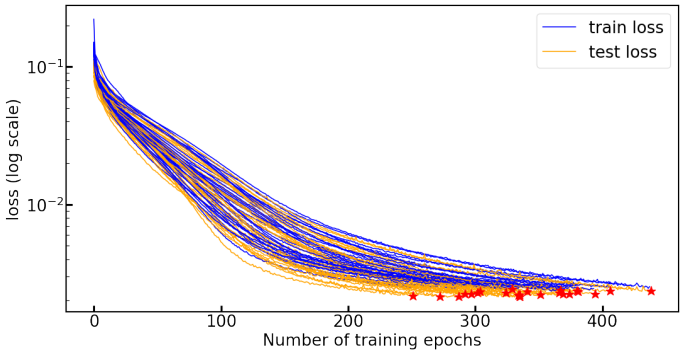


**Fig. 4.** Value of the loss functions for the train (blue) and test (orange) sets for the 30 CNN runs as a function of the epoch. The red stars identify the selected 24 models.

Appendix A.2, monitors the metric and the best model weights are saved. We trained an ensemble of 30 models[7][8], where for each model, weights were randomly initialized. The training for the models stopped at different epochs due to the stochastic nature of the learning algorithm.

In Fig. 4, we show the progress of the training by plotting the evolution of the loss functions of the training (blue) and test (orange) sets for the 30 models. The loss curves show that the training was smooth and provides a good fit as the training and test loss decreases to a point of stability, with a small gap between the two final loss values.

The models with higher test loss than the 80th percentile value are discarded, and the predictions from the selected 24 models are averaged as the final result. The dispersion is provided as the label uncertainties (see Sect. 4.3 for more on uncertainties).

#### 3.2.1. Result of the training

In Fig. 5, we show a comparison of the input GES-iDR6 labels to the CNN prediction for the train and test sets. The figure shows a well-behaved 1-to-1 relation with no apparent systematic trends. The bias and scatter values represent the mean and the standard deviation of the residuals. The results show no bias (negligible for $T_{\rm eff}$). The scatter is comparable for the train and test sets, with slightly higher scatter for scarcely populated label regions such as $\log(g) < 2.0$ dex and [Fe/H] $< -0.5$ dex. Overall, the test set follows the train set, showing that the trained models do not over-fit. Even though the wavelength range in the GIRAFFE HR15N setup is not optimal for determination of atmospheric parameters (Lanzafame et al. 2015), and despite masking the H$\alpha$ line, which is an important spectral feature for the estimation of $T_{\rm eff}$ and $\log(g)$, the CNN shows very good performances. This indicates that the trained CNN models have learned significantly from the available spectral features.

In Fig. 6, we present Kiel diagrams ($T_{\rm eff}$ vs. $\log(g)$ for the train (top panels) and test (bottom panels) sets. The left columns show the input iDR6 labels and the right columns show the labels as predicted by the CNN. We see that the main features of the Kiel diagram are well recovered. The dwarfs and giants are clearly separated with a smooth transition from main-sequence turn-off to the subgiants and the metallicity gradient in the giant branch is very well described for both the train and test sets. The dwarfs, which span a large $T_{\rm eff}$ range from 7000 K to 4000 K, are adequately parametrized even for the very hot and the very cool regime. The metal-poor giants, around 5000 K, show much less scatter for the CNN output compared to the GES-iDR6. Two distinct issues can explain this difference: 1. This region is very sparsely populated in the training data, so the one way to improve CNN prediction would be to add more training data in this region. 2. No benchmark stars are present in this region, namely, there are no metal-poor giants (see Sect. 5.1 for details). Similar lower scatter, at the metal-poor end for giants when predicted by the ML methods have been reported by Ness et al. (2015, see Fig. 12 and Ting et al. 2019, see Fig. 7); both studies compared their results with isochrones to find their ML results at this region in better agreement with stellar isochrones compared to the surveys, suggesting discrepancies due to calibration issues.

In Fig. 7, we present the lithium abundance trends, colored by $T_{\rm eff}$, for both train and test sets. The main features are also very well recovered. The most metal-poor globular cluster NGC 6752 with [Fe/H] $< -1.5$ dex and A(Li) $\sim 2.2$ dex is well located for both train and test sets. We also find good agreements for globular clusters such as NGC 1281 and NGC 2808, seen around $-1.5 <$ [Fe/H] $< -1.0$ dex and A(Li) $\sim 1.2$ dex. The $T_{\rm eff}$ dependence for Li, with higher Li abundance for hotter stars and lower Li abundance for cooler stars, is also seen. The highest Li abundances, at the metal-rich regime, seen for the hottest stars and the coolest PMS stars, are also recovered for both train and test sets. It is consistent, for instance, with Romano et al. (2021), who use GES iDR6 to infer the highest, undepleted Li abundances for both field (hot stars) and cluster (hot MS and cool PMS) stars.

#### 3.2.2. Examining if the CNN can learn from spectral features

Treating our neural network as a mathematical function that maps input spectra to output labels, it is desirable to check how each part of the input spectrum influences the output labels. In other words, if we can calculate the sensitivity of output labels to each of the input fluxes, we can understand whether the CNN is learning from the spectral features. Calculating gradients is one such method for generating a sensitivity map for a spectrum by performing partial derivatives of each of $T_{\rm eff}$, $\log(g)$, [Fe/H], and A(Li) with respect to every input neuron (or wavelength), namely, $\partial\text{Label}/\partial\lambda$. The gradient-based optimizing algorithm Adam (Kingma & Ba 2014) calculates a negative gradient of the weight matrix at each iteration to reduce the loss function, hence, calculating gradients are inherent to neural networks. The gradient of an output label is a kind of back-propagation of the model through the CNN and is obtained by using the simple chain rule of derivative moving backward from output to the hidden layers and finally to the input layer. This is achieved via a set of techniques called automatic differentiation[9], which makes it possible to evaluate the derivative of the function represented by the CNN. We used the GradientTape function from Tensorflow to calculate the gradients.

In Fig. 8, we show as an example, the gradients of $\log(g)$ and A(Li) for the 13 solar twins in our training sample. We make following representative observations: First, the gradient of the lithium label with respect to $\lambda$ is only active at the

[7] The training of the models required a time period of 16–26 min using only normal CPU on the COLAB cloud service at AIP for compute and storage.

[8] We adopted 30 models for the Ensemble method as a good trade-off between the reliable statistics and computational load.

[9] For further details on automatic differentiation and gradients, see https://www.tensorflow.org/guide/autodiff.

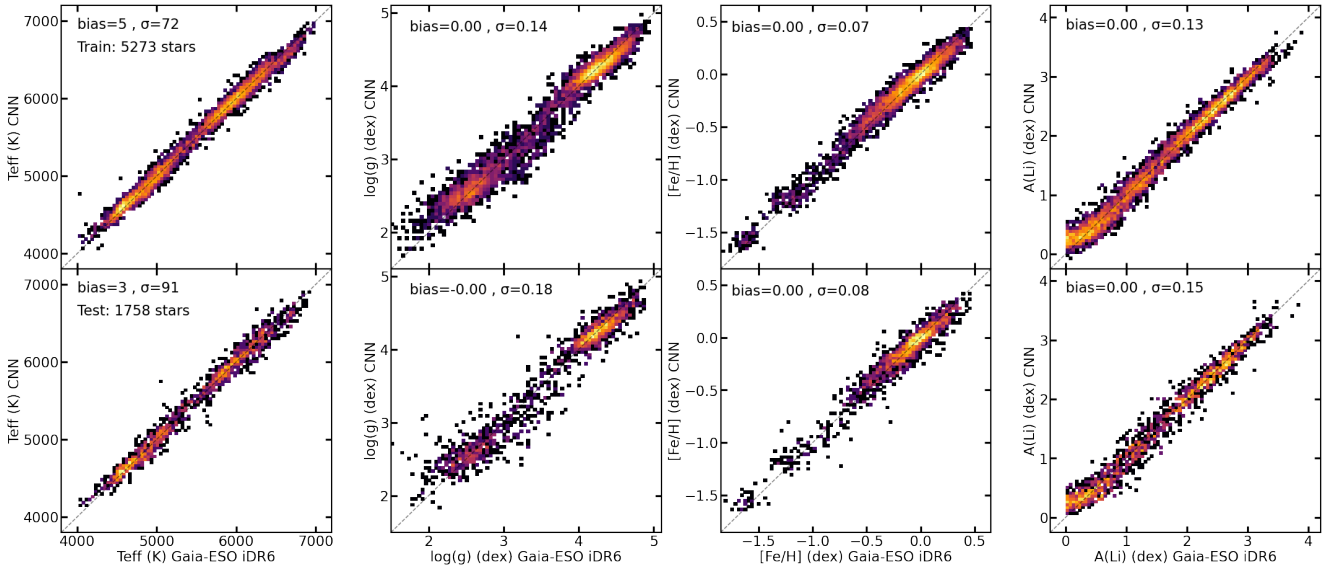


**Fig. 5.** 2D histograms showing 1-to-1 comparison between the GES-iDR6 labels (CNN input, $x$-axis) and CNN predictions ($y$-axis) for the train (top row) and test (bottom row) sets. The bias = mean(CNN-iDR6) and $\sigma$ = std(CNN-iDR6) are also calculated.

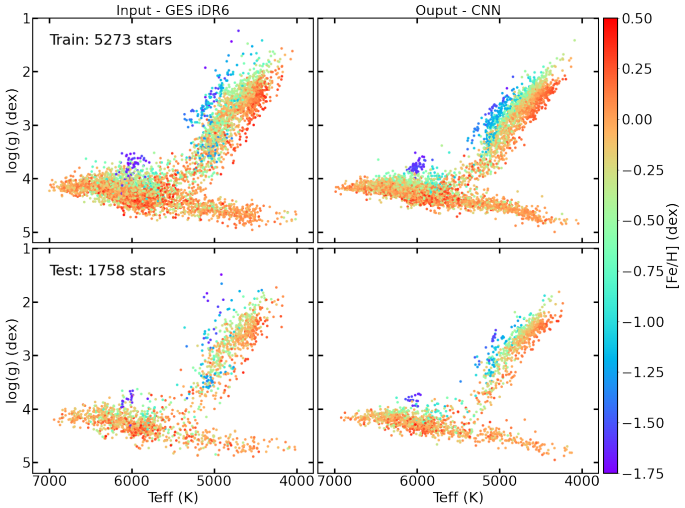


**Fig. 6.** Kiel Diagrams for the input and CNN output colored by [Fe/H]: top two panels show the train set stars using iDR6 input labels on the left and CNN output on the right. Bottom two panels show the same for the test set.

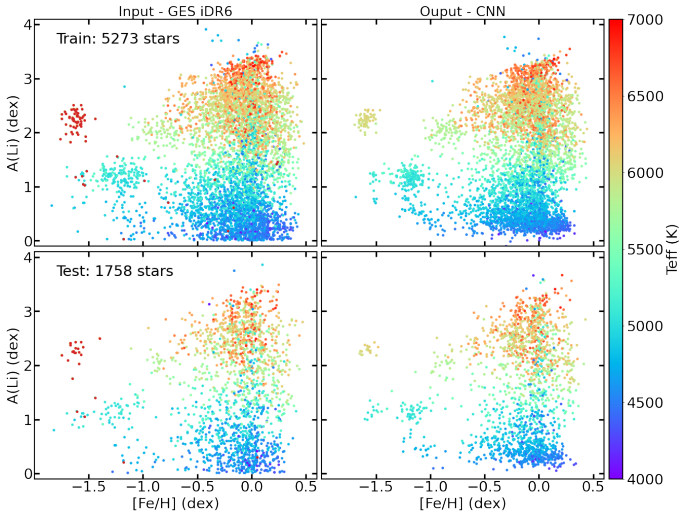


**Fig. 7.** [Fe/H] vs. A(Li) for the iDR6 input and CNN output colored by $T_{\rm eff}$: top two panels show the train set stars using iDR6 input labels on the left and CNN output on the right. Bottom two panels show the same for the test set.

lithium line and almost flat elsewhere. This shows the ability of our CNN to discard all other wavelengths and learn from this singular feature. The CNN then properly measures lithium abundances, instead of simply inferring them from correlations among the labels. Second, Damiani et al. (2014) showed that the quintet feature, between 6490–6500Å consisting of blended FeI, CaI, BaII, and TiI lines, is highly sensitive to gravity. The TiII 6491.56 Å line, on the bluer side of the quintet, was also considered as an important line for their spectral indices. Here, the CNN gradients $\partial \log(g)/\partial \lambda$ show that these wavelength regions are indeed very sensitive to $\log(g)$. Finally, Jofré et al. (2015) listed the ionized Scandium, ScII, line at 6604.6 Å as a Golden Line for FGK dwarfs and giants but not for metal-poor stars and M giants. Our $\log(g)$ gradients also show very high response at this wavelength region.

Such diagnostic checks confirmed that CNN properly learns from spectral features and these gradients could allow for the identification of new sensitive spectral features that are presently not used by standard classical pipelines. Then, the classical pipelines and the CNN could be used in a sort of feedback manner to improve their mutual output.

### 3.2.3. Sensitivity to the radial velocity

Accurate and precise radial velocities are crucial for obtaining a reliable estimate of the atmospheric parameters and chemical abundances, as it matches the observed spectrum to the line-list which is the ground truth for any EW or spectral fitting methods. The radial velocities (and associated uncertainties) of the GIRAFFE HR15N spectra were estimated by GES, by spectral fitting of the observations to model spectra

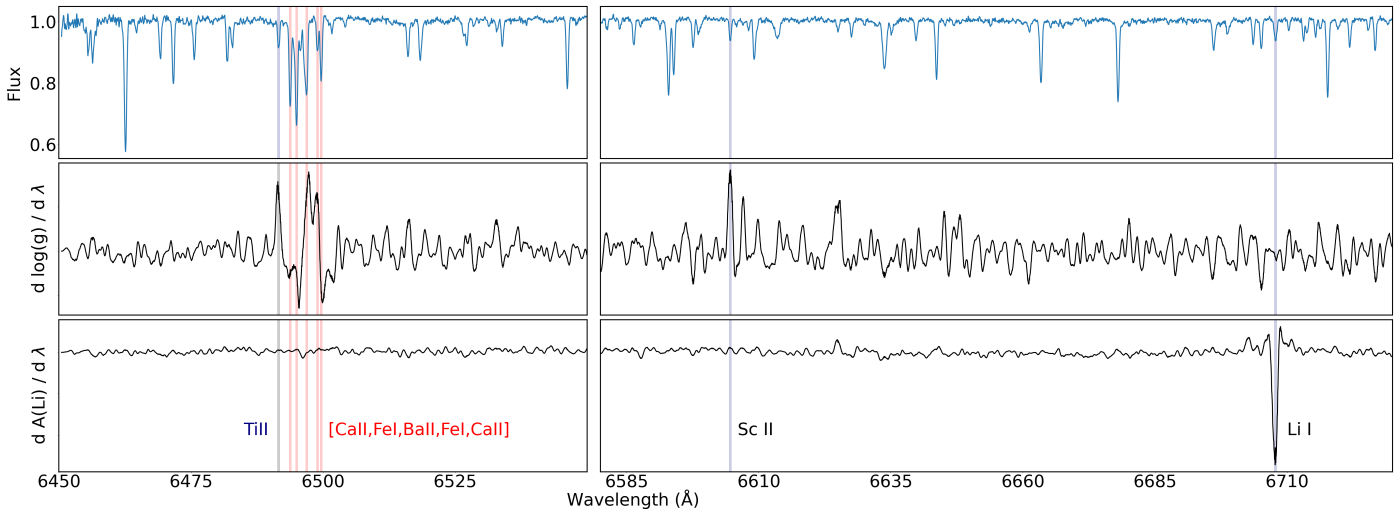


**Fig. 8.** Gradients of the output labels with respect to input pixels for the solar twins in the training sample. Selected as $T_{\rm eff}$ = 5777 ± 25 K, $\log(g) = 4.44 \pm 0.10$ dex and [Fe/H] = 0.0±0.05 dex, there are 13 stars. The top row shows the mean input spectrum and the second and third row represent the gradient/response for $\log(g)$ and A(Li), respectively. Left column shows wavelength region [6450–6550] Å and right shows [6580–6730] Å as we mask the H$\alpha$ region. Various spectral features that are discussed in the text are labeled.

(Gilmore et al. 2022). The radial velocity is measured using the HR15N spectra, but an offset is applied to it during the homogenization process to bring radial velocities measured from different setups to the same scale. The offsets are measured considering HR10 (5340 Å–5620 Å) setup as a zero-point of the radial velocity scale; GES made sure that HR10 radial velocities are in good agreement with *Gaia* radial velocity standards. However, such a combination of different setups can be a source of small systematics. While GES reports the highest $V_{\rm rad}$ precision achieved to be on the order of 0.25 km s$^{-1}$ (see Gilmore et al. 2022), over 80% of the HR15N sample have Vrad errors larger than 0.25 km s$^{-1}$ and with a third of the sample above 0.55 km s$^{-1}$.

Figure 9 shows the residual (CNN-iDR6) plots for the selected observed sample, colored in bins of GES radial velocity uncertainties. We clearly see that the dispersion increases with increasing $V_{\rm rad}$ uncertainties and a large bias is visible for stars with large E_VRAD, for instance, as shown by the red dots. Due to such results, we apply a cut at E_VRAD < 0.5 km s$^{-1}$ in our training sample. Jackson et al. (2015) report that $V_{\rm rad}$ precision for GIRAFFE spectra worsens for $T_{\rm eff}$ > 5200 K, as a result of paucity of strong narrow lines in hotter stars. We also observe that E_VRAD > 0.5 km s$^{-1}$ are mostly for stars hotter than 5500 K in iDR6. The HR10 re-calibration is a function of $T_{\rm eff}$, $\log(g)$, and [Fe/H], and this could create tiny $V_{\rm rad}$ corrections that the CNN is able to detect. We avoid a deeper investigation as it is outside the scope of this paper.

However, we showed that ML pipelines can be very sensitive to small wavelength shifts in the input data. For upcoming surveys such as 4MOST and WEAVE, which will observe in multiple setups, a precise radial velocity estimation will be more important as ML techniques will be extensively used due to the larger volume of observations. Also, another source of $V_{\rm rad}$ errors for GES could be the fact that the different wavelength ranges were calibrated independently (Randich et al. 2022). The expected accuracy of 4MIDABLE-HR radial velocities is expected to be <1.0 km s$^{-1}$(de Jong et al. 2019). Further tests on real 4MOST spectra will be necessary in order to estimate the CNN sensitivity to $V_{\rm rad}$.

#### 3.2.4. Inferring lithium abundances without a lithium line

ML algorithms are efficient at learning astrophysical correlations, for example, inferring oxygen abundances from spectra with no oxygen feature (Ting et al. 2017, 2018). Lithium abundance is highly correlated to the $T_{\rm eff}$, and depends a lot on the surface gravity (see e.g., Fig. 2). To test whether it is possible to infer lithium based on pure astrophysical correlations, we trained a CNN with the same GIRAFFE training sample, but masking the 6707.8 Å lithium line. In Fig. 10, we compare the CNN Li abundance with GES-iDR6 Li abundance, finding very poor performance compared to Fig. 5, with a large scatter of over 0.5 dex throughout the label range for both the train and the test sets. Here, we note that the A(Li) output by CNN comes purely from the correlations among labels and it is not a measurement from the spectral feature. Hence, we see an underprediction at higher values and an overprediction at lower values, also known as regression dilution. The Li-rich giants (see Sect. 6.2) are completely missed when inferring lithium solely from astrophysical correlations. We visually inspected the Li sensitivity map, as we did in Sect. 3.2.2, and most of the HR15N features are used to infer Li. Then, Li must be then measured from Li spectral feature instead of being inferred based on correlations.

## 4. Catalog of stellar parameters and Li abundances

### 4.1. CNN parametrization of the GES GIRAFFE spectra

We used CNN models to predict the atmospheric parameters and lithium abundances for the observed sample spectra. Prediction using a trained model is very fast and takes only ∼20 s for the four labels, $T_{\rm eff}$, $\log(g)$, [Fe/H], and A(Li), for all 33 119 observed sample spectra. The prediction for the selected 24 models takes only about nine minutes. An average of the 24 predictions is computed as the final result and the dispersion as an uncertainty.

For the stars within the training set limits, a typical Kiel diagram is seen, similar to Fig. 11a, with clear distinction between the main sequence and the giants, along with the metallicity gradient for the giants as well as the turn-off stars. At the cool end, we see few stars with $\log(g) \sim 4.0$: we checked the spectra for these stars and found the presence of emission lines. An

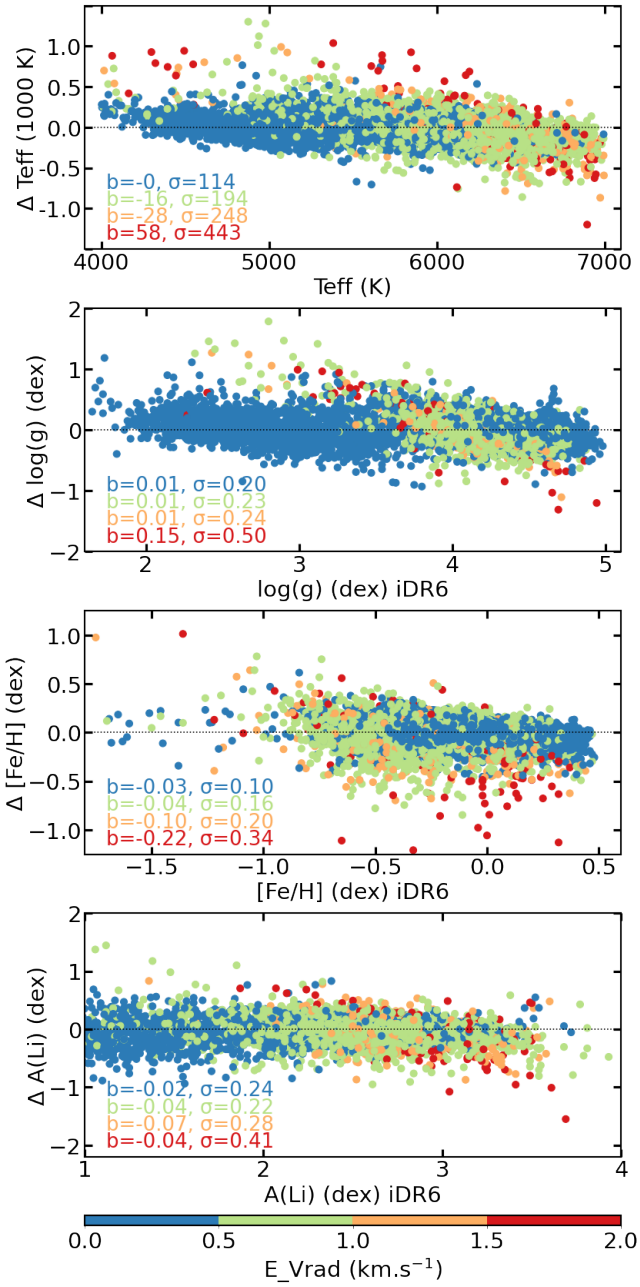


**Fig. 9.** Residuals (CNN-iDR6) as a function of labels for the selected observed sample. The stars are color-coded into four bins based on their reported uncertainties in radial velocities. For each label, the bias = mean(CNN-iDR6) and $\sigma$ = std(CNN-iDR6) in the four VRAD uncertainty bins are listed.

example of a HR15N spectrum with emission lines and molecular bands is shown in Fig. 12. For the second column Kiel diagram in Fig. 11, we see similar trends as in the case of training limits, except there is a cool dwarf clump. The group consists of very young clusters members, with emission lines and TiO molecular bands (M dwarfs). As there were no cool M dwarfs ($T_{\rm eff}$ < 3500 K) in the training set, some systematics may be present in the parametrization of these stars. However, GES is still refining the flags, thus further exploration of the particular flags is out of the scope of this project. In the third column of the Kiel diagram, the observed sample with radial velocity uncertainties >0.5 km s$^{-1}$ are presented. Most of these stars lie in the

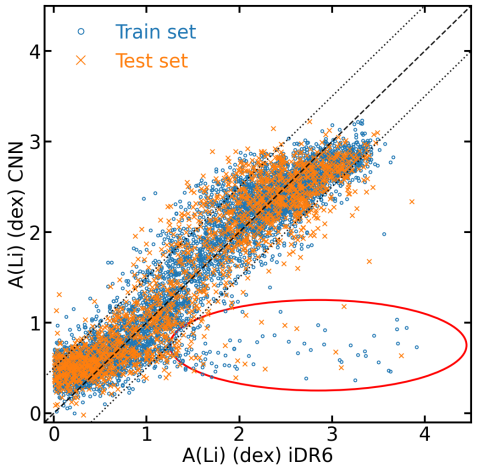


**Fig. 10.** CNN vs. GES-iDR6 A(Li) for the CNN trained using3 spectra masked at 6707.8 Å Li line. Blue and orange represent train and test sets respectively. The dashed line is the 1-to-1 line, and two dotted lines are at ±0.5 dex. The red ellipse shows the incorrectly inferred Li-rich giants.

warm dwarf region, as uncertainties in VRAD increase with $T_{\rm eff}$ (as discussed in Sect. 3.2.3). The metallicity gradient is also seen for these warm dwarf stars.

In Fig. 11d–f, we also present lithium abundance trends with respect to [Fe/H]. We see that most of the stars in the panels d and e are cool Li-poor stars, with a peak at solar [Fe/H]. For the observed sample stars in the training set limits, we see a clear trend with $T_{\rm eff}$, with only a few cool stars with A(Li) > 3.0 dex. In plot e, an increase of cool stars with high lithium is seen. These are young cluster members, for which the Li depletion has not been completed. In plot f we see the stars with GES flags and E_VRAD > 0.5 km s$^{-1}$. Most of these stars are hotter stars with $T_{\rm eff}$ > 5500 K (see Sect. 3.2.3). Some of these warm, lithium-rich stars are likely to represent the warm group of stars on the left side of lithium dip.

In Fig. 13, we present the comparison of CNN predicted labels with iDR6 labels for a selection of the observed sample with S/N > 20 pix$^{-1}$, E_VRAD < 1.0 km s$^{-1}$ and no TECH and PECULI flags. In the first row, we show 4481 observed sample stars with iDR6 Li abundance with the flag UPPER_COMBINED_LI1 = 0. The second row shows comparison for 3099 stars, with Li upper limits given by UPPER_COMBINED_LI1 = 1. There is an upper limit provided by GES on the Li abundance when the 6707.8 Å Li line is undetected (too low S/N or too low lithium). For stars with GES Li measurement, we see a very good one to one match with no bias. There is a scatter of 162 K for $T_{\rm eff}$, 0.22 dex for log($g$), 0.13 dex for [Fe/H] and 0.23 dex for A(Li). For the stars with GES Li upper limit, a very good one to one match with iDR6 measurement is seen with a small bias of 13 K for $T_{\rm eff}$ and no bias for log($g$) and [Fe/H]. A larger bias and scatter for A(Li) is observed, but this is expected as the iDR6 values are upper limits, and we provide lithium measurement for these stars. The scatter for $T_{\rm eff}$, log($g$), and [Fe/H] is higher for the Li measurement stars as most of these spectra (~80%) have S/N < 40 pix$^{-1}$, while the most of the Li upper limits have higher S/N; this is because stars with higher S/N and Li measurements, that is, those without a limit, are included in the training set. Also, most of the stars with an upper limit for lithium are giants that have already evolved past their Li depletion phase (defined in Sect. 6.1).

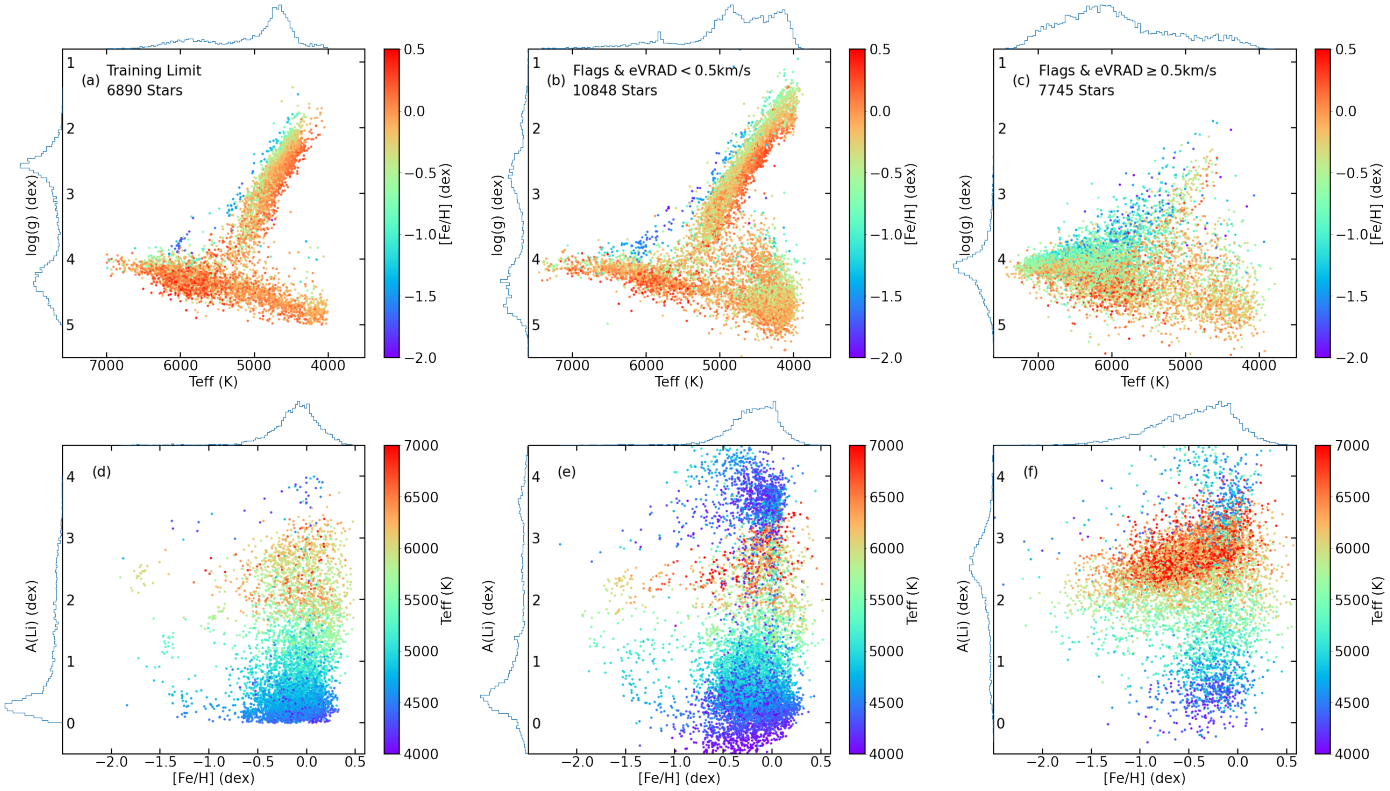


**Fig. 11.** Results for the observed sample. Top row: Kiel diagram for the observed sample stars with $S/N > 10$ dex and labels within training limits color-coded with [Fe/H]. (b) Same plot as (a) but for stars with $S/N > 10$ pix, GES-iDR6 flags and E_VRAD $< 0.5$ km s$^{-1}$. (c) Same selection as (b) but for E_VRAD $\geq 0.5$ km s$^{-1}$. Each subplot shows a histogram of the labels on the left and top axis. Bottom row: A(Li) vs. [Fe/H] color-coded with $T_{\rm eff}$ for the same stars as the Kiel diagram on top.

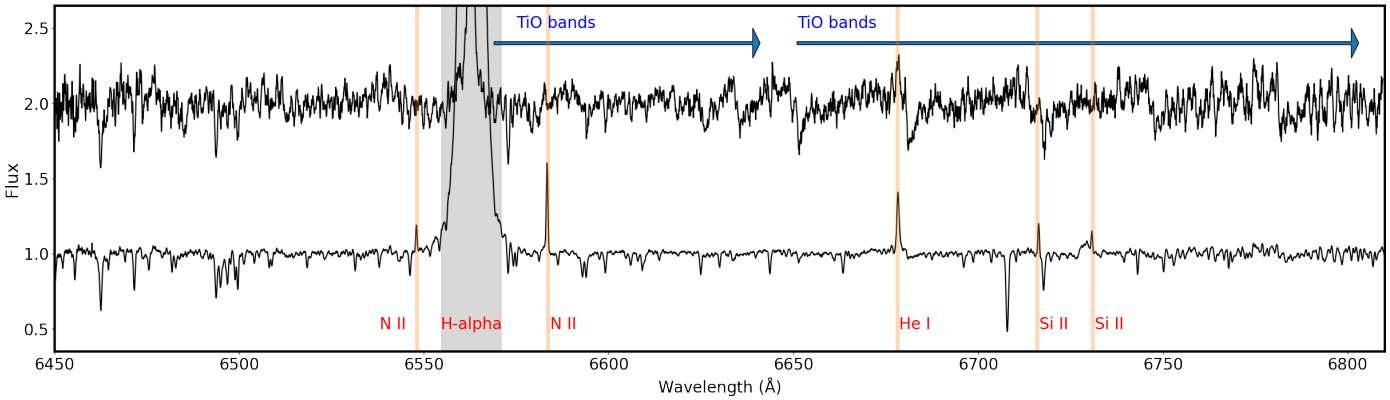


**Fig. 12.** HR15N spectra with emission lines highlighted in yellow. From left to right: lines: 6548 Å NII, 6563 Å $H_\alpha$, 6583 Å NII, 6678 Å HeI, 6716 Å SiII, and 6731 Å SiII. For the upper spectrum, the region for the strong molecular bands of TiO starting at 6569 Å and 6651 Å are seen. The relative flux values for top spectrum are increased by a unit for the ease of plotting.

Our catalog of atmospheric parameters ($T_{\rm eff}$, log($g$)), [Fe/H], and lithium abundances for ∼40 000 stars is summarized in Table 1. Of course, the apt use of this catalog will depend on the scientific application, but we encourage the reader to use lithium abundances within the training set limits (flag_li = 1), and Li uncertainties below 0.15 dex (S/N > 20). Similarly, atmospheric parameters are reliable only within the training set limits (flag_x = 1). In addition, we make the CNN code, spectra and labels available to the community online via GitHub[10].

[10] https://github.com/SamirNepal/Li_CNN_2022

### 4.2. Effects of noise and rotation on CNN predictions

The CNN was trained with spectra with S/N > 40 pix$^{-1}$, as this provides a balance in the training sample size and good quality. Noise is an unavoidable aspect of observational data (see Appendix A.3). In poor S/N spectra, the spectral features can be affected by the noise and can lead to a poor training performance as the CNN starts to learn the unwanted correlations due to noise. We find the mean difference between GES input and CNN output is uniform for different S/N ranges and do not see any significant increase with decreasing S/N (for both the training and observed samples). We conclude that CNN does not show any significant

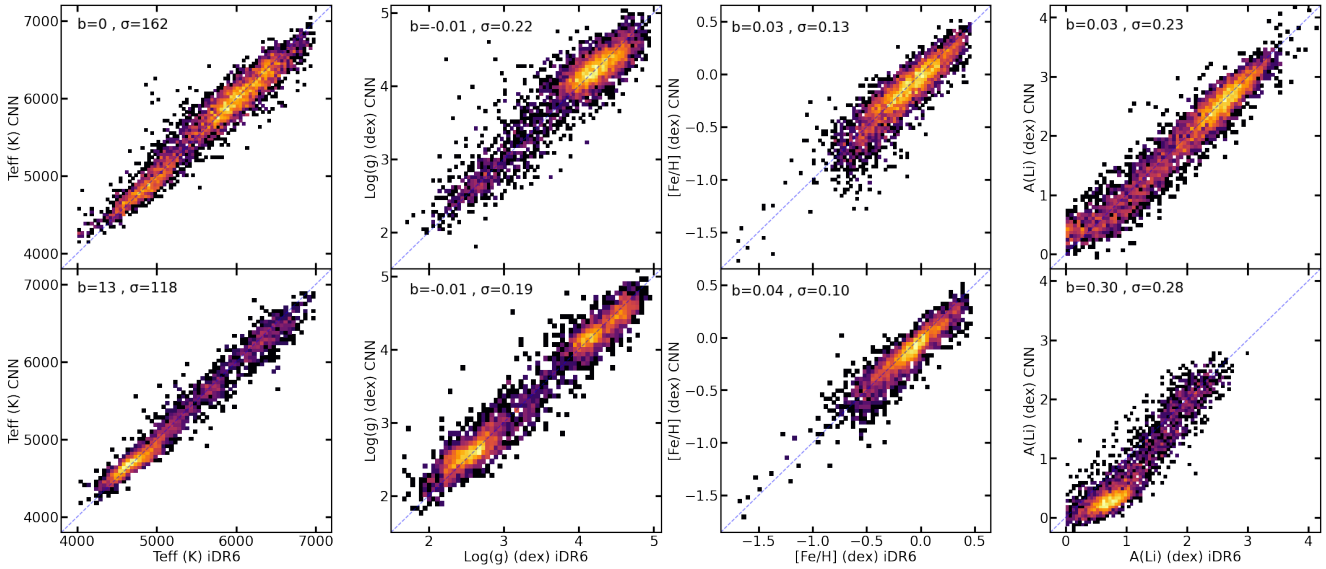


**Fig. 13.** One-to-one comparison for observed sample stars with, $S/N > 20\,\mathrm{pix}^{-1}$, eVrad less than $1\,\mathrm{km\,s}^{-1}$, no PECULI or TECH flags and within the training label range. Here, bias = mean(CNN-iDR6) and $\sigma$ = std(CNN-iDR6). Top row: stars with Li measurements. Bottom row: stars with Li upper limit. Most of the stars in the observed sample with Li measurement have low S/N spectra, hence, the higher scatter for $T_{\rm eff}$, log($g$), and [Fe/H].

**Table 1.** Atmospheric parameters, Li abundances, and boundary flags of the publicly available online catalog for ~40 000 stars.

| Col | Format | Units | Label | Description |
|---|---|---|---|---|
| 1 | char | – | cname | GES ID |
| 2 | char | – | spectra_name | Name of Spectrum |
| 3 | float | K | teff | Effective temp. ($T_{\rm eff}$) |
| 4 | float | K | eteff | Uncertainty of $T_{\rm eff}$ |
| 5 | int | – | flag_teff | Boundary flag for $T_{\rm eff}$ |
| 6 | float | cm s$^{-2}$ | logg | Surface gravity |
| 7 | float | cm s$^{-2}$ | elogg | Uncertainty of log($g$) |
| 8 | int | – | flag_logg | Boundary flag for log($g$) |
| 9 | float | dex | feh | [Fe/H] ratio |
| 10 | float | dex | efeh | Uncertainty of [Fe/H] |
| 11 | int | – | flag_feh | Boundary flag for [Fe/H] |
| 12 | float | dex | li | Li abundance |
| 13 | float | dex | eli | Uncertainty of Li |
| 14 | int | – | flag_li | Boundary flag for Li |
| 15 | int | pixel$^{-1}$ | snr | Signal-to-noise ratio |

bias as a function of S/N (see Appendix B and Fig. B.1 for further details).

Another important aspect concerns the stellar rotational velocity. As the projected rotational velocity ($v \sin i$) increases, the spectral lines get wider and shallower and there is an increase in the line blending (with conserved EW). Classical spectroscopic pipelines must take into account rotational broadening during analysis of a spectrum.

Our training sample of 7031 spectra has a distribution of rotational velocities (in km s$^{-1}$) as follows: $[v \sin i \leq 10] = 62\%$, $[10 < v \sin i \leq 30] = 34\%$, $[30 < v \sin i \leq 50] = 3\%$, and $[v \sin i > 50] = 1\%$. Assuming that stars with $v \sin i > 10\,\mathrm{km\,s}^{-1}$ are fast-rotators, the training sample has a significant number of such spectra. In fact, the CNN can learn from spectral features about the rotational broadening effects, even if $v \sin i$ is not used as a stellar label. As shown in Fig. 14, for $v \sin i < 50\,\mathrm{km\,s}^{-1}$, there is no significant change in dispersion (between input and output labels) and we observe no visible trends with the increasing rotation, even for hot stars with $T_{\rm eff} > 6000$, indicating an excellent CNN performance. For very fast rotators at $v \sin i > 50\,\mathrm{km\,s}^{-1}$, the line shapes are significantly altered; we

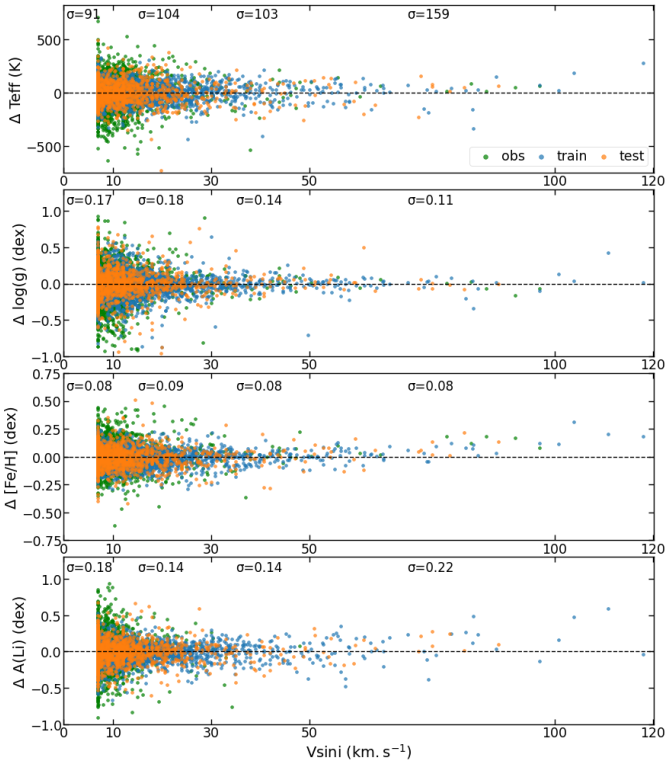


**Fig. 14.** Residuals ($\Delta$label = GES − CNN) as a function of $v \sin i$ (km s$^{-1}$) for the train (blue), test (orange), and selected observed sample (green) stars. The observed sample is selected within training label limits, $S/N > 10\,\mathrm{pix}^{-1}$, E_VRAD < 0.5 km s$^{-1}$, with no GES flags and with Li measurement. The mean scatter of the residuals ($\sigma$) in the $v \sin i$ bins (≤10, (10,30], (30,50] and >50) is also shown for each label.

see an increase in dispersion, to 159 K and 0.22 dex, for $T_{\rm eff}$ and A(Li). Also for [Fe/H], for $v \sin i > 70\,{\rm km\,s^{-1}}$, we see a trend of under-prediction by CNN. We conclude that CNN does not suffer from significant systematics due to rotational broadening, thus it allows us to accurately parametrize fast-rotating stars.

### 4.3. CNN internal uncertainty and estimation of precision and accuracy

The CNN internal uncertainties are calculated as the dispersion of the predictions from 24 selected models and is representative of the internal precision of the CNN. In Fig. 15, we present the uncertainty distributions for atmospheric parameters and Li abundance for the 31 272 observed sample stars with S/N > 10 pix$^{-1}$. Overall, the uncertainties are low and similar to the training sample and reflect that our models provide stable results. We find larger uncertainties for lower S/N spectra and for stars with labels outside the training limits.

The train, test, and observed sets show similar uncertainties, if the observed sample is restrained to the training sample limits. The uncertainties are very low, with medians of about 19 K for $\sigma T_{\rm eff}$, 0.03 dex for $\sigma\log(g)$, 0.017 dex for $\sigma$[Fe/H] and 0.035 dex for $\sigma$A(Li) for the train, test, and observed sets (within the training sample limits). It comes from the fact that the training sample covers a higher S/N range and also includes spectra without any TECH or PECULI flags. The increased error for the whole observed sample is simply the irreducible uncertainty due to the sampling of the noise in the training set. We note that nearly 60% of the observed sample have S/N below the training minimum of 40 per pixel. The train, test, and observed sets follow each other well, meaning that the CNN models are able to generalize properly.

The CNN internal uncertainties may, however, be underestimated. To show a realistic approximation of the accuracy and precision of the method, in Fig. 16 we present the bias (running mean difference) and sigma (running mean dispersion) curves for our train, test, and observed sample predictions, compared to GES-iDR6 labels. The observed sample is selected within the training set limits, with S/N > 20 pix$^{-1}$ and no GES flags, and GES lithium detection. The bias curves corresponds to the accuracy and the sigma curves correspond to the precision of CNN.

For $T_{\rm eff}$, between $4400 < T_{\rm eff} < 6600$ K, the accuracy is within 25 K and increases only at the edges of the training set limits due to sparse training data. We report a good precision within 100 for the train and test sets and within 120 for the observed sample, affected by the lower S/N data. Similarly, for $\log(g)$, an excellent accuracy is seen within 0.1 dex across the label range except at the edges, due to the low statistics. A similar effect is seen in the precision curves within 0.2 dex across the range except $\log(g) < 2.0$ dex and $3.0 < \log(g) < 4.0$ dex, which are less populated. For [Fe/H] < −1.0 dex, with just 19 stars that have available GES-iDR6 values in the observed sample, the bias and $\sigma$ curves cannot be adequately interpreted. For [Fe/H] > −1.0 dex, we achieve a very good accuracy within 0.05 dex and precision within 0.1 dex. For A(Li), the observed sample bias curve follows the train set, with an excellent accuracy within 0.05 dex except at A(Li) > 3.5 dex, where we have very few stars. The precision of the train and test sets are within 0.2 dex, while the observed sample is within 0.3 dex as ∼90% of the stars have S/N < 40 pix$^{-1}$. For future applications, such sigma and bias curves could be used to provide realistic precision and accuracy estimates.

## 5. Validation of CNN predictions

### 5.1. Validation with Gaia benchmark Stars

The *Gaia* benchmark star (GBS; Heiter et al. 2015; Blanco-Cuaresma et al. 2014; Jofré et al. 2014) sample provides precise stellar parameters and chemical abundances, derived from the best available spectra with very high-resolution and S/N along with the requirements of having accurate parallaxes, angular diameters from interferometry, bolometric flux, and stellar masses. The GBS are selected to represent typical Milky Way FGK stars covering different regions of the Hertzsprung–Russell diagram and a wide range of metallicities. Benchmark stars are commonly used as validators or calibrators by large spectroscopic surveys, such as GES (Pancino et al. 2017). In Fig. 17, we compare CNN predictions with the GBS catalog Version 2.1 (Jofré et al. 2018) which contains 36 benchmark stars in total. The benchmarks stars were excluded from the training sample. There were 26 benchmark stars from the GBS in GES-iDR6, with high S/N, for which we compare the $T_{\rm eff}$, $\log(g)$, and [Fe/H] to the CNN predictions. As the GBS catalog does not provide lithium abundances, we used the AMBRE Li abundances from Guiglion et al. (2016), which has 15 stars in common between the GBS and GES-iDR6. The AMBRE Li catalog provides Li abundances derived from high-resolution ($R$ = 40 000) ESO spectra using an optimization pipeline GAUGUIN, based on a synthetic spectra grid and a Gauss-Newton algorithm.

The benchmark stars in Fig. 17, are sorted by increasing $T_{\rm eff}$, and most of the stars are within the training set limits. We find that for most of the GBS, the CNN results compare very well. The cool giants alf_Cet, gam_Sge and alf_Tau have $T_{\rm eff}$ and $\log(g)$ outside the training limits, hence, we see a spread in $\log(g)$ and [Fe/H]. The GBS catalog also reports higher uncertainty for these three stars and the CNN [Fe/H] measurements are within the uncertainty limits. There are three metal-poor stars, HD 122563, HD 140283, and HD 84937, with [Fe/H] less than −2.0 dex. HD 122563 is the most metal-poor star with [Fe/H] = −2.62 dex for which we see the highest differences in $T_{\rm eff}$, $\log(g)$ and [Fe/H], although CNN estimate for A(Li) agrees with the AMBRE value. For HD 140283, with [Fe/H] = −2.36 dex, we see a difference of ∼500 for $T_{\rm eff}$ and 0.7 dex in [Fe/H], while the estimates for $\log(g)$ and A(Li) are in a good match. For HD 84937, CNN predictions for $T_{\rm eff}$, $\log(g)$ and A(Li) are in a very good agreement with GBS and AMBRE measurements, but we note a difference of 0.5 dex for [Fe/H]. In the case of lithium, for most of the GBS stars, CNN predictions compare well with AMBRE abundances within 1-$\sigma$. For stars with A(Li) below the training set limit of 0.0 dex, we see a difference of up to 0.8 dex in CNN and AMBRE/iDR6 predictions; for stars that are within the training limit and have A(Li) < 1.5 dex, a small difference (∼0.25 dex) in CNN, iDR6, and AMBRE measurements are seen. Overall, the CNN performs very well across the training label range and differences are seen only for stars outside the training range. Future spectroscopic surveys should be careful to target more metal-poor stars and cool giants. Also, the benchmark stars should include more metal-poor stars and cool giants.

In Fig. 18, we present the HR15N spectra around the 6707.8 Å lithium line for some solar twins, in different A(Li) regimes. The solar twins are selected from the training sample with S/N > 90 pix$^{-1}$ and with $T_{\rm eff}$ = 5 777 ± 150 K, $\log(g) = 4.44 \pm 0.15$ dex and [Fe/H] = 0.0±0.15 dex. CNN provides robust measurements for A(Li) ≥ 1.25 dex. Below this

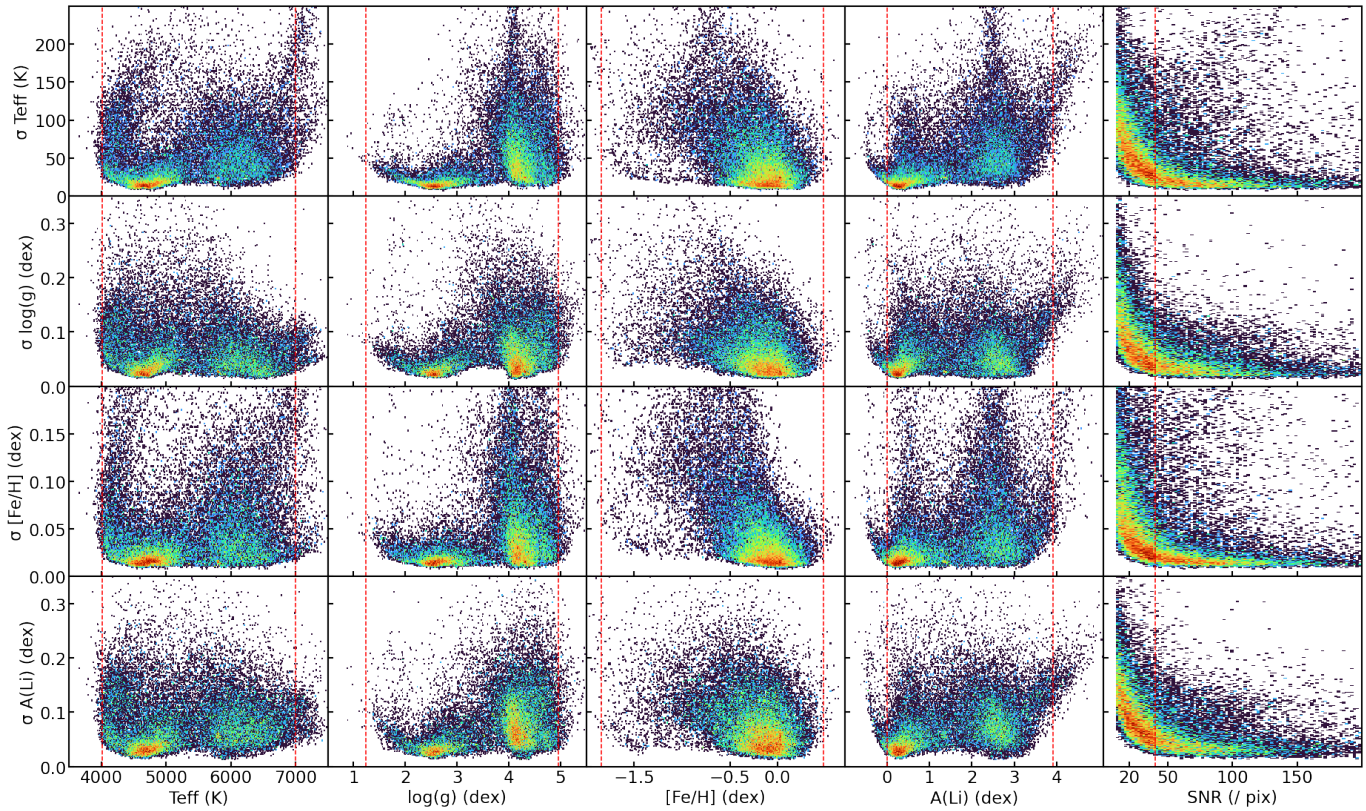


**Fig. 15.** 2D histograms showing CNN uncertainties (internal precision) as a function of four labels ($T_{\rm eff}$, log($g$), [Fe/H], A(Li)) and S/N for the observed sample with $S/N > 10\,{\rm pix}^{-1}$, i.e., 31 272 spectra. The red dashed line shows the limits of the training labels. The $x$-axis represents the labels and the $y$-axis shows the uncertainty ($\sigma$).

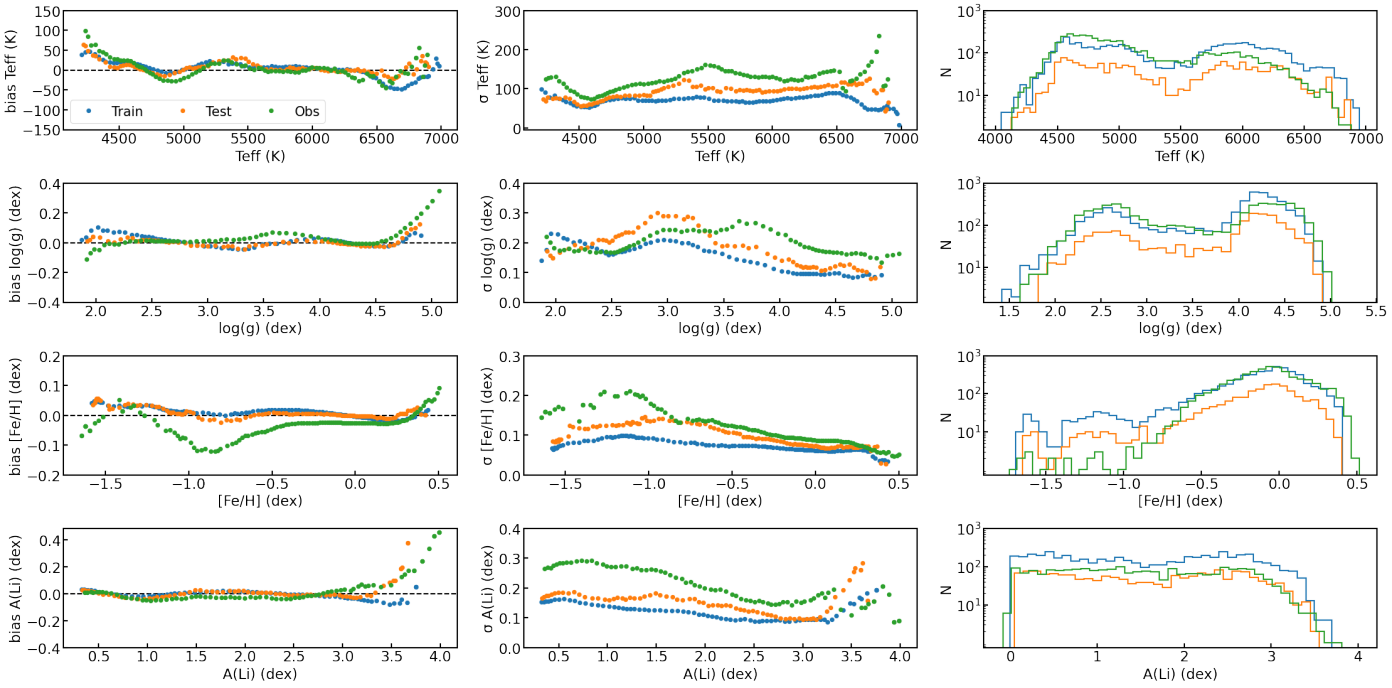


**Fig. 16.** Running mean bias and mean dispersion as a function of labels for the train (blue), test (orange), and observed (green) sets calculated in bin sizes: 250 K for $T_{\rm eff}$, 0.3 dex for log($g$), [Fe/H], and A(Li). The curves are representative of the real accuracy and precision of our CNN predictions. Bias = mean(CNN-iDR6) and $\sigma$ = std(CNN-iDR6) for each bin. On the right column we present the distribution of the train, test and observed sets in logarithmic $y$-axis. The observed sample is selected within the training set, with $S/N > 20\,{\rm pix}^{-1}$ and no GES flags; for A(Li), we selected only stars with Li measurements, instead of those with upper limit Li estimates.

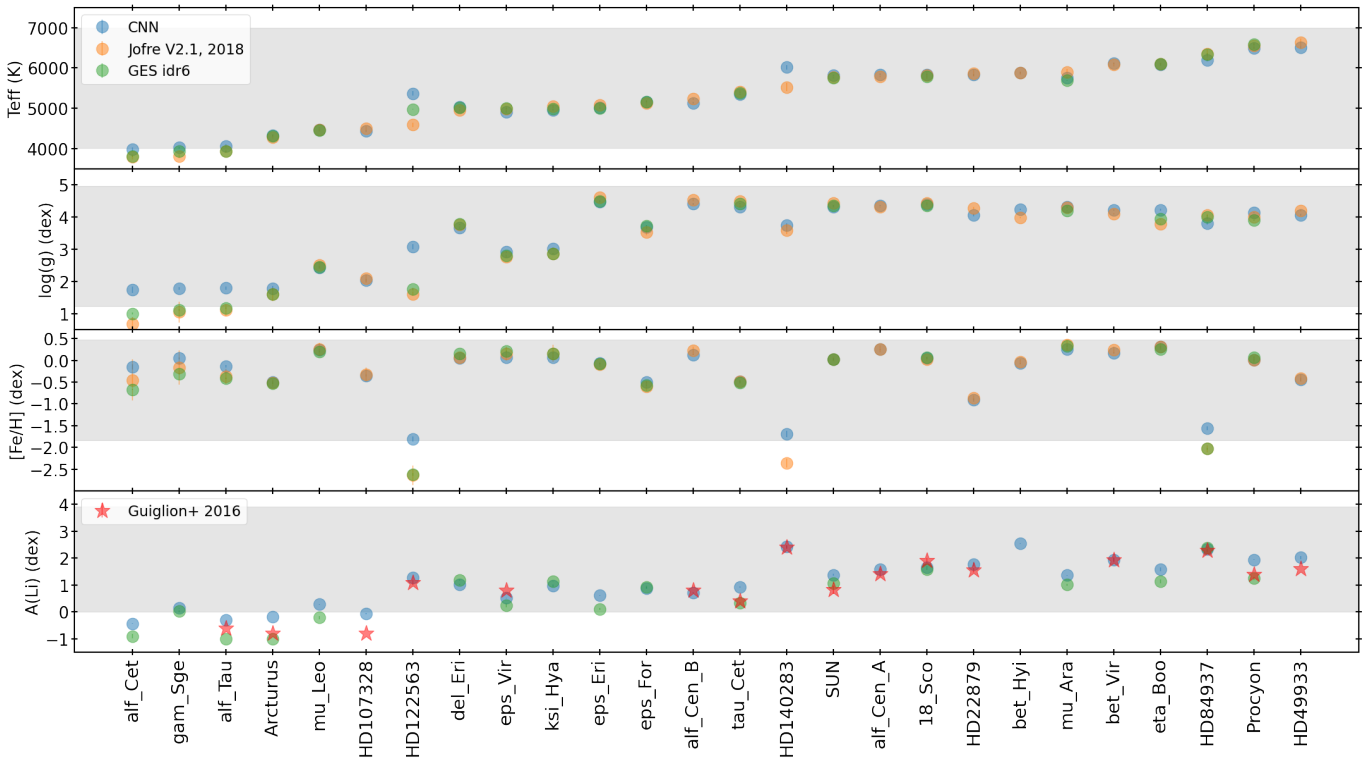


**Fig. 17.** Comparison of CNN prediction for the *Gaia* Benchmarks Stars (GBS). The reference $T_{\rm eff}$, log($g$), and [Fe/H] come from Jofré et al. (2018) and A(Li) from Guiglion et al. (2016). The GES-iDR6 values are also shown for comparison. On the $x$-axis, we present the GBS names sorted by increasing $T_{\rm eff}$ and on the $y$-axis, we present the four labels. The shaded region for each label represents the training set limits. The CNN predictions and error bars are mean of the estimates for the multiple spectra. CNN error bars are too small to be seen.

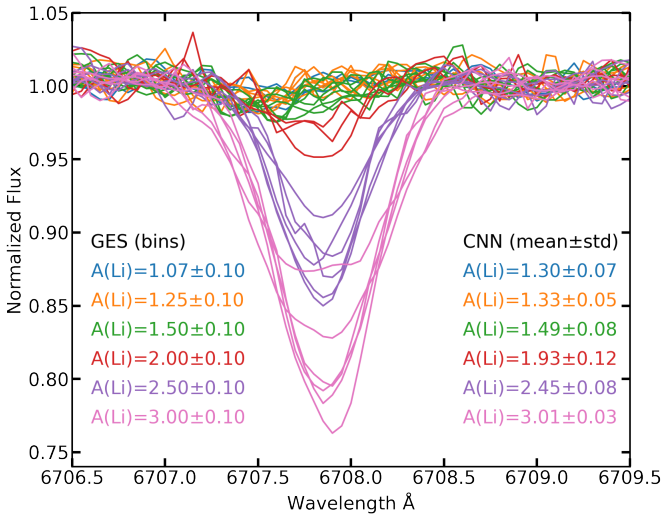


**Fig. 18.** Li features for the "solar twins" with varying Li abundance. The solar twins in the training sample are selected with $S/N > 90\,{\rm pix}^{-1}$ and with GES $T_{\rm eff}$ = 5777 ± 150 K, log($g$) = 4.44±0.15 dex and [Fe/H] = 0.0±0.15 dex. The colors represent the A(Li) bins, as listed on the left. On the right, we show the mean of CNN prediction for the shown spectra in each bin.

limit, CNN suffers from a positive bias, namely, the Solar abundance reported by GES is A(Li) = 1.07, while CNN measures 1.3 dex. For A(Li) of 1.07 dex (blue) and 1.25 dex (orange), the spectral features look almost identical within the noise. For these spectra, we see that the maximum flux absorption is ∼1.5% and most of the signal comes from an Fe blend.

An accurate measurement for lithium below 1.25 dex in Solar twins at resolution $R \sim 20\,000$ with CNN is then challenging and basically Li < 1.25 dex should be considered as limit in the dwarf regime. This could explain the difference in CNN, iDR6 and AMBRE measurements for the lithium measured in some of the benchmark stars. We carried out the same exercise for a typical RC star (around Solar [Fe/H]), and given the line is deeper, the CNN performs with no significant bias up to Li = 0 dex. It is representative of the well-known temperature dependence of the lithium line-shape. For 4MOST-LR/HR, it will be important to generalize this type of detection limit to the whole parameter space of the sample.

### 5.2. Validation with GALAH-DR3

The Galactic Archaeology with HERMES (GALAH, Buder et al. 2021) survey provides stellar parameters and chemical abundances, including lithium, using the spectrum synthesis code Spectroscopy Made Easy (SME) and 1D MARCS model atmospheres, along with additional photometry and astrometry. GALAH spectra are obtained at a higher resolution of $R \sim 28\,000$, compared to the GIRAFFE at $R \sim 20\,000$, and in four non-contiguous spectral bands between 4700 Å and 7900 Å. In Fig. 19, we present a comparison of CNN results for GES-iDR6 HR15N stars in common with the third data release GALAH-DR3 (Buder et al. 2021). The selected GES/CNN sub-sample has 73 HR15N stars in common with GALAH with available $T_{\rm eff}$, log($g$), [Fe/H], and A(Li). For GES/CNN we only consider the stars within the training set limits, S/N > 30 pix$^{-1}$, eVRAD < 0.5 km s$^{-1}$, and no GES flags. For GALAH stars, we followed the GALAH recommended S/N and

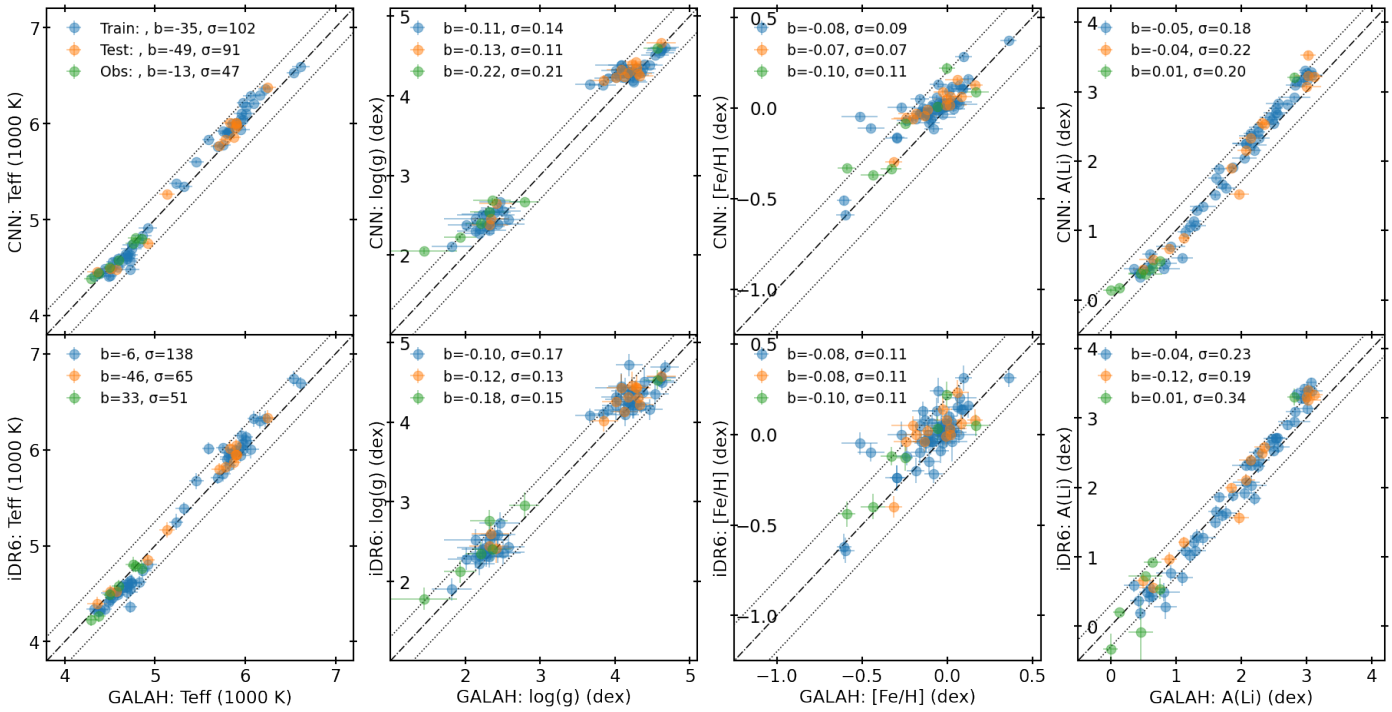


**Fig. 19.** Comparison of CNN results for stars in common with GALAH-DR3 (Buder et al. 2021). GES-iDR6 sample has stars selected with $S/N > 30\,\mathrm{pix}^{-1}$, within the training label limits, eVRAD $< 0.5\,\mathrm{km\,s}^{-1}$ and no GES flags and GALAH stars are selected with snr_c3_iraf $>30\,\mathrm{pix}^{-1}$, flag_sp = 0, flag_fe_h = 0, and flag_Li_fe = 0. The dash-dot line is the 1-to-1 line and two dotted lines are at ± 250 for $T_{\rm eff}$, ±0.3 dex for log($g$), ±0.2 dex for [Fe/H], ±0.3 dex for A(Li). The error bars show the errors reported in GES-iDR6 and GALAH-DR3; CNN uncertainties are too small to be seen.

flags, namely, snr_c3_iraf $> 30\,\mathrm{pix}^{-1}$, flag_sp = 0, flag_fe_h = 0, and flag_Li_fe = 0 (the flags = 0 represent no identified problems with determination of stellar parameters and iron and lithium abundances, respectively). The CNN atmospheric parameters and lithium predictions agree very well with GALAH, within 250 for $T_{\rm eff}$, 0.3 dex for log($g$), 0.2 dex for [Fe/H], 0.3 dex for A(Li). For the case of A(Li) < 1.0 dex, the spread in 1-to-1 relation is less for the case of CNN versus GALAH, indicating that CNN results are in better agreement with GALAH than the iDR6 measurements. Given the higher resolution for GALAH, it should be able to capture weaker lithium lines, hence providing more precise lithium values at A(Li) < 1.0 dex. We see that CNN works better at low lithium than standard pipelines in the cool regime (see also Fig. 13). Also, CNN can also efficiently deal with the noise. We see systematic $T_{\rm eff}$ offsets in GALAH vs. iDR6 with lower iDR6 measurements for cooler stars, and higher for hotter stars. This is also seen in the GALAH vs. CNN comparison. A similar systematic offset is seen for lithium, with lower CNN/iDR6 measurements for A(Li) < 2.5 dex and higher CNN/iDR6 measurements for A(Li) > 2.5 dex. Overall, GALAH, and CNN are in a good agreement and the offsets seen are systematic between GALAH and GES-iDR6.

### 5.3. Validation with Asteroseismic gravities

Here, we are aiming to compare CNN surface gravities with precise asteroseismic gravities. In Fig. 20, we present a comparison of log($g$) for 32 stars present in the CoRoT-GES sample of Valentini et al. (2016) with the CNN predictions. We selected only stars with good asteroseismic results given by flag OFLAG_GIR=0 from Valentini et al. (2016) and CNN/iDR6 stars are selected within the training label limits, $S/N > 30\,\mathrm{pix}^{-1}$, eVRAD $< 0.5\,\mathrm{km\,s}^{-1}$, and no GES flags. Figure 20 shows that there is an intrinsic bias between GES-iDR6 and CoRoT labels due to the different methods for deriving log($g$). The CNN results are consistent with the GES-iDR6 values, and they show a similar trend. The comparison shows presence of some outliers and we discuss two such outliers below.

For the star CNAME=19264480+0032497, with $T_{\rm eff}$ = 4815 K and log($g$) = 3.59 dex in iDR6, the CNN results (4635 K and 2.83 dex) agree better with CoRoT-GES values (4550 K and 2.71 dex). The star has a high projected rotational velocity ($v \sin i$) of $27.6\,\mathrm{km\,s}^{-1}$, which can be a cause behind this difference. About 35% of our training sample have stars with $v \sin i > 10\,\mathrm{km\,s}^{-1}$, hence, CNN are able to learn about the rotationally broadened spectral features.

For the star CNAME = 19240528+0152010, the iDR6 predictions are $T_{\rm eff}$ = 4663 K, log($g$) = 3.27 dex, and [Fe/H] = 0.01 dex, which is in agreement with CNN output (4872 K, 3.2 dex, and 0.04 dex), while there is a discrepancy with Corot predictions (4514 K, 1.77 dex, and −0.46 dex). A significantly lower log($g$) and [Fe/H] is provided by CoRoT-GES. We compare the spectrum of this star with another star for which the atmospheric parameters are similar to our CNN result and for which the CNN, iDR,6 and CoRoT-GES results agree. Both spectra look similar (besides the slightly lower log($g$) of the second spectrum), showing that Corot atmospheric parameters for this star should be taken with caution.

Such a comparison between the CNN predictions and Corot tells us that CNN is able to properly parametrize giants, while considering the HR15N is not an optimal setup for precisely constraining log($g$)s. We also show that CNN can correct inaccurate labels that are misclassified by standard pipelines; it is illustrative of the anomaly detection capability of CNNs.

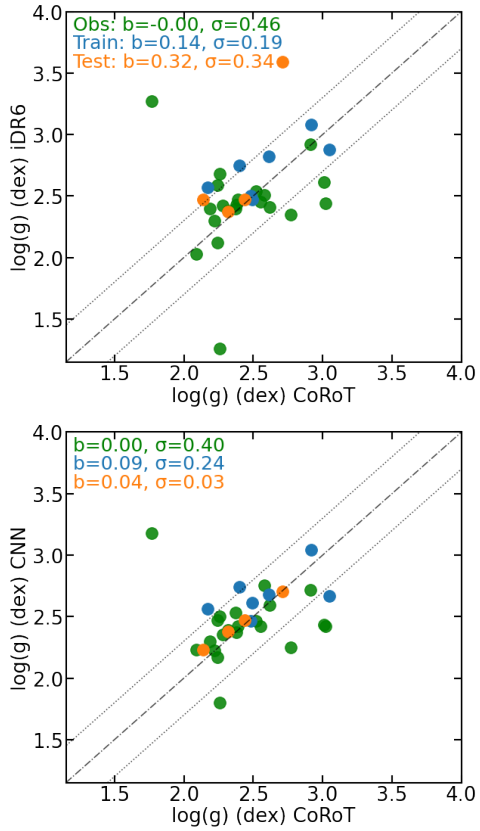


**Fig. 20.** Comparison with asteroseismic results. Left: CoRoT-GES vs. GES-iDR6 labels, Right: CoRoT-GES vs CNN predictions. Blue, orange, and green symbols represent the train, test, and observed sample selected within the training set limits ($S/N > 30\,\mathrm{pix}^{-1}$, $\mathrm{eVRAD} < 0.5\,\mathrm{km\,s}^{-1}$, no GES flags) and with CoRoT-GES flag OFLAG_GIR=0. The bias=mean(CNN-CoRoT) and $\sigma$ = std(CNN-CoRoT) are provided. The dash-dotted line is the 1-to-1 line and the two dotted lines are at ±0.3 dex.

## 6. Constraining the chemical evolution of lithium in the Milky Way

### 6.1. Galactic evolution of lithium

Recently, a number of studies have challenged the possibility to use main-sequence stars ($T_{\rm eff} > 5500$ K) to trace the lithium ISM abundance. Guiglion et al. (2019) suggested that the upper boundary of lithium in the super-solar metallicity main-sequence stars do not reflect the original ISM content – but, rather, lithium depletion due to an interplay between stellar evolution and radial-migration (see also Miglio et al. 2021 and references therein). Randich et al. (2020) investigated this Li decrease using GES stars both on the warm side of the lithium dip ($T_{\rm eff} > 6800$ K) in metal-rich open clusters together with PMS stars from very young clusters[11] (age < 100 Myr). They showed a lithium plateau of A(Li) ~ 3.4 dex at $0.1 < [\mathrm{Fe/H}] < 0.3$ dex. Their conclusion supported the scenario of Guiglion et al. (2019) which has recently been confirmed by Dantas et al. (2022).

Stars on the hot side of this dip have not undergone any Li depletion and they are the best candidates for the study of the galactic evolution of lithium with metallicities, ages, and galactocentric distances. However, atomic diffusion might have changed the original Li abundances in the atmospheres of (some) solar-metallicity stars (Romano et al. 2021; Charbonnel et al. 2021). Indeed, the lithium-dip (Li-dip), namely, the drop in A(Li) observed in the main sequence stars in temperature range of 6400–6800 K, has been confirmed in both cluster and field stars (e.g., Boesgaard & Tripicco 1986; Deliyannis et al. 2019). The origin of the Li-dip at this narrow $T_{\rm eff}$ range has been attributed to an interplay of mass-temperature dependent processes, most importantly, shallow surface convective zone and higher atmospheric mixing due to significant spin-down of initial PMS rotational velocity. Charbonnel et al. (2021) recently showed that hot metal-rich field stars do not exhibit any lithium decrease using GALAH and AMBRE data. This finding is in agreement with the result in Gao et al. (2020) using warm field stars from GALAH, and Randich et al. (2020) using OC stars, and Romano et al. (2021) using both.

In Fig. 21, we further investigate the Li ISM, with a sample of stars on the warm side of the Li-dip (warm group). To select these stars we adopted the following criteria: $S/N > 75\,\mathrm{pix}^{-1}$, $T_{\rm eff} > 6800$ K, $3.8 < \log(g) < 4.25$ dex, $-1 < [\mathrm{Fe/H}] < 0$ dex, A(Li) > 1.0 dex, $\mathrm{eVRAD} < 1.0\,\mathrm{km\,s}^{-1}$, $\mathrm{e}T_{\rm eff} < 200$ K, $\mathrm{elog}(g) < 0.1$ dex, $\mathrm{e[Fe/H]} < 0.2$ dex, and eA(Li) < 0.2 dex and also avoid peculiar stars and stars with emissions. We find stars with Li around 3.4 dex at [Fe/H] ~ 0.2 dex, consistently with the peak at A(Li) ~ 3.4 dex reported by Randich et al. (2020). We note the presence of super-solar [Fe/H] stars with lithium between 2.2 and 3.0 dex. These stars could be old (>6–7 Gyr) and have depleted their lithium. To be able to confirm these stars have indeed migrated from inner regions, an estimate of their birth-radii would be needed (e.g., Minchev et al. 2018).

We further investigate the ISM evolution in the metallicity regime $-1 < [\mathrm{Fe/H}] < 0$ dex. All of these stars have Li abundance above the Spite plateau value and there is a clear increase of lithium with metallicity from 2.2 to 3.2 dex. Given the small sample size, we cannot reliably confirm the presence/absence of a warm plateau at A(Li) = 2.69 dex (see GALAH survey, Gao et al. 2020), in the region of $-1.0 < [\mathrm{Fe/H}] < -0.5$ dex. However, the mean A(Li) for the 33 stars present in that metallicity range is lower at A(Li) = 2.44 ± 0.12 dex and show a gradient with metallicity. If we trust that the hot stars on the hot side of the dip are accurate tracers of the lithium ISM, we do not measure the usually reported steep rise of the ISM in the domain $-1.0 < [\mathrm{Fe/H}] < -0.5$ dex (based on cool dwarfs), but, instead, a shallow increase.

The consequence of such finding for the modeling of the lithium ISM on the domain $-1 < [\mathrm{Fe/H}] < -0.5$ dex would be to take into account earlier Li production by more massive sources and a longer delay in the production of lithium by the long-lived sources (as suggested by the chemical evolution model of Cescutti & Molaro 2019). Romano et al. (2021) arrived to the same conclusion based on GES-iDR6 data, suggesting a shorter delay in the production of lithium, claiming that nova white-dwarf progenitors must be in the range 3–8 $M_{\odot}$ rather than 1–8 $M_{\odot}$, as usually assumed (see Fig. 8 of Romano et al. 2021).

### 6.2. Search for lithium-rich giants

Standard stellar evolution models predict that the surface Li abundances of low-mass red giants after the first dredge-up decreases by ~60 times to below A(Li) ~ 1.50 dex (e.g., Lagarde et al. 2012) when starting from an initial A(Li) = 3.3 dex (solar meteoritic value). Lithium-rich giants are rare objects and confirm that lithium can be produced in stellar interiors (see e.g., Magrini et al. 2021b, and references therein); this results from the Cameron & Fowler (1971) mechanism.

[11] An updated list of clusters comprising also the OCs released in iDR6 can be found in Table 2 of Romano et al. (2021).

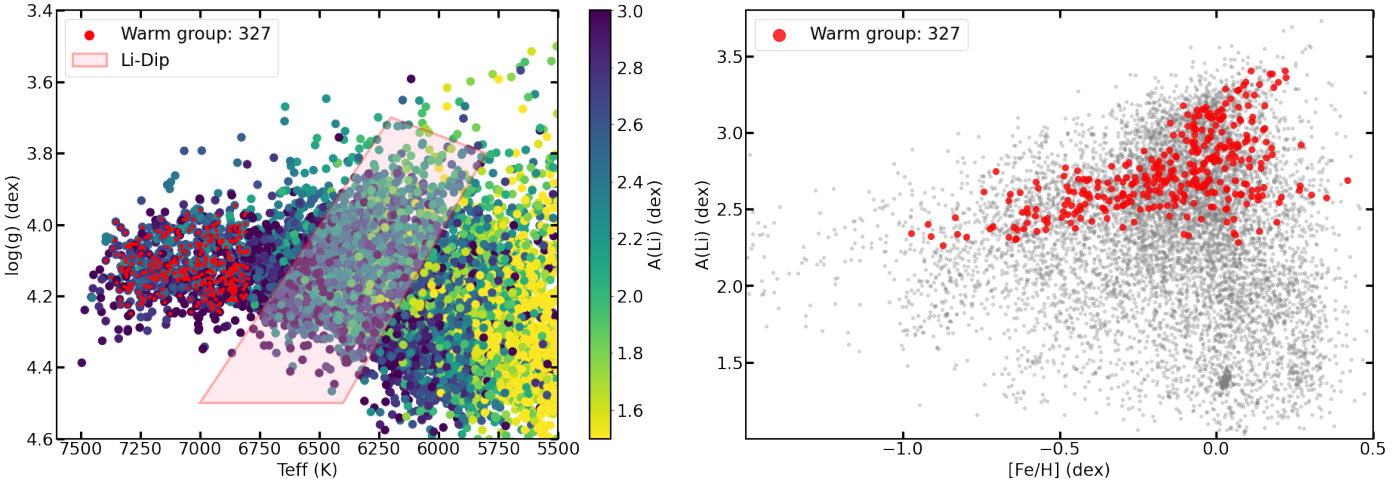


**Fig. 21.** Effective temperature vs. surface gravity diagram, with the stars color-coded according to their Li abundance (left). The approximate location of the Li-dip region according to Gao et al. (2020) is highlighted in pink. The red points represent the warm stars, $T_{\rm eff} > 6800$ and $S/N > 75$ pix$^{-1}$. The [Fe/H] vs. Li abundance trend for the warm stars shown as red points. Gray dots represent the other stars shown in the left plot (right).

These authors proposed that the reaction $^3$H + $\alpha$ $\rightarrow^7$ Be + $\gamma$ produces $^7$Be, which is then rapidly transported outwards by convection and non-standard mixing processes to lower temperatures, where it decays into $^7$Li. The Li-rich giants are believed to play a role in the enrichment of the ISM (Romano et al. 2001). Stellar Li enrichment is also possible due to external sources such as the measured over-abundance of Li as a result of a mass transfer process in a binary system, where the companion produces Li through the Cameron–Fowler mechanism. Planet engulfment was also proposed to explain such high lithium abundance in giants, although it seems this mechanism can increase the abundance only up to A(Li) ~ 2.2 dex (Aguilera-Gómez et al. 2016). We refer to Casey et al. (2016) for a review on the enrichment processes in Li-rich giants.

Our training sample contains just 38 lithium rich giants, considering a strict condition of $\log(g) < 3.2$ dex and A(Li) > 2.0 dex. It is important that the CNN is able to identify these rare objects, as they are of a great scientific interest. The Li-rich giants have previously been reported in earlier *Gaia*-ESO papers (Casey et al. 2016; Smiljanic et al. 2018; Sanna et al. 2020) and some of them are present in our training sample. In addition, we report the discovery of 31 new lithium rich giants by CNN in the observed sample (see Fig. 22). These stars were not reported in previous *Gaia*-ESO papers. We also checked the GALAH survey catalog in the southern sky of Li-rich giants by Martell et al. (2021) and found no match.

To identify the Li-rich giants, we selected stars with $T_{\rm eff} < 5500$ K, $\log(g) < 3.5$ dex and A(Li) > 2.0 dex, for which GES-iDR6 has not provided either one or any of the labels. To assure a reliable parameter estimation, we further selected spectra with low CNN uncertainties of e$T_{\rm eff} < 50$ K, e$\log(g) < 0.1$ dex, e[Fe/H] < 0.1 dex and eA(Li) < 0.1 dex, and S/N > 25 pix$^{-1}$ and E_VRAD < 0.5 km s$^{-1}$. We also checked for good photometry in *Gaia* EDR3 (Gaia Collaboration 2021) by selecting RUWE ≤ 1.4. The CNAME and atmospheric parameters for the 33 stars are listed in Table 2. Out of the 31 Li-rich giants, half of the stars have A(Li) between 2.0 and 3.0 dex with half have A(Li) > 3.0 dex with a maximum lithium abundance of 3.88 dex. One of the Li-rich giants is a fast-rotator with $v \sin i$ = 12.1 km s$^{-1}$; giants with high $v \sin i$ and A(Li) can indicate planetary engulfment and needs further study. We additionally confirmed that our Li-rich giants are not misclassified objects (e.g., PMS stars) using the $\gamma$-index of Damiani et al. (2014).

As seen in Fig. 22, our new Li-rich giants seem to be distributed along the whole giant branch, although a clear concentration is seen at the position of the red clump. However, in recent years, a view has emerged staring that Li-rich giants can be found only in the He-core burning red clump phase (Deepak & Reddy 2019; Deepak & Lambert 2021; Martell et al. 2021). Further analyses of our new sample is essential for investigating their properties and evaluating the possible mechanisms for their Li enrichment. Further investigations on these 31 Li-rich giants could be complemented by very precise asteroseismic $\log(g)$ (see for instance Zhou et al. 2022 with LAMOST data), if available with surveys such as TESS and PLATO (Singh et al. 2021).

## 7. Summary and future prospects

To prepare the ground for the future 4MOST and WEAVE spectroscopic surveys, we developed a convolutional neural network approach for determining atmospheric parameters ($T_{\rm eff}$, $\log(g)$, [Fe/H]) and lithium abundances from GES stellar spectra. We built a training set of 7031 stars, based on high-quality stellar labels from GES iDR6. The main results are summarized as follows:

1. Our CNN shows very good performance, even though we masked H$\alpha$ and despite the fact that the wavelength range in GIRAFFE HR15N setup is not considered optimal for determinations of atmospheric parameters (Lanzafame et al. 2015). These results indicate that our trained CNN models are competent and have learned the available spectral features. The CNN is able to provide results with typical uncertainties of ~35 K for $T_{\rm eff}$, 0.05 dex for $\log(g)$, 0.03 dex for [Fe/H], and 0.06 dex for A(Li).
2. Overall, the CNN predictions show a very good agreement in comparison with the GES-iDR6 input labels. The CNN achieves a good performance for all S/N values, including the low S/N (≈20 pix$^{-1}$) spectra. Thanks to the large variety of rotational velocities in the training sample, the CNN is able to accurately predict atmospheric parameters, even for the fast rotators for which the spectral features are broadened and can be blended with neighbouring lines. As CNN is sensitive to even small systematics in the input data, we found

**Table 2.** 31 newly discovered GES Li-rich giants and their CNN associated atmospheric parameters: $T_{\rm eff}$ (K), log($g$) (dex), [Fe/H] (dex); and lithium abundances, A(Li) (dex).

| CNAME | $T_{\rm eff}$ | log($g$) | [Fe/H] | A(Li) |
|---|---|---|---|---|
| 07434938−3841399 | 4841.0 | 2.84 | −0.31 | 3.88 |
| 10495937−6345553 | 4805.0 | 2.72 | −0.16 | 3.83 |
| 07464933−3750081 | 4948.0 | 2.90 | −0.20 | 3.62 |
| 08064077−4736441 | 4797.0 | 2.66 | −0.10 | 3.56 |
| 16271097−2455213 | 4920.0 | 2.82 | −0.45 | 3.55 |
| 06410348+0905141 | 5071.0 | 3.13 | −0.18 | 3.50 |
| 07493206−3759457 | 4799.0 | 2.69 | −0.24 | 3.48 |
| 10430727−6456318 | 4619.0 | 2.47 | −0.16 | 3.43 |
| 08110435−4853491 | 4831.0 | 2.68 | −0.30 | 3.42 |
| 10400095−6419586 | 4525.0 | 2.34 | 0.01 | 3.29 |
| 08084532−4701292 | 4836.0 | 2.74 | −0.19 | 3.26 |
| 07462219−3712141 | 4862.0 | 2.82 | −0.20 | 3.22 |
| 06273069−0440141 | 4714.0 | 2.53 | −0.68 | 3.21 |
| 08512566−4135067 | 4331.0 | 2.20 | 0.23 | 3.15 |
| 08102172−4845417 | 4514.0 | 2.36 | −0.06 | 3.14 |
| 06255393−0457404 | 4981.0 | 2.93 | −0.29 | 3.03 |
| 08083354−4711111 | 4441.0 | 2.31 | 0.10 | 3.00 |
| 07442999−3812166 | 4857.0 | 2.65 | −0.24 | 2.98 |
| 10350175−6405092 | 4469.0 | 2.35 | 0.11 | 2.88 |
| 07475310−3733040 | 4853.0 | 2.80 | −0.21 | 2.86 |
| 10483936−6327542 | 4383.0 | 2.21 | 0.08 | 2.74 |
| 10420066−6421333 | 4397.0 | 2.22 | 0.08 | 2.73 |
| 11130526−7617396 | 4815.0 | 2.67 | −0.32 | 2.73 |
| 11123294−7727006 | 4315.0 | 2.31 | 0.16 | 2.61 |
| 10575316−7634459 | 4858.0 | 2.70 | −0.21 | 2.42 |
| 07472841−3850499 | 5276.0 | 3.47 | −0.12 | 2.35 |
| 10513847−6335341 | 4352.0 | 2.24 | 0.31 | 2.29 |
| 07472390−3856376 | 5049.0 | 2.93 | −0.24 | 2.27 |
| 06272996−0518528 | 4522.0 | 2.44 | −0.02 | 2.20 |
| 08075108−4744027 | 4719.0 | 2.57 | −0.13 | 2.19 |
| 07483625−3724338 | 4939.0 | 3.01 | −0.15 | 2.03 |

**Notes.** Table is ordered by A(Li).

that large uncertainties in $V_{\rm rad}$ (>0.5 km s$^{-1}$) can degrade the CNN performances.

3. *Gaia* benchmark stars within the training label range are accurately predicted within 1-sigma by CNN while those outside show some systematics. The origin of such a discrepancy could be a lack of metal-poor stars (both dwarfs and giants) in the training set. It could also come from the fact that metal-poor stars are more difficult to parametrize due to weaker lines and possible NLTE effect.
4. The catalog of atmospheric parameters and Li abundances for ∼40 000 stars is publicly available at CDS. In addition, we have made the CNN code, spectra and labels available to the community[12].
5. The CNN atmospheric parameters and lithium predictions agree very well with GALAH DR3, within 250 K for $T_{\rm eff}$, 0.3 dex for log($g$), 0.2 dex for [Fe/H], 0.3 dex for A(Li). Systematic offsets are present between the GALAH DR3 and CNN (also with respect to input GES-iDR6 labels) due to the different instrument setup, spectroscopic pipelines, and calibration strategies. We show that the CNN atmospheric parameters match up nicely with asteroseismic results from CoRoT. We also demonstrate that CNN can correct wrongly assigned labels.
6. We have verified that the CNN is learning from relevant spectral features for the atmospheric parameters (e.g., the Quintet is sensitive to log($g$)) and found that CNN is able to single out the lithium line among hundreds of other lines, for precisely determining lithium. Using correlations for inferring elemental abundances without spectral features should be avoided.
7. We investigated the ISM chemical evolution of lithium, with the stars on the hot side of the lithium dip (more representative of the ISM). Our findings suggest that the usually reported steep rise of the upper boundary of lithium is not visible on the domain $-1 <$ [Fe/H] $< 0$ dex, exhibiting a shallower rise of the ISM. This suggests that earlier Li production by more massive sources and a longer delay in the production of Li by the long-lived sources for enriching the ISM should be taken in account, as claimed by recent chemical evolution modeling (Cescutti & Molaro 2019; Romano et al. 2021). In addition, there is no decrease in the lithium boundary with [Fe/H] > 0 dex, but we report the presence of stars with lithium between 2.2 and 3.0 dex which are likely to have depleted their lithium content.
8. We report the discovery of 31 new Li-rich giants. A follow-up study using asteroseimic data for these stars could provide an insight on stellar Li production and mixing mechanisms. 4MOST is expected to discover thousands of these objects, making it possible to study these peculiar stars over a large Galactic volume, for instance, in the bulge, and metallicity range.

Our work confirms that CNNs are efficient for deriving lithium abundances based on HR15N spectra, namely, very similar data as 4MOST and WEAVE. It gives excellent perspectives for data analysis with CNN in the context of these two surveys. However, several improvements could be made in order to refine CNN performance. For instance, in order to increase the diversity in the training sample, adding the spectra of binary stars and properly dealing with emission features could be helpful.

For the future use of CNNs, it will be crucial to build the training sets proactively, namely, not only relying on sets we build for a given survey, but carefully filling in regions of the HR diagram with proper targets. In particular, attention should be paid to populating the metal-poor tail of the training set in order to avoid biases. In such a way, the training set limits would be extended and a larger label space could be probed – as the current application is clearly limited within the available training set limits. In a future work, it would be interesting to explore Bayesian NNs and different types of loss functions such as the negative log likelihood to provide better uncertainty estimates.

One important aspect of spectroscopy that was not taken into account in this project are the NLTE effects coupled with a 3D structure of the atmosphere that can affect lithium abundance measurements. Several studies have published grids of NLTE corrections for lithium abundances, such as Lind et al. (2009), and more recently Wang et al. (2021). This NLTE-3D corrections affect mainly the cool-giants (up to +0.3 dex) in the high-lithium regime. For metal-rich dwarfs, the typical correction is on the order of −0.1 dex, for 5000 < $T_{\rm eff}$ < 6500 K (see also Figs. 1 and 2 of Magrini et al. 2021a). A potential future task could be to include these NLTE corrections to the training-set lithium label, but we expect no major change in the results presented in this work. In the context of future surveys, 3D NLTE measurements should be performed homogeneously for as many elements as possible. For instance, $\alpha$-elements such as O, Mg,

[12] https://github.com/SamirNepal/Li_CNN_2022

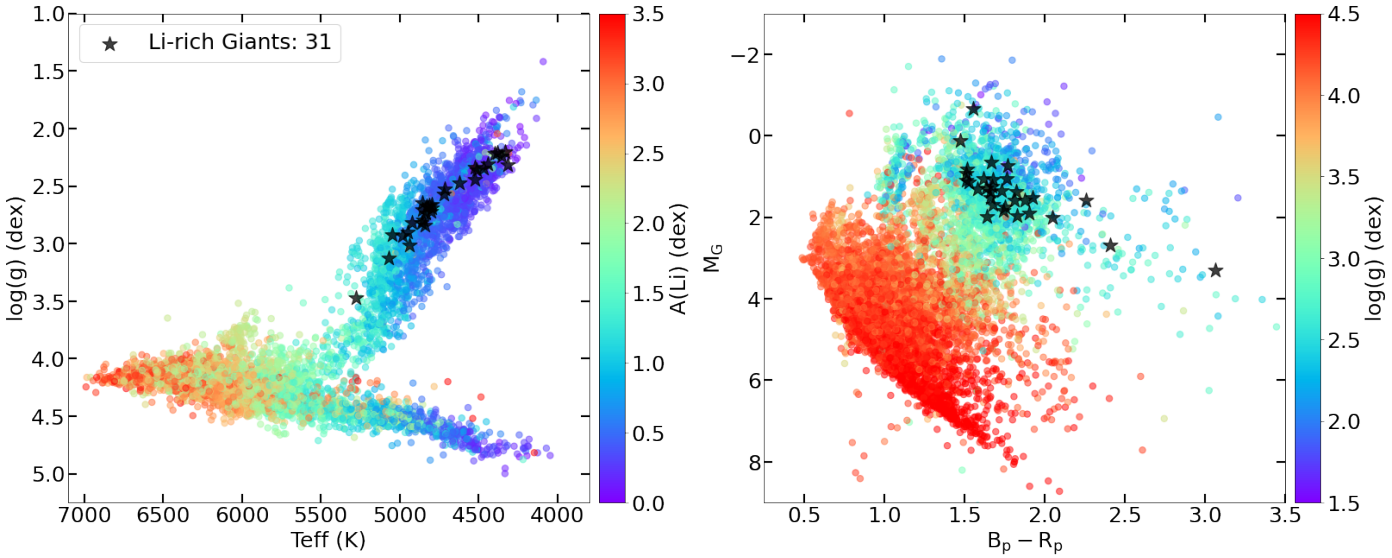


**Fig. 22.** Kiel-diagram showing the newly-discovered Li-rich giants (black stars) along with the training sample color-coded according to their Li abundance (left). *Gaia* color–magnitude diagram for the same stars (right). The training sample stars are colored by their surface gravities.

and Ti will be measurable by 4MIDABLE-HR and are affected by 3D NLTE in a non-negligible way (Bergemann et al. 2021, 2017, 2012; Sitnova et al. 2018).

Concerning the optimization of the training set, properly including M stars with strong TiO bands in the training set will allow us to accurately parametrize this type of object. It will be a necessity for 4MOST, which is planned to observe (among other targets) open-clusters. Regarding the sensitivity of CNN to $V_{rad}$, future surveys observing with multiple spectrographs should take care to provide accurate radial velocities in order to minimize the possible systematics during the training phase.

On this study, we show that lithium abundances in solar-type stars with lithium lower than 1.25 dex can not be measured precisely at the GIRAFFE HR15 resolution (~20 000). For the future use of CNN (or ML in general) for stellar abundance measurements, it will be necessary to develop an objective criterion for deciding whether an abundance is a real detection or an upper[13] limit.

*Acknowledgements.* We thank the anonymous referee for the very constructive comments and suggestions. This work was partly supported by the European Union FP7 programme through ERC grant number 320360 and by the Leverhulme Trust through grant RPG-2012-541. We acknowledge the support from INAF and Ministero dell' Istruzione, dell' Università' e della Ricerca (MIUR) in the form of the grant "Premiale VLT 2012". The results presented here benefit from discussions held during the *Gaia*-ESO workshops and conferences, supported by the ESF (European Science Foundation) through the GREAT Research Network Programme. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. S.N. is grateful for the unwavering support of his family. M. L. L. Dantas and R. Smiljanic acknowledge support by the National Science Centre, Poland, project 2019/34/E/ST9/00133. T.B. was supported by grant No. 2018-04857 from the Swedish Research Council. M.B. is supported through the Lise Meitner grant from the Max Planck Society. We acknowledge support by the Collaborative Research centre SFB 881 (projects A5, A10), Heidelberg University, of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 949173). This work made use of OVERLEAF for preparing this document, and of the following PYTHON packages (not previously mentioned): MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), PANDAS (McKinney 2010), SEABORN (Waskom 2021). This work also benefited from TOPCAT (Taylor 2005).

[13] https://www.overleaf.com/

## References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org.
Aguilera-Gómez, C., Chanamé, J., Pinsonneault, M. H., & Carlberg, J. K. 2016, ApJ, 829, 127
Ambrosch, M., Guiglion, G., Mikolaitis, Š., et al. 2023, A&A, in press https://doi.org/10.1051/0004-6361/202244766
Anders, F., Chiappini, C., Santiago, B. X., et al. 2018, A&A, 619, A125
Bailer-Jones, C. A. L., Irwin, M., Gilmore, G., & von Hippel, T. 1997, MNRAS, 292, 157
Bailer-Jones, C. A. L., Irwin, M., & von Hippel, T. 1998, MNRAS, 298, 361
Bensby, T., & Lind, K. 2018, A&A, 615, A151
Bergemann, M., Lind, K., Collet, R., Magic, Z., & Asplund, M. 2012, MNRAS, 427, 27
Bergemann, M., Collet, R., Amarsi, A. M., et al. 2017, ApJ, 847, 15
Bergemann, M., Hoppe, R., Semenova, E., et al. 2021, MNRAS, 508, 2236
Bialek, S., Fabbro, S., Venn, K. A., et al. 2020, MNRAS, 498, 3817
Bishop, C. M. 1995, Neural Networks for Pattern Recognition (Oxford: Oxford University Press, Inc.)
Blanco-Cuaresma, S., Soubiran, C., Jofré, P., & Heiter, U. 2014, A&A, 566, A98
Boesgaard, A. M., & Tripicco, M. J. 1986, ApJ, 302, L49
Bonifacio, P., & Molaro, P. 1997, MNRAS, 285, 847
Bragaglia, A., Alfaro, E. J., Flaccomio, E., et al. 2022, A&A, 659, A200
Brown, J. A., Sneden, C., Lambert, D. L., & Dutchover, E., Jr 1989, ApJS, 71, 293
Buder, S., Sharma, S., Kos, J., et al. 2021, MNRAS, 506, 150
Cameron, A. G. W., & Fowler, W. A. 1971, ApJ, 164, 111
Casey, A. R., Ruchti, G., Masseron, T., et al. 2016, MNRAS, 461, 3336
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, A&A, 635, A45
Cescutti, G., & Molaro, P. 2019, MNRAS, 482, 4372
Charbonnel, C., & Balachandran, S. C. 2000, A&A, 359, 563
Charbonnel, C., Borisov, S., de Laverny, P., & Prantzos, N. 2021, A&A, 649, L10
Chollet, F. 2015, Keras, https://keras.io
Čotar, K., Zwitter, T., Traven, G., et al. 2021, MNRAS, 500, 4849
Dalton, G. 2016, ASP Conf. Ser., 507, 97
Damiani, F., Prisinzano, L., Micela, G., et al. 2014, A&A, 566, A50

Dantas, M. L. L., Guiglion, G., Smiljanic, R., et al. 2022, A&A, 668, L7
D'Antona, F., & Matteucci, F. 1991, A&A, 248, 62
de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, The Messenger, 175, 3
Deepak, & Lambert, D. L. 2021, MNRAS, 507, 205
Deepak, & Reddy, B. E. 2019, MNRAS, 484, 2000
Delgado Mena, E., Bertrán des Lis, S., Adibekyan, V. Z., et al. 2015, A&A, 576, A69
Deliyannis, C. P., Anthony-Twarog, B. J., Lee-Brown, D. B., & Twarog, B. A. 2019, AJ, 158, 163
Fabbro, S., Venn, K. A., O'Briain, T., et al. 2018, MNRAS, 475, 2978
Fields, B. D. 2011, Ann. Rev. Nucl. Particle Sci., 61, 47
Filipi Gonçalves dos Santos, C., & Papa, J. P. 2022, ACM Computing Surveys, 54, 213
Franciosini, E., Randich, S., de Laverny, P., et al. 2022, A&A, 668, A49
Fu, X., Romano, D., Bragaglia, A., et al. 2018, A&A, 610, A38
Fukushima, K., & Miyake, S. 1982, Competition and Cooperation in Neural Nets (New York: Springer), 267
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2021, A&A, 649, A1
Gao, Q., Shi, J.-R., Yan, H.-L., et al. 2019, ApJS, 245, 33
Gao, X., Lind, K., Amarsi, A. M., et al. 2020, MNRAS, 497, L30
Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25
Gilmore, G., Randich, S., Worley, C. C., et al. 2022, A&A, 666, A120
Gratton, R. G., & D'Antona, F. 1989, A&A, 215, 66
Grevesse, N., Asplund, M., & Sauval, A. J. 2007, Space. Sec. Rev., 130, 105
Guiglion, G., de Laverny, P., Recio-Blanco, A., et al. 2016, A&A, 595, A18
Guiglion, G., Chiappini, C., Romano, D., et al. 2019, A&A, 623, A99
Guiglion, G., Matijevič, G., Queiroz, A. B. A., et al. 2020, A&A, 644, A168
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357
Heiter, U., Jofré, P., Gustafsson, B., et al. 2015, A&A, 582, A49
Heiter, U., Lind, K., Bergemann, M., et al. 2021, A&A, 645, A106
Hong-liang, Y., & Jian-rong, S. 2022, Chinese. Astron. Astrophys, 46, 1
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Izzo, L., Della Valle, M., Mason, E., et al. 2015, ApJ, 808, L14
Jackson, R. J., Jeffries, R. D., Lewis, J., et al. 2015, A&A, 580, A75
Jofré, P., Heiter, U., Soubiran, C., et al. 2014, A&A, 564, A133
Jofré, P., Heiter, U., Soubiran, C., et al. 2015, A&A, 582, A81
Jofré, P., Heiter, U., Tucci Maia, M., et al. 2018, Res. Notes Am. Astron. Soc., 2, 152
Kingma, D. P., & Ba, J. 2014, ArXiv e-prints [arXiv:1412.6980]
Kusakabe, M., Cheoun, M.-K., Kim, K. S., et al. 2019, ApJ, 872, 164
Lagarde, N., Decressin, T., Charbonnel, C., et al. 2012, A&A, 543, A108
Lambert, D. L., & Reddy, B. E. 2004, MNRAS, 349, 757
Lanzafame, A. C., Frasca, A., Damiani, F., et al. 2015, A&A, 576, A80
LeCun, Y., & Bengio, Y. 1995, in The Handbook of Brain Theory and Neural Networks, ed. M. A. Arbib (MIT Press)
LeCun, Y., Boser, B., Denker, J. S., et al. 1989, Neural Comput., 1, 541
Lecun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436
Leung, H. W., & Bovy, J. 2019, MNRAS, 483, 3255
Lima, E. V. R., Sodré, L., Bom, C. R., et al. 2022, Astron. Comput., 38, 100510
Lin, Y.-C., & Wu, J.-H. P. 2021, Phys. Rev. D, 103, 063034
Lind, K., Asplund, M., & Barklem, P. S. 2009, A&A, 503, 541
Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021, A&A, 649, A2
Lodders, K., & Palme, H. 2009, Meteorit. Planet. Sci. Suppl., 72, 5154
Magrini, L., Lagarde, N., Charbonnel, C., et al. 2021a, A&A, 651, A84
Magrini, L., Smiljanic, R., Franciosini, E., et al. 2021b, A&A, 655, A23
Margalef-Bentabol, B., Huertas-Company, M., Charnock, T., et al. 2020, MNRAS, 496, 2346
Martell, S. L., Simpson, J. D., Balasubramaniam, A. G., et al. 2021, MNRAS, 505, 5340
Matijevič, G., Chiappini, C., Grebel, E. K., et al. 2017, A&A, 603, A19
Matteucci, F., D'Antona, F., & Timmes, F. X. 1995, A&A, 303, 460
McKellar, A. 1940, PASP, 52, 407
McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, eds. S. van der Walt, & J. Millman, 56
Miglio, A., Chiappini, C., Mackereth, J. T., et al. 2021, A&A, 645, A85
Minchev, I., Anders, F., Recio-Blanco, A., et al. 2018, MNRAS, 481, 1645
Ness, M., Hogg, D. W., Rix, H. W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16
O'Briain, T., Ting, Y.-S., Fabbro, S., et al. 2021, ApJ, 906, 130
Pancino, E., Lardo, C., Altavilla, G., et al. 2017, A&A, 598, A5
Pasquini, L., Avila, G., Blecha, A., et al. 2002, The Messenger, 110, 1
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129
Pinsonneault, M. 1997, ARA&A, 35, 557
Pitrou, C., Coc, A., Uzan, J.-P., & Vangioni, E. 2018, Phys. Rep., 754, 1
Prantzos, N., de Laverny, P., Guiglion, G., Recio-Blanco, A., & Worley, C. C. 2017, A&A, 606, A132
Ramírez, I., Fish, J. R., Lambert, D. L., & Allende Prieto, C. 2012, ApJ, 756, 46
Randich, S., & Gilmore, G. 2013, The Messenger, 154, 47
Randich, S., & Magrini, L. 2021, Front. Astron. Space Sci., 8, 6
Randich, S., Pasquini, L., Franciosini, E., et al. 2020, A&A, 640, L1
Randich, S., Gilmore, G., Magrini, L., et al. 2022, A&A, 666, A121
Reeves, H., Fowler, W. A., & Hoyle, F. 1970, Nature, 226, 727
Romano, D., Matteucci, F., Molaro, P., & Bonifacio, P. 1999, A&A, 352, 117
Romano, D., Matteucci, F., Ventura, P., & D'Antona, F. 2001, A&A, 374, 646
Romano, D., Magrini, L., Randich, S., et al. 2021, A&A, 653, A72
Sackmann, I. J., & Boothroyd, A. I. 1999, ApJ, 510, 217
Sanna, N., Franciosini, E., Pancino, E., et al. 2020, A&A, 639, L2
Singh, R., Reddy, B. E., Campbell, S. W., Kumar, Y. B., & Vrard, M. 2021, ApJ, 913, L4
Sitnova, T. M., Mashonkina, L. I., & Ryabchikova, T. A. 2018, MNRAS, 477, 3343
Smiljanic, R., Korn, A. J., Bergemann, M., et al. 2014, A&A, 570, A122
Smiljanic, R., Franciosini, E., Bragaglia, A., et al. 2018, A&A, 617, A4
Sneden, C., Bean, J., Ivans, I., Lucatello, S., & Sobeck, J. 2012, Astrophysics Source Code Library [record ascl:1202.009]
Spite, F., & Spite, M. 1982, A&A, 115, 357
Stonkutė, E., Koposov, S. E., Howes, L. M., et al. 2016, MNRAS, 460, 1131
Taylor, M. B. 2005, ASP Conf. Ser., 347, 29
Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2017, ApJ, 843, 32
Ting, Y.-S., Conroy, C., Rix, H.-W., & Asplund, M. 2018, ApJ, 860, 159
Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2019, ApJ, 879, 69
Valenti, J. A., & Piskunov, N. 1996, A&AS, 118, 595
Valentini, M., Chiappini, C., Miglio, A., et al. 2016, Astron. Nachr., 337, 970
Van der Maaten, L., & Hinton, G. 2008, J. Mach. Learn. Res., 9, 2579
Wang, E. X., Nordlander, T., Asplund, M., et al. 2021, MNRAS, 500, 2159
Waskom, M. L. 2021, J. Open Source Softw., 6, 3021
Woosley, S. E., & Weaver, T. A. 1995, ApJS, 101, 181
Xiang, M., Ting, Y.-S., Rix, H.-W., et al. 2019, ApJS, 245, 34
Zhang, X., Zhao, G., Yang, C. Q., Wang, Q. X., & Zuo, W. B. 2019, PASP, 131, 094202
Zhou, Y., Wang, C., Yan, H., et al. 2022, ApJ, 931, 136

## Appendix A: CNN model technical details

In the following sections, we detail the technical aspect of the CNN, in particular, the architecture, the choice of hyperparameters, and model generalization.

### A.1. Convolution and the fully connected layers

Convolution layers are the central part of the CNN class of neural networks, as they are key to identifying patterns and features in input data (Fukushima & Miyake 1982; LeCun et al. 1989). The 1D stellar spectra we use are characterized by absorption features governed by the physical properties of the stellar atmosphere. The CNN's goal is then to learn how these spectral features correlate with the stellar labels. The convolution layer, consisting of a collection of filters, when convolved with the 1D input from the previous layer, is able to extract the features. During the learning process, these filter parameters are optimized. After extensive tests, we adopted the model with 3 Conv1D layers with eight, six, and four filters, respectively. Using multiple filters in each convolution layer is similar to looking at the same object with different perspectives.

After the first and second convolution layers, we apply a Maxpooling process, which reduces the feature map size by half. This is very useful to reduce the overall training parameters, which also reduces training time, while the network focuses on important features. Maxpooling isn't applied after the third convolution layer to avoid losing too much information.

At the heart of every neural network are the fully connected layers (or dense layers) (Lecun et al. 2015). They make up the central component that adds complexity and meaning to the functional approximation of the relationship between (in our case) the input spectrum and the output labels. As shown in Fig. 3, the features learned from the input spectrum by the convolution layers are passed to the dense layers. This combination of convolution and dense layers ensures that the model learns from the whole spectral range instead of just the individual spectral features.

Our architecture contains three dense layers and one output layer (also a dense layer). The four feature maps from the last Conv1D layers are flattened before being fed to the first dense layer. The first dense layer has 64 neurons and receives input from the 6 788 neurons of the flattened layer. The second and third dense layers have 128 and 32 neurons, respectively. The output dense layer is naturally composed of four neurons corresponding to the four training labels. Our choice of the number of layers and neurons is based on a good deal of experimentation, with the goal of having a CNN that is complex enough, without mitigating the training performance.

### A.2. Choice of hyperparameters

Hyperparameters are set at the beginning of the training and remain the same throughout the training, as opposed to the learnable model parameters such as the weights and biases. Here we discuss some important hyperparameters:

1. **Weight initialization:** The weights of all parameters in the model have to be initialized before the training and neural networks are very sensitive to the initial weight values, as poor initialization can lead to a non-convergence. We adopted the intensively used "golrot uniform" that initializes weights from a uniform distribution within a certain range.
2. **Activation functions:** Activation functions are the mathematical functions that decide whether a neuron is activated or not. It adds non-linearity to the network and decides the output of any node or layer depending on the input. Each layer is activated using the "Leaky-ReLu" activation function and for the output layer, we use the "linear" activation.
3. **Epochs:** One complete pass of the training data through the network is called an epoch. Multiple epochs are needed for a good training. We allow for large number of training epochs until the training and test loss curves flatten out and stopped by using the EarlyStopping process (see Fig. 4).
4. **Batch size:** This refers to the number of data items used for one update of the model parameters during a single training epoch. The "mini batch stochastic gradient descent" learning algorithm updates the model weights multiple times depending on the batch size in a single training epoch. It is an excellent way to lower the training time. A good choice for the batch size also provides regularization and stability during the training. We adopted a batch size of 64 as a balance between good approximation of the training set and faster training time.
5. **Learning rate:** The learning rate ($\eta$) is the amount by which the weights are updated during the training and affect both the smooth convergence and training time. We tested several values of $\eta$ and found that the best performances, for our model, are achieved for $\eta = 0.0001$.

### A.3. Model generalization: Avoiding over- or under-fitting

The generalization and proper convergence of the model during the training is important to avoid over or underfitting and to ensure that the training progresses smoothly (Filipi Gonçalves dos Santos & Papa 2022). Our choice of convolution and dense layers ensures that the model does not underfit the training data, hence, attention must be paid to avoiding overfitting the model. For this purpose, we employ the following regularization, dropouts, and early-stopping procedures detailed below.

In each of the three convolution layers, the L2 Regularization function is applied, allowing for a penalization of the loss function (see Sect. 3.2) by adding to it a squared magnitude of model weights as a penalty term. The penalty term minimizes the model weights and ascertains that less significant features in the spectrum do not significantly affect the label prediction.

We applied a dropout layer on the inputs of the three inner dense layers. At each training epoch (explained below in Sect. A.2), a certain number of neurons are randomly selected and their contribution to the activation of neurons in subsequent layers is temporally removed. This forces the network to learn from the whole wavelength range of the spectrum as the model weights do not rely only on a very few spectral features and do not neglect less significant features. In Fig. 3, we can see that 20% of the neurons are dropped prior to the dense layers.

While training the CNN model, it is recommended that the training stop once the validation performance starts to degrade. For this task, we employed a callback called EarlyStopping in the model. This callback monitors the validation and test loss at the end of each training epoch and once the loss degrades or stagnates, over the last 25 epochs, the training is stopped and the model weights of the best training epoch are saved.

Besides these techniques, the noise in the real observational data also plays an important role: noise in the training data acts as a regularizer and reduces over-fitting Bishop (1995), allowing for a faster training. Model based networks that do not use real observations but synthetic data instead, such as The Payne (Ting et al. 2019) using noise free spectra and StarNet

(Fabbro et al. 2018) with added Gaussian noise, are usually not representative of the inherent correlated noise of real spectra. Interstellar extinction, atmospheric extinction, and instrumental signatures are not simulated in the synthetic spectra and can lead to a significant synthetic gap. Furthermore, these synthetic data are also normally homogeneous in terms of labels, which is also not a true representation of the observations. The data-driven CNN employed in our study is able to deal with the real noise efficiently. The noise in the data lead to a more efficient regularization and reduced generalization errors.

## Appendix B: Performance at different S/N

Here, we investigate the robustness of CNN predictions for different S/N regimes. As illustrated in Fig. B.1, the mean bias and mean scatter between CNN and GES predictions remains constant across the different S/N bins for our four labels, in the training sample. For the observed sample, even though there are fewer stars in the higher S/N bins, CNN performances are similar compared to the training sample for the atmospheric parameters. In the bins S/N $\leq$ 30, the values are slightly higher for the observed sample, but are expected considering the level of noise and spectral resolution. The bias and $\sigma$ for Lithium show an increasing trend for S/N > 50 in the observed sample due to low statistics for these bins. Also, the iDR6 A(Li) values for these stars mostly lie at the edges of our training A(Li) range, namely, above 3.0 dex or below 1.0 dex. We conclude that is robust to the noise in both the training and observed samples.

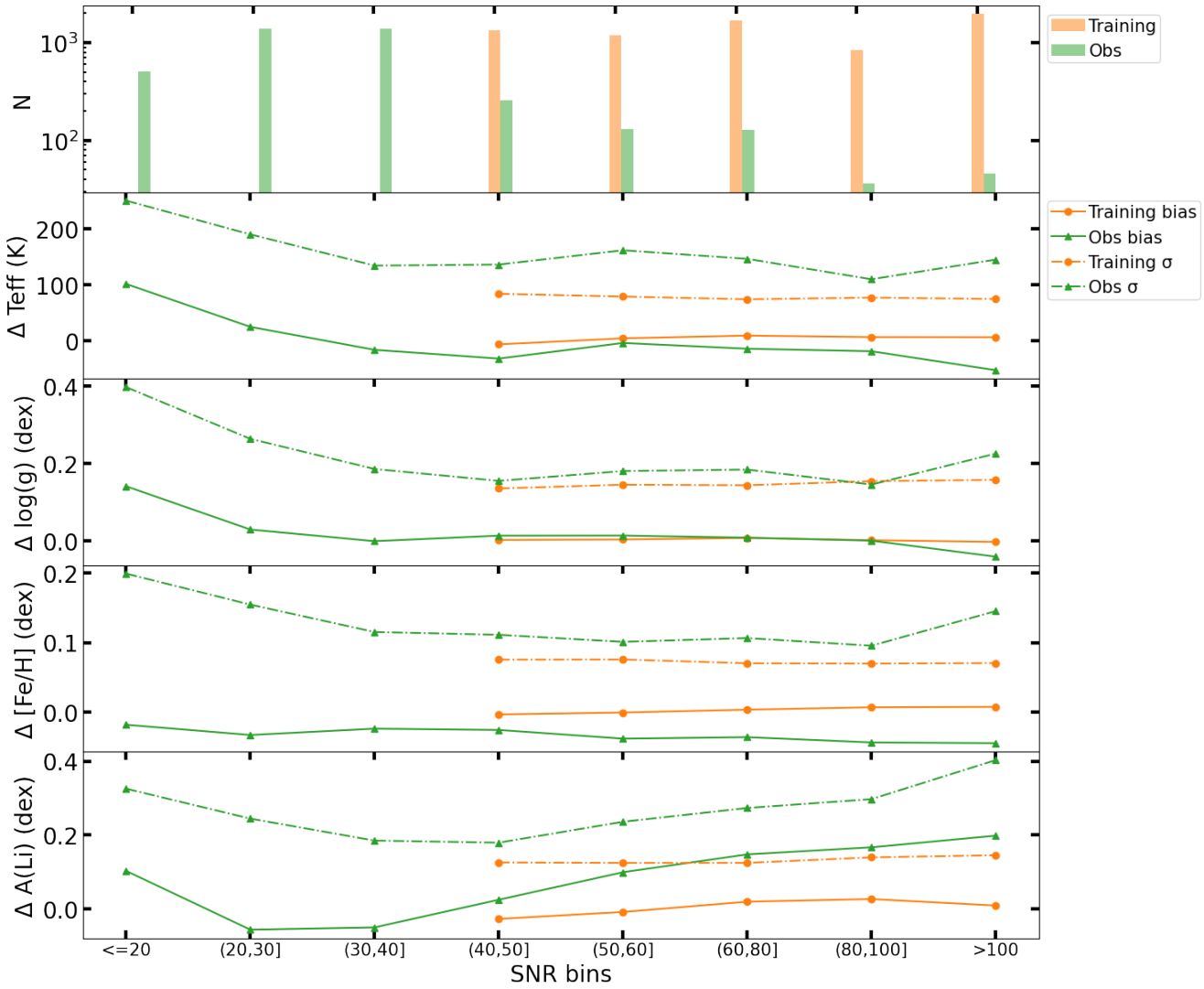


**Fig. B.1.** Histogram showing the number of spectra in each S/N bin for the training (yellow, 7031 stars) and observed (green, 5096 stars) samples (top panel). Bias (Bias = mean (CNN-iDR6), solid) and dispersion ($\sigma$ = std(CNN-iDR6), dash-dot) as a function of S/N (bottom 4 panels). The observed sample is selected within training label limits, eVRAD < 1.0 km/s, with no GES flags and with UPPER_COMBINED_LI1 = 0.0. For the observed sample, we have very few stars in the two highest S/N bins in comparison to the lower S/N bins.