

# Hierarchically modelling *Kepler* dwarfs and subgiants to improve inference of stellar properties with asteroseismology

Alexander J. Lyttle,<sup>1,2\*</sup> Guy R. Davies,<sup>1,2</sup> Tanda Li,<sup>1,2</sup> Lindsey M. Carboneau,<sup>1,2</sup> Oliver J. Hall,<sup>1,2</sup> Hin Leung,<sup>1,2,3</sup> Harry Westwood,<sup>1,2</sup> William J. Chaplin,<sup>1,2</sup> Daniel Huber,<sup>1</sup> Andrea Miglio,<sup>1,2</sup> Martin B. Nielsen,<sup>1,2</sup> and To B. Confirmed

<sup>1</sup>School of Physics and Astronomy, University of Birmingham, Birmingham, B15 2TT, UK

<sup>2</sup>Stellar Astrophysics Centre (SAC), Department of Physics and Astronomy, Aarhus University, Ny Munkegade 120, DK-8000 Aarhus C, Denmark

<sup>3</sup>School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews, KY16 9SS, UK

Accepted XXX. Received YYY; in original form ZZZ

## ABSTRACT

With recent advances in modelling stars using high-precision asteroseismology, the systematic effects associated with our assumptions of stellar helium abundance ( $Y$ ) and the mixing-length theory parameter ( $\alpha_{\text{MLT}}$ ) are becoming more important. We apply a new method to improve the inference of stellar parameters for a sample of *Kepler* dwarfs and subgiants across a narrow mass range ( $0.8 < M < 1.2 \text{ M}_\odot$ ). In this method, we include a statistical treatment of  $Y$  and the  $\alpha_{\text{MLT}}$ . We develop a hierarchical Bayesian model to encode information about the distribution of  $Y$  and  $\alpha_{\text{MLT}}$  in the population, fitting a linear helium enrichment law including an intrinsic spread around this relation and normal distribution in  $\alpha_{\text{MLT}}$ . We test various levels of pooling parameters, with and without solar data as a calibrator. We are able to constrain  $\alpha_{\text{MLT}}$  in the population with a mean of  $1.90^{+0.10}_{-0.09}$  with, and  $1.74^{+0.08}_{-0.07}$  without solar data. We find the gradient for the enrichment law,  $\Delta Y / \Delta Z$  also varies with and without the solar data, obtaining  $1.05^{+0.28}_{-0.25}$  and  $1.60^{+0.45}_{-0.42}$  respectively. While accounting for the uncertainty in  $Y$  and  $\alpha_{\text{MLT}}$ , we are still able to report statistical uncertainties of 2.5 per cent in mass, 1.2 per cent in radius, and 12 per cent in age. Our method can also be applied to larger samples which will lead to improved constraints on both the population level inference and the star-by-star fundamental parameters. We show that more work is needed to study the validity of using the Sun as a calibrator for  $Y$  and  $\alpha_{\text{MLT}}$ , perhaps by including asteroseismic signatures of helium abundance.

## Key words:

asteroseismology – stars: fundamental parameters – stars: low-mass – stars: oscillations – stars: solar-type – stars: statistics

## 1 INTRODUCTION

The inference of stellar ages, masses and radii have improved through the use of asteroseismology in recent years (e.g. see the review by Chaplin & Miglio 2013). Measuring the oscillation modes in stars using photometric time series data, from missions such as *CoRoT* (Baglin et al. 2006), *Kepler* (Borucki et al. 2010) and *TESS* (Ricker et al. 2015) has given us new insights into the structure and evolution of stars. Recent examples include a deeper understanding of stellar structure (Verma et al. 2017), chronology of a Milky Way merger (Chaplin et al. 2020) and classifying exoplanetary systems

(Huber et al. 2019). Several studies used grids of stellar models with constraints from asteroseismology to produce catalogues of precise stellar parameters (Pinsonneault et al. 2014; Silva Aguirre et al. 2017). However, with increasing precision on fundamental parameters inferred from stellar models with asteroseismology, extra care should be taken to ensure that we are accounting for uncertainty in our choice of stellar physics.

Typically, stars are modelled on a star-by-star basis with estimates of stellar properties including some assumptions handled by simple scaling relations and solar calibrations. For instance, a helium ( $Y$ ) to heavy-element ( $Z$ ) ratio ( $\Delta Y / \Delta Z$ ) and mixing-length theory parameter ( $\alpha_{\text{MLT}}$ ) are often assumed. However, there has been little effort to account for the population distribution of such

\* E-mail: ajl573@student.bham.ac.uk

quantities. Assuming values for  $Y$  and  $\alpha_{\text{MLT}}$  can result in inaccurate inference and systematics on the inferred stellar parameters (Valle et al. 2015). Independently modelling  $Y$  and  $\alpha_{\text{MLT}}$  can also be computationally demanding and requires high-precision observations in order to return meaningful stellar properties.

In this work, we apply the method of Davies et al. (in prep.) to determine stellar properties for a sample of *Kepler* dwarfs and subgiants using a hierarchical Bayesian model (HBM). With an HBM, we introduce population-level distributions for  $Y$  and  $\alpha_{\text{MLT}}$  to encode prior information throughout the sample. We will show that when an HBM is used, we can increase the precision of inferred masses, radii and ages despite increasing the number of free parameters.

The use of HBMs has been demonstrated in other areas of astrophysics to reduce individual parameter uncertainties by encoding prior information about the distribution of the parameter in a population. For example, HBMs have been used with data from *Gaia*, improving distance measures (Leistedt & Hogg 2017; Anderson et al. 2018) and calibrating the red clump as a standard candle (Hawkins et al. 2017; Chan & Bovy 2020) using asteroseismology (Hall et al. 2019). In other instances, HBMs have been used to infer binary-star and exoplanet eccentricities (Hogg et al. 2010), obliquity of transit systems (Morton & Winn 2014), and stellar inclination with asteroseismology (Campante et al. 2016; Kuszlewicz et al. 2019).

To describe the distribution of  $Y$ , we assume a linear helium enrichment law characterised by freely varied population parameters: the gradient ( $\Delta Y/\Delta Z$ ), the primordial helium abundance at  $Z = 0$  ( $Y_P$ ) and an intrinsic spread in helium ( $\sigma_Y$ ). There have been many studies to determine a linear enrichment law, from modelling eclipsing binaries (Ribas et al. 2000) to spectroscopy of galactic H II regions (Balser 2006). Values of  $\Delta Y/\Delta Z$  have also been determined for samples of main sequence stars (Casagrande et al. 2007), open clusters (Brogaard et al. 2012) and more recently with asteroseismology using individual oscillation frequencies (Silva Aguirre et al. 2017) and the glitch due to the second helium ionisation zone (Verma et al. 2019). In these studies the value of the enrichment ratio was typically inferred to be  $1 \lesssim \Delta Y/\Delta Z \lesssim 3$ .

The widely used mixing-length theory of convection, parametrised by  $\alpha_{\text{MLT}}$ , has been tested throughout the Hertzsprung-Russell diagram with 3D hydrodynamical simulations (Trampedach et al. 2014; Magic et al. 2015) and asteroseismology (Tayar et al. 2017; Viani et al. 2018; Li et al. 2018) with values in the range  $0.8 \lesssim \alpha_{\text{MLT}}/\alpha_{\text{MLT},\odot} \lesssim 1.2$  where  $\alpha_{\text{MLT},\odot}$  is the value calibrated to the Sun. However, in many grids of stellar models, a constant value calibrated to reproduce the Sun is assumed. This can lead to systematic uncertainties in stellar ages due to the effects of variable mixing depending on the mixing length. In this work, we experiment with two prior assumptions for  $\alpha_{\text{MLT}}$ , one assuming the best-fitting  $\alpha_{\text{MLT}}$  is normally distributed in our sample, and the other assuming it is constant throughout.

Our HBM requires a way to map from the stellar initial (or bulk) properties to predict observables. We can achieve this with a large grid of stellar evolutionary models. However, a discrete grid can produce inaccurate posterior distributions, limited to the grid resolution. Increasing the resolution is computationally demanding, especially when scaling to higher input dimensions. One solution is to interpolate the stellar models, as is common in the isochrone-fitting method (see e.g. Berger et al. 2020). However, interpolation can become computationally expensive at high input dimensions and grid size, and evaluating the likelihood using modern Bayesian sampling techniques is slow.

We use machine learning to map stellar inputs to observables to provide a fast way to sample the HBM. One example of machine learning in asteroseismology is a random forest to map observables to fundamental parameters (Bellinger et al. 2016). In this work, we train an artificial neural network (ANN) on a large grid of stellar models. There have also been similar applications of ANNs in asteroseismology (Verma et al. 2016; Hendriks & Aerts 2019) but not yet in the context of an HBM. Using the machine learning speed-up, we demonstrate a scalable method for obtaining fundamental stellar parameters.

In Section 2, we describe the data for the sample of 81 *Kepler* dwarfs and subgiants studied in this work. We then present the methods in Section 3 for which we produced a large grid of stellar evolutionary models to use as training data for an ANN (see Sections 3.1 and 3.2 respectively). We use the ANN as an emulator in a series of statistical models described in Section 3.3 to model the sample and present our results in Section 4. Finally, in Section 5 we discuss the results from each model and compare them to results for the sample of stars in the literature.

## 2 DATA

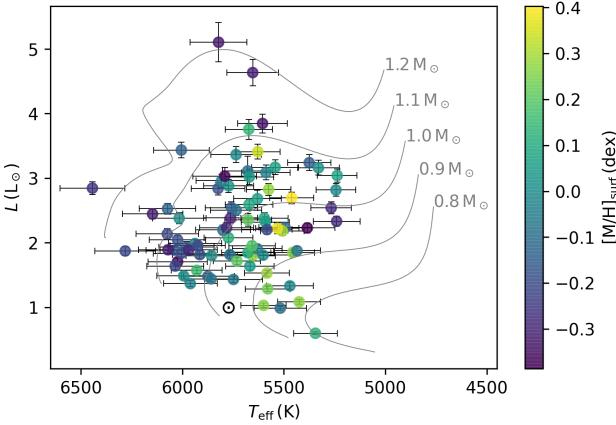
For this study, we selected the sample of 415 stars from the first APOKASC catalogue of dwarfs and subgiants (Serenelli et al. 2017, hereafter S17). This sample provides an extensive set of dwarfs and subgiant stars with asteroseismic detections observed by the *Kepler* mission. S17 used grid-based modelling to determine the ages ( $\tau$ ) masses ( $M$ ) radii ( $R$ ) and surface gravity ( $\log g$ ) of stars in the sample, using global asteroseismic parameters, effective temperature ( $T_{\text{eff}}$ ) and metallicity ([M/H]) as inputs.

Using five independent pipelines, S17 determined values for global asteroseismic parameters – the large frequency separation  $\Delta\nu$  and the frequency at maximum power,  $\nu_{\text{max}}$  with median uncertainties of 1.7 per cent and 4 per cent respectively. We chose to adopt the  $\Delta\nu$  determined in their work as inputs for our method. They also used [M/H] published in Data Release 13 (DR13; Alberti et al. 2017) of the APOGEE stellar abundances pipeline (ASPCAP; García Pérez et al. 2016) with uncertainties of 0.1 dex. For their preferred set of results, they adopted a  $T_{\text{eff}}$  scale from the Sloan Digital Sky Survey (SDSS) *griz*-band photometry (Pinsonneault et al. 2012) with a median uncertainty of 70 K.

We removed more evolved stars from the APOKASC sample by cutting those with  $\log g < 3.8$  dex. We then kept stars within  $1-\sigma$  of  $-0.5 < [\text{M}/\text{H}] < 0.5$  dex to remove metal-poor and -rich stars. Main sequence stars with  $M \gtrsim 1.2 M_{\odot}$  are understood to have a convective, hydrogen-burning core, with some dependence on the chemical composition and choice of stellar physics (Appourchaux et al. 2015). Stellar models with a convective core require the treatment of extra stellar physics such as overshooting, which is beyond the scope of this work. Therefore, we keep only stars with masses within  $1-\sigma$  of 0.8 to  $1.2 M_{\odot}$  from the preferred set of results of S17.

Following cuts to the sample, we adopted updated ASPCAP spectroscopic metallicities, [M/H], from Data Release 14 (DR14; Blanton et al. 2017) which had a median uncertainty of 0.07 dex. We also chose to adopt  $T_{\text{eff}}$  from the same catalogue to be internally consistent. We note that our chosen effective temperature scale is offset from the photometric temperature scale of S17 by approximately  $-170$  K with a dispersion of  $\sim 120$  K. The median uncertainty in our adopted ASPCAP  $T_{\text{eff}}$  was 125 K which is compatible with the dispersion observed.

To provide a means of calculating luminosities, we used *Gaia*



**Figure 1.** The luminosity,  $L$  against effective temperature,  $T_{\text{eff}}$  of the sample of 81 *Kepler* dwarfs and subgiants studied in this work. Each star is coloured according to metallicity. The grey lines depict evolutionary tracks with  $[\text{M}/\text{H}]_{\text{init}} = 0.0 \text{ dex}$ ,  $Y_{\text{init}} = 0.28$  and  $\alpha_{\text{MLT}} = 1.9$  for different stellar masses. The current position of the Sun is shown by the  $\odot$  symbol.

Data Release 2 (DR2) parallaxes (Gaia Collaboration et al. 2016, 2018). We cross-matched the remaining sample with the DR2 catalogue, taking the nearest neighbours within a  $4''$  radius. Although DR2 parallaxes have improved upon the DR1 values at the time of S17, there was still evidence for a zero-point offset (Lindegren et al. 2018). We adopted a global offset of 0.05 mas, in the sense that DR2 parallaxes were underestimated, representative of values obtained in the literature (see e.g. Riess et al. 2018; Zinn et al. 2019; Hall et al. 2019; Chan & Bovy 2020). We then cross-matched our sample with the Two-Micron All Sky Survey (2MASS; Skrutskie et al. 2006) to obtain  $K_S$ -band ( $2.16 \mu\text{m}$ ) photometry.

We determined luminosities,  $L$  for the sample using the direct method of ISOCLASSIFY (Huber et al. 2017; Berger et al. 2020) with  $K_S$ -band photometry, Gaia DR2 parallaxes, ASPCAP  $[\text{M}/\text{H}]$  and  $T_{\text{eff}}$  and asteroseismic  $\log g$  as inputs. This involved computing absolute  $K_S$ -band magnitudes using the Gaia DR2 parallaxes and extinctions determined by the 3D galactic reddening maps of Green et al. (2018). We determined absolute bolometric magnitudes by interpolating the MIST bolometric correction tables as a function of  $T_{\text{eff}}$ ,  $\log g$  and  $[\text{M}/\text{H}]$  (Dotter 2016; Choi et al. 2016). We adopted the uncertainty of 0.02 mag assumed by ISOCLASSIFY for both the extinctions and bolometric corrections, representative of typical systematics during interpolation (Huber et al. 2017). We obtained luminosities for the sample with a median uncertainty of 3.4 per cent.

The final sample comprised 81 stars for which we had complete data for  $T_{\text{eff}}$ ,  $[\text{M}/\text{H}]$ ,  $\Delta v$  and  $L$  to use as inputs for our stellar modelling method – see Table 1. In Figure 1, we show the Hertzsprung-Russell diagram for the sample plot in context with a series of stellar evolutionary tracks at solar metallicity.

### 3 METHODS

Our principle goal was to improve inference of fundamental stellar parameters for our sample of low-mass stars. To achieve this, we constructed a hierarchical Bayesian model (HBM) which utilises a prior assumption of the population distribution to share information between the stars. Based on the work of Davies et al. (in prep.), our

HBM is a generative model which requires a function to map bulk stellar parameters to their observables.

Firstly, we used a stellar evolutionary code to compute a grid of models to predict observable quantities (see 3.1). For a given initial mass,  $M$ , metallicity,  $[\text{M}/\text{H}]_{\text{init}}$ , helium fraction,  $Y_{\text{init}}$  and mixing-length theory parameter,  $\alpha_{\text{MLT}}$  our stellar models output  $T_{\text{eff}}$ ,  $L$  and chemical composition as a function of age,  $\tau$ . Calls to models of stellar evolution are slow and the grid produced is discrete. This makes it difficult to robustly evaluate an HBM using the grid alone. We could interpolate the grid of stellar models. However, interpolation does not scale well with the number of input dimensions and points on the grid, reducing the scalability of our method.

In Section 3.2, we describe a method to replace the grid of stellar models with a smooth function approximation using machine learning. In particular, we trained an artificial neural network (ANN) on the grid of stellar models to map bulk stellar parameters to observables. Fast evaluation of the ANN gradient is required during training. Consequently, estimating the gradient of the model likelihood is possible with an ANN. Hence, we open up the possibility of using the Hamiltonian Monte Carlo (HMC) algorithm which requires the gradient to sample the model posterior – for example, using the No-U-Turn Sampler (NUTS; Homan & Gelman 2014).

Finally, we constructed three Bayesian models in Section 3.3 which each used the trained ANN to estimate stellar fundamental parameters. We then tested the models on a set of synthetic stars generated by the stellar evolutionary code. Once we had tested the model accuracy with the synthetic stars, we evaluated each model on the subset of the APOKASC catalogue selected in Section 2.

#### 3.1 Grid of stellar models

We built a stellar model grid to use in training the ANN. The grid includes four independent model inputs: stellar mass ( $M$ ), initial helium fraction ( $Y_{\text{init}}$ ), initial metallicity ( $[\text{M}/\text{H}]_{\text{init}}$ ), and the mixing-length parameter ( $\alpha_{\text{mlt}}$ ). Ranges and grid steps of the four model inputs are summarised in Table 2. We increased the resolution at higher  $[\text{M}/\text{H}]$  to give a more consistent resolution in  $Z$  which is used as an ANN input in Section 3.2. We computed each stellar evolutionary track from the Hayashi line to the base of red-giant branch defined here by  $\log g = 3.6 \text{ dex}$ . We also computed evolutionary tracks with input values at the midpoint between points on the grid for validating the ANN.

##### 3.1.1 Stellar models and input physics

We used Modules for Experiments in Stellar Astrophysics (MESA, version 12115) to establish a grid of stellar models. MESA is an open-source stellar evolution package which is undergoing active development. Descriptions of input physics and numerical methods can be found in Paxton et al. (2011, 2013, 2015). We adopted the solar chemical mixture,  $(Z/X)_\odot = 0.0181$ , provided by Asplund et al. (2009). The initial chemical composition was calculated by:

$$\log(Z_{\text{init}}/X_{\text{init}}) = \log(Z/X)_\odot + [\text{M}/\text{H}]_{\text{init}}. \quad (1)$$

We used the MESA  $\rho - T$  tables based on the 2005 update of OPAL EOS tables (Rogers & Nayfonov 2002) and OPAL opacity supplemented by low-temperature opacity (Ferguson et al. 2005). The MESA ‘simple’ photosphere were used as the set of boundary conditions for modelling the atmosphere. The mixing-length theory of convection was implemented, where  $\alpha_{\text{MLT}} = \ell_{\text{MLT}}/H_p$  is the

**Table 1.** The observables and their respective uncertainties for the 10 stars in sample of 81 stars. The whole table is available online.

Name	$T_{\text{eff}}$ (K)	$\sigma_{T_{\text{eff}}}$ (K)	$L$ ( $L_{\odot}$ )	$\sigma_L$ ( $L_{\odot}$ )	$\Delta\nu$ ( $\mu\text{Hz}$ )	$\sigma_{\Delta\nu}$ ( $\mu\text{Hz}$ )	$[\text{M}/\text{H}]_{\text{surf}}$ (dex)	$\sigma_{[\text{M}/\text{H}]}$ (dex)
KIC10079226	5928.84	124.84	1.57	0.05	116.04	0.73	0.16	0.07
KIC10215584	5666.92	119.33	1.64	0.06	115.16	2.83	0.04	0.07
KIC10319352	5456.17	106.65	1.85	0.06	78.75	1.73	0.27	0.06
KIC10322381	6146.79	148.58	2.44	0.08	86.64	6.57	-0.32	0.08
KIC10417911	5628.26	109.99	3.41	0.12	56.14	2.10	0.34	0.07

**Table 2.** Stellar model grid parameters for training and test datasets.  $N_{\text{track}}$  are the numbers of stellar evolutionary tracks computed for each dimension of the grid, multiplied to a total of 17,220 tracks.

Stellar model grid			
Input Parameter	Range	Increment	$N_{\text{track}}$
$M$ ( $M_{\odot}$ )	0.80 – 1.20	0.01	41
$[\text{M}/\text{H}]$ (dex)	-0.5 – 0.2/0.25 – 0.5	0.1/0.05	14
$Y_{\text{init}}$	0.24 – 0.32	0.02	5
$\alpha_{\text{milt}}$	1.5 – 2.5	0.2	6
<b>Total</b>			17,220

mixing-length parameter. We also applied the MESA predictive mixing scheme (Paxton et al. 2018, 2019) in the model computation.

Atomic diffusion of helium and heavy elements was also taken into account. MESA calculates particle diffusion and gravitational settling by solving Burger's equations using the method and diffusion coefficients of Thoul et al. (1994). We considered eight elements ( $^1\text{H}$ ,  $^3\text{He}$ ,  $^4\text{He}$ ,  $^{12}\text{C}$ ,  $^{14}\text{N}$ ,  $^{16}\text{O}$ ,  $^{20}\text{Ne}$ , and  $^{24}\text{Mg}$ ) for diffusion calculations, and had the charge calculated by the MESA ionization module, which estimates the typical ionic charge as a function of  $T$ ,  $\rho$ , and free electrons per nucleon from Paquette et al. (1986).

### 3.1.2 Oscillation models and asteroseismic quantities

Theoretical stellar oscillations were calculated with the GYRE code (version 5.1), which was developed by Townsend & Teitler (2013). We computed radial modes (for  $\ell = 0$ ) for 42 radial orders by solving the adiabatic stellar pulsation equations with the structural models generated by MESA. We determined the asteroseismic large separation ( $\Delta\nu$ ) for each model with theoretical radial modes to avoid the systematic offset of the scaling relation. We derived  $\Delta\nu$  with the approach given by White et al. (2011), which is a weighted least-squares fit to the radial frequencies as a function of  $n$ .

We chose to ignore the well known, yet poorly characterised impact of modelled oscillation mode inaccuracies in the near-surface region of the star (Kjeldsen et al. 2008; Ball & Gizon 2014; Sonoi et al. 2015). This typically presents only a small effect when considering the average large frequency spacing,  $\Delta\nu$ .

## 3.2 Artificial neural network

Once we constructed our grid of models, we needed a way in which we could continuously sample the grid for use in our statistical model. Interpolation would not scale well to the size of our model grid and input dimensionality. Moreover, evaluating the gradient of an interpolated function is slow. In this work, we utilised machine learning to approximate the grid of stellar models via an artificial neural network (ANN). The ANN is advantageous over interpolation

due to scaling well with dimensionality, fast training and evaluation, and easy gradient evaluation due to its roots in linear algebra (Haykin 2007).

We trained an ANN on the data generated by the grid of stellar models to map fundamentals to observables. We split the grid into a *train* and *test* dataset for tuning the ANN, as described in Section 3.2.1. We then tested a multitude of ANN configurations and training data inputs, repeatedly evaluating them with the test dataset in Section 3.2.2. In Section 3.2.3, we reserved a set of off-grid stellar models as our final *validation* dataset to evaluate the approximation ability of the best-performing ANN. Firstly, we briefly describe the theory and motivation behind the ANN.

An ANN is a network of artificial *neurons* which each transform some input vector,  $\mathbf{x}$  based on trainable weights,  $\mathbf{w}$  and a bias,  $b$ . The weights are represented by the connections between neurons and the bias is a unique scalar associated with each neuron. A multi-layered ANN is where neurons are arranged into a series of layers such that any neuron in layer  $k - 1$  is connected to at least one of the neurons in layer  $k$ .

In this work, we considered a fully-connected ANN, where each neuron in layer  $k - 1$  is connected to every neuron in layer  $k$ . The output of the  $i$ -th neuron in layer  $k$  is,

$$x_{i,k} = f_k(\mathbf{w}_{i,k} \cdot \mathbf{x}_{k-1} + b_{i,k}) \quad (2)$$

where  $f_k$  is the *activation* function for the  $k$ -th layer,  $\mathbf{w}_{i,k}$  are the weights connecting all the neurons in layer  $k - 1$  to the current neuron, and  $b_{i,k}$  is the bias. This generalises such that the output of the  $k$ -th layer is,

$$\mathbf{x}_k = f_k(\mathbf{W}_k \cdot \mathbf{x}_{k-1} + \mathbf{b}_k), \quad (3)$$

where  $\mathbf{W}_k$  is the matrix of weights leading to all neurons in the  $k$ -th layer. For a regression ANN, we typically have a linear activation function applied to the output of the final layer. Therefore, the output of a network of  $H$  hidden layers with initial input  $X$  is,

$$\mathbf{Y} = \mathbf{W}_H \cdot f_{H-1}(\dots f_1(\mathbf{W}_1 \cdot f_0(\mathbf{W}_0 \cdot X + \mathbf{b}_0) + \mathbf{b}_1)) + \mathbf{b}_M \quad (4)$$

We also restricted our configuration to an ANN with the same number of neurons,  $N$  in each hidden layer. Hereafter, we refer to our choice of neurons per layer,  $N$  and hidden layers,  $H$  as the *architecture*.

To fit the ANN, we used a set of training data,  $\mathbf{D}_{\text{train}} = \{(\mathbf{X}_1, \mathbf{Y}_1) \dots (\mathbf{X}_{N_{\text{train}}}, \mathbf{Y}_{N_{\text{train}}})\}$  comprising  $N_{\text{train}}$  input-output pairs. We split the training data into pseudo-random batches,  $\mathbf{D}_{\text{batch}}$  because this has been shown to improve ANN stability and computational efficiency (Masters & Luschi 2018). The set of predictions made for each batch is evaluated with an error function,  $E(\mathbf{D}_{\text{batch}})$ , also known as the *loss* which quantifies the difference between the training data and predictions. We also considered an addition to the loss called *regularisation* which helps reduce over-fitting (Goodfellow et al. 2016). During fitting, the weights are updated after each batch using an algorithm called the *optimizer*, back-propagating the error with the goal of minimising the loss.

We trained the ANN using TensorFlow (Abadi et al. 2016). We varied the architecture, number of batches, choice of loss function, optimizer and regularisation during the optimisation phase. For each set of ANN parameters, we initialised the ANN with a random set of weights and biases and minimized the loss over a given number of *epochs*. An epoch is defined as one iteration through the entire training dataset,  $D_{\text{train}}$ . We tracked the loss for each ANN using an independent test dataset to determine the most effective choice of ANN parameters (see Section 3.2.2).

### 3.2.1 Train, test and validation data

We built the train and test dataset from the outputs of the grid of stellar models in Section 3.1. This included the input parameters:  $M$ ,  $\alpha_{\text{MLT}}$ ,  $Y_{\text{init}}$  and the initial heavy-elements fraction,  $Z_{\text{init}}$ . We also included the  $T_{\text{eff}}$ ,  $\log g$ ,  $\Delta\nu$ , stellar age ( $\tau$ ), radius ( $R$ ), surface metallicity ( $[\text{M}/\text{H}]_{\text{surf}}$ ) and other chemical composition information generated by the models. We determined the fractional main sequence (MS) lifetime,  $f_{\text{MS}} = \tau/\tau_{\text{MS}}$ , of each evolutionary track by taking  $\tau_{\text{MS}}$  as the age when the central hydrogen fraction,  $X_c < 0.01$ . We then cut data where  $f_{\text{MS}} < 0.01$  to remove points on the grid prior to the MS.

Once we had refined the data from the grid of models, we randomly sampled  $7.736 \times 10^6$  points to use as the training dataset, with the remaining  $\sim 2 \times 10^6$  points given to the test dataset. We then varied our choice of ANN input and output parameters among those available in the training dataset during tuning (see Section 3.2.2).

We produced a validation dataset of  $\sim 2 \times 10^6$  stellar models evolved using MESA. Values for the initial  $M$ ,  $[\text{M}/\text{H}]$ ,  $Y$  and  $\alpha_{\text{MLT}}$  were chosen at the midpoint of the grid parameters described in Table 2. We prepared this dataset in the same way as the training set, but also constrained it to  $\tau < 15$  Gyr because ages above  $\sim 15$  Gyr are considered unphysical and such points are sparse in the training data. The validation dataset was set aside and evaluated on the final ANN.

### 3.2.2 Tuning

We needed to train an ANN which would reproduce stellar observables according to our choice of physics with greater accuracy than typical observational precisions. We experimented with a variety of ANN parameter choices, such as the architecture, activation function, optimization algorithm and loss function. We tuned the ANN parameters by varying them in both a grid-based and heuristic approach, each time evaluating the accuracy using the test dataset.

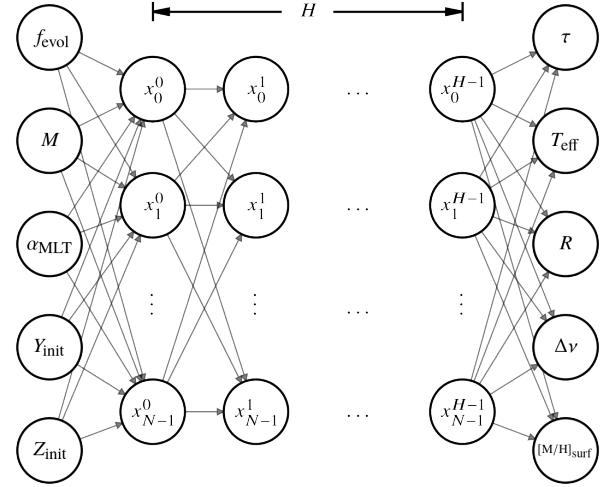
During initial tuning, we found that having stellar age as an input was unstable, because it varied heavily with the other input parameters. We mitigated this by introducing an input to describe the fraction of time a star had spent in a given evolutionary phase,  $f_{\text{evol}}$ .

$$f_{\text{evol}} = \begin{cases} f_{\text{MS}}, & f_{\text{MS}} \leq 1 \\ 1 + \frac{\tau - \tau_{\text{MS}}}{\tau_{\text{end}} - \tau_{\text{MS}}}, & f_{\text{MS}} > 1 \end{cases} \quad (5)$$

where  $\tau_{\text{end}}$  is the age of the star at the end of the track,

$$f_{\text{MS}} = \frac{\tau}{\tau_{\text{MS}}}, \quad (6)$$

and  $\tau_{\text{MS}}$  is the MS lifetime. In other words, a star with  $f_{\text{evol}} \in (0, 1]$  is in its MS phase, burning hydrogen in its core, and  $f_{\text{evol}} \in (1, 2]$  has left the MS. Consequently,  $f_{\text{evol}}$  gives the ANN information



**Figure 2.** An artificial neural network comprising  $H$  hidden layers with  $N$  neurons per layer. Arrows connecting the nodes represent tunable weights.

about the internal state of the star which affects the output observables. Otherwise,  $f_{\text{evol}}$  is a meaningless parameter, although it could loosely be interpreted as a measure of the evolutionary phase of the star.

We also observed the ANN struggled to fit areas with a high rate of change in observables, partly because of poor grid coverage. To bias training to such areas, we calculated the gradient in  $T_{\text{eff}}$  and  $\log g$  between each point for each stellar evolutionary track and used them as optional weights to the loss during tuning. These weights multiplied the difference between the ANN prediction and the training data in our chosen loss function.

After preliminary tuning, we chose the ANN input and output parameters to be  $X = \{f_{\text{evol}}, M, \alpha_{\text{MLT}}, Y_{\text{init}}, Z_{\text{init}}\}$  and  $Y = \{\log(\tau), T_{\text{eff}}, R, \Delta\nu, [\text{M}/\text{H}]_{\text{surf}}\}$  respectively. A generalised form of our neural network is depicted in Figure 2. The inputs corresponded to initial conditions in the stellar modelling code and the outputs corresponded to surface conditions throughout the lifetime of the star, with the exception of age which is mapped from  $f_{\text{evol}}$ .

We standardised the training dataset by subtracting the median,  $\mu_{1/2}$  and dividing by the standard deviation,  $\sigma$ . We found that the ANN performed better when the training data was scaled in this way. In Table A1, we show the locations and scales of the standardisation for our chosen input and output parameters.

We found that the optimal choice of architecture ( $N$  and  $H$ ) varied depending on our choice of other ANN parameters. Therefore, each time we explored a new parameter, we trained an ANN with a grid of  $(N, H)$  ranging from  $(32, 2)$  to  $(512, 10)$ .

We evaluated the performance of three activation functions: the hyperbolic-tangent, the rectified linear unit (ReLU; Hahnloser et al. 2000; Glorot et al. 2011) and the exponential linear unit (ELU; Clevert et al. 2015). Although the ReLU activation function outperformed the other two in speed and accuracy, the resulting ANN output was not smooth. The discontinuity in the ReLU function,  $f(x) = \max(0, x)$  in turn caused the output to be discontinuous. This made the ANN difficult to sample for our choice of statistical model (see Section 3.3). Out of the remaining activation functions, ELU performed the best, providing a smooth output which was well-suited to our probabilistic sampling methods.

We compared the performance of two optimisers: Adam

**Table 3.** The median error,  $\mu_{1/2}$  and median absolute deviation of the error,  $\sigma_{\text{MAD}} = 1.4826 \cdot \text{median}(|E(x) - \mu_{1/2}|)$  for a given parameter,  $x$ , derived from the output of the ANN. The error,  $E(x)$ , is given in the first column and  $\delta x$  is the truth provided by the validation dataset subtracted from the ANN prediction. Fractional errors are given as percentages (%).

Error	$\mu_{1/2}$	$\sigma_{\text{MAD}}$
$\delta\tau/\tau$ (%)	-0.003	0.178
$\delta T_{\text{eff}}$ (K)	-0.100	1.595
$\delta R/R$ (%)	0.002	0.071
$\delta L/L$ (%)	0.060	0.146
$\delta\Delta\nu$ ( $\mu\text{Hz}$ )	-0.007	0.084
$\delta[\text{M}/\text{H}]_{\text{surf}}$ (dex)	0.000	0.001

(Kingma & Ba 2014) and stochastic gradient descent (SGD; see e.g. Ruder 2016) with and without momentum (Qian 1999). Both optimizers required a choice of *learning rate* which determined the rate at which the weights were adjusted during training. We found that Adam performed well but the test loss was noisy as a function of epochs as it struggled to converge. The SGD optimizer was less noisy than Adam, but it was difficult to tune the learning rate. However, SGD with momentum allowed for more adaptive weight updates and out-performed the other configurations.

There are several ways to reduce over-fitting, from minimising the complexity of the architecture, to increasing the size and coverage of the training dataset. One alternative is to introduce weight regularisation. So-called L2 regularisation adds a term,  $\sim \lambda_k \sum_i w_{i,k}^2$  to the loss function for a given hidden layer,  $k$  which acts to keep the weights small. We varied the magnitude of  $\lambda_k$  and found that if it was too large it would dominate the loss function, but if it was too small then performance on the test dataset was poorer.

We compared the choice of two error functions: mean squared error (MSE) and mean absolute error (MAE). The former is widely used among ANNs because it is more sensitive to large errors. However, we tracked both metrics regardless of which was added to the loss function and found that MAE converged faster. Although MAE is less effective at large errors, we found that these were typically at the edges of the grid and the accuracy was good enough everywhere else.

After extensive tuning, we opted for an ANN with  $N = 128$  neurons in each of  $H = 6$  hidden layers. Each of the hidden layers used an ELU activation function and L2 weight regularisation with  $\lambda = 1 \times 10^{-6}$ . We trained the ANN for 50,000 epochs with a 500 training data batches each containing 15,472 input-output pairs. To fit the ANN, we used an SGD optimiser with an initial learning rate of  $1 \times 10^{-4}$  and momentum of 0.999 with an MAE loss function. Training took  $\sim 48$  h on an NVidia Tesla V100 graphics processing unit (GPU).

### 3.2.3 Validation

The validation dataset contained  $\sim 2 \times 10^6$  models evolved in the same way as the training dataset but with initial conditions at the midpoint of those in the grid. We made predictions for the validation dataset, deriving luminosity from the output radius and effective temperature, using the final trained ANN as described in Section 3.2.2. We then evaluated the accuracy of the ANN by taking the difference between the validation truth and prediction,  $x_{\text{true}} - x_{\text{pred}}$ .

We found good agreement between the validation dataset and ANN predictions, within typical observational uncertainties. We

noted that the largest errors lay at the boundaries of the training data and in areas sparsely populated by the grid. This is apparent in Figure 3 where we plot the validation error against each parameter. For example, the spread in error increases at high temperatures which may be attributed to poor sampling of the high-mass MS turn-off ‘hook’ by the stellar evolutionary models. Otherwise, the accuracy is very good within the observed range covered by our sample of 81 dwarfs and subgiants. Hence, we chose the median absolute deviation (MAD) as an estimator of the spread in error, because it is less sensitive to outliers than the standard deviation.

To represent the accuracy of the ANN, we present the median,  $\mu_{1/2}$  and MAD estimator,  $\sigma_{\text{MAD}} = 1.4826 \cdot \text{median}(|E(x) - \mu_{1/2}|)$  of the error ( $E(x)$ ) in Table 3. The median is close to zero for all parameters, showing little systematic bias in the ANN. The MAD is also lower than observational uncertainties quoted in Section 2. Although the error in  $\Delta\nu$  is  $\sim 0.1 \mu\text{Hz}$  is comparable to observations with the best signal-to-noise, this error is random throughout the validation data and should not produce any systematic bias.

### 3.3 Statistical models

We devised three Bayesian models, each with varying levels of parameter sharing (pooling) between stars in the population. Initially, we tested the models and demonstrated reduction of statistical uncertainties in the stellar fundamental parameters by analysing a random sample of 100 synthetic stars modelled using MESA. Then, we applied the models to the sample of stars in Table 1 (with and without Solar data for two of the models) and compared the results with that of S17.

Our first model was equivalent to modelling each star individually and featured no pooling; henceforth, we refer to it as the no-pooled (NP) model (see Section 3.3.1). We then derived two hierarchical Bayesian models (HBMs) which use population-level parameters to describe their distribution in the sample. Both of these models partially-pooled helium using a linear enrichment law. We drew the initial helium fraction for each star from a normal distribution with a mean described by the enrichment law and standard deviation representing its spread. Similarly, we partially-pooled the mixing-length theory parameter,  $\alpha_{\text{MLT}}$  in one model, whereas we maximally-pooled  $\alpha_{\text{MLT}}$  in the other, such that it assumes the same value for the entire sample. Hence, we refer to the former as the partial-pooled (PP) model and the latter as the max-pooled (MP) model, described in Sections 3.3.2 and 3.3.3 respectively.

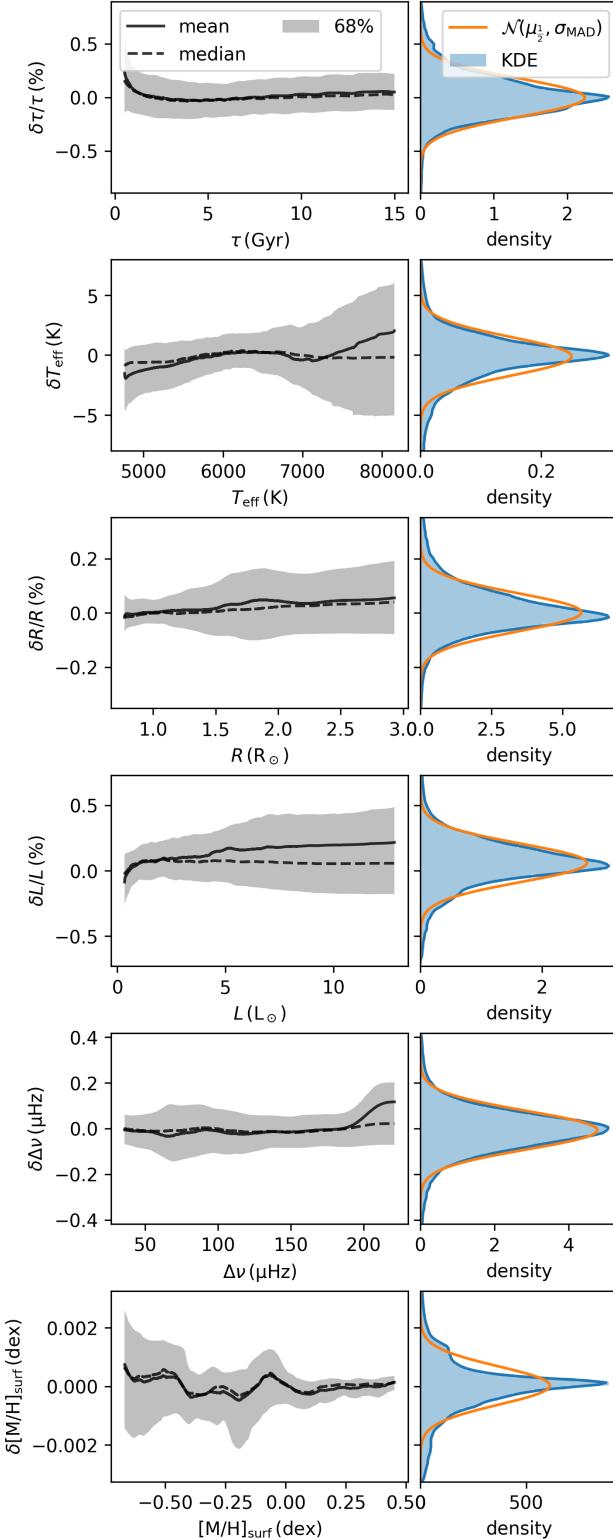
#### 3.3.1 No-pooled model

Firstly, we constructed a model comprising independent parameters  $\theta_i = \{f_{\text{evol},i}, M_i, \alpha_{\text{MLT},i}, Y_i, Z_i\}$  for a given star,  $i$ . Using Bayes’ theorem, the *posterior* probability density function (PDF) of the model parameters given a set of observed data,  $\mathbf{d}_i$  is,

$$p(\theta_i | \mathbf{d}_i) \propto p(\theta_i) p(\mathbf{d}_i | \theta_i), \quad (7)$$

where  $p(\theta_i)$  is the *prior* PDF of the model parameters and  $p(\mathbf{d}_i | \theta_i)$  is the *likelihood* of observing the data given the model.

We chose weakly-informative, bounded priors for the independent parameters, restricting them to their respective ranges in the ANN training data. Although the neural network is able to make predictions outside the training data range, these have not been tested and may be unreliable. Therefore, we used a beta distribution with  $\alpha = \beta = 1.2$  as the prior PDF on the independent parameters, transformed such that the probability is null outside the chosen range,



**Figure 3.** Left: the rolling error between the validation dataset (*true*) and the ANN predictions (*pred*) plotted against each parameter, where  $\delta x = x_{\text{pred}} - x_{\text{true}}$  for a given output  $x$ . Right: a kernel density estimate (KDE) of the validation error and a normal distribution centred on the median,  $\mu_{1/2}$  with an estimator for the standard deviation from the median absolute deviation,  $\sigma_{\text{MAD}}$ .

$$p(\boldsymbol{\theta}_i) = \prod_{k=1}^{N_{\theta}} [\theta_{k,\min} + (\theta_{k,\max} - \theta_{k,\min}) \mathcal{B}(\theta_{k,i}|1.2, 1.2)], \quad (8)$$

where the beta distribution is defined as,

$$\mathcal{B}(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du}. \quad (9)$$

The beta distribution was preferred over a bounded uniform distribution because our sampler evaluates the gradient of the posterior and hence sensitive to discontinuities.

Using notation which represents the distribution of some parameter  $x \sim q$  as equivalent to  $p(x) \propto q(x)$  where  $q(x)$  is some non-normalised probability density function, we write the priors for  $\boldsymbol{\theta}_i$  as,

$$\begin{aligned} f_{\text{fevol},i} &\sim 0.01 + 1.99 \cdot \mathcal{B}(1.2, 1.2), \\ M_i &\sim 0.8 + 0.4 \cdot \mathcal{B}(1.2, 1.2), \\ \alpha_{\text{MLT},i} &\sim 1.5 + \mathcal{B}(1.2, 1.2), \\ Y_{\text{init},i} &\sim 0.24 + 0.08 \cdot \mathcal{B}(1.2, 1.2), \\ Z_{\text{init},i} &\sim 0.005 + 0.035 \cdot \mathcal{B}(1.2, 1.2), \end{aligned}$$

where each beta distribution is scaled to cover the boundaries of the grid of stellar models computed in Section 3.1.

We made predictions for each star using the trained ANN,  $\{\log(\tau)_i, T_{\text{eff},i}, R_i, \Delta\nu_i, [\text{M}/\text{H}]_{\text{surf},i}\} = f_{\text{ANN}}(\boldsymbol{\theta}_i)$ , from which we derived the luminosity,  $L_i$  using the Stefan-Boltzmann law. Any of the model parameters may be passed as an observable. Hereafter, we denote the set of model observables as  $\boldsymbol{\mu}_i = f(\boldsymbol{\theta}_i)$ . Thus, we write the likelihood we observe any  $\mathbf{d}_i$  with known uncertainty,  $\sigma_i$  given the model as,

$$p(\mathbf{d}_i | \boldsymbol{\theta}_i) = \prod_{k=1}^{N_{\text{obs}}} \frac{1}{\sigma_{k,i} \sqrt{2\pi}} \exp \left[ -\frac{(d_{k,i} - \mu_{k,i})^2}{2\sigma_{k,i}^2} \right], \quad (10)$$

where  $N_{\text{obs}}$  is the number of observed variables. We chose to use observed  $T_{\text{eff}}$ ,  $L$ ,  $\Delta\nu$  and  $[\text{M}/\text{H}]$  collated for our sample as described in Section 2.

It follows that the posterior PDF for a population of  $N_{\text{stars}}$  stars for the NP model is,

$$p(\boldsymbol{\Theta} | \mathbf{D}) = \prod_{i=1}^{N_{\text{stars}}} p(\boldsymbol{\theta}_i | \mathbf{d}_i), \quad (11)$$

where  $\boldsymbol{\Theta}$  is the matrix of model parameters and  $\mathbf{D}$  is the matrix of observables. A probabilistic graphical model (PGM) of the NP model can be seen inside the grey box of Figure 4, without the arrow connecting  $Z_{\text{init}}$  to  $Y_{\text{init}}$ . We ignore the nodes outside the box because these correspond to the the PP model described next.

### 3.3.2 Partial-pooled model

Sharing, or pooling parameters between stars in a population can improve the uncertainties on stellar fundamentals by encoding our prior knowledge of their distribution in a population. We constructed a hierarchical model, which builds upon the NP model by introducing population-level *hyperparameters*. Specifically, we chose to describe initial helium and  $\alpha_{\text{MLT}}$  by partially-pooling them.

We constructed the PP model such that each of the initial helium,  $Y_{\text{init}}$  and mixing-length theory parameter,  $\alpha_{\text{MLT}}$  are drawn

from a common distribution characterised by the set of hyperparameters,  $\phi$ . Thus, Bayes' theorem becomes,

$$p(\phi, \Theta | D) \propto p(\phi) p(Y_{\text{init}}, \alpha_{\text{MLT}} | \phi) p(f_{\text{evol}}, M, Z) p(D | \Theta), \quad (12)$$

where  $\Theta$  is the same as in the NP model, i.e. each object-level parameter,  $\theta_j = \{\theta_{j,i}\}_{i=1}^{N_{\text{stars}}}$ , and  $\phi = \{\Delta Y / \Delta Z, Y_P, \sigma_Y, \mu_\alpha, \sigma_\alpha\}$ . The hyperparameters for  $Y_{\text{init}}$  comprise the helium enrichment ratio,  $\Delta Y / \Delta Z$ , primordial helium abundance fraction,  $Y_P$  and the spread in helium,  $\sigma_Y$ . The remaining hyperparameters for  $\alpha_{\text{MLT}}$  comprise the mean,  $\mu_\alpha$  and spread,  $\sigma_\alpha$ .

We assumed the initial helium and the mixing-length parameter are each drawn from a normal distribution characterised by a population mean and standard deviation. The probability of  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$  given  $\phi$  is,

$$p(Y_{\text{init}}, \alpha_{\text{MLT}} | \phi) = p(Y_{\text{init}} | \mu_Y, \sigma_Y) p(\alpha_{\text{MLT}} | \mu_\alpha, \sigma_\alpha), \quad (13)$$

where  $\mu_Y$  and  $\sigma_Y$  is the mean initial helium fraction as described by the linear helium enrichment law,

$$\mu_Y = Y_P + \frac{\Delta Y}{\Delta Z} Z_{\text{init}}. \quad (14)$$

Therefore, we may write the prior PDF of initial helium given its population-level hyperparameters as,

$$p(Y_{\text{init}} | Z_{\text{init}}, \Delta Y / \Delta Z, Y_P, \sigma_Y) = \prod_{i=1}^{N_{\text{stars}}} \mathcal{N}(Y_{\text{init},i} | \mu_{Y,i}, \sigma_Y). \quad (15)$$

Similarly, for the second term of Equation 13, we chose to partially-pool the mixing-length parameter. We assume that convection in stars of a similar mass, evolutionary stage and area of the HR diagram may be approximated using a similar value of  $\alpha_{\text{MLT}}$ , but the accuracy of the mixing-length theory may vary from star-to-star. There is theoretical evidence for such a variation with [M/H],  $T_{\text{eff}}$  and  $\log g$  in 3D hydrodynamical stellar models (Magic et al. 2015; Viani et al. 2018). However, investigating such dependencies are beyond the scope of this paper. Given the small range of our sample, any such variation will be absorbed by the spread parameter,  $\sigma_\alpha$ . Therefore, we decided to describe the prior on  $\alpha_{\text{MLT}}$  as,

$$p(\alpha_{\text{MLT}} | \mu_\alpha, \sigma_\alpha) = \prod_{i=1}^{N_{\text{stars}}} \mathcal{N}(\alpha_{\text{MLT},i} | \mu_\alpha, \sigma_\alpha) \quad (16)$$

We gave all of the hyperparameters weakly informative priors, with the exception of  $Y_P$  for which we adopt a recent measurement of the primordial helium abundance from big bang nucleosynthesis (BBN) as the mean (Pitrou et al. 2018), with a standard deviation representative of the range of values in the literature (Aver et al. 2015; Peimbert et al. 2016; Cooke & Fumagalli 2018). Hence, we assumed priors on the hyperparameters as follows,

$$\Delta Y / \Delta Z \sim 4.0 \cdot \mathcal{B}(1.2, 1.2),$$

$$Y_P \sim \mathcal{N}(0.247, 0.001),$$

$$\sigma_Y \sim \mathcal{LN}(0.01, 1.0),$$

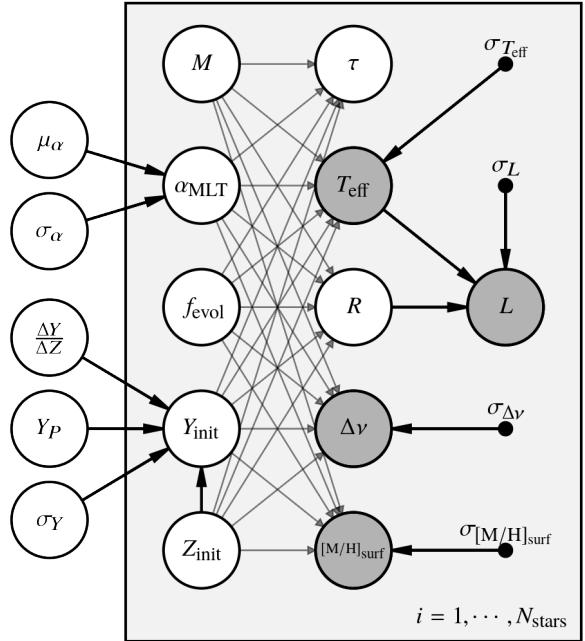
$$\mu_\alpha \sim 1.5 + \mathcal{B}(1.2, 1.2),$$

$$\sigma_\alpha \sim \mathcal{LN}(0.1, 1.0),$$

where,  $x \sim \mathcal{LN}(m, \sigma)$  represents a random variable drawn from the log-normal distribution,

$$\mathcal{LN}(x | m, \sigma) = \frac{1}{x \sigma \sqrt{2\pi}} \exp \left[ -\frac{\ln(x/m)^2}{2\sigma^2} \right]. \quad (17)$$

We produced a PGM for the model, depicted in Figure 4. The



**Figure 4.** A probabilistic graphical model (PGM) of the partially-pooled (PP) hierarchical model. Nodes outside of the grey rectangle represent the hyperparameters in the model. Nodes inside the grey rectangle represent individual stellar parameters. Dark grey nodes represent observables which each have their respective observational uncertainties given by the solid black nodes. The direction of the arrows represent the dependencies in the generative model.

hyperparameters are shown outside of the grey box containing the individual stellar parameters to represent the hierarchical aspect of the model.

### 3.3.3 Max-pooled model

We built another hierarchical model similar to the PP model except that  $\alpha_{\text{MLT}}$  is max-pooled. In other words, we assumed that the mixing length must be the same value for every star in the sample, but still allowed it to freely vary. Thus the hyperparameters are now,  $\phi = \{\Delta Y / \Delta Z, Y_P, \sigma_Y, \alpha_{\text{MLT}}\}$ . The posterior distribution of the model takes the same form as in Equation 12 except that the mixing-length theory parameter for the  $i$ -th star is,

$$\alpha_{\text{MLT},i} = \alpha_{\text{MLT}}, \quad (18)$$

where,

$$\alpha_{\text{MLT}} \sim 1.5 + \mathcal{B}(1.2, 1.2). \quad (19)$$

chosen such that  $\alpha_{\text{MLT}}$  is confined to the boundaries of the grid of stellar models ( $1.5 < \alpha_{\text{MLT}} < 2.5$ ). In other words,  $\alpha_{\text{MLT}}$  is a free parameter in the model but is assumed to be the same in all stars.

### 3.4 Solar calibrator

Pooling parameters in an HBM allows us to use the Sun as a calibrator in a unique way. Rather than calibrating our model physics to the Sun and then assuming the calibrated parameters across our sample, we can include the Sun as a part of the same population. In other words, if we assume  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$  for the Sun are a part of

**Table 4.** Solar input data where  $\mu$  and  $\sigma$  are the means and standard deviations respectively.

Input	$\mu$	$\sigma$	Reference
$M$ ( $M_\odot$ )	1.000	0.001	—
$\tau$ (Gyr)	4.6	0.1	Connelly et al. (2012)
$T_{\text{eff}}$ (K)	5777	20	Scott et al. (2015)
$R$ ( $R_\odot$ )	1.000	0.001	—
$L$ ( $L_\odot$ )	1.00	0.01	—
$\Delta\nu$ (μHz)	135.1	0.2	Huber et al. (2011)
[M/H]surf (dex)	0.00	0.01	Asplund et al. (2009)

the same prior distribution as for the sample, then we can add Solar data to our model inputs.

For both the PP and MP models, we ran versions with and without data for the Sun included in the APOKASC population, referred to as PPS and MPS respectively. We adopted standard solar data as shown in Table 4 with uncertainties conservatively limited to the accuracy of the ANN for  $R$ ,  $L$  and representative of variation in the literature for  $T_{\text{eff}}$ . We also adopted  $\Delta\nu = 135.1 \pm 0.2$  μHz from Huber et al. (2011) with a standard deviation representative of variations in measurements of the solar  $\Delta\nu$  (Broomhall et al. 2011).

### 3.5 Sampling

We obtained results for each of the models described above by sampling their posterior distributions using a Markov chain Monte Carlo (MCMC) algorithm. In particular, we used the NUTS algorithm implemented in TENSORFLOW PROBABILITY (TFP; Abadi et al. 2016; Dillon et al. 2017)<sup>1</sup>. For each model, we produced 20,000 samples split across 10 MCMC chains and computed summary statistics for the marginalised posteriors of each parameter in the model. We removed stars with problems during tuning using the Gelman-Rubin diagnostic ( $\hat{r}$ ; Gelman & Rubin 1992). We used results from each model once  $\hat{r} < 1.04$  for all parameters, indicating good model convergence.

Initially, we created a random synthetic population of stars using MESA to test the ability of the method to recover stellar properties according to our choice of model physics and population priors. We tested the NP, PP and MP models. Since our sample was fictitious, it would not have been appropriate to include real solar data. We summarise the results for the synthetic stellar parameters and hyperparameters in Appendix B. We found that the models were able to recover the true synthetic properties accurately, with increased precision when pooling parameters.

Once we had tested the method on synthetic stars, we obtained results for the sample of 81 dwarfs and subgiants described in Section 2. Here, we included the PPS and MPS to test the effects of adding the Sun as a star in our population. For the purpose of comparison, we fit the hyperparameters of the PP model ( $\Delta Y/\Delta Z, Y_P, \sigma_Y, \mu_\alpha, \sigma_\alpha$ ) to the results from the NP model. We show the results from these models in the following section.

<sup>1</sup> We interacted with the TFP using the now deprecated PyMC4 package, developed as a successor to PyMC3 (Salvatier et al. 2016)

## 4 RESULTS

In this section, we present the results for each of the NP, PP and MP models ran with the sample of 81 APOKASC dwarfs and subgiants as inputs. We also present the results for the PPS and MPS models ran with Solar data as a calibrator. Firstly, we show the reduction in age, mass and radius uncertainty with the addition of pooling in Section 4.1. We then show the results for model hyperparameters in Section 4.2 where we infer the initial helium abundance and mixing-length parameter distribution in the sample.

### 4.1 Stellar parameter results

During an initial run of the NP model, we dropped 16 of the 81 stars from the sample. Of the removed stars, we found the posteriors in  $M$  for 6 skewed towards the prior upper mass limit of  $1.2 M_\odot$ . The remaining 10 removed stars suffered poor convergence during sampling ( $\hat{r} >> 1.04$ ) which could be because of poor step-size tuning and sampling at the prior boundary.

In Table 5, we preset results for the remaining 65 stars from the NP model. Running the NP model with synthetic stars resulted in unreliable uncertainties (see Appendix B). This was because the boundary of the priors in  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$  truncated the posterior distribution leading to underestimated uncertainties and skewing their posterior means towards the centre of their priors. Therefore, we present the NP results only for comparison purposes, but we exclude them from further discussion.

Out of the 65 stars with results from the NP model, 2 stars were dropped from the PP model with problems during sampling. A consequence of partially pooling parameters is that a population spread,  $\sigma$  allows for individual parameters to vary outside of the prior range given to the population mean,  $\mu$ . In this case, individual stellar  $\alpha_{\text{MLT}}$  was allowed to vary outside the range for which the ANN was valid if  $\sigma_\alpha$  was large. The 2 removed stars happened to have high likelihoods outside of the valid  $\alpha_{\text{MLT}}$  range. We found that removing the same 2 stars from the other models made negligible difference to the results, so we leave a solution to this problem to future work. In Tables 6 and 7 we present the results for the remaining 63 stars from the PP and PPS model respectively.

When we max-pooled  $\alpha_{\text{MLT}}$  we did not see the same issue with the 2 removed stars from the PP model because we did not include any population spread. Therefore, we kept these stars in the sample and present results for 65 stars in Tables 8 and 9 for the MP and MPS models respectively.

Figure 5 compares the uncertainties in mass, radius and age for each of the models in this work, with those of the results for the same stars from S17. We saw a similar improvement in uncertainty between the NP and pooled models as with the synthetic stars in Figure B1. We found that the statistical uncertainties in mass from the pooled models were reduced by a factor of  $\sim 2$  over S17 with a median of 2.5 per cent. We also obtained smaller uncertainties in radius and age of 1.2 and 12 per cent respectively when pooling the stellar parameters.

### 4.2 Population parameter results

We obtained values for the hyperparameters for each of the models and present them in Table 10 along with their upper and lower 68 per cent credible regions. We omitted the results for  $Y_P$  because its posterior was the same as the prior,  $Y_P = 0.247 \pm 0.001$  for all the models. We fit the same hyperparameters from the PP model to the NP model results for the purpose of comparison. However, the NP

**Table 5.** The median of the marginalised posterior samples for each parameter output by the NP model, with their respective upper and lower 68 per cent credible intervals. For the full table, see online.

Name	$f_{\text{evol}}$	$M (\text{M}_\odot)$	$\alpha_{\text{MLT}}$	$Y_{\text{init}}$	$Z_{\text{init}}$	$\tau (\text{Gyr})$	$T_{\text{eff}} (\text{K})$	$R (\text{R}_\odot)$	$\Delta\nu (\mu\text{Hz})$	[M/H] <sub>surf</sub> (dex)
KIC10079226	$0.44^{+0.16}_{-0.20}$	$1.14^{+0.04}_{-0.04}$	$2.07^{+0.26}_{-0.30}$	$0.28^{+0.02}_{-0.02}$	$0.021^{+0.003}_{-0.003}$	$2.5^{+1.2}_{-1.3}$	$5990^{+51}_{-52}$	$1.16^{+0.01}_{-0.02}$	$116.0^{+0.7}_{-0.7}$	$0.16^{+0.07}_{-0.07}$
KIC10215584	$0.50^{+0.21}_{-0.21}$	$1.13^{+0.04}_{-0.05}$	$1.92^{+0.33}_{-0.26}$	$0.27^{+0.02}_{-0.02}$	$0.018^{+0.002}_{-0.002}$	$2.9^{+1.6}_{-1.3}$	$5949^{+64}_{-65}$	$1.18^{+0.02}_{-0.02}$	$112.5^{+2.6}_{-2.6}$	$0.08^{+0.06}_{-0.06}$
KIC10319352	$1.51^{+0.15}_{-0.28}$	$1.09^{+0.05}_{-0.05}$	$1.87^{+0.33}_{-0.23}$	$0.28^{+0.03}_{-0.02}$	$0.028^{+0.004}_{-0.003}$	$9.6^{+1.7}_{-1.5}$	$5507^{+57}_{-56}$	$1.49^{+0.03}_{-0.03}$	$78.6^{+1.6}_{-1.6}$	$0.28^{+0.06}_{-0.06}$
KIC10322381	$0.98^{+0.23}_{-0.28}$	$1.07^{+0.07}_{-0.07}$	$2.04^{+0.29}_{-0.31}$	$0.28^{+0.02}_{-0.02}$	$0.010^{+0.002}_{-0.002}$	$4.7^{+1.5}_{-1.7}$	$6132^{+94}_{-94}$	$1.39^{+0.05}_{-0.04}$	$85.6^{+4.9}_{-4.6}$	$-0.30^{+0.07}_{-0.08}$
KIC10732098	$1.60^{+0.14}_{-0.19}$	$1.12^{+0.04}_{-0.05}$	$1.86^{+0.33}_{-0.23}$	$0.28^{+0.02}_{-0.02}$	$0.017^{+0.002}_{-0.002}$	$6.7^{+0.8}_{-0.8}$	$5720^{+67}_{-66}$	$1.77^{+0.04}_{-0.04}$	$62.1^{+1.8}_{-1.7}$	$0.06^{+0.07}_{-0.07}$

**Table 6.** The same as Table 5, but for the PP model.

Name	$f_{\text{evol}}$	$M (\text{M}_\odot)$	$\alpha_{\text{MLT}}$	$Y_{\text{init}}$	$Z_{\text{init}}$	$\tau (\text{Gyr})$	$T_{\text{eff}} (\text{K})$	$R (\text{R}_\odot)$	$\Delta\nu (\mu\text{Hz})$	[M/H] <sub>surf</sub> (dex)
KIC10079226	$0.22^{+0.10}_{-0.09}$	$1.16^{+0.02}_{-0.03}$	$1.75^{+0.11}_{-0.09}$	$0.28^{+0.01}_{-0.01}$	$0.020^{+0.003}_{-0.002}$	$1.2^{+0.6}_{-0.5}$	$5962^{+44}_{-42}$	$1.17^{+0.01}_{-0.01}$	$115.9^{+0.7}_{-0.7}$	$0.15^{+0.07}_{-0.07}$
KIC10215584	$0.37^{+0.15}_{-0.13}$	$1.14^{+0.03}_{-0.03}$	$1.74^{+0.10}_{-0.09}$	$0.27^{+0.01}_{-0.01}$	$0.018^{+0.002}_{-0.002}$	$2.1^{+1.0}_{-0.8}$	$5941^{+57}_{-56}$	$1.18^{+0.02}_{-0.02}$	$112.5^{+2.6}_{-2.7}$	$0.07^{+0.07}_{-0.07}$
KIC10319352	$1.41^{+0.11}_{-0.27}$	$1.08^{+0.03}_{-0.03}$	$1.73^{+0.10}_{-0.09}$	$0.29^{+0.02}_{-0.01}$	$0.028^{+0.004}_{-0.004}$	$8.6^{+1.1}_{-1.0}$	$5512^{+45}_{-46}$	$1.49^{+0.02}_{-0.02}$	$78.6^{+1.6}_{-1.6}$	$0.28^{+0.06}_{-0.07}$
KIC10322381	$0.78^{+0.23}_{-0.19}$	$1.14^{+0.03}_{-0.06}$	$1.75^{+0.10}_{-0.09}$	$0.26^{+0.01}_{-0.01}$	$0.011^{+0.002}_{-0.002}$	$3.6^{+1.7}_{-1.1}$	$6081^{+95}_{-92}$	$1.41^{+0.05}_{-0.05}$	$86.2^{+4.8}_{-5.2}$	$-0.31^{+0.07}_{-0.07}$
KIC10732098	$1.50^{+0.13}_{-0.14}$	$1.14^{+0.03}_{-0.04}$	$1.74^{+0.10}_{-0.09}$	$0.28^{+0.01}_{-0.01}$	$0.018^{+0.002}_{-0.002}$	$6.4^{+0.6}_{-0.6}$	$5701^{+59}_{-58}$	$1.78^{+0.03}_{-0.03}$	$62.2^{+1.7}_{-1.7}$	$0.06^{+0.06}_{-0.06}$

**Table 7.** The same as Table 5, but for the PPS model.

Name	$f_{\text{evol}}$	$M (\text{M}_\odot)$	$\alpha_{\text{MLT}}$	$Y_{\text{init}}$	$Z_{\text{init}}$	$\tau (\text{Gyr})$	$T_{\text{eff}} (\text{K})$	$R (\text{R}_\odot)$	$\Delta\nu (\mu\text{Hz})$	[M/H] <sub>surf</sub> (dex)
KIC10079226	$0.35^{+0.11}_{-0.12}$	$1.17^{+0.02}_{-0.03}$	$1.95^{+0.14}_{-0.15}$	$0.27^{+0.01}_{-0.01}$	$0.020^{+0.003}_{-0.002}$	$2.1^{+0.8}_{-0.8}$	$5962^{+44}_{-43}$	$1.17^{+0.01}_{-0.01}$	$116.0^{+0.7}_{-0.7}$	$0.15^{+0.06}_{-0.07}$
KIC10215584	$0.47^{+0.16}_{-0.16}$	$1.14^{+0.03}_{-0.03}$	$1.90^{+0.15}_{-0.17}$	$0.27^{+0.01}_{-0.01}$	$0.018^{+0.002}_{-0.002}$	$2.7^{+1.2}_{-1.1}$	$5943^{+56}_{-58}$	$1.18^{+0.02}_{-0.02}$	$112.6^{+2.6}_{-2.6}$	$0.07^{+0.06}_{-0.07}$
KIC10319352	$1.51^{+0.10}_{-0.22}$	$1.09^{+0.03}_{-0.03}$	$1.88^{+0.16}_{-0.16}$	$0.28^{+0.01}_{-0.01}$	$0.028^{+0.004}_{-0.004}$	$9.6^{+1.1}_{-1.2}$	$5507^{+47}_{-48}$	$1.49^{+0.02}_{-0.02}$	$78.6^{+1.6}_{-1.6}$	$0.28^{+0.06}_{-0.06}$
KIC10322381	$0.89^{+0.21}_{-0.22}$	$1.12^{+0.05}_{-0.06}$	$1.93^{+0.15}_{-0.16}$	$0.26^{+0.01}_{-0.01}$	$0.010^{+0.002}_{-0.002}$	$4.3^{+1.7}_{-1.2}$	$6093^{+92}_{-89}$	$1.41^{+0.04}_{-0.04}$	$86.1^{+5.0}_{-4.9}$	$-0.31^{+0.07}_{-0.08}$
KIC10732098	$1.60^{+0.11}_{-0.14}$	$1.14^{+0.03}_{-0.04}$	$1.90^{+0.15}_{-0.17}$	$0.27^{+0.01}_{-0.01}$	$0.017^{+0.002}_{-0.002}$	$6.9^{+0.6}_{-0.6}$	$5704^{+62}_{-61}$	$1.78^{+0.04}_{-0.03}$	$62.2^{+1.8}_{-1.7}$	$0.06^{+0.06}_{-0.06}$

model suffered from boundary effects which made the resulting fit unreliable, so it is left here only for comparison.

Figure 6 shows the joint and marginal distributions (corner plot) output by the PP and PPS model. We saw an anti-correlation between  $\Delta Y/\Delta Z$  and  $\mu_\alpha$ , expected due to the degeneracy between the two parameters in the stellar evolutionary models. In Figure 7, we also show the corner plot for the MP and MPS model output. Similarly, we see an anti-correlation between  $\Delta Y/\Delta Z$  and  $\alpha_{\text{MLT}}$ .

We present 100 random samples from the posterior for the helium enrichment relation from the PPS model in Figure 8. In this figure, we also plot the individual results for  $Y_{\text{init}}$  and  $Z_{\text{init}}$  for each of the stars in the sample. This is an example of *shrinkage* in the HBM; the estimates for individual stellar parameters move towards the mean of the population.

## 5 DISCUSSION

So far, we have shown that we can add parameters to stellar models without sacrificing statistical uncertainties through the application of an HBM. We freed the  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$  using pooling to encode our prior knowledge of their distribution in the population. We also tested the impact of including the Sun in our population as a calibrator. We first discuss the impact of pooling and our choice of population priors for  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$  in Section 5.1 and 5.2.

To assess the accuracy of our model with respect to the literature, we compare our results to those of S17 in Section 5.3. We found good agreement between this work and their results, despite some differences in observables and stellar model physics which we discuss further.

Finally, we discuss sources of systematic uncertainties in Section 5.4. Although we have accounted for uncertainties in 5.1 and 5.2 in our model, there are still differences between stellar modelling codes and other model physics which should be considered. In Section 5.5, we highlight a possible outlier in our dataset. Then, we discuss the future scalability of this method in Section 5.6.

### 5.1 Helium enrichment

We found the value for the helium enrichment ratio,  $\Delta Y/\Delta Z$  to be the same in both the PP and MP models,  $\Delta Y/\Delta Z = 1.6^{+0.5}_{-0.4}$ . This is consistent with values of  $\sim 1.4$  in the literature for stellar models which include heavy element diffusion (Broggaard et al. 2012; Verma et al. 2019).

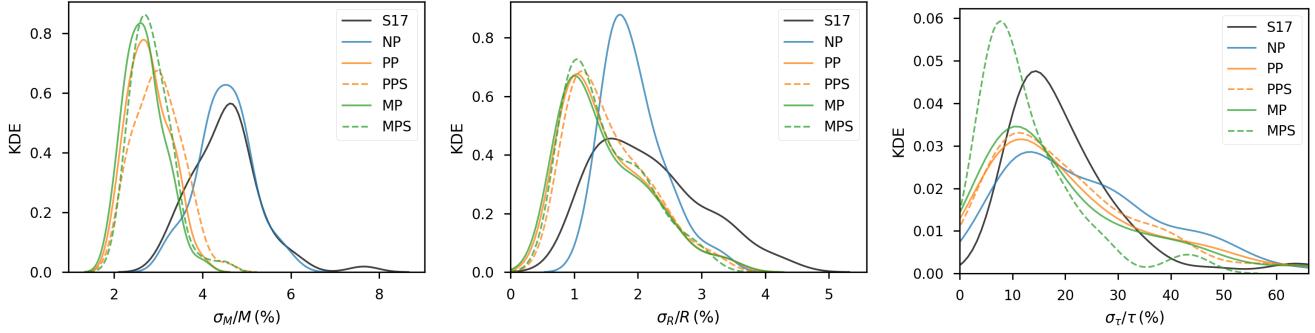
When we added the Sun to the pooled models, PPS and MPS, obtained  $\Delta Y/\Delta Z$  up to  $2\sigma$  lower than the models without the Sun. In both models, the resulting  $\Delta Y/\Delta Z$  of approximately 0.8 to 1.0 was consistent with the initial helium fraction expected from solar models with our choice of Asplund et al. (2009) abundances (Serenelli & Basu 2010). However, such solar models have been shown to not recover helioseismic measurements of helium in the Sun (Basu & Antia 2004; Serenelli et al. 2009; Villante et al. 2014). Solar models with the older Grevesse & Sauval (1998) abundances typically yield higher helium fractions more in-line with helioseismology. Interestingly, the  $\Delta Y/\Delta Z$  from the PP and MP models are higher than those including the Sun, in closer agreement with those calibrated with the older abundances. We could extend our model to include asteroseismic indicators of helium to improve the uncertainties on  $Y$  and test whether this difference becomes more significant.

**Table 8.** The same as Table 5, but for the MP model.

Name	$f_{\text{evol}}$	$M (\text{M}_\odot)$	$Y_{\text{init}}$	$Z_{\text{init}}$	$\tau$ (Gyr)	$T_{\text{eff}}$ (K)	$R (\text{R}_\odot)$	$\Delta\nu$ ( $\mu\text{Hz}$ )	[M/H]surf (dex)
KIC10079226	$0.20^{+0.08}_{-0.08}$	$1.17^{+0.02}_{-0.03}$	$0.28^{+0.01}_{-0.01}$	$0.019^{+0.003}_{-0.002}$	$1.1^{+0.5}_{-0.4}$	$5961^{+42}_{-41}$	$1.17^{+0.01}_{-0.01}$	$115.9^{+0.7}_{-0.7}$	$0.15^{+0.06}_{-0.07}$
KIC10215584	$0.36^{+0.14}_{-0.13}$	$1.14^{+0.03}_{-0.03}$	$0.27^{+0.01}_{-0.01}$	$0.018^{+0.002}_{-0.002}$	$2.0^{+0.9}_{-0.8}$	$5941^{+57}_{-57}$	$1.18^{+0.02}_{-0.02}$	$112.5^{+2.6}_{-2.7}$	$0.07^{+0.06}_{-0.07}$
KIC10319352	$1.41^{+0.10}_{-0.25}$	$1.08^{+0.03}_{-0.03}$	$0.29^{+0.02}_{-0.01}$	$0.028^{+0.004}_{-0.004}$	$8.6^{+1.0}_{-0.9}$	$5512^{+44}_{-45}$	$1.49^{+0.02}_{-0.02}$	$78.6^{+1.7}_{-1.6}$	$0.28^{+0.06}_{-0.07}$
KIC10322381	$0.77^{+0.23}_{-0.19}$	$1.14^{+0.04}_{-0.06}$	$0.27^{+0.01}_{-0.01}$	$0.011^{+0.002}_{-0.002}$	$3.5^{+1.6}_{-1.0}$	$6076^{+96}_{-91}$	$1.41^{+0.05}_{-0.05}$	$86.1^{+4.7}_{-5.3}$	$-0.32^{+0.07}_{-0.07}$
KIC10732098	$1.50^{+0.13}_{-0.13}$	$1.14^{+0.03}_{-0.04}$	$0.28^{+0.01}_{-0.01}$	$0.018^{+0.002}_{-0.002}$	$6.4^{+0.6}_{-0.6}$	$5702^{+56}_{-58}$	$1.78^{+0.03}_{-0.03}$	$62.2^{+1.7}_{-1.7}$	$0.06^{+0.06}_{-0.06}$

**Table 9.** The same as Table 5, but for the MPS model.

Name	$f_{\text{evol}}$	$M (\text{M}_\odot)$	$Y_{\text{init}}$	$Z_{\text{init}}$	$\tau$ (Gyr)	$T_{\text{eff}}$ (K)	$R (\text{R}_\odot)$	$\Delta\nu$ ( $\mu\text{Hz}$ )	[M/H]surf (dex)
KIC10079226	$0.44^{+0.07}_{-0.06}$	$1.16^{+0.02}_{-0.03}$	$0.26^{+0.01}_{-0.01}$	$0.021^{+0.003}_{-0.002}$	$2.7^{+0.5}_{-0.4}$	$5965^{+40}_{-40}$	$1.17^{+0.01}_{-0.01}$	$116.0^{+0.7}_{-0.7}$	$0.15^{+0.06}_{-0.06}$
KIC10215584	$0.59^{+0.11}_{-0.13}$	$1.13^{+0.03}_{-0.03}$	$0.26^{+0.01}_{-0.01}$	$0.018^{+0.002}_{-0.002}$	$3.6^{+0.9}_{-0.9}$	$5952^{+55}_{-56}$	$1.18^{+0.02}_{-0.02}$	$112.7^{+2.7}_{-2.7}$	$0.08^{+0.06}_{-0.07}$
KIC10319352	$1.61^{+0.04}_{-0.06}$	$1.08^{+0.03}_{-0.03}$	$0.27^{+0.01}_{-0.01}$	$0.028^{+0.004}_{-0.003}$	$10.8^{+0.7}_{-0.8}$	$5516^{+46}_{-47}$	$1.49^{+0.02}_{-0.02}$	$78.6^{+1.7}_{-1.6}$	$0.28^{+0.07}_{-0.06}$
KIC10322381	$0.98^{+0.19}_{-0.20}$	$1.10^{+0.06}_{-0.05}$	$0.26^{+0.01}_{-0.01}$	$0.010^{+0.002}_{-0.001}$	$5.1^{+1.3}_{-1.5}$	$6106^{+94}_{-80}$	$1.40^{+0.04}_{-0.04}$	$85.8^{+5.6}_{-4.3}$	$-0.30^{+0.08}_{-0.08}$
KIC10732098	$1.69^{+0.06}_{-0.09}$	$1.14^{+0.03}_{-0.04}$	$0.26^{+0.01}_{-0.01}$	$0.017^{+0.002}_{-0.002}$	$7.4^{+0.5}_{-0.5}$	$5715^{+61}_{-61}$	$1.77^{+0.04}_{-0.03}$	$62.3^{+1.8}_{-1.8}$	$0.07^{+0.06}_{-0.07}$


**Figure 5.** Kernel density estimates (KDEs) of the uncertainties in the results from each model compared with that of (S17) for the sample of APOKASC dwarfs and subgiants.

We assumed a linear helium enrichment law dependent only on the initial heavy element abundance,  $Z_{\text{init}}$ . However, helium enrichment could vary non-linearly depending on other chemical abundances (West & Heger 2013) or the location of the star in the Milky Way (Frebel 2010). Our model has the advantage of being adaptable to different population priors, stellar inputs and outputs. Future work will explore the helium enrichment relation further, with the inclusion of metal-poor stars and a dependence on different chemical abundances.

Our models assumed a prior of  $Y_P = 0.247 \pm 0.001$  for the primordial helium fraction which dominated its posterior. This was a sensible assumption to make when using a linear enrichment law, because measurements of the primordial helium correspond to the abundance at the epoch of BBN according to current cosmological theory (Cyburt et al. 2016). However, if we used a less informative prior for  $Y_P$  we might yield more uncertain results, or even a different value for  $Y_P$ . In previous work fitting a linear enrichment law, some results for  $Y_P$  suggested a value below the BBN value (Casagrande et al. 2007; Silva Aguirre et al. 2017). It is more probable that the assumption of a linear enrichment law is inaccurate than a sample of stars could contradict independent  $Y_P$  from cosmology. We justify our prior on  $Y_P$  as in-line with the assumption of a linear enrichment law, but highlight the need to investigate other ways of describing helium in a population of stars.

## 5.2 Mixing-length theory

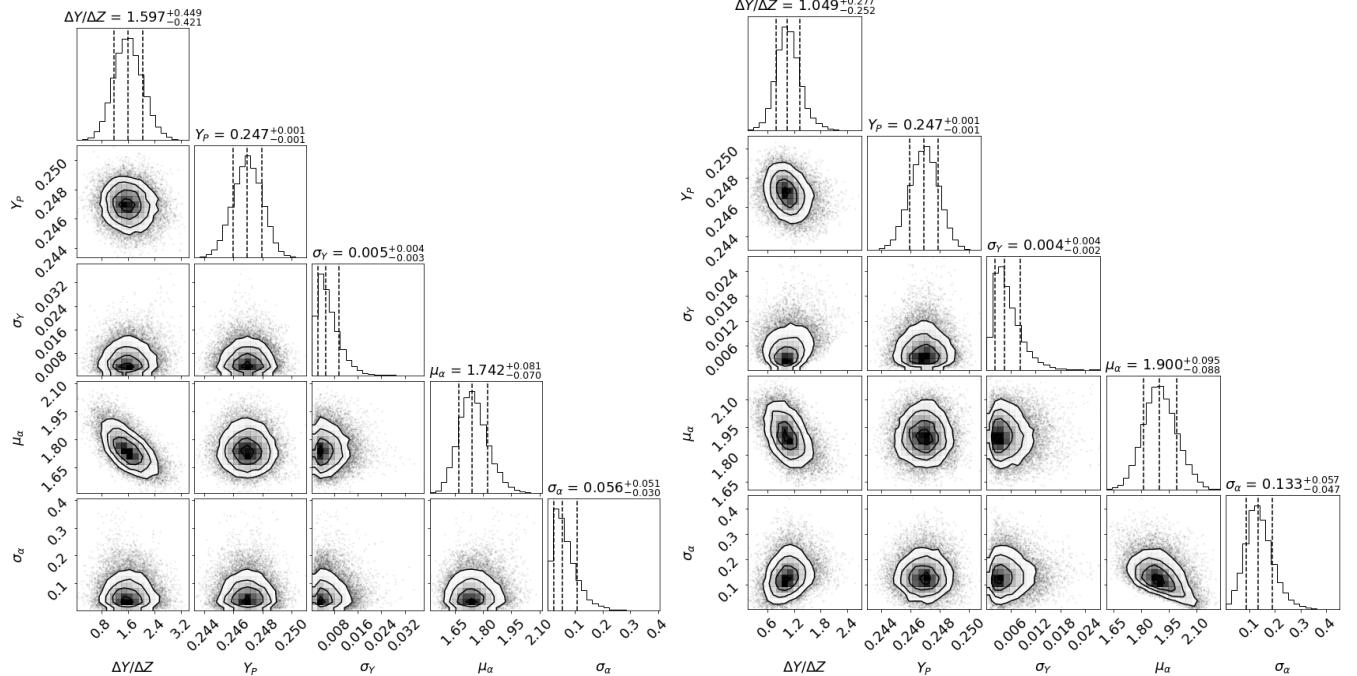
To a greater degree than chemical composition, the best-fitting  $\alpha_{\text{MLT}}$  depends on the choice of model physics and stellar modelling code. The mixing-length theory is an approximation of convection which is often calibrated to the Sun and then assumed for all stars in a model. However, studies of 3D hydrodynamical simulations suggest that the degree in which  $\alpha_{\text{MLT}}$  approximates convection varies across the HR diagram (Magic et al. 2015) and this is confirmed when modelling stars with asteroseismology (Tayar et al. 2017).

Our fit to the NP model results obtained a result for  $\mu_\alpha$  close to 2.0 – the midpoint of its prior distribution. This was expected because the tests on synthetic stars showed that  $\alpha_{\text{MLT}}$  was biased in this way due to boundary effects from truncating the distribution at 1.5 and 2.5. Again, the NP hyperparameter fit is used for comparison, but is ultimately unreliable.

The PP model (without the Sun) favoured a mean mixing-length parameter of  $\mu_\alpha \approx 1.7$ . Whereas, the PPS model yielded a higher value of  $\mu_\alpha \approx 1.9$  by  $\sim 2\sigma$ . We found this was attributed to the addition of the Sun. The individual solar results for the PPS model yielded a value of  $\alpha_{\text{MLT}\odot} = 2.12 \pm 0.03$  which was considerably higher than the  $\alpha_{\text{MLT}}$  obtained for the other stars in the sample (see Appendix C). The solar value also exceeds typical solar calibrated values of  $\sim 1.9$  for the same stellar evolution code (Stancliffe et al. 2016). The reason for this is unclear, except that the difference

**Table 10.** Hyperparameter results for each model with the omission of  $Y_P$ .

Model	$\Delta Y/\Delta Z$	$\sigma_Y$	$\mu_\alpha$	$\sigma_\alpha$	$\alpha_{\text{mlt}}$
NP	$1.69^{+0.21}_{-0.21}$	$0.0074^{+0.0026}_{-0.0022}$	$1.954^{+0.040}_{-0.041}$	$0.065^{+0.030}_{-0.024}$	—
PP	$1.60^{+0.45}_{-0.42}$	$0.0051^{+0.0045}_{-0.0027}$	$1.742^{+0.081}_{-0.070}$	$0.056^{+0.051}_{-0.030}$	—
PPS	$1.05^{+0.28}_{-0.25}$	$0.0045^{+0.0038}_{-0.0023}$	$1.900^{+0.095}_{-0.088}$	$0.133^{+0.057}_{-0.047}$	—
MP	$1.60^{+0.45}_{-0.42}$	$0.0051^{+0.0044}_{-0.0027}$	—	—	$1.728^{+0.077}_{-0.066}$
MPS	$0.76^{+0.24}_{-0.27}$	$0.0049^{+0.0039}_{-0.0025}$	—	—	$2.088^{+0.031}_{-0.029}$

**Figure 6.** Corner plots showing the joint and marginalised sampled posterior distributions for the hyperparameters for the PP (left) and PPS (right) models. The vertical dashed lines give the 16th, 50th and 84th percentiles.

we see between  $\alpha_{\text{MLT}, \odot}$  and the rest of the sample is not unique. Previous work on the LEGACY sample of *Kepler* dwarfs found the best fitting  $\alpha_{\text{MLT}}$  for their sample approximately 90 per cent of the solar calibrated value for one of their pipelines with similar model physics to this work (Silva Aguirre et al. 2017).

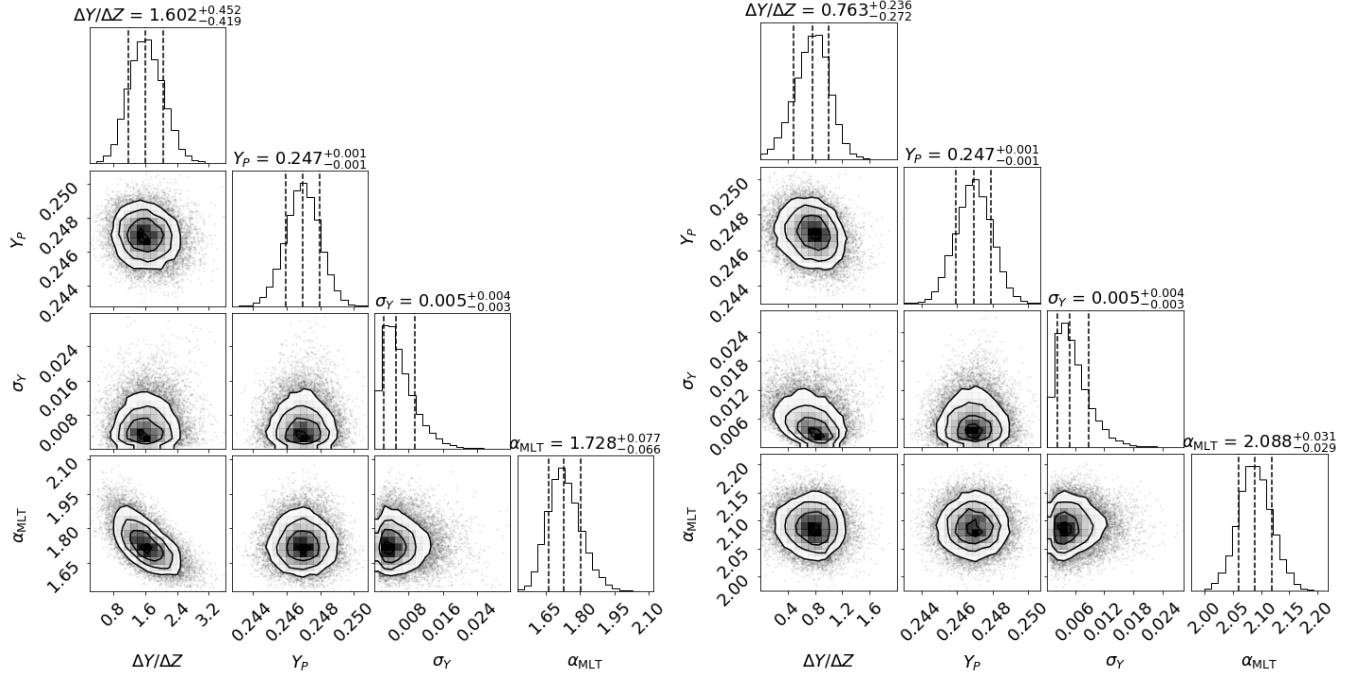
Despite the difference in  $\mu_\alpha$ , the resulting spread in mixing-length for the PPS model  $\sigma_\alpha \approx 0.13$  was double that of the PP model to cope with the high solar value. This implies that a large population spread in  $\alpha_{\text{MLT}}$  could explain the difference we see. In other words, if we assume that the best-fitting  $\alpha_{\text{MLT}}$  is normally distributed in our population, then the Sun lies within  $2\sigma$  of the mean, among 95 per cent of all stars in the population.

There are a few prior studies which look at the spread in  $\alpha_{\text{MLT}}$  for a population of stars, typically by fitting  $\alpha_{\text{MLT}}$  as a function of  $[\text{M}/\text{H}]$ ,  $T_{\text{eff}}$  and  $\log g$ . For example, results from Viani et al. (2018) for stellar models including diffusion, predict  $\alpha_{\text{MLT}}$  in the range 1.5 to 2.3 across our sample. This dispersion would be more compatible with the larger spread obtained by our PPS model. However, in future work we should further investigate how  $\alpha_{\text{MLT}}$  varies with stellar parameters, as our assumption of a normal distribution may not be accurate.

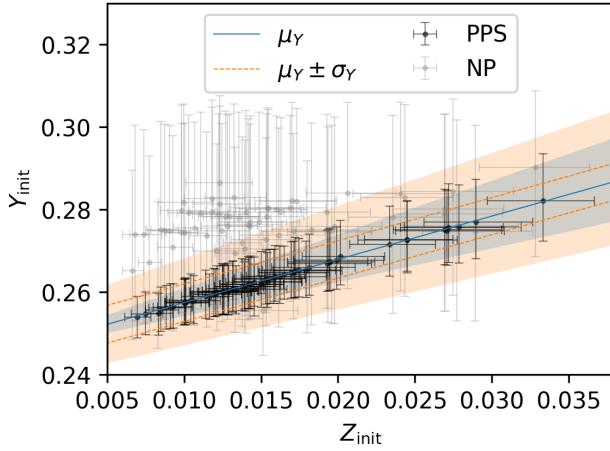
We found a greater difference in  $\alpha_{\text{MLT}}$  between the models with and without the Sun when we max-pooled  $\alpha_{\text{MLT}}$ . The MP

models yielded a global  $\alpha_{\text{MLT}}$  in-line with  $\mu_\alpha$  from the PP model. However, when we added the Sun, the model yielded  $\alpha_{\text{MLT}} \approx 2.1$  which is in common with the solar results (see Appendix C). This had a similar affect as assuming a solar calibrated value, because the model favoured fitting to data with the best observational precision. The change in  $\alpha_{\text{MLT}}$  between the MP and MPS models resulted in a mean difference of  $\sim 20$  per cent between the individual stellar ages. This is an example of how adopting a solar calibrated value can bias stellar ages. We argue that carefully including the Sun as a part of the population with an intrinsic spread is a better way to calibrate the stellar models.

In all observables except for  $L$ , the Sun is near the centre of our distribution of stars. However, we found no relationship between  $L$  and  $\alpha_{\text{MLT}}$  in both our NP and PP models. A possible explanation for the difference in  $\alpha_{\text{MLT}}$  with and without the Sun could be some systematic offset in our observational data for the sample. Here, we point to our choice of spectroscopic  $T_{\text{eff}}$  which typically underestimates  $T_{\text{eff}}$  compared to photometric scales, as noted in S17. There is a positive correlation between  $\alpha_{\text{MLT}}$  and  $T_{\text{eff}}$ , when holding all other variables constant. It is possible that the lower  $\alpha_{\text{MLT}}$  obtained without the Sun as a calibrator could be caused by underestimated effective temperatures.



**Figure 7.** The same as Figure 6 but for the MP (left) and MPS (right) models.



**Figure 8.** The results for initial helium fraction ( $Y_{\text{init}}$ ) against initial heavy-element fraction ( $Z_{\text{init}}$ ) for each star from the PPS model are shown by the black markers. The mean helium enrichment,  $\mu_Y = Y_P + (\Delta Y / \Delta Z) Z_{\text{init}}$  with its 68 per cent credible interval are shown in blue by a solid line and shaded region respectively. The population spread,  $\mu_Y \pm \sigma_Y$  and its 68 per cent credible interval are shown in orange by a dashed line and shaded region respectively. The individual results from the NP model are shown by the light grey markers.

### 5.3 Comparison with APOKASC results

Before we compare our results to S17, we should highlight some key differences between our data and methodology. The results from S17 were determined using a grid-based-modelling technique, which estimates the likelihood across a dense grid of stellar models. They used results from several pipelines to estimate the systematic uncertainties. For the central values of their results, they used the Bayesian stellar algorithm (BASTA; Silva Aguirre et al. 2015) using

a grid computed with GARSTEC (Weiss & Schlattl 2008). Their choice of stellar physics was similar to this work, except for two major differences.

Firstly, the results of S17 were determined using stellar models calculated without heavy-element diffusion. The inclusion of diffusion when modelling the Sun has been commonplace over the last few decades, with good agreement between models and helioseismic observations (Christensen-Dalsgaard et al. 1993; Bahcall et al. 1995). More recent work explored the diffusion in cluster stars (Korn et al. 2007; Önehag et al. 2014) and another demonstrated the impact of including diffusion of stellar ages (Dotter et al. 2017). Our stellar models were computed with heavy-element diffusion. Recently, work by Nsamba et al. (2018) on a similar sample of stars showed models without diffusion compared to those including diffusion can lead to, on average, underestimated radii and masses and overestimated ages by 1, 3 and 16 per cent respectively.

Secondly, our choice of Asplund et al. (2009) solar chemical mixture differs from the Grevesse & Sauval (1998) mixtures adopted by S17. The former leads to a solar heavy-element to hydrogen ratio of  $(Z/X)_\odot = 0.0181$ , and the latter,  $(Z/X)_\odot = 0.0230$ . Typically, Grevesse & Sauval (1998) abundances are favoured in asteroseismic modelling because they are better able to reproduce measurements of helium in the Sun from helioseismology (Serenelli et al. 2009). An effect of using the Asplund et al. (2009) abundances, is that it favours lower  $Z_{\text{init}}$  for a given  $[\text{M}/\text{H}]_{\text{surf}}$ . As a result, models using Grevesse & Sauval (1998) abundances on average underestimate radii and mass compared to those without by about 1 and 0.5 per cent respectively (Nsamba et al. 2018).

Although updated, much of our observable data is comparable to that of S17, with the exception of  $T_{\text{eff}}$ . The preferred results from S17 were determined using a photometric  $T_{\text{eff}}$  scale which we found to be on average  $\sim 170$  K greater than our spectroscopic scale from DR14. In S17, they saw a similar offset between the DR13  $T_{\text{eff}}$  available at the time. They found a median difference in mass, radius and age of approximately  $-6$ ,  $-2$  and  $+35$  per cent

respectively with results from the photometric  $T_{\text{eff}}$  scale subtracted from the spectroscopic scale.

In Figure 5 we compare our statistical uncertainties with results for the equivalent stars from S17 for mass, radius and age. We found that the NP model yielded comparable uncertainties to S17 but note that these are likely underestimated due the influence of the prior boundaries for  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$ . We expected larger uncertainties because we included additional free parameters ( $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$ ) over the work of S17. However, when we treat these parameters hierarchically, we saw a reduction in uncertainties from all of the pooled models. This is because our prior assumptions about the population allows for the sharing of information between the stars. This uncertainty reduction scales with the number of stars in our sample, demonstrated by the results for the synthetic stars in Figure B1. Thus, hierarchically modelling our population resulted in improved statistical uncertainties in stellar fundamental parameters.

In the following subsections, we compare the results between our PPS model with that of S17 with reference to Figure 9. In particular, we compare the mass, radius and age results and find good agreement overall.

### 5.3.1 Mass

In the left panel of Figure 9, we compare the masses obtained by the PPS model with S17 and found a dispersion of around 2 per cent. Our masses were on average 1 per cent above the results from S17. Although we might expect the lower  $T_{\text{eff}}$  scale in this work to underestimate the mass, we attribute this overall effect to our choice of stellar model physics. As previously discussed, the use of Asplund et al. (2009) solar abundances and heavy-element diffusion has the cumulative effect of overestimating stellar masses compared to the physics adopted by S17. We also found that the results from all the pooled models returned similar masses, with or without the Sun.

### 5.3.2 Radius

In the central panel of Figure 9, we show that our radii are similar to S17 with a spread of 1 per cent. We also found radii on average 1 per cent greater than the APOKASC results. Similarly to with mass, this contradicts what would be expected from a lower  $T_{\text{eff}}$  scale and could also be explained by model physics. Our radii also varied little between models with and without the Sun.

### 5.3.3 Age

Our ages were also consistent with those from S17. The right-most panel of Figure 9 shows the spread in the relative age differences to be about 18 per cent, slightly underestimated by 4 per cent. We would expect the lower  $T_{\text{eff}}$  scale to overestimate the ages as found in S17, but instead they are comparable. However, as discussed previously, including diffusion has been shown to reduce age estimates compared to those without. Since we have included diffusion, this could explain the similar ages despite the difference in  $T_{\text{eff}}$  scale.

Including the Sun in our pooled models affected the resulting ages more than mass and radius. Including the Sun typically overestimated the ages compared to models without the Sun. This is expected given the higher  $\alpha_{\text{MLT}}$  for the models including solar data, because a larger mixing-length leads to more efficient nuclear burning and more time spent during the main sequence phase.

## 5.4 Systematic uncertainties

We have already accounted for systematics due to the choice of helium enrichment and mixing-length parameter by marginalising over their uncertainties assuming their population distributions. However, there are other model physics which we have not freely varied, including diffusion and choice of solar mixture. Although our method can be adapted to different stellar evolutionary codes and choice of physics, an in-depth analysis of systematic uncertainties is left to future work.

In previous work studying stars in the APOKASC sample, several pipelines used a range of stellar evolutionary codes and model physics are employed to evaluate systematic uncertainties from the models (Serenelli et al. 2017; Silva Aguirre et al. 2017). Using a hierarchical model in this work enabled us to reduce median statistic uncertainties to 2.5 per cent in mass, 1.2 per cent in radius and 12 per cent in age. The systematic uncertainty analysis of S17 found median systematics of 3, 1 and 13 per cent in mass, radius and age respectively. Reducing statistical uncertainties highlights the importance of understanding systematics.

Other systematics could arise from observational data. For example, we chose the ASPCAP DR14  $T_{\text{eff}}$  scale which was systematically lower than the photometric scale of choice in S17. However, our method was still able to recover similar masses, radii and ages. This could be explained by our choice of stellar model physics, as discussed previously.

## 5.5 Outliers

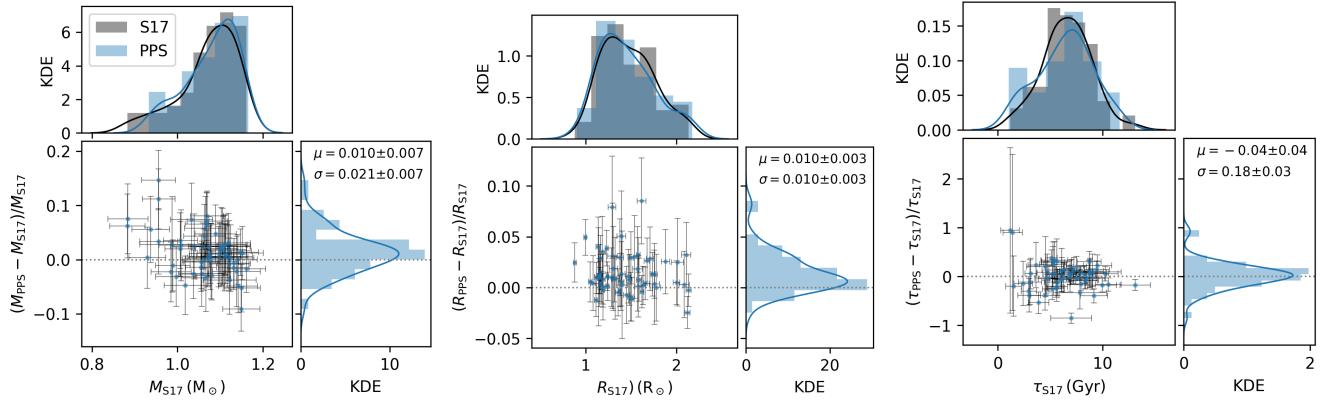
We identified KIC 9025370 as a possible outlier. Consistent across all our models, its output effective temperature,  $T_{\text{eff}} = 5934 \pm 50$  K was about  $4\sigma$  greater than its observed  $T_{\text{eff}}$ , and its modelled  $L$  was about  $2\sigma$  dimmer than its observed luminosity. Only  $\Delta\nu$  and  $[\text{M}/\text{H}]_{\text{surf}}$  were consistent between modelled and observed values. The difference was also apparent in our comparison of ages with S17 where we obtained an age of  $1.5^{+0.7}_{-0.6}$  Gyr compared to their value of  $7.0^{+2.0}_{-1.6}$  Gyr.

KIC 9025370 turned out to be a double-lined spectroscopic binary (Nissen et al. 2017), discovered after S17 and hence included in the original sample. The brighter observed luminosity and possibly unreliable spectroscopic  $T_{\text{eff}}$  compared to our models were consistent with a spectroscopic binary. We calculated a photometric  $T_{\text{eff}}$  using the IRFM method (Casagrande et al. 2010) with the available 2MASS photometry for the target and obtained  $T_{\text{eff}} = 5983 \pm 120$  K, more consistent with our modelled effective temperature and inconsistent with its spectroscopic  $T_{\text{eff}}$ . Thus, our inferred  $T_{\text{eff}}$  was within the dispersion between different observed  $T_{\text{eff}}$  scales. Running the model without KIC 9025370 did not affect the resulting inferred hyperparameters, demonstrating the robustness of our model. Therefore, we present KIC 9025370 in our results but suggest that further investigation should be carried out.

## 5.6 The potential of pooling

Using HBM has allowed us to introduce more free parameters without sacrificing statistical uncertainties. We used an ANN to approximate stellar models, a method which can be extended to higher input dimensions with little impact on training and evaluation time. Our model also scales well with the number of stars, making use of GPU parallel processing when sampling the posterior.

As shown in tests with synthetic stars (Appendix B) and apparent in Figure 5, increasing the number of stars decreases the



**Figure 9.** The mean and standard deviation in age, mass and radius results from the PPS model compared with the results (using the photometric temperature scale) from S17.

statistical uncertainties when parameters are pooled. The theoretical limit to this improvement is  $\sqrt{N_1/N_2}$  for two populations of size  $N_1$  and  $N_2$ . For example, if we increase our sample to 300 stars, we would expect the uncertainties to reduce by up to a factor of 2. Naturally, the uncertainty is still limited by observational precision. However, hierarchical modelling as demonstrated in this work, allows us to get the most out of our data and paves the way for a data-driven analysis of model systematics.

Including all-sky data from *TESS* and in anticipation of *Plato* (Rauer et al. 2014) we can expect our sample size of asteroseismic dwarfs and subgiants only to increase. There is also scope to extend our grid of models to include red giants, for which there are vast catalogues of stars already studied with *Kepler* (Pinsonneault et al. 2018).

## 6 CONCLUSION

We have shown that modelling  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$  to improve inference of fundamental parameters can be done through the use of an HBM, whilst still improving statistical uncertainties. Our results were in good agreement with S17 with small changes in mass and radii expected from our choice of model physics and updated observables. Taking our partially-pooled model including the Sun (PPS) as our preferred set of results, we obtained median statistical uncertainties on  $M$ ,  $R$  and  $\tau$  of 2.5, 1.2 and 12 per cent respectively. Furthermore, we demonstrated that the uncertainties reduced with increasing sample size in a population of synthetic stars, giving scope to further improve our inference on larger sample sizes from *TESS*.

We found that the gradient,  $\Delta Y/\Delta Z$  of the linear helium enrichment law ranged from 0.8 to 1.6 depending on the level of parameter pooling and the inclusion of the Sun in our sample, with  $\Delta Y/\Delta Z = 1.1^{+0.3}_{-0.3}$  from our preferred PPS model. Consistent across our models was the spread in initial helium about the enrichment law,  $\sigma_Y = 0.005^{+0.004}_{-0.003}$ . The mean  $\alpha_{\text{MLT}}$  in the population was  $\mu_\alpha = 1.90^{+0.10}_{-0.09}$  for the PPS model, with values from 1.7 to 2.1 depending on the level of pooling and whether or not solar data was included. We also found the spread in  $\alpha_{\text{MLT}}$  doubled to  $\sigma_\alpha = 0.13^{+0.06}_{-0.05}$  to account for the addition of the Sun in our sample. We conclude that there are still discrepancies between the best-fitting  $\alpha_{\text{MLT}}$  in our population and that of the Sun which need to be investigated further. Perhaps, the addition of asteroseismic signatures of helium abundance (see e.g. Verma et al. 2017)

would improve our constraints on  $Y_{\text{init}}$  and thus reduce star-by-star uncertainties in  $\alpha_{\text{MLT}}$ .

## ACKNOWLEDGEMENTS

This work is a part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (CartographY; grant agreement ID 804752).

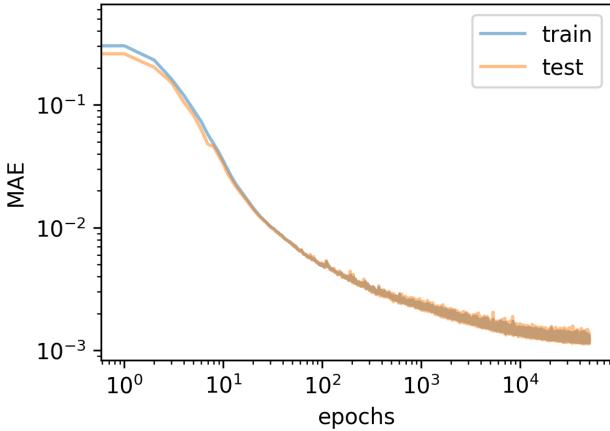
## REFERENCES

- Abadi M., et al., 2016, arXiv e-prints, 1605, arXiv:1605.08695
- Albareti F. D., et al., 2017, *The Astrophysical Journal Supplement Series*, 233, 25
- Anderson L., Hogg D. W., Leistedt B., Price-Whelan A. M., Bovy J., 2018, *AJ*, 156, 145
- Appourchaux T., et al., 2015, *Astronomy and Astrophysics*, 582, A25
- Asplund M., Grevesse N., Sauval A. J., Scott P., 2009, *ARA&A*, 47, 481
- Aver E., Olive K. A., Skillman E. D., 2015, *Journal of Cosmology and Astroparticle Physics*, 07, 011
- Baglin A., et al., 2006, in 36th COSPAR Scientific Assembly, p. 3749
- Bahcall J. N., Pinsonneault M. H., Wasserburg G. J., 1995, *Reviews of Modern Physics*, 67, 781
- Ball W. H., Gizon L., 2014, *Astronomy and Astrophysics*, 568, A123
- Balser D. S., 2006, *AJ*, 132, 2326
- Basu S., Antia H. M., 2004, *The Astrophysical Journal Letters*, 606, L85
- Bellinger E. P., Angelou G. C., Hekker S., Basu S., Ball W. H., Guggenberger E., 2016, *The Astrophysical Journal*, 830, 31
- Berger T. A., Huber D., van Saders J. L., Gaidos E., Tayar J., Kraus A. L., 2020, *The Astronomical Journal*, 159, 280
- Blanton M. R., et al., 2017, *The Astronomical Journal*, 154, 28
- Borucki W. J., et al., 2010, *Science*, 327, 977
- Brogaard K., et al., 2012, *Astronomy and Astrophysics*, 543, A106
- Broomhall A.-M., Chaplin W. J., Elsworth Y., New R., 2011, *MNRAS*, 413, 2978
- Campante T. L., et al., 2016, *ApJ*, 819, 85
- Casagrande L., Flynn C., Portinari L., Girardi L., Jimenez R., 2007, *MNRAS*, 382, 1516
- Casagrande L., Ramirez I., Meléndez J., Bessell M., Asplund M., 2010, *Astronomy and Astrophysics*, 512, A54
- Chan V. C., Bovy J., 2020, *MNRAS*, 493, 4367
- Chaplin W. J., Miglio A., 2013, *Annual Review of Astronomy and Astrophysics*, 51, 353
- Chaplin W. J., et al., 2020, *Nature Astronomy*, 4, 382

- Choi J., Dotter A., Conroy C., Cantiello M., Paxton B., Johnson B. D., 2016, *The Astrophysical Journal*, 823, 102
- Christensen-Dalsgaard J., Proffitt C. R., Thompson M. J., 1993, *The Astrophysical Journal Letters*, 403, L75
- Clevert D.-A., Unterthiner T., Hochreiter S., 2015, arXiv e-prints, 1511, arXiv:1511.07289
- Connelly J. N., Bizzarro M., Krot A. N., Nordlund Å., Wielandt D., Ivanova M. A., 2012, *Science*, 338, 651
- Cooke R. J., Fumagalli M., 2018, *Nature Astronomy*, 2, 957
- Cyburt R. H., Fields B. D., Olive K. A., Yeh T.-H., 2016, *Reviews of Modern Physics*, 88, 015004
- Dillon J. V., et al., 2017, arXiv e-prints, 1711, arXiv:1711.10604
- Dotter A., 2016, *The Astrophysical Journal Supplement Series*, 222, 8
- Dotter A., Conroy C., Cargile P., Asplund M., 2017, *The Astrophysical Journal*, 840, 99
- Ferguson J. W., Alexander D. R., Allard F., Barman T., Bodnarik J. G., Hauschildt P. H., Heffner-Wong A., Tamanai A., 2005, *The Astrophysical Journal*, 623, 585
- Frebel A., 2010, *Astronomische Nachrichten*, 331, 474
- Gaia Collaboration et al., 2016, *A&A*, 595, A1
- Gaia Collaboration et al., 2018, *Astronomy and Astrophysics*, 616, A1
- García Pérez A. E., et al., 2016, *The Astronomical Journal*, 151, 144
- Gelman A., Rubin D. B., 1992, *Statistical Science*, 7, 457
- Glorot X., Bordes A., Bengio Y., 2011, Proc. 14th Int. Conf. Artif. Intell. Statistics AISTATS 2011, 15, 315
- Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press
- Green G. M., et al., 2018, *Monthly Notices of the Royal Astronomical Society*, 478, 651
- Grevesse N., Sauval A. J., 1998, *Space Sci. Rev.*, 85, 161
- Hahnloser R. H. R., Sarpeshkar R., Mahowald M. A., Douglas R. J., Seung H. S., 2000, *Nature*, 405, 947
- Hall O. J., et al., 2019, *MNRAS*, 486, 3569
- Hawkins K., Leistedt B., Bovy J., Hogg D. W., 2017, *MNRAS*, 471, 722
- Haykin S., 2007, Neural Networks: A Comprehensive Foundation (3rd Edition). Prentice-Hall, Inc., USA
- Hendriks L., Aerts C., 2019, *PASP*, 131, 108001
- Hogg D. W., Myers A. D., Bovy J., 2010, *ApJ*, 725, 2166
- Homan M. D., Gelman A., 2014, *J. Mach. Learn. Res.*, 15, 1593
- Huber D., et al., 2011, *ApJ*, 743, 143
- Huber D., et al., 2017, *The Astrophysical Journal*, 844, 102
- Huber D., et al., 2019, *The Astronomical Journal*, 157, 245
- Kingma D. P., Ba J., 2014, arXiv e-prints, 1412, arXiv:1412.6980
- Kjeldsen H., Bedding T. R., Christensen-Dalsgaard J., 2008, *The Astrophysical Journal Letters*, 683, L175
- Korn A. J., Grundahl F., Richard O., Mashonkina L., Barklem P. S., Collet R., Gustafsson B., Piskunov N., 2007, *The Astrophysical Journal*, 671, 402
- Kuszlewicz J. S., Chaplin W. J., North T. S. H., Farr W. M., Bell K. J., Davies G. R., Campante T. L., Hekker S., 2019, *MNRAS*, 488, 572
- Leistedt B., Hogg D. W., 2017, *AJ*, 154, 222
- Li T., Bedding T. R., Huber D., Ball W. H., Stello D., Murphy S. J., Bland Hawthorn J., 2018, *Monthly Notices of the Royal Astronomical Society*, 475, 981
- Lindegren L., et al., 2018, *A&A*, 616, A2
- Magic Z., Weiss A., Asplund M., 2015, *A&A*, 573, A89
- Masters D., Luschi C., 2018, arXiv e-prints, 1804, arXiv:1804.07612
- Morton T. D., Winn J. N., 2014, *ApJ*, 796, 47
- Nissen P. E., Silva Aguirre V., Christensen-Dalsgaard J., Collet R., Grundahl F., Slumstrup D., 2017, *Astronomy and Astrophysics*, 608, A112
- Nsamba B., Campante T. L., Monteiro M. J. P. F. G., Cunha M. S., Rendle B. M., Reese D. R., Verma K., 2018, *Monthly Notices of the Royal Astronomical Society*, 477, 5052
- Önehag A., Gustafsson B., Korn A., 2014, *Astronomy and Astrophysics*, 562, A102
- Paquette C., Pelletier C., Fontaine G., Michaud G., 1986, *The Astrophysical Journal Supplement Series*, 61, 177
- Paxton B., Bildsten L., Dotter A., Herwig F., Lesaffre P., Timmes F., 2011, *ApJS*, 192, 3
- Paxton B., et al., 2013, *The Astrophysical Journal Supplement Series*, 208, 4
- Paxton B., et al., 2015, *The Astrophysical Journal Supplement Series*, 220, 15
- Paxton B., et al., 2018, *The Astrophysical Journal Supplement Series*, 234, 34
- Paxton B., et al., 2019, *The Astrophysical Journal Supplement Series*, 243, 10
- Peimbert A., Peimbert M., Luridiana V., 2016, *Revista Mexicana de Astronomía y Astrofísica*, 52, 419
- Pinsonneault M. H., An D., Molenda-Żakowicz J., Chaplin W. J., Metcalfe T. S., Bruntt H., 2012, *ApJS*, 199, 30
- Pinsonneault M. H., et al., 2014, *The Astrophysical Journal Supplement Series*, 215, 19
- Pinsonneault M. H., et al., 2018, *The Astrophysical Journal Supplement Series*, 239, 32
- Pitrou C., Coc A., Uzan J.-P., Vangioni E., 2018, *Physics Reports*, 754, 1
- Qian N., 1999, *Neural Networks*, 12, 145
- Rauer H., et al., 2014, *Experimental Astronomy*, 38, 249
- Ribas I., Jordi C., Torra J., Giménez Á., 2000, *MNRAS*, 313, 99
- Ricker G. R., et al., 2015, *Journal of Astronomical Telescopes, Instruments, and Systems*, 1, 014003
- Riess A. G., et al., 2018, *The Astrophysical Journal*, 861, 126
- Rogers F. J., Nayfonov A., 2002, *The Astrophysical Journal*, 576, 1064
- Ruder S., 2016, arXiv e-prints, 1609, arXiv:1609.04747
- Salvatier J., Wiecki T. V., Fonnesbeck C., 2016, *PeerJ Comput. Sci.*, 2, e55
- Scott P., et al., 2015, *Astronomy and Astrophysics*, 573, A25
- Serenelli A. M., Basu S., 2010, *ApJ*, 719, 865
- Serenelli A. M., Basu S., Ferguson J. W., Asplund M., 2009, *The Astrophysical Journal Letters*, 705, L123
- Serenelli A., et al., 2017, *ApJS*, 233, 23
- Silva Aguirre V., et al., 2015, *MNRAS*, 452, 2127
- Silva Aguirre V., et al., 2017, *ApJ*, 835, 173
- Skrutskie M. F., et al., 2006, *The Astronomical Journal*, 131, 1163
- Sonoi T., Samadi R., Belkacem K., Ludwig H.-G., Caffau E., Mosser B., 2015, *Astronomy and Astrophysics*, 583, A112
- Stancilffe R. J., Fossati L., Passy J.-C., Schneider F. R. N., 2016, *Astronomy and Astrophysics*, 586, A119
- Tayar J., et al., 2017, *The Astrophysical Journal*, 840, 17
- Thoul A. A., Bahcall J. N., Loeb A., 1994, *The Astrophysical Journal*, 421, 828
- Townsend R. H. D., Teitler S. A., 2013, *Monthly Notices of the Royal Astronomical Society*, 435, 3406
- Trampedach R., Stein R. F., Christensen-Dalsgaard J., Nordlund Å., Asplund M., 2014, *MNRAS*, 445, 4366
- Valle G., Dell’Omodarme M., Prada Moroni P. G., Degl’Innocenti S., 2015, *Astronomy and Astrophysics*, 575, A12
- Verma K., Hanasoge S., Bhattacharya J., Antia H. M., Krishnamurthi G., 2016, *MNRAS*, 461, 4206
- Verma K., Raodeo K., Antia H. M., Mazumdar A., Basu S., Lund M. N., Silva Aguirre V., 2017, *The Astrophysical Journal*, 837, 47
- Verma K., Raodeo K., Basu S., Silva Aguirre V., Mazumdar A., Mosumgaard J. R., Lund M. N., Ranadive P., 2019, *MNRAS*, 483, 4678
- Viani L. S., Basu S., Ong J. M. J., Bonaca A., Chaplin W. J., 2018, *The Astrophysical Journal*, 858, 28
- Villante F. L., Serenelli A. M., Delahaye F., Pinsonneault M. H., 2014, *The Astrophysical Journal*, 787, 13
- Weiss A., Schlattl H., 2008, *Astrophysics and Space Science*, 316, 99
- West C., Heger A., 2013, *The Astrophysical Journal*, 774, 75
- White T. R., Bedding T. R., Stello D., Christensen-Dalsgaard J., Huber D., Kjeldsen H., 2011, *The Astrophysical Journal*, 743, 161
- Zinn J. C., Pinsonneault M. H., Huber D., Stello D., 2019, *ApJ*, 878, 136

## APPENDIX A: TRAINING THE ANN

In Table A1 we give the scaling parameters used to standardise our training dataset. We determined the median,  $\mu_{1/2}$  and standard



**Figure A1.** The MAE as a function of epochs for the train and test dataset.

deviation,  $\sigma$  to 3 decimal places for each of the input and output dimensions. We then standardised the data by subtracting  $\mu_{1/2}$  and dividing by  $\sigma$ .

In Figure A1 we show the train and test MAE as a function of epochs for the final ANN configuration. The train and test loss were comparable throughout training.

## APPENDIX B: THE SYNTHETIC POPULATION

In this section, we present the results for the NP, PP and MP models run on a synthetic sample of 100 stars with the following initial conditions. We randomly generated initial  $M$  and  $[M/H]_{\text{init}}$  uniformly. We drew initial values for  $Y_{\text{init}}$  from a normal distribution centred on the helium enrichment law from Equation 14 with  $\Delta Y/\Delta Z = 1.8$  and  $Y_P = 0.247$ , and scaled by  $\sigma_Y = 0.008$ . We also generated initial values for  $\alpha_{\text{MLT}}$  from a normal distribution centred on  $\mu_\alpha = 2.0$  and scaled by  $\sigma_\alpha = 0.08$ .

We evolved the synthetic stars to randomly chosen ages using MESA. We then took the output  $\tau$ ,  $T_{\text{eff}}$ ,  $L$ ,  $\Delta\nu$  and  $[M/H]_{\text{surf}}$  from the models and used these as true values for each of the stars. We added random noise to the observed quantities centred on the true values with a standard deviation of 2.2 per cent in  $T_{\text{eff}}$ , 3.5 per cent in  $L$ , 0.9  $\mu\text{Hz}$  in  $\Delta\nu$  and 0.07 dex in  $[M/H]_{\text{surf}}$  chosen to be representative of the APOKASC sample.

### B1 Stellar parameters

We found that the NP model recovered the true values for the individual stellar parameters, but the uncertainties were unreliable. The observational quantities alone were not good enough to constrain  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$ . As a result, their distributions were truncated at the bounds of their priors. These boundary effects skewed the marginalised posterior means for  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$  towards the centre of the prior (0.28 and 2.0 respectively).

The PP model recovered true values for the synthetic stars with more reliable uncertainty than the NP model. The addition of pooling  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$  between the stars improved their uncertainty which reduced the effects of the prior as seen in the NP model. We

We found little difference between the results of the PP and MP models.

We reran the PP model with 10 and 50 stars chosen randomly from the sample of synthetic stars. In Figure B1, we show the

uncertainties in the several parameters from the results of each of the models. For the two pooled parameters,  $Y_{\text{init}}$  and  $\alpha_{\text{MLT}}$ , the uncertainty reduction due to pooling is most obvious. We see the PP model repeatedly improves on the uncertainties from the NP model when  $N_{\text{stars}}$  is increased.

In Figure B1 we also see a similar reduction in uncertainty for  $\tau$ ,  $M$  and  $R$ , with all models improve upon the NP model. However, we do not see the same effect in  $Z_{\text{init}}$  for which the uncertainty appears dominated by observations of  $[M/H]_{\text{surf}}$ .

### B2 Population parameters

In Figure B2, we see that the PP model also recovers the hyperparameter truths well, with some noise due to random realisation error. Fitting the model this way has the added benefit over the NP model of improving the inference of the individual stellar parameters, as shown in the previous two sections. We also found that when we ran the PP model with 10 and 50 stars, the uncertainties on the hyperparameters also shrank with increasing  $N_{\text{stars}}$ .

Figure B3 shows the hyperparameter results for the MP model. Here,  $\alpha_{\text{MLT}}$  was assumed to be the same for all stars. This model also recovers the true hyperparameters for helium well, and the assumed value for  $\alpha_{\text{MLT}}$  is within uncertainty of the true  $\mu_\alpha$ .

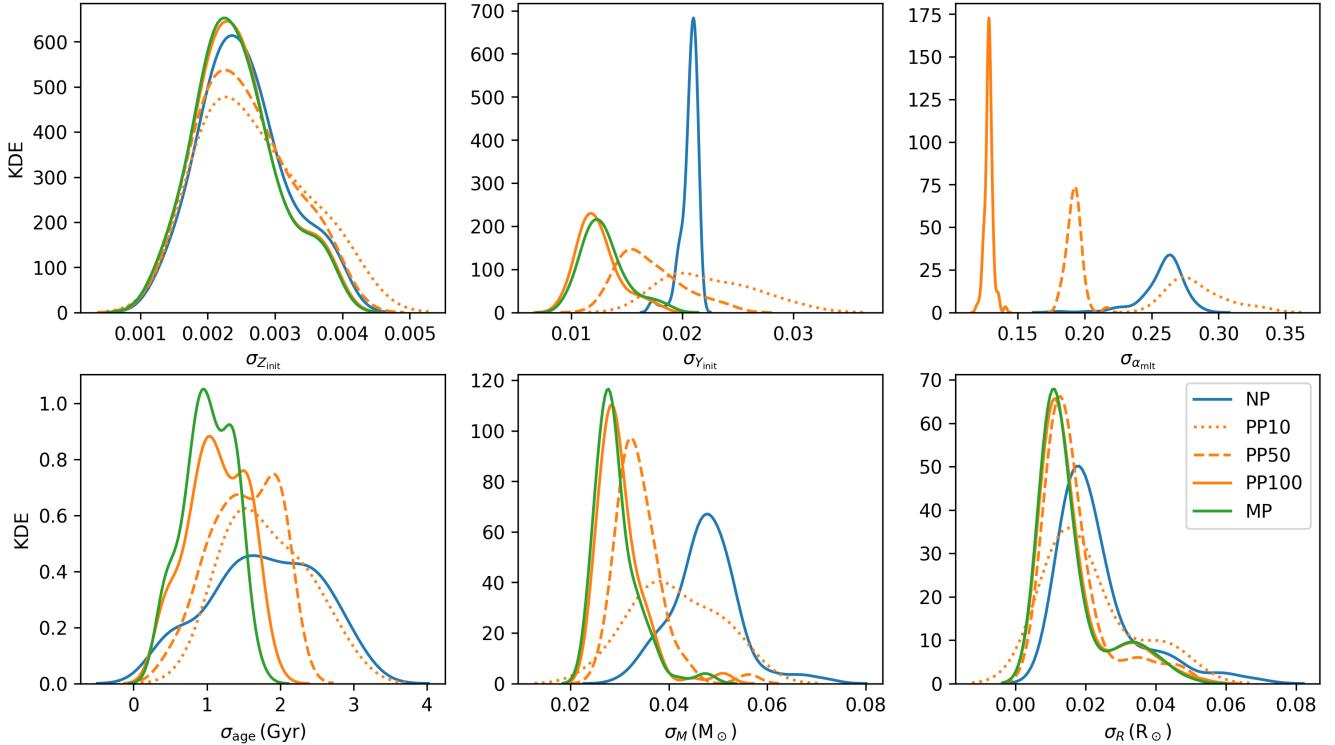
## APPENDIX C: THE SOLAR CALIBRATOR

We found that our model consistently recovers the Sun when modelled in each of the NP, PP and MP models. In Table C1 we present the results for the Sun as a star from the NP model to show what we obtain without the influence of any other stars in the sample. We show the marginal and joint posterior distributions for the solar parameters from the NP model in the corner plot in Figure C1.

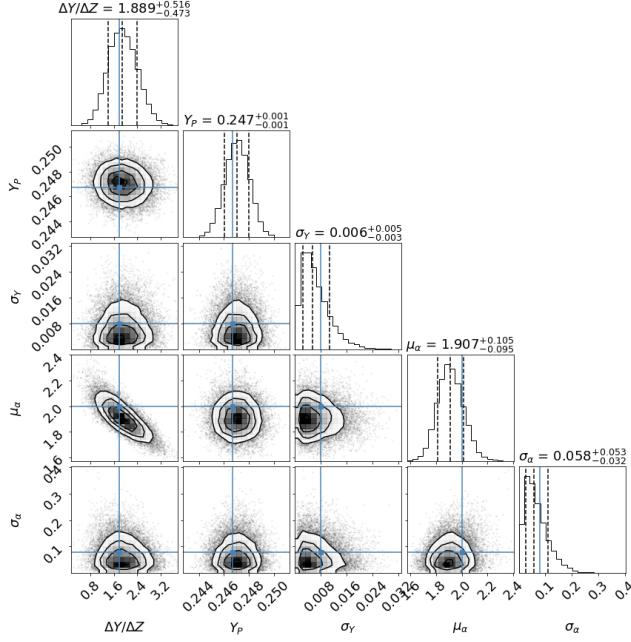
This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.

**Table A1.** The median,  $\mu_{1/2}$  and standard deviation,  $\sigma$  for each parameter in the training data, used to standardise the dataset.

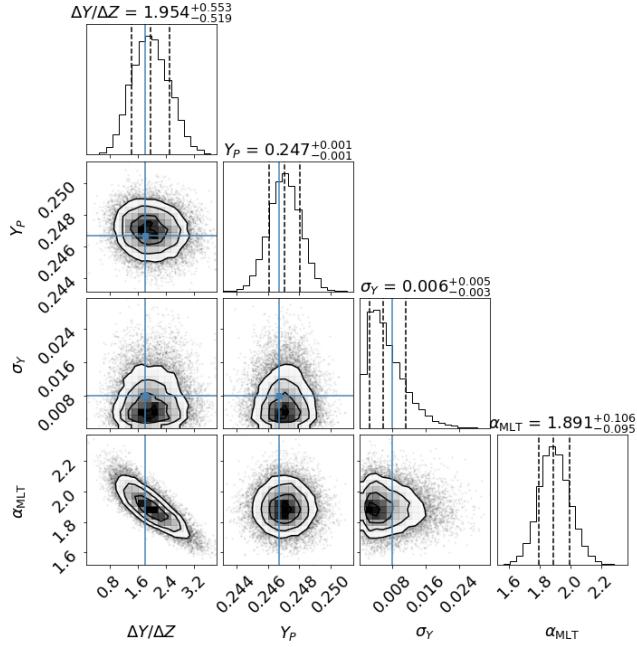
	Input					Output				
	$f_{\text{evol}}$	$M (\text{M}_\odot)$	$\alpha_{\text{MLT}}$	$Y_{\text{init}}$	$Z_{\text{init}}$	$\log(\tau/\text{Gyr})$	$T_{\text{eff}} (\text{K})$	$R (\text{R}_\odot)$	$\Delta\nu (\mu\text{Hz})$	$[\text{M}/\text{H}]_{\text{surf}} (\text{dex})$
$\mu_{1/2}$	0.865	1.000	1.900	0.280	0.017	0.790	5566.772	1.224	100.720	0.081
$\sigma$	0.651	0.118	0.338	0.028	0.011	0.467	601.172	0.503	42.582	0.361

**Figure B1.** Kernel density estimates (KDEs) showing the reduction of statistical uncertainties between models of the sample of synthetic stars. The PP model was run with 10, 50 and 100 stars and is denoted PP10, PP50, and PP100 respectively. The NP and MP models were both run with the full set of 100 stars.**Table C1.** Solar results from the NP model. The second column shows the median marginalised posterior samples for each parameter with their respective upper and lower 68 per cent credible intervals.

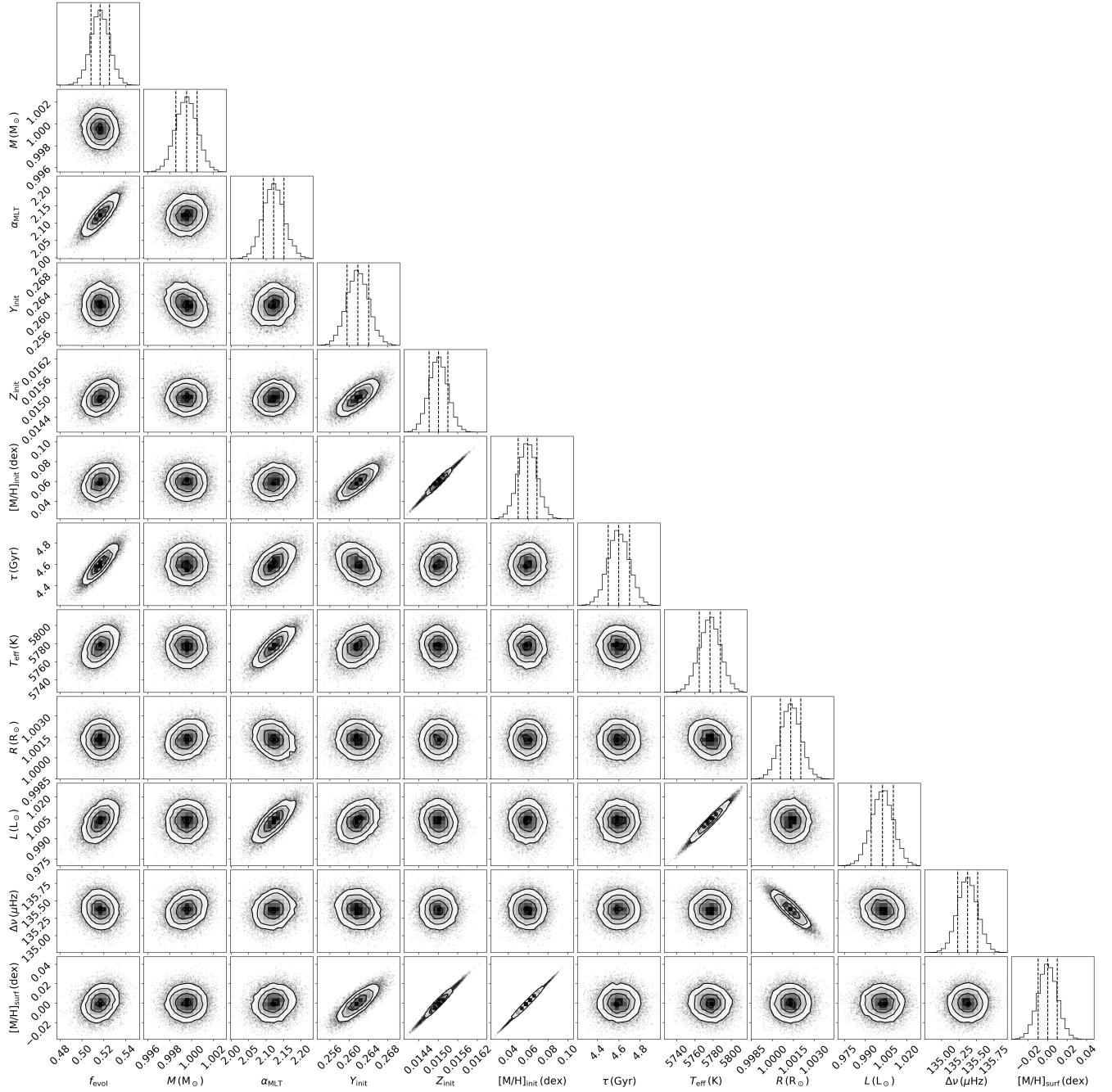
Output	Value
$f_{\text{evol}}$	$0.517^{+0.009}_{-0.008}$
$M (\text{M}_\odot)$	$1.000^{+0.001}_{-0.001}$
$\alpha_{\text{MLT}}$	$2.12^{+0.03}_{-0.03}$
$Y_{\text{init}}$	$0.262^{+0.002}_{-0.002}$
$Z_{\text{init}}$	$0.0150^{+0.0003}_{-0.0003}$
$\tau (\text{Gyr})$	$4.6^{+0.1}_{-0.1}$
$T_{\text{eff}} (\text{K})$	$5777^{+12}_{-12}$
$R (\text{R}_\odot)$	$1.001^{+0.001}_{-0.001}$
$\Delta\nu (\mu\text{Hz})$	$135.37^{+0.14}_{-0.14}$
$[\text{M}/\text{H}]_{\text{surf}} (\text{dex})$	$0.00^{+0.01}_{-0.01}$



**Figure B2.** Corner plot showing the marginalised and joint posterior distributions between the PP model hyperparameters for 100 synthetic stars. The true values are shown by the blue lines.



**Figure B3.** The same as Figure B2 but for the MP model.



**Figure C1.** A corner plot showing the sampled marginal and joint posterior distributions for the Sun as a part of the NP model.