The Gaia-ESO Survey: Preparing the ground for 4MOST & WEAVE galactic surveys - Chemical Evolution of Lithium with Machine–Learning[⋆]

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

[1] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
e-mail: snepal@aip.de, gguiglion@aip.de
[2] Institut für Physik und Astronomie, Universität Potsdam, Karl-Liebknecht-Str. 24/25, 14476 Potsdam, Germany
[3] Institute of Theoretical Physics and Astronomy, Vilnius University, Sauletekio av. 3, 10257 Vilnius, Lithuania
[4] INAF, Osservatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, 50125 Firenze, Italy
[5] Keele University, Keele, Staffs ST5 5BG, UK
[6] Lund Observatory, Department of Astronomy and Theoretical Physics, Box 43, SE-22100 Lund, Sweden
[7] INAF, Osservatorio di Astrofisica e Scienza dello Spazio, Via Gobetti 93/3, 40129 Bologna, Italy
[8] Nicolaus Copernicus Astronomical Center, Polish Academy of Sciences, ul. Bartycka 18, 00-716, Warsaw, Poland
[9] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom
[10] Spanish Virtual Observatory, Centro de Astrobiología (INTA-CSIC), 28691 Villanueva de la Cañada, Madrid, Spain
[11] Max Planck Institute for Astronomy, Könnigstuhl 17, 69117, Heidelberg, Germany
[12] Niels Bohr International Academy, Niels Bohr Institute, University of Copenhagen Blegdamsvej 17, DK-2100 Copenhagen, Denmark
[13] Núcleo de Astronomía, Facultad de Ingeniería y Ciencias, Universidad Diego Portales (UDP), Santiago de Chile
[14] INAF, Osservatorio Astronomico di Padova, vicolo dell'Osservatorio, 5 - 35122 PADOVA, Italy

Received ... / Accepted ...

ABSTRACT

Context. Originating from several sources (Big Bang, stars, cosmic rays) and being strongly depleted during stellar lifetime, the lithium element is of great interest as its chemical evolution in the Milky Way is not yet well understood. 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 industrial parametrization, 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 grounds for future spectroscopic surveys like 4MOST and WEAVE, we aim to apply machine–learning techniques for lithium study/measurement.

Methods. We train 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 $\sim 40\,000$ stars.

Results. We show that the CNN properly learns the physics of the stellar labels, from relevant spectral features, over a large 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 like lithium-rich giants are found in our sample. Such performances are 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 aiming 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 the machine–learning applications should be properly investigated along with realistic label uncertainties and upper limits for abundances.

Key words. techniques: spectroscopic – methods: data analysis – Surveys – Catalogs – Stars: fundamental parameters – Stars: abundances – Galaxy: stellar content – Galaxy: evolution

[⋆] Based on observations made with the ESO/VLT, at Paranal Observatory, under program 188.B-3002 (The Gaia-ESO Public Spectroscopic Survey, PIs G. Gilmore and S. Randich). Also based on observations under programs 171.0237 and 073.0234

1. Introduction

The element lithium[1] is of great interest in Astrophysics due to 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). The attempts of 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, independent of metallicity and effective temperature, called the "Spite plateau" with A(Li) $\sim$ 2.2 dex (Spite & Spite 1982; Bonifacio & Molaro 1997). This factor of three difference between the theoretical prediction and observation presents the famous cosmological lithium problem (e.g. Fields 2011).

At later times Li is produced at two distinct sources; in the Inter Stellar Medium (ISM) via a spallative interaction of galactic cosmic rays and the ISM through the p + C,N,O or α+C,N,O reaction channels (Reeves et al. 1970) and in stellar sources like Asymptotic Giant Branch (AGB) stars (McKellar 1940), Red Giants (Sackmann & Boothroyd 1999), 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) where, first, ^{7}Be is formed in temperatures hotter than 4×10^7 K by the reaction $^3{\rm He} + \alpha \rightarrow {}^7{\rm Be} + \gamma$. 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). Li could also be produced via the $\nu-$process happening in the external shells of collapsing massive stars (Woosley & Weaver 1995; Kusakabe et al. 2019). Additionally, Li can also be easily destroyed in stars by the proton capture reaction $^7{\rm Li(p,}\alpha)^4$He at temperatures as low as 2.5×10^6 K already in the pre-main sequence (PMS) and in later stages, whenever that temperature is reached (Pinsonneault 1997). For example, the meteoritic A(Li) is $\sim$3.26 dex (Lodders & Palme 2009), which represents the initial ISM Li for the Sun while the Solar photospheric abundance of only A(Li) $\sim$ 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, one needs 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 became 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) homogeneously built from ESO high resolution spectra a Li catalog composed of 7300 stars, and studied the lithium evolution in the Milky Way. In very recent years, the number of stars with available Li abundances has rapidly increased thanks to large scale Milky Way spectroscopic surveys such as the 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) and have contributed to our understanding of the evolution of Li.

A 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 are detectable and measurable only 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 an Fe I line. It is thus challenging for classical spectroscopic pipelines to provide precise Li abundances at intermediate and low-resolution, or in the presence of noise.

1.1. The Machine–Learning approach

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), several 10^7 spectra will be gathered and supplemented by the wealth of astrometric and photometric data provided by the Gaia satellite (Gaia Collaboration et al. 2016, 2020). 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). They become very popular and have been successfully applied in many other astronomy fields, for instance, gravitational lensing (Petrillo et al. 2017), search for open clusters in Gaia data (Castro-Ginard et al. 2020), detection of gravitational waves (Lin & Wu 2021), photometric redshift prediction (Lima et al. 2022) and many others. Although with a very simple architecture compared to modern networks, NNs actually have been long used in astrophysical applications: 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.

ML approaches also started to play an important role in the derivation of stellar labels. Such methods transfer the

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

[2] $A(Li) = log(N_{Li}/N_H) + 12$

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

knowledge from a reference set of data, so-called *training sample* to a larger set of data, in order 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. The Payne (Ting et al. 2019) demonstrated that one can combine physical stellar models using Neural Networks as a function to generate spectra, instead 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 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 network 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. 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 spectra by training the CNN with various synthetic spectral grids while mitigating the “synthetic gap". Leung & Bovy (2019) developed the astroNN tool, able to handle missing labels, trained on observational data to derived 22 stellar parameters and chemical abundances based of APOGEE DR14 spectra and labels. Zhang et al. (2019) used StarNet to estimate 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, see for instance O’Briain et al. (2021); Čotar et al. (2021). These research efforts and the developments in future spectroscopic surveys, computational power and better ML techniques are the motivation to prepare 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: effective temperature T_{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 Ambrosch et al. (sub) that focuses on the chemical evolution of Al and Mg abundances with CNN from GES GIRAFFE HR10&21 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 goal is to prepare the ground for 4MOST and WEAVE Li analysis; 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 et al. 2013) data. GES gathered spectra for all major Galactic components (halo, bulge, thin and thick disks), including a large number of open and globular clusters, and calibration observations such as benchmark stars, radial velocity (V_{rad}) standards, asteroseismic CoRoT/K2 fields (see Bragaglia et al. 2022; Pancino et al. 2017a; Stonkutė et al. 2016; Valentini et al. 2016). For this study, we use the spectra and parameters+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 use the H665.0/HR15N setup that includes the Li doublet at 6 708 Å. The HR15N setup is centered 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, has not been used in this work.

The spectroscopic analysis within GES was performed by multiple nodes which use different spectroscopic tools, but adopting the same line list and model atmospheres (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 homogenised 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 χ^2 fit to the observed spectra, and iii) multi-linear regression method that derives atmospheric parameters and abundances by projecting the observed spectrum into vector functions which are constructed as the best linear combination of synthetic spectra from a grid. GES-iDR6 atmospheric parameters T_{eff}, and $\log(g)$, as well as [Fe/H] abundance ratio were adopted for this project.

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 section 2.1 of Romano et al. 2021, and Franciosini et al., in prep.). For 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

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

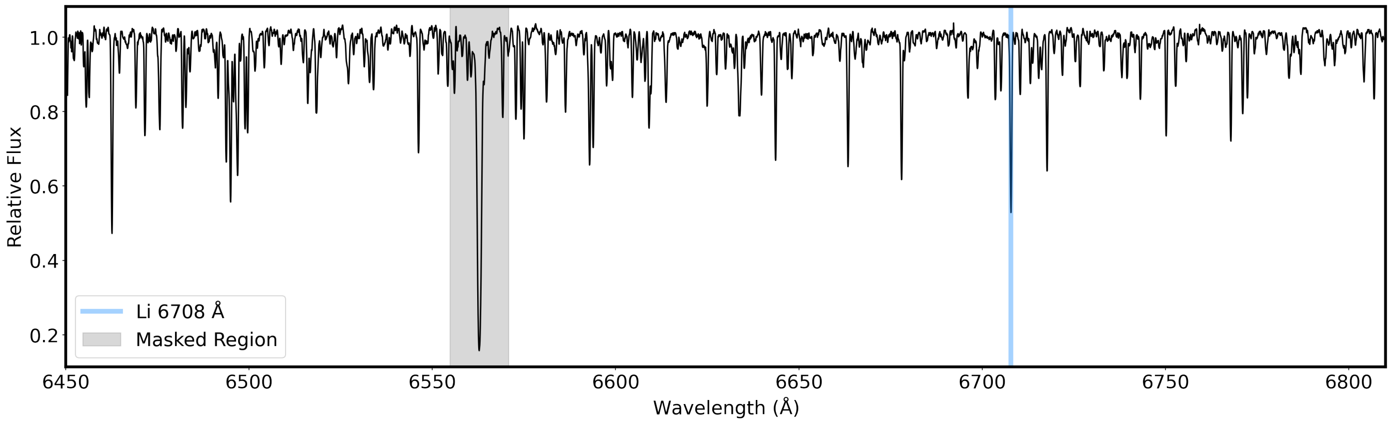

Fig. 1: An example GIRAFFE HR15N spectrum. This spectrum is of a star with labels: $T_{\rm eff}$ 4897 K, log(g) 2.55 dex, [Fe/H] -0.11 dex and A(Li) 2.63 dex. Lithium spectral feature is shaded with blue while the gray shaded region centered at $\rm H_{\alpha}$ is masked and not used in the spectral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

Li abundances (UPPER_COMBINED_LI1, 0=detection, 1=upper limit); upper limit is provided when the 6707.8 Å Li line is undetected (too low S/N or too low lithium) (see Franciosini et al., in prep. for details).

2.1. Training and Observed Sample

To build the training set, we apply several selection criteria. Starting with the total of 41 710 HR15N spectra, we selected objects with signal-to-noise ratio (S/N) > 40 (see Sect. 4.2 below) and apply the following cuts for labels: $4\,000 < T_{\rm eff} < 7\,000\,{\rm K}$, $1.0 < \log(g) < 5.0\,{\rm dex}$, $-2.0 < {\rm [Fe/H]} < 0.5\,{\rm dex}$ and $0 < {\rm A(Li)} < 4.0\,{\rm dex}$. We further cleaned the training set by applying uncertainty cuts of ${\rm e}\,T_{\rm eff} < 100\,{\rm K}$, ${\rm elog}(g) < 0.3\,{\rm dex}$, ${\rm e[Fe/H]} < 0.2\,{\rm dex}$ and ${\rm eA(Li)} < 0.5\,{\rm dex}$. We rejected stars with Li upper limits. We also apply an uncertainty cut on the radial velocity E_VRAD $< 0.5\,{\rm km\,s^{-1}}$ (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. GES provides same homogenized labels for these multiple observations; these objects were subsequently removed from the training. The training set is then composed of 7 031 spectra. The remaining 33 119 spectra, not included in the training set, comprise the observed sample. We do not provide labels for 1560 spectra due to missing $\rm V_{rad}$ or very high $\rm V_{rad}$ which shifts a spectrum out of the desired wavelength range after correction.

Next we apply radial velocity correction to the GES normalized spectra and remove 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 value is 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 which have strong $\rm H_{\alpha}$ emission lines. As dealing with this is out of the scope of the current work, we mask the region of 16 Å around $\rm H_{\alpha}$.

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 Sect. 3.1.3). GES provides repeat observations, hence some stars have multiple spectra available with varying S/N values. These repeat 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 adopt 25% of the total training data as the *test set* (leading to 1 758 spectra). The test set is not directly used for training of the CNN model but is only used to monitor the performance of the trained models. The *train set* is then composed of 5 273 spectra. Train and test samples are uniformly distributed across the label range, as homogeneity is crucial to help the CNN generalizing instead of over/under-fitting. See Sect. 2.3 for further discussion on homogeneity.

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

2.3. t-SNE for homogeneity check and outlier detections

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

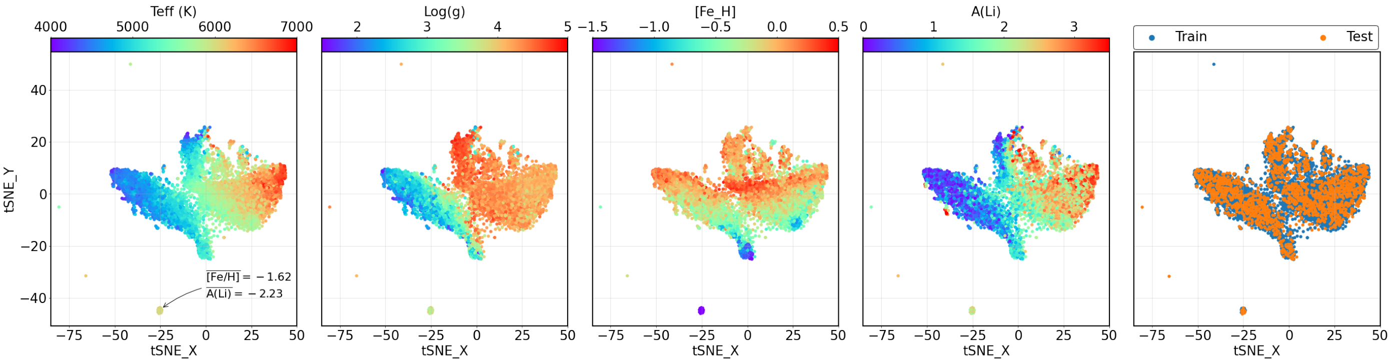

Fig. 2: 2D projection of t-SNE output for the 7 031 spectra of the training sample, colored by the labels $\mathrm{T_{eff}}$, $\log(g)$, [Fe/H] and A(Li) respectively. The right-most plot shows the t-SNE as the train and test samples to highlight their similar distribution across the label range. In the left subplot we show the mean [Fe/H] and A(Li) for the highlighted island which consists of Spite plateau-like stars in the globular cluster NGC 6752.

We plot the t-SNE maps (perplexity = 50.0) for the whole training data (7 031 spectra) 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 samples 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 to have low S/N and are affected by bad cosmic ray removal. The island at tSNE_X = −25 and tSNE_Y = −45, consist of Spite plateau-like stars ($\overline{[\mathrm{Fe/H}]} = -1.62$, $\overline{\mathrm{A(Li)}} = -2.23$) 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 correlate. 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 build our CNN model with the open source Deep Learning library Keras (Chollet et al. 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. TENSORFLOW, developed by Google Brain Team, is an opensource software library for ML.

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. Implementation of various architectures for stellar spectra parametrization can be found in literature, we refer readers to the work referenced in Sect. 1 for details. For this project we built on work from 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_{eff}, $\log(g)$, [Fe/H] and A(Li). The model has 3 convolution layers and 4 (3 + 1) dense layers including the output layer (discussed below). 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.

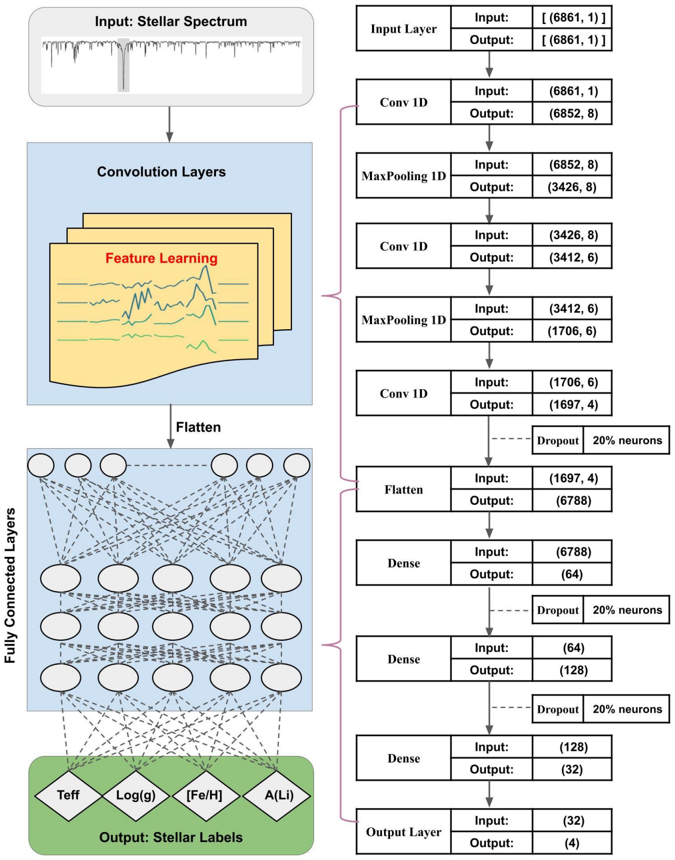

Fig. 3: Architecture of the CNN adopted for this study is shown as a block diagram on the left and its detailed structure with layers is shown on the right panel. The model can be divided into 4 distinct sections: Input Layer, Convolution Layers, Fully Connected Layers and Output Layer and has a total 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.

3.1.1. The Convolution and the Fully Connected Layers

Convolution Layers are the central part of the CNN class of Neural Networks as they are the 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. 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 previous layer, extract the features. During the learning process, these filter parameters are optimized. After extensive tests, we adopted the model with 3 Conv1D layers with 8, 6 and 4 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 Maxpooling which reduces the feature map size by half. This is very useful to reduce the overall training parameters, which also reduces training time, while 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 lies the fully connected layers (or dense layers) (Lecun et al. 2015). It is 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 4 feature maps from the last Conv1D layers are flattened before being fed to the 1st dense layer. The 1st dense layer has 64 neurons and receives input form the 6788 neurons of the flattened layer. The 2nd and 3rd dense layers have 128 and 32 neurons respectively. The output dense layer is naturally composed of 4 neurons corresponding to the four training labels. Our choice of the number of layers and neurons is based on many experimentation, with the goal of having a CNN complex enough, without mitigating the training performance.

3.1.2. The choice of hyperparameters

Hyperparameters are set at the beginning of the training and remain the same throughout the training, as opposed to the learn-able 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 "LeakyReLu" activation function and for the output layer we use "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 large number of training epochs until the training and test loss curves flatten out and stopped by using the EarlyStopping. See Fig. 4.
4. **Batch size:** It is 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 of batch size also provides regularization and stability during the training. We adopt a batch size of 64 as balance between good approximation of the training set and faster training time.
5. **Learning rate:** The learning rate (η) 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 η, and found that the best performances, for our model, are achieved for $\eta = 0.0001$.

3.1.3. Model Generalization: Avoiding Over/Under-fitting

Generalization and proper convergence of the model during the training is important to avoid over/under-fitting and to ensure that the training progresses smoothly. Our choice of convolution and dense layers ensures that the model does not under-fit the training data, hence attention is needed to avoid over-fitting. For this we employ *regularization*, *dropouts* and *early-stopping* procedures detailed below.

In each of the three convolution layers the L2 Regularization function is applied, allowing to penalise 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 makes sure that less significant features in the spectrum do not significantly affect the label prediction.

We apply Dropout layer on the inputs of the 3 inner dense layers. At each training epoch (explained below in Sect. 3.1.2), a certain number of neurons are randomly selected and their contribution to the activation of neurons in subsequent layers is temporally removed. This forces 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. As shown in Fig. 3, 20% of the neurons are dropped before the dense layers.

While training the CNN model, it is recommended to stop the training once the validation performance starts to degrade. For this task, we employ a callback called EarlyStopping in the model. This callback monitors the validation/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 in preventing over-fitting and allow a faster training. Model based networks that do not use real observations but synthetic data, 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. The data-driven CNN employed in our study is efficiently able to deal with the real noise. The noise in the data lead to a more efficient regularization, and reduced generalization errors.

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 metric on the test data. We use "Mean Squared Error (MSE)" as loss function as well as the metric. The EarlyStopping callback, defined in Sect. 3.1.2, monitors the metric and the best model weights are saved. We trained an ensemble of 30 models[5][6], 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) samples for the 30 models. The loss curves shows 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 Sec. 4.3 for more on uncertainties.)

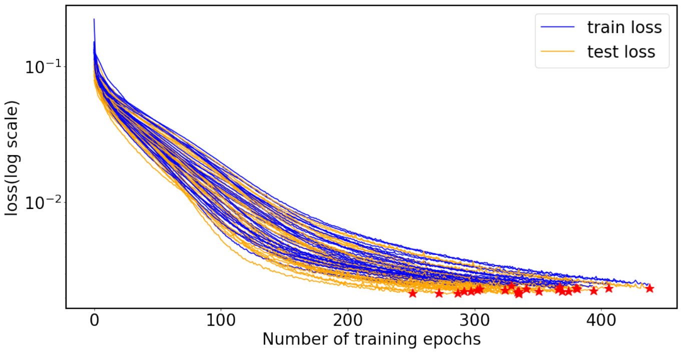

Fig. 4: Value of the loss functions for the train (blue) and test (orange) samples for the 30 CNN runs as a function of the epoch. The red stars identify the selected 24 models.

3.2.1. Result of the Training

In Fig. 5, we show comparison of the input GES-iDR6 labels to the CNN prediction for the train and test samples. 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 samples, 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 sample follows the train sample, showing that the trained models do not over-fit. Even though the wavelength range in the GIRAFFE HR15N setup is not optimal for atmospheric parameters determination (Lanzafame et al. 2015), and despite masking H_α 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 ($T_{\rm eff}$ v.s. $\log(g)$) for the train (top panels) and test (bottom panels) samples. The left columns show the input iDR6 labels and 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 samples. The dwarfs which span a large $T_{\rm eff}$ range from 7000 K to 4000 K are well parametrized even for the very hot and the very cool regime. The metal-poor giants, around 5 000 K, show much less scatter for 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 i.e., 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 and A(Li)∼2.2 is well located for both train and test samples. We also find good agreements for globular clusters like NGC 1281 and NGC 2808, seen around -1.5<[Fe/H]<-1.0 and A(Li)∼1.2. 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 samples. 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. Does the CNN learn from spectral features?

Considering our neural network as a mathematical function which 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 sensi-

[5] The training of the models required a time period of 16 to 26 minutes using only CPU on the COLAB cloud service at AIP for compute and storage.

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

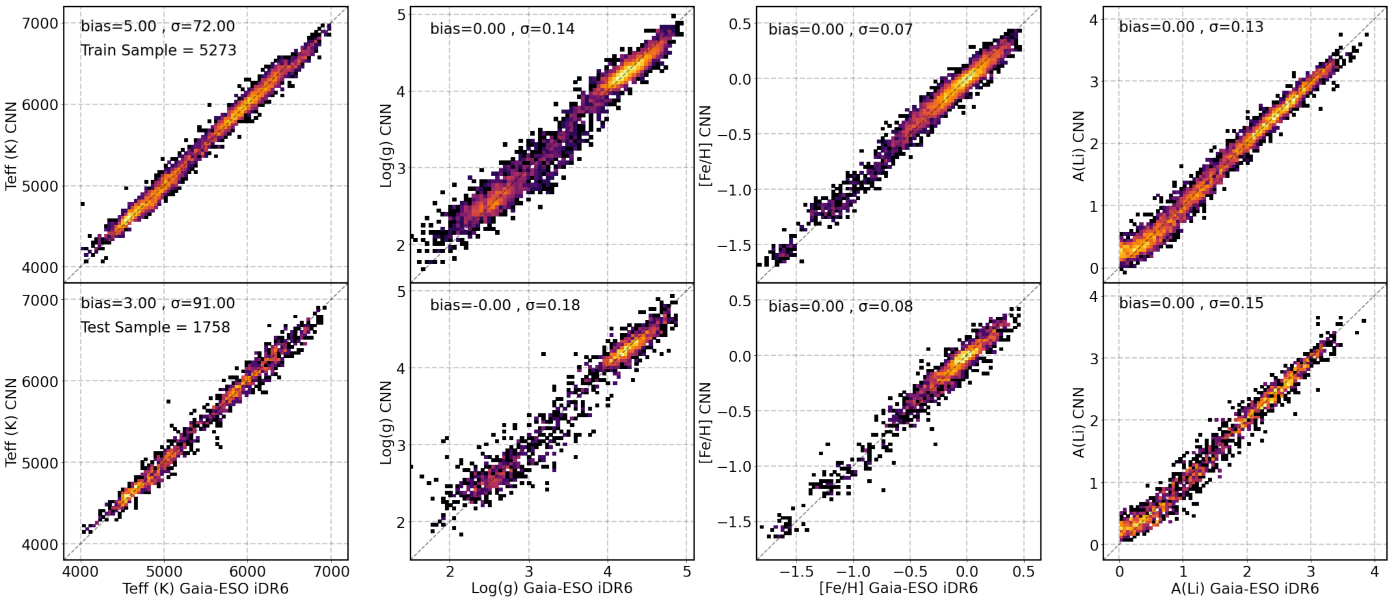

Fig. 5: 2d histograms showing 1-to-1 comparison between the GES-iDR6 labels (CNN input, x-axis) and CNN predictions (y-axis) for the train (top row) and test (bottom row) samples. The bias=mean(CNN-iDR6) and σ=std(CNN-iDR6) are also calculated.

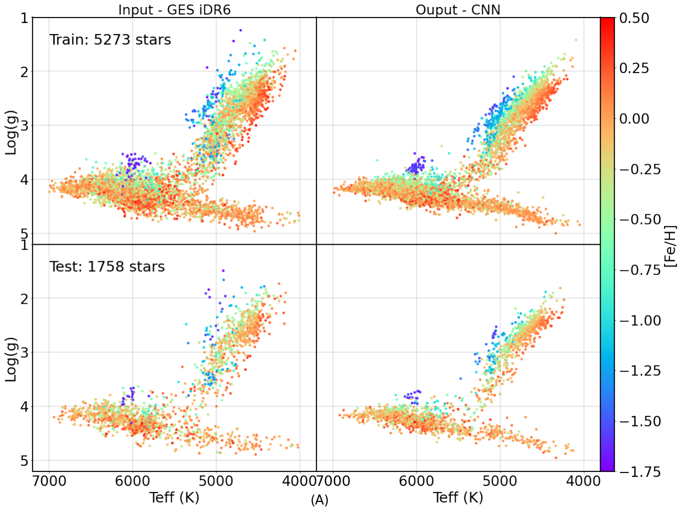

Fig. 6: Kiel Diagrams for the input and CNN output colored by [Fe/H]: Top two panels show the train sample stars using iDR6 input labels on the left and CNN output on the right. Bottom two panels show the same for the test sample.

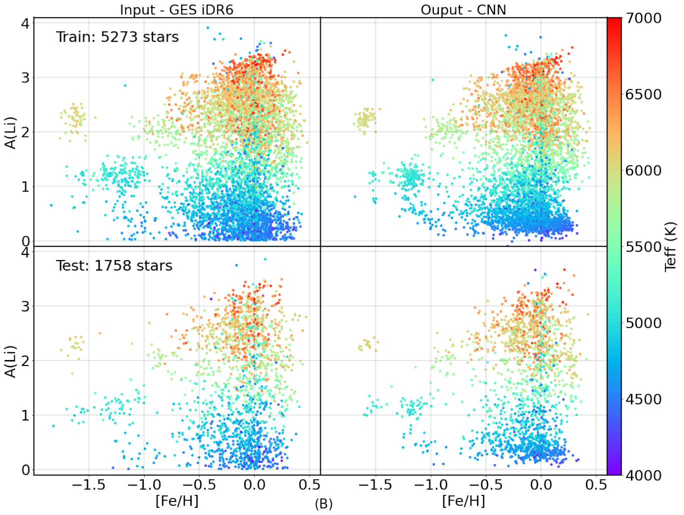

Fig. 7: [Fe/H] vs A(Li) for the iDR6 input and CNN output colored by T_{eff}: Top two panels show the train sample stars using iDR6 input labels on the left and CNN output on the right. Bottom two panels show the same for the test sample.

tivity of output labels to each of the input fluxes we can understand if the CNN is learning from the spectral features. This is easily achieved by calculating partial derivatives of each of T_{eff}, $\log(g)$, [Fe/H] and A(Li) with respect to every input neuron (or wavelength), i.e., $\partial Label/\partial\lambda$. The gradient of an output label is a sort of back propagation of the model through the CNN. In Fig. 8, we show the gradients of $\log(g)$ and A(Li) for the 13 solar twins in our training sample. We can make several observations:

1. The gradient of the lithium label with respect to λ is only active at the 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.
2. Damiani et al. (2014) showed that the quintet feature, between 6490-6500 Å consisting of blended FeI, CaI, BaII and TiI lines, is highly sensitive to gravity. The TiII 6491.56 Å line, on bluer side of the quintet, was also considered as an important line for their spectral indices. Here, the CNN gradients $\partial\, log(g)/\partial\,\lambda$ shows that these wavelength regions are indeed very sensitive to $\log(g)$.
3. Jofré et al. (2015) listed the ionised 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.

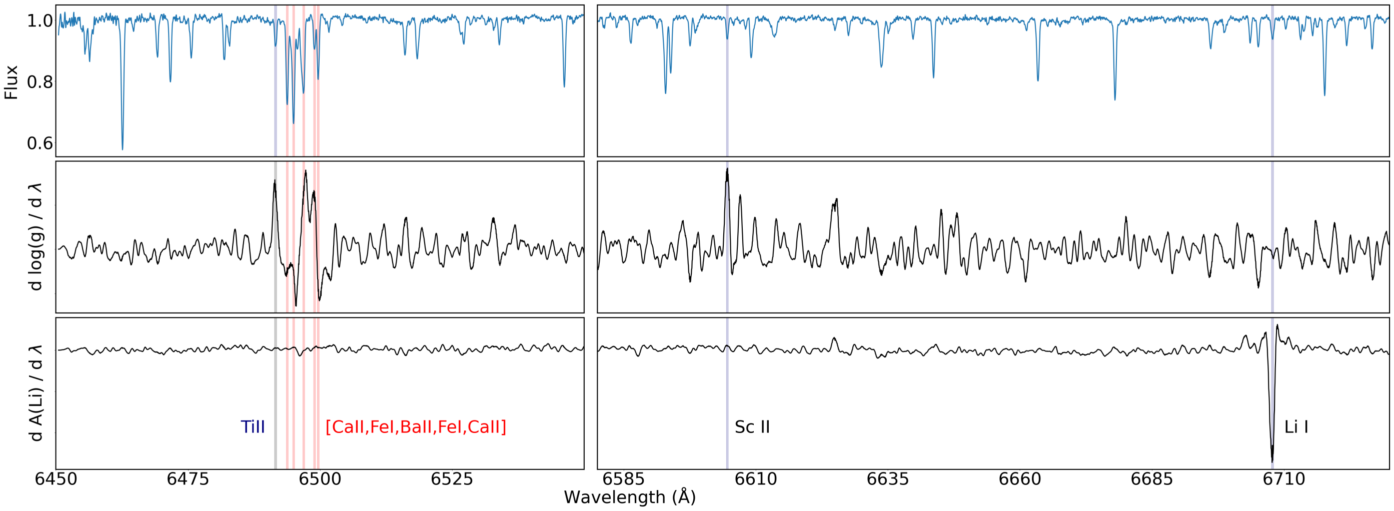

Fig. 8: Gradients of the output labels with respect to input pixels for the solar twins in the training sample. Selected as $T_{\rm eff} = 5\,777 \pm 25$ K, $\log(g) = 4.44{\pm}0.10$ dex and [Fe/H] $= 0.0{\pm}0.05$ dex, there are 13 stars. Top row shows mean input spectrum and the second and third row 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 $\rm H_\alpha$ region. The various spectral features that are discussed in the text are labelled.

Such diagnostic checks confirmed that CNN properly learns from spectral features, and these gradients could allow to identify 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

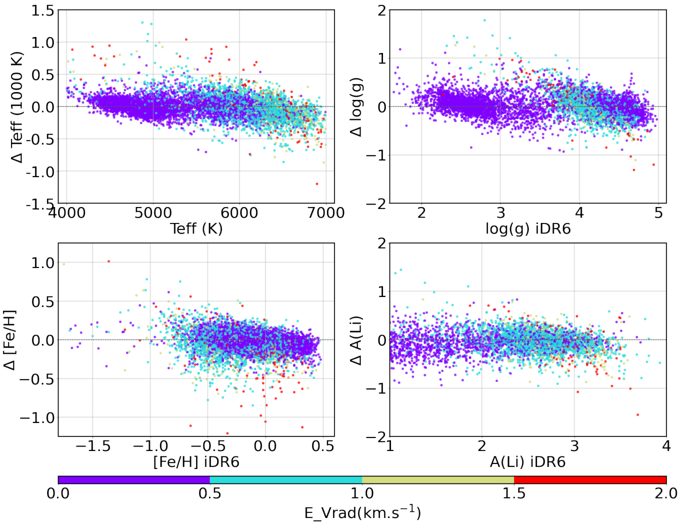

Fig. 9: Scatter plots showing residuals as a function labels for the selected observed sample stars color-coded by the reported uncertainties in radial velocities. The trends in residuals show the sensitivity of CNN to the uncertainties in radial velocity.

Accurate and precise radial velocities are crucial for 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 (Gilmore et al. 2022). The radial velocity is measured using the HR15N spectra, but an offset is applied to it during 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 have a good agreement with Gaia radial velocity standards. Yet, such a combination of different setups can be a source of small systematics. While GES reports highest Vrad precision achieved to be of 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 $\rm V_{rad}$ uncertainties and a very clear trend is seen for E_VRAD > 0.5 km s^{-1}. 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 $\rm V_{rad}$ precision for GIRAFFE spectra worsens for $T_{\rm eff} > 5\,200$ 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 5 500 K in iDR6. The HR10 re-calibration is a function of $T_{\rm eff}$, $\log(g)$, [Fe/H], and could create tiny $\rm V_{rad}$ corrections that the CNN is able to detect. We avoid 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 like 4MOST and WEAVE, which will observe in multiple setups, 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 $\rm V_{rad}$ errors for GES could be the fact that the different wavelength ranges were not observed at the same time and were calibrated independently (Randich et al. 2022). The expected accuracy of 4MIDABLE-HR radial velocities is expected 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 Vrad.

3.2.4. Can CNN infer lithium abundances without lithium line?

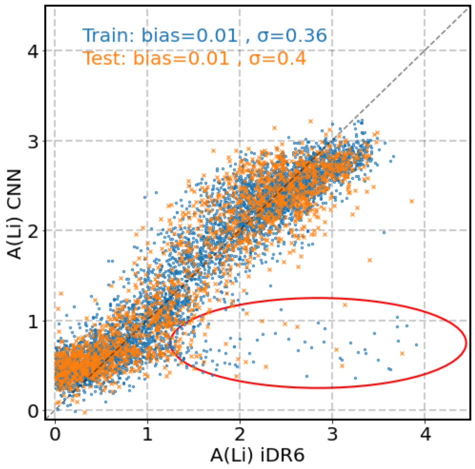

Fig. 10: CNN vs GES-iDR6 A(Li) for the CNN trained using spectra masked at 6707.8 Å Li line. Blue and orange represent train and test samples respectively. The bias=mean(CNN-iDR6) and σ=std(CNN-iDR6) are also calculated. The dashed line represents the 1-to-1 line. The red ellipse shows the incorrectly inferred Li-rich giants.

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 of the surface gravity, see for instance Fig. 2. To test whether one can 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 to find very poor performance compared to Fig. 5, with a large scatter for both the training and the test samples. Li rich giants (see Sect. 6.2) are completely missed when inferring lithium only from astrophysical correlations. Li must be then measured from spectral features, and not inferred based on correlations.

4. Catalog of Stellar Parameters & Li Abundance

4.1. CNN parametrization of 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 seconds for the 4 labels, $T_{\rm eff}$, $\log(g)$, [Fe/H] and A(Li), of all 33 119 observed sample spectra. The prediction for the selected 24 models then takes only ∼9 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. 11 (a) with clear distinction between the main sequence and the giants, along with the metallicity gradient for the giants as well as turn-off stars. At the cool end, we see few stars with $\log(g)$ ∼4.0: we checked the spectra for these stars and found the presence of emission lines. An example of a HR15N spectrum with nebular 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, and have emission lines and TiO molecular bands (M dwarfs). As there were no cool M dwarfs ($T_{\rm eff} <$ 3500 K) in the training set, CNN will provide un-reliable parameters for these stars. GES is still refining the flags and thus further exploration of the particular flags is out of the scope of this project. In the third column Kiel diagram, the observed sample with radial velocity uncertainties >0.5 km s^{-1} are presented. Most of these stars lie in the warm dwarf region as uncertainties in VRAD increases with $T_{\rm eff}$, as discussed in 3.2.3. The metallicity gradient is also seen for these warm dwarf stars.

In Fig. 11 d-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 $T_{\rm eff}$ > 5 500 K (see Sect. 3.2.3). Some of these warm lithium rich stars probably 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, E_VRAD<1.0 km s^{-1} and no TECH and PECULI flags. In the first row, we show 4 481 observed sample stars with iDR6 Li abundance with the flag UPPER_COMBINED_LI1 = 0.0. The second row shows comparison for 3 099 stars, with Li upper limits given by UPPER_COMBINED_LI1 = 1.0. GES provides an upper limit 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 while the most of the Li upper limits have higher S/N; this is because stars with higher S/N and Li measurement, i.e. not a limit, are included in the training set. Also, most of the stars with lithium upper limit are giants having already evolved past their Li depletion phase (defined in Sec. 6.1).

Our catalog of atmospheric parameters ($T_{\rm eff}$, $\log(g)$), [Fe/H], and lithium abundances for ∼ 40 000 stars is summarised in Table 1. The data table is available at: doi:// *to be added upon paper acceptance.*

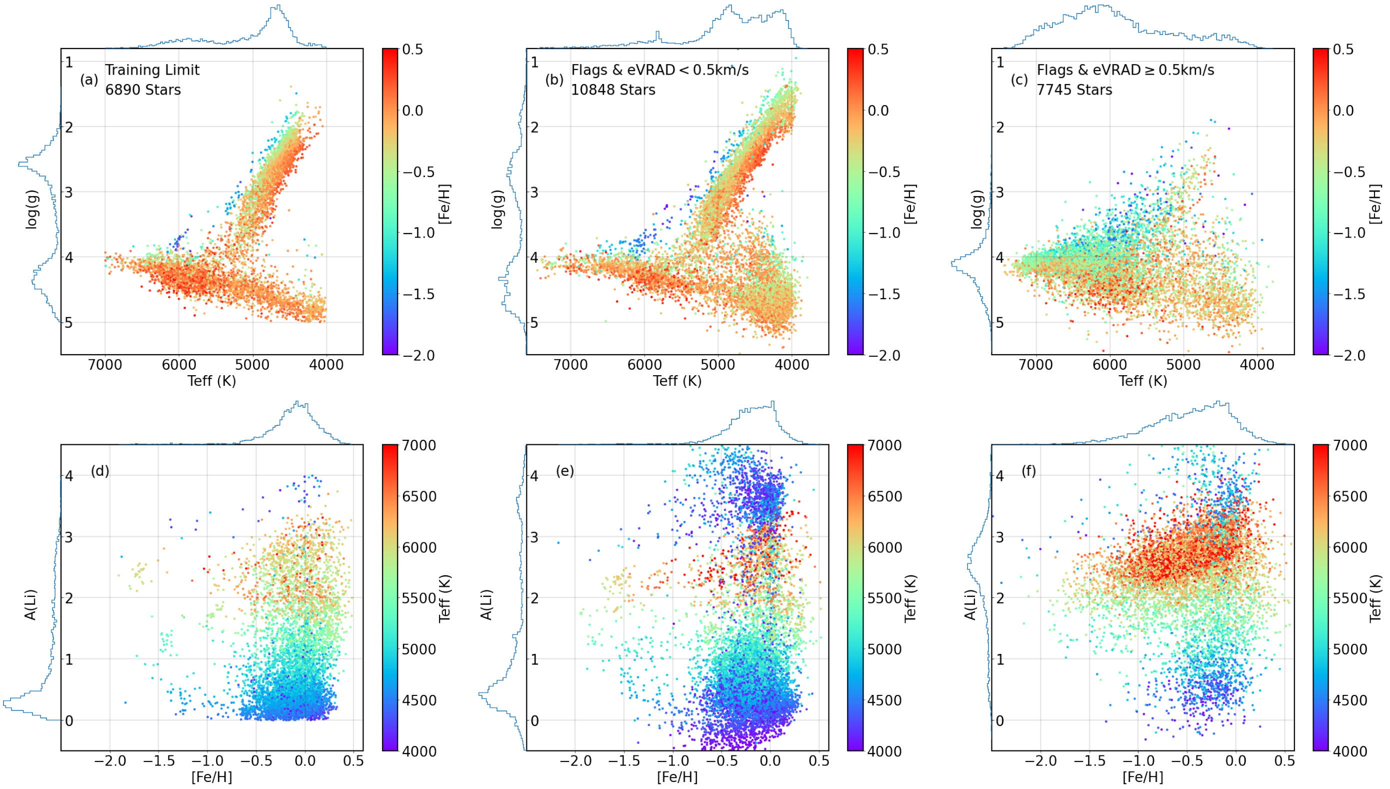

Fig. 11: Results for the Observed Sample. *Top Row:* **(a)** Kiel diagram for the observed sample stars with S/N>10 and labels within training limits color-coded with [Fe/H]. **(b)** Same plot as (a) but for stars with S/N>10, GES-iDR6 flags and E_VRAD $< 0.5\,\mathrm{km\,s^{-1}}$. **(c)** Same selection as (b) but for E_VRAD $\geq 0.5\ \mathrm{km\,s^{-1}}$. 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_{eff} for the same stars as the Kiel diagram on top.

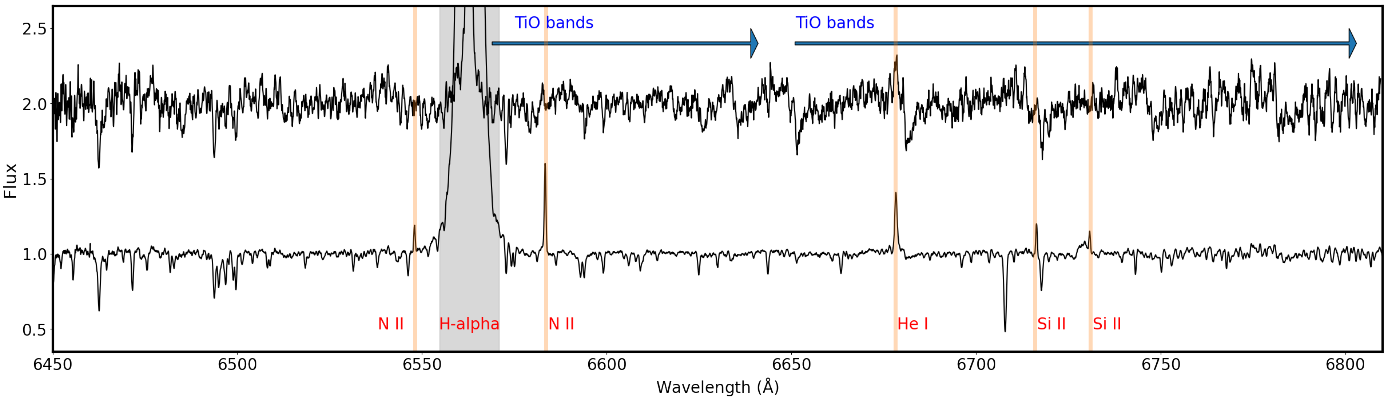

Fig. 12: HR15N spectra with nebular emission lines highlighted in yellow. From left to right the lines are; 6548 Å NII, 6563 Å H_{α}, 6583 Å NII, 6678 Å HeI, 6716 Å SiII, 6731 Å SiII. For the upper spectrum, the region for the strong molecular bands of TiO starting at 6569 Å and 6651 Å are seen. The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

4.2. Effects of noise and rotation on CNN predictions

The CNN was trained with spectra with $\mathrm{S/N} > 40$ per pixel as it provided a balance in the training sample size and good quality. Noise is an unavoidable aspect of observational data (see Sect. 3.1.3 above). 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 do not show any significant bias as a function of S/N.

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 increased blending conserving the EW. Classical spectroscopic pipelines must take into account rotational broadening during analysis of a spectrum.

Our training sample of 7 031 spectra has a distribution of rotational velocities (in $\mathrm{km\,s^{-1}}$) as follows: $v\sin i \leq 10 = 62\%$, $10 < v\sin i \leq 30 = 34\%$, $30 < v\sin i \leq 50 = 3\%$ and $v\sin i > 50 = 1\%$. Considering stars with $v\sin i$

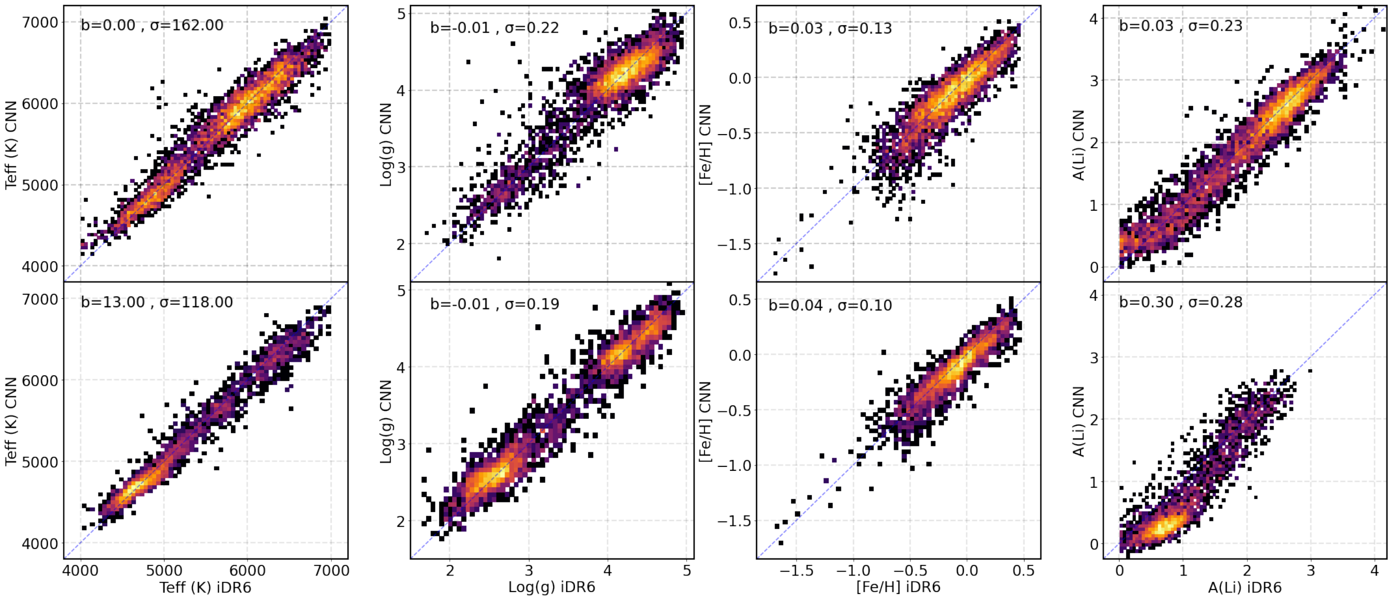

Fig. 13: One-to-One comparison for observed sample stars with, S/N > 20, eVrad less than 1 $\mathrm{km\,s^{-1}}$, no PECULI and TECH flags and within training label range. Here bias=mean(CNN-iDR6) and σ=std(CNN-iDR6). Top row: stars with Li measurement, bottom row: stars with Li upper limit. Most of 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
3	float	K	teff	Effective temp. ($T_{\rm eff}$)
4	float	K	eteff	Error of $T_{\rm eff}$
5	int	-	flag_teff	Boundary flag for $T_{\rm eff}$
6	float	cm $\mathrm{s^{-2}}$	logg	Surface gravity
7	float	cm $\mathrm{s^{-2}}$	elogg	Error on log(g)
8	int	-	flag_logg	Boundary flag for log(g)
12	float	dex	feh	[Fe/H] ratio
13	float	dex	efeh	Error on [Fe/H]
14	int	-	flag_feh	Boundary flag for [Fe/H]
15	float	dex	li	Li abundance
16	float	dex	eli	Error on Li
17	int	-	flag_li	Boundary flag for Li

> 10 $\mathrm{km\,s^{-1}}$ as 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 K, 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 see an increase in dispersion, to 159 K and 0.22 dex, for $T_{\rm eff}$ and A(Li). Also for [Fe/H], for $v\sin i > 70\mathrm{km\,s^{-1}}$, we see a trend of underprediction by CNN. We conclude that CNN do not suffer from significant systematics due to rotational broadening, and allows 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 per pixel. 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 samples show similar uncertainties, if the observed sample is restrained to the training set limits. The uncertainties are very low with medians of about 19 K for $\sigma T_{\rm eff}$, 0.03 dex for σlog(g), 0.017 dex for σ[Fe/H] and 0.035 dex for σA(Li) for the train, test and observed samples (within the training set 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 samples follow each other well, meaning that the CNN models are able to generalize properly.

The CNN internal uncertainties could however be under-estimated. 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 and no GES flags, and GES lithium detection.

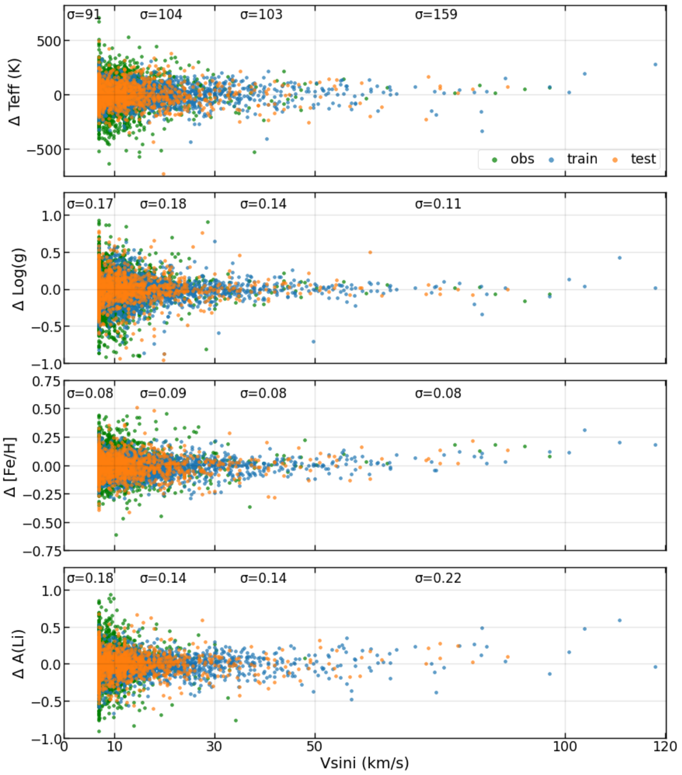

Fig. 14: Residuals (Δlabel = GES - CNN) as a function $v\sin i$ (km/s) for the train (blue), test (orange) and selected observed sample (green) stars. The observed sample is selected within training label limits, $S/N > 10$, E_VRAD $< 0.5\,\mathrm{km\,s^{-1}}$, with no GES flags and with Li measurement.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

The bias curves corresponds to the accuracy and the sigma curves correspond to the precision of CNN.

For T_{eff}, between 4 400 K < T_{eff} < 6 600 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 K for the train and test samples and within 120 K 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 $\log(g) < 2.0$ and $3.0 < \log(g) < 4.0$ which are less populated. For [Fe/H]<-1.0, with just 19 stars with available GES-iDR6 values in the observed sample, the bias and σ curves cannot be well interpreted. For [Fe/H]>-1.0, 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 where we have very few stars. The precision of the train and test samples are within 0.2 dex, while the observed sample is within 0.3 dex as ~90% of the stars have S/N<40. 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 stars (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 metallicity. Benchmark stars are commonly used as validators/calibrators by large spectroscopic surveys, such as GES (Pancino et al. 2017b). 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_{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_{eff}, and most of the stars are within the training set limits. We find that for most of the GBS, CNN results compare very well. The cool giants alf_Cet, gam_Sge and alf_Tau have T_{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, HD122563, HD140283 and HD84937, with [Fe/H] less than -2.0 dex. HD122563 is the most metal-poor star with [Fe/H] = -2.62 for which we see the highest differences in T_{eff}, $\log(g)$ and [Fe/H], although CNN estimate for A(Li) agrees with the AMBRE value. For HD140283, with [Fe/H] = -2.36, we see a difference of ~500 K for T_{eff} and 0.7 dex in [Fe/H] while the estimates for $\log(g)$ and A(Li) are in a good match. For HD84937, CNN predictions for T_{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 case of lithium, for most of the GBS stars, CNN predictions compare well with AMBRE abundances within $1-\sigma$. For the 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, as well as for stars which are within training limit and have A(Li)<1.5, 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 and with $T_{\mathrm{eff}} = 5\,777 \pm 150\,\mathrm{K}$, $\log(g)$

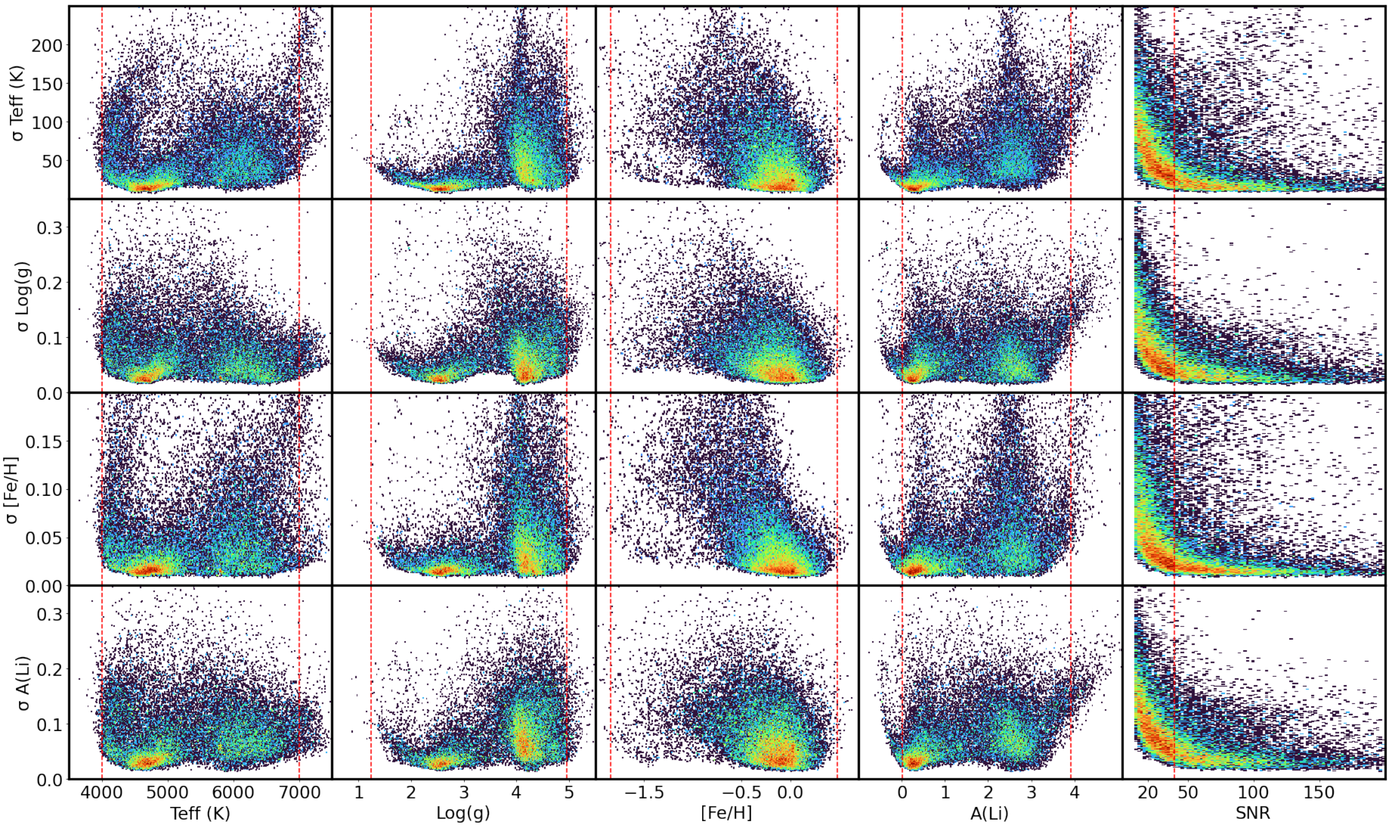

Fig. 15: 2D histograms showing CNN uncertainties (internal precision) as a function of 4 labels ($T_{\rm eff}$, log(g), [Fe/H], A(Li)) and S/N for the observed sample with S/N>10, i.e., 31,272 spectra. The red dashed line shows the limits of the training labels. The x-axis represents the labels, and the y-axis the uncertainty (σ).

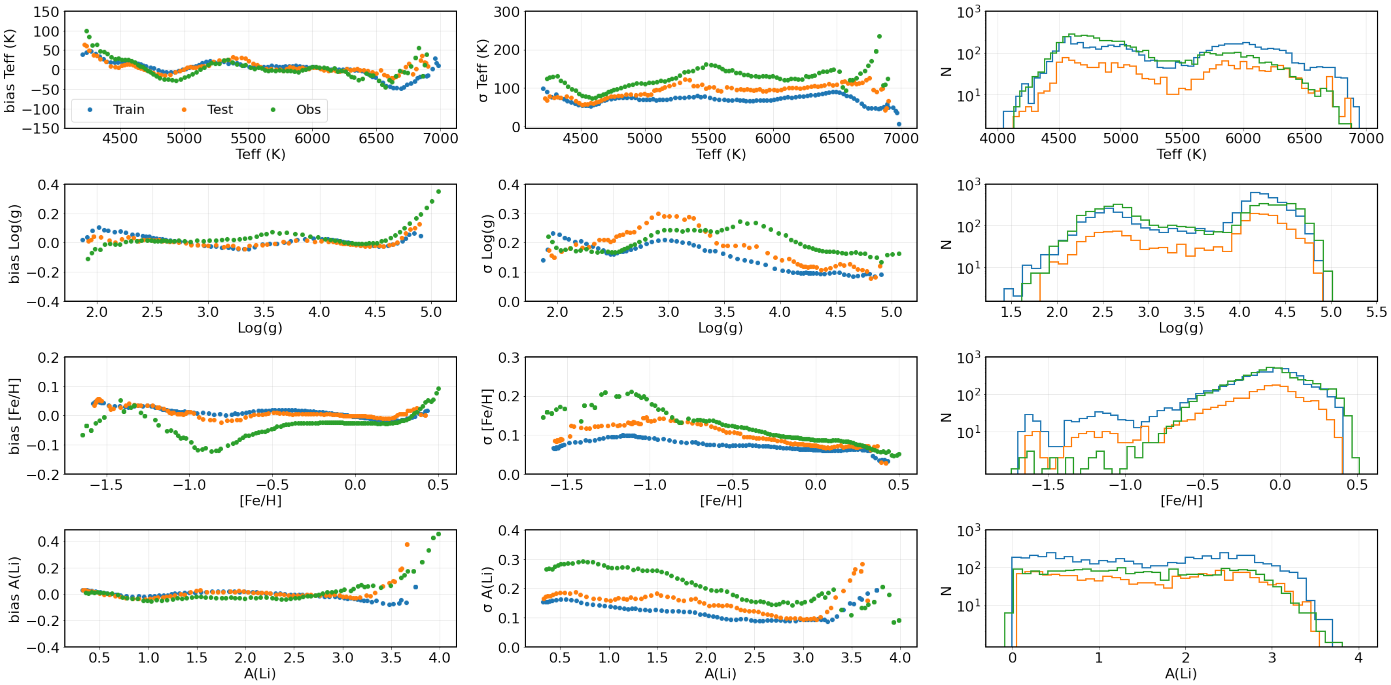

Fig. 16: Running mean bias and mean dispersion as a function of labels for the train (blue), test (orange), and observed (green) samples calculated in bin sizes: 250 K for $T_{\rm eff}$, 0.3 dex for log(g), [Fe/H] and A(Li). The curves are representative of the real accuracy and precision of our CNN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

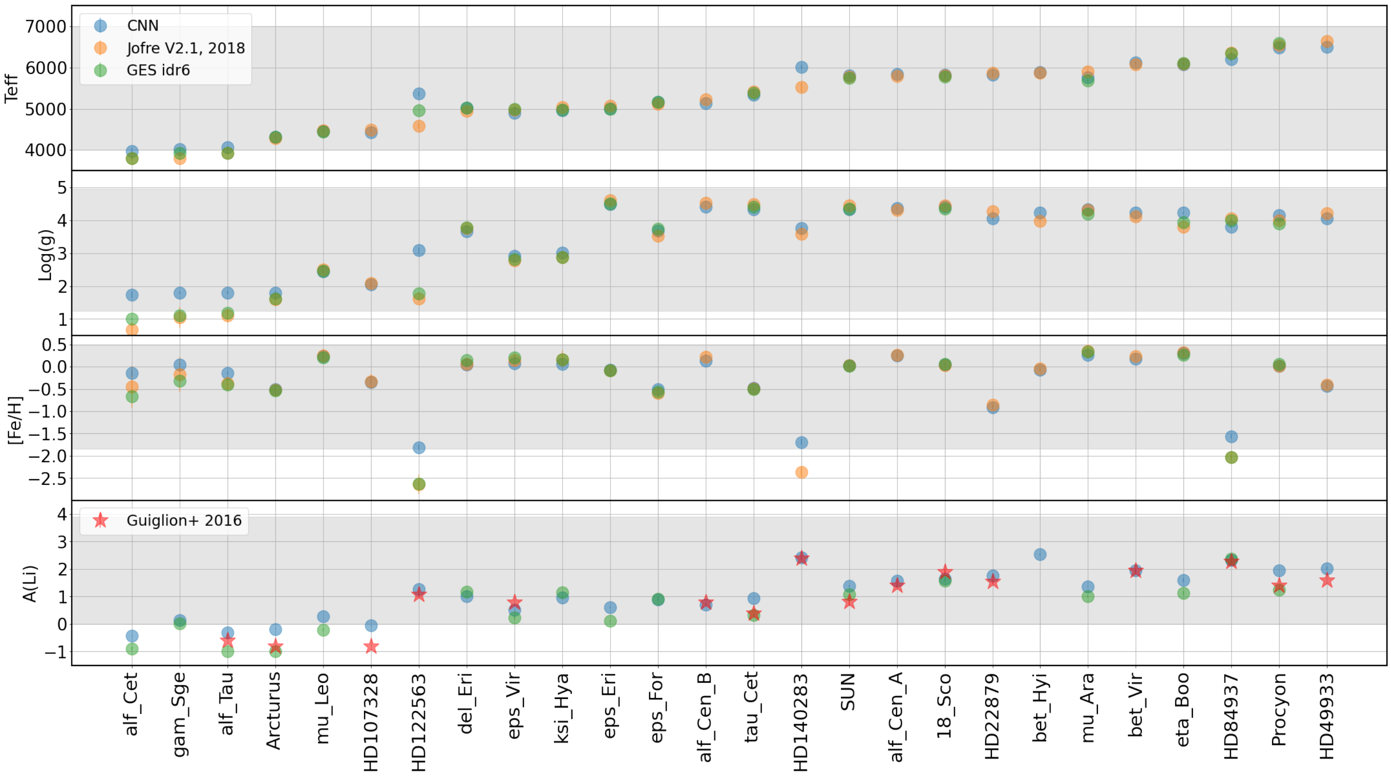

Fig. 17: Comparison of CNN prediction for the Gaia Benchmarks Stars (GBS). The reference T_{eff}, $\log(g)$ and [Fe/H] are from Jofré et al. (2018) and A(Li) from Guiglion et al. (2016). GES-iDR6 values are also shown for comparison. On x-axis we present the GBS names sorted by increasing T_{eff} and on y-axis we present the 4 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.

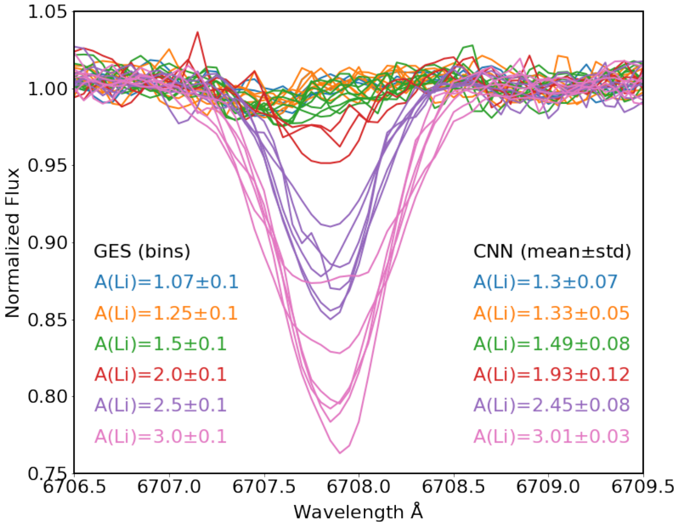

Fig. 18: Li feature for Solar Twins with varying Li abundance. The solar twins in the training sample are selected with S/N > 90 and with GES $T_{\text{eff}} = 5\,777 \pm 150\,\text{K}$, $\log(g) = 4.44{\pm}0.15$ and $[\text{Fe/H}] = 0.0{\pm}0.15$. The different colors represent the GES Li bins as listed on the left. On right we show the mean of CNN prediction for the shown spectra in each bins.

$= 4.44{\pm}0.15$ and $[\text{Fe/H}] = 0.0{\pm}0.15$. CNN provides robust measurements for A(Li)≥1.25. Below this limit, CNN suffers from a positive bias, *i.e.* 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 $\sim 1.5\%$ 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. This could explain the difference in CNN, iDR6 and AMBRE measurements for the lithium measured in some of the benchmark stars. We did the same exercise for a typical RC stars (around Solar [Fe/H]), and the line being deeper, the CNN performs with no significant bias up to Li=0. For 4MOST-LR/HR, it will be important to generalise 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∼28 000, compared to the GIRAFFE at R∼20 000, and in four non-contiguous spectral bands

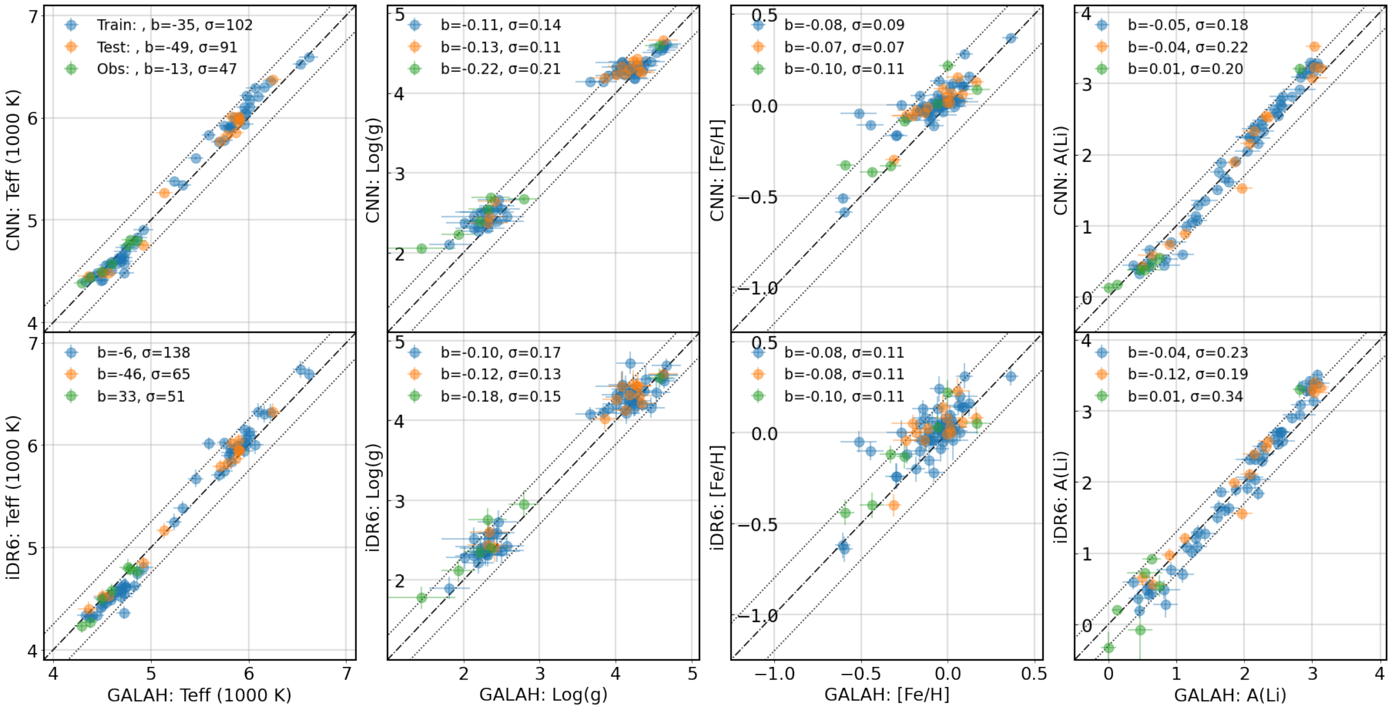

Fig. 19: Comparison of CNN results for stars in common with GALAH-DR3 Buder et al. (2021). GES-iDR6 sample has stars selected with S/N>30, within the training label limits, eVRAD< 0.5km s^{-1} and no GES flags and GALAH stars are selected with snr_c3_iraf > 30, flag_sp = 0, flag_fe_h = 0 and flag_Li_fe = 0. The dash-dot line is the 1-to-1 line and two dotted lines are at ± 250 K 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.

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, eVRAD < 0.5 km s^{-1} and no GES flags. For GALAH stars, we followed the GALAH recommended S/N and flags, i.e. snr_c3_iraf > 30, flag_sp = 0, flag_fe_h = 0 and flag_Li_fe = 0 (the flags = 0 represent no identified problems with determination of stellar parameters, iron and lithium abundances respectively). The CNN atmospheric parameters and lithium predictions agree very well with GALAH, 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). For the case of A(Li) < 1.0, the spread in 1-to-1 relation is less for the case of CNN vs 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 at A(Li) < 1.0. CNN works better at low lithium than standard pipelines, because it can 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 and higher CNN/iDR6 measurements for A(Li) > 2.5. 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

We aim here at comparing 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, eVRAD<0.5 km s^{-1}, no GES flags. Fig. 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 similar trend. The comparison shows presence of some outliers, below we discuss two of such outliers:

For the star CNAME=19264480+0032497, with $T_{\rm eff}$ = 4815 K and log(g) = 3.59 in iDR6, 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 km s^{-1}, which can be a cause of this difference. About 35% of our training sample have stars with $v$sin$i$ > 10 km s^{-1}, hence, the CNN can 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 and [Fe/H] = 0.01, 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 CNN, iDR6 and CoRoT-

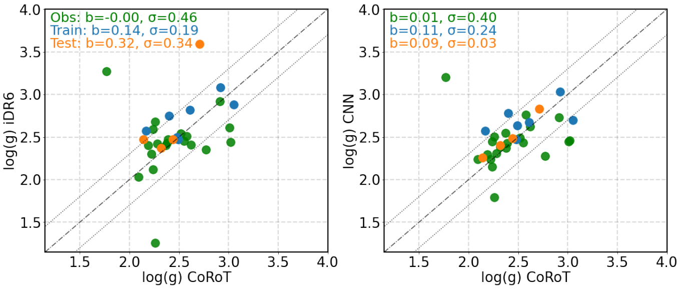

Fig. 20: Comparison with asteroseismic results. Left: CoRoT-GES vs GES-iDR6 labels, Right: CoRoT-GES vs CNN predictions. Blue, orange and green symbols represent the train, test, and observed sample selected within the training set limits (S/N>30, eVRAD<0.5 km s^{-1}, no GES flags) and with CoRoT-GES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

GES results agree. Both spectra looks 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 CNN predictions and Corot tells us that CNN is able to properly parametrize giants, considering the HR15N is not an optimal setup for precisely constraining log(g)s. We also showed that CNN can correct inaccurate labels that are miss-classified by standard pipelines.

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

6.1. Galactic Evolution of lithium

Recently, many studies challenged the possibility to use main-sequence stars ($T_{\rm eff}$ > 5 500 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 ($T_{\rm eff}$ > 6 800 K) in metal-rich open clusters together with PMS stars from very young clusters[7] (age < 100 Myr). They showed a lithium plateau of A(Li)∼3.4 at 0.1<[Fe/H]<0.3. Their conclusion supported the scenario of Guiglion et al. (2019) which has recently been confirmed by Dantas et al., in prep.

Stars on the hot side of this dip have not undergone any Li depletion, and 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), 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 (*eg.* 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). We find stars with Li around 3.4 dex at [Fe/H]∼ 0.2 dex, consistently with the peak at A(Li)∼3.4 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 investigate further the ISM evolution in the metallicity regime -1.0<[Fe/H]<0.0. 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 (see GALAH survey, Gao et al. 2020), in the region of -1.0<[Fe/H]<-0.5. But the mean A(Li) for the 29 stars present in that metallicity range is lower at A(Li) = 2.46±0.11 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<[Fe/H]<-0.5 (based on cool dwarfs), but a shallow increase.

The consequence of such finding for the modelling the lithium ISM on the domain -1<[Fe/H]<-0.5 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 MSun rather than 1-8 MSun, 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 (e.g. Lagarde et al. 2012) when starting from an initial A(Li) = 3.3 (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). The responsible mechanism is the Cameron & Fowler (1971) mechanism. 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. Li-rich giants are believed to play a role in the enrichment of the ISM (Romano et al. 2001). Stellar Li

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

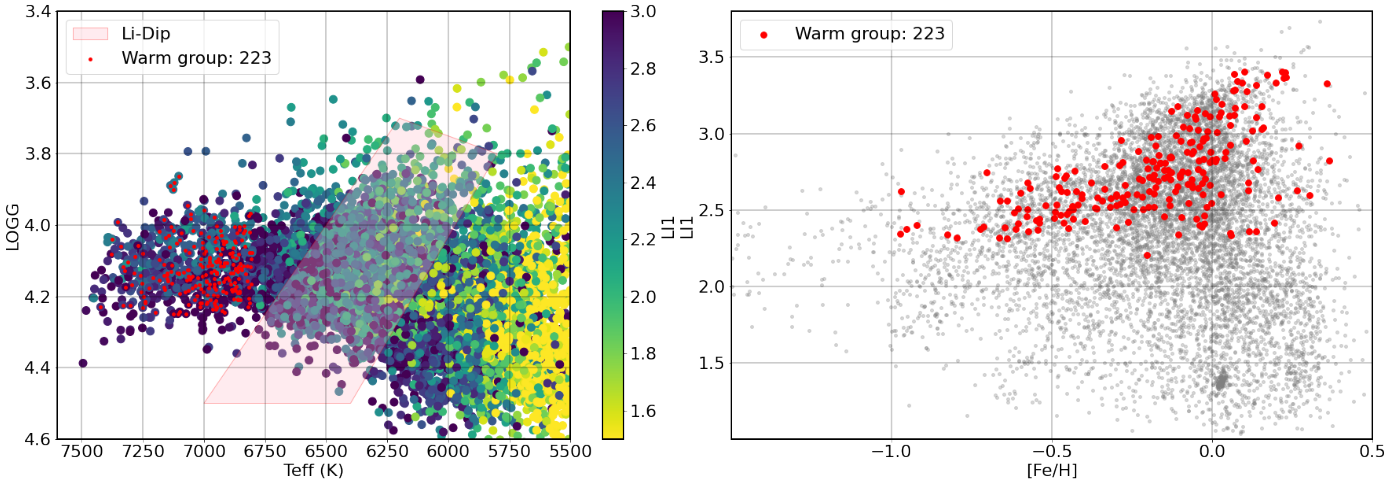

Fig. 21: Left: Effective temperature vs. surface gravity diagram with the stars color-coded according to their Li abundance. The approximate location of the Li-dip region according to Gao et al. (2020) is highlighted in pink. The red points represent the warm stars, $T_{\rm eff}$ >6800 K and S/N>50.0. Right: [Fe/H] vs. Li abundance trend for the warm stars shown as red points. Gray dots represent the stars over-plotted in left plot color-coded with A(Li).

enrichment is also possible due to external sources such as the measured over-abundance of Li as a result of 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 $\mathrm{A(Li)} \sim 2.2$ (Aguilera-Gómez et al. 2016). We refer the readers 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 $\log(g) < 3.2$ and $\mathrm{A(Li)} > 2.0$. It is important that the CNN is able to identify these rare objects as they are of a great scientific interest. 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 on previous Gaia-ESO papers. We also check the GALAH survey's catalog, in the southern sky, of Li-rich giants by Martell et al. (2021) to find no match.

To identify the Li-rich giants, we select stars with $T_{\rm eff} < 5500\,\mathrm{K}$, $\log(g) < 3.5$ and $\mathrm{A(Li)} > 2.0$ for which GES-iDR6 has not provided either one or all of the labels. To assure a reliable parameter estimation, we further select spectra with low CNN uncertainties of $\mathrm{e}\,T_{\rm eff} < 50\,\mathrm{K}$, $\mathrm{e}\log(g) < 0.1$, $\mathrm{e[Fe/H]} < 0.1$ and $\mathrm{eA(Li)} < 0.1$ and $\mathrm{S/N} > 25$ and $\mathrm{E\_VRAD} < 0.5\,\mathrm{km\,s^{-1}}$. We also check for good photometry in Gaia EDR3 by selecting $\mathrm{RUWE} \leq 1.4$. 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 2.0-3.0 dex with half have A(Li)>3.0 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 miss-classified objects (for instance 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, the view that Li-rich giants can be found only in the He-core burning red clump phase has emerged (Deepak & Reddy 2019; Deepak & Lambert 2021; Martell et al. 2021). Further analysis of our new sample is essential to investigate their properties and evaluate 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).

CNAME	$T_{\rm eff}$	$\log(g)$	[Fe/H]	A(Li)
08512566-4135067	4331	2.20	0.23	3.15

Table 2: The 31 newly discovered GES Li-rich giants and their CNN associated atmospheric parameters and lithium abundances. The table is ordered by A(Li).

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 7 031 stars, based on high-quality stellar labels from GES iDR6. The main results are summarized here:

- Our CNN shows very good performances, even though we mask H_α and despite the wavelength range in GIRAFFE HR15N setup is not considered optimal for atmospheric parameters determination (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).

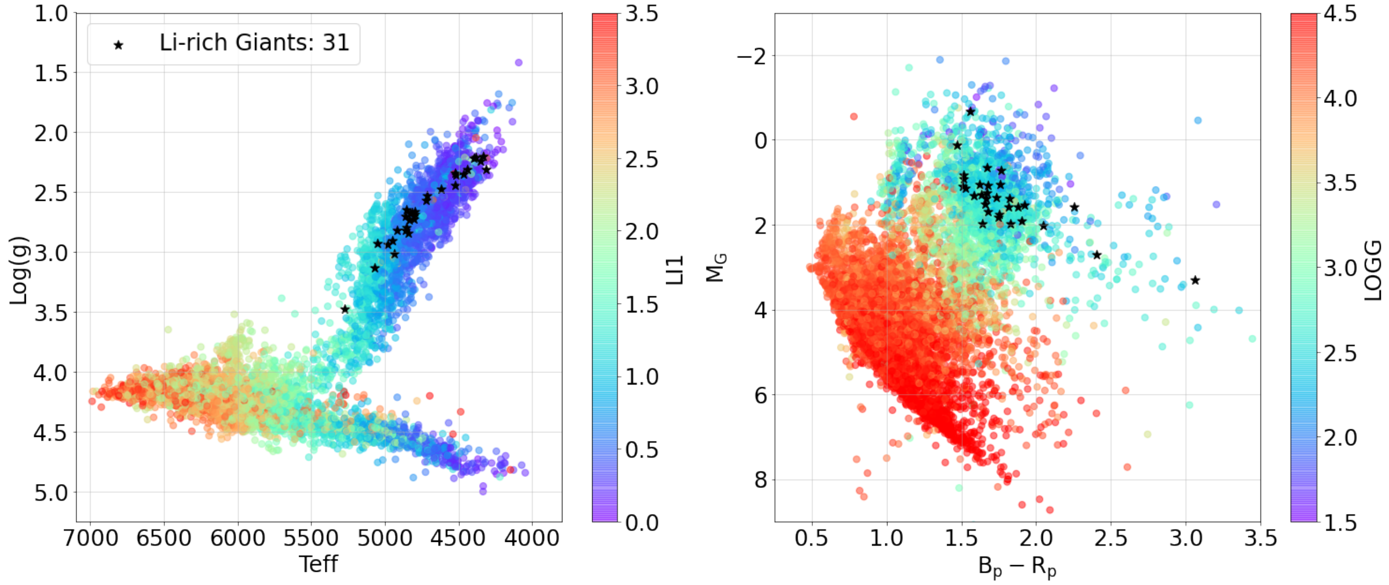

Fig. 22: Left: Kiel-diagram showing the newly-discovered Li-rich giants (black stars) along with the training sample color-coded according to their Li abundance. Right: Gaia Color-Magnitude diagram for the same stars. The training sample stars are colored by their surface gravities.

- Overall, the CNN predictions compare very well with the GES-iDR6 input labels. The CNN achieves a good performance for all S/N values including the low S/N ($\approx$ 20) 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 that large uncertainties in $V_{\rm rad}$ ($>$0.5 km s^{-1}) can degrade the CNN performances.

- 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 also could come from the fact that metal-poor stars are more difficult to parametrize due to weaker lines, and possible NLTE effect.

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 also show that the CNN atmospheric parameters match nicely with asteroseismic results from CoRoT and also demonstrated that CNN can correct wrongly assigned labels.

- We also verify that the CNN is learning from relevant spectral features for the atmospheric parameters (for example, the Quintet are 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.

- 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 suggested that the usually reported steep rise of the upper-boundary of lithium is not visible on the domain -1 $<$ [Fe/H] $<$ 0, exhibiting a more 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 modelling (Cescutti & Molaro 2019; Romano et al. 2021). In addition, there is no decrease of lithium boundary with [Fe/H] $>$ 0, but we report the presence of stars with lithium between 2.2 and 3.0 dex, likely to have depleted their lithium content.

- We report the discovery of 33 new Li-rich giants. 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, i.e. very similar data as 4MOST and WEAVE. It gives excellent perspectives for data analysis with CNN in the context of these 2 surveys. In order to increase the diversity in the training sample, one could think about adding spectra of binary stars, and properly dealing with emission features.

For the future use of CNNs, it will be crucial to build the training sets pro-actively, i.e. not only relying on sets we build for a given survey, but carefully filling-in regions of the HR diagram with proper targets. Especially, attention should be paid for populating the metal-poor tail of the training set, in order to avoid biases.

In future work, it would be interesting to explore Bayesian NNs, and different types of loss functions like negative log likelihood , in order to provide a 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 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 in of the order of -0.1 dex, for 5 000< $T_{\rm eff}$<6 500 K (see also Figures 1 & 2 of Magrini et al. 2021a). 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, and Ti will be measurable by 4MIDABLE-HR and are affected by 3D NLTE in a non-negligible way (Bergemann et al. 2021, 2017; Sitnova et al. 2018; Bergemann et al. 2012).

Concerning the optimization of the training set, properly including M stars with strong TiO bands in the training set will allow to accurately parametrize this type of objects. It will be a necessity for 4MOST, that plans to observe among other targets, open-clusters.

Regarding the sensitivity of CNN to $\mathrm{V_{rad}}$, the future surveys observing with multiple spectrographs should pay attention in providing accurate radial velocities, to minimize the possible systematics during the training phase.

We have seen, in this study, 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 in general ML for stellar abundances measurements, one will have to develop an objective criterion to decide whether an abundance is a real detection or an upper limit.

Acknowledgements. 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. 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[8] 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).

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
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., Collet, R., Amarsi, A. M., et al. 2017, ApJ, 847, 15
Bergemann, M., Hoppe, R., Semenova, E., et al. 2021, MNRAS, 508, 2236
Bergemann, M., Lind, K., Collet, R., Magic, Z., & Asplund, M. 2012, MNRAS, 427, 27
Bialek, S., Fabbro, S., Venn, K. A., et al. 2020, MNRAS, 498, 3817
Blanco-Cuaresma, S., Soubiran, C., Jofré, P., & Heiter, U. 2014, A&A, 566, A98
Boesgaard, A. M. & Tripicco, M. J. 1986, Astrophysical Journal Letters, 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, Edward, J. 1989, ApJS, 71, 293
Buder, S., Sharma, S., Kos, J., et al. 2021, MNRAS
Cameron, A. G. W. & Fowler, W. A. 1971, The Astrophysical Journal, 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. et al. 2015, Keras, https://keras.io
Dalton, G. 2016, in Astronomical Society of the Pacific Conference Series, Vol. 507, Multi-Object Spectroscopy in the Next Decade: Big Questions, Large Surveys, and Wide Fields, ed. I. Skillen, M. Balcells, & S. Trager, 97
Damiani, F., Prisinzano, L., Micela, G., et al. 2014, A&A, 566, A50
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 de 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, Annual Review of Nuclear and Particle Science, 61, 47
Fu, X., Romano, D., Bragaglia, A., et al. 2018, A&A, 610, A38
Fukushima, K. & Miyake, S. 1982, in Competition and cooperation in neural nets (Springer), 267–285
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2020, arXiv e-prints, arXiv:2012.01533
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
Gao, Q., Shi, J.-R., Yan, H.-L., et al. 2019, The Astrophysical Journal Supplement Series, 245, 33
Gao, X., Lind, K., Amarsi, A. M., et al. 2020, Monthly Notices of the Royal Astronomical Society, 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, arXiv e-prints, arXiv:2208.05432
Gratton, R. G. & D'Antona, F. 1989, A&A, 215, 66
Grevesse, N., Asplund, M., & Sauval, A. J. 2007, Space Sci. Rev., 130, 105
Guiglion, G., Chiappini, C., Romano, D., et al. 2019, A&A, 623, A99
Guiglion, G., de Laverny, P., Recio-Blanco, A., et al. 2016, A&A, 595, A18
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

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

Hong-liang, Y. & Jian-rong, S. 2022, Chinese Astron. Astrophys., 46, 1

Hunter, J. D. 2007, Computing in Science and Engineering, 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. 2015, A&A, 582, A81

Jofré, P., Heiter, U., Soubiran, C., et al. 2014, A&A, 564, A133

Jofré, P., Heiter, U., Tucci Maia, M., et al. 2018, Research Notes of the American Astronomical Society, 2, 152

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., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436

LeCun, Y., Boser, B., Denker, J. S., et al. 1989, Neural Computation, 1, 541

Leung, H. W. & Bovy, J. 2019, MNRAS, 483, 3255

Lima, E. V. R., Sodré, L., Bom, C. R., et al. 2022, Astronomy and Computing, 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

Lodders, K. & Palme, H. 2009, Meteoritics and Planetary Science Supplement, 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

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, ed. Stéfan van der Walt & Jarrod Millman, 56 – 61

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. 2017a, A&A, 598, A5

Pancino, E., Lardo, C., Altavilla, G., et al. 2017b, 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, Journal of Machine Learning Research, 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., & Gaia-ESO Consortium. 2013, The Messenger, 154, 47

Randich, S., Gilmore, G., Magrini, L., et al. 2022, arXiv e-prints, arXiv:2206.02901

Randich, S. & Magrini, L. 2021, Frontiers in Astronomy and Space Sciences, 8, 6

Randich, S., Pasquini, L., Franciosini, E., et al. 2020, A&A, 640, L1

Reeves, H., Fowler, W. A., & Hoyle, F. 1970, Nature, 226, 727

Romano, D., Magrini, L., Randich, S., et al. 2021, A&A, 653, A72

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

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., Franciosini, E., Bragaglia, A., et al. 2018, A&A, 617, A4

Smiljanic, R., Korn, A. J., Bergemann, M., et al. 2014, A&A, 570, A122

Sneden, C., Bean, J., Ivans, I., Lucatello, S., & Sobeck, J. 2012, MOOG: LTE line analysis and spectrum synthesis

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, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29

Ting, Y.-S., Conroy, C., Rix, H.-W., & Asplund, M. 2018, ApJ, 860, 159

Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2017, ApJ, 843, 32

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, Astronomische Nachrichten, 337, 970

Van der Maaten, L. & Hinton, G. 2008, Journal of machine learning research, 9

Čotar, K., Zwitter, T., Traven, G., et al. 2021, MNRAS, 500, 4849

Wang, E. X., Nordlander, T., Asplund, M., et al. 2021, MNRAS, 500, 2159

Waskom, M. L. 2021, Journal of Open Source Software, 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