Draft version April 25, 2020
Typeset using LaTeX **twocolumn** style in AASTeX63

Modelling the first APOKASC sample of *Kepler* dwarfs and subgiants using machine learning I. Project plan

Alex Lyttle[1]

[1] *University of Birmingham, Edgbaston, B15 2TT*

ABSTRACT

Using grid based modelling to obtain fundamental stellar parameters is... However, such modelling methods suffer from systematic uncertainties... Interpolation can provide finer grid, but it brings about further systematics from the interpolation method. I present a project plan which aims to train an artificial neural network on at least one grid of stellar models. (neural networks trump interpolation because they learn the relationships between parameters by the adjusting of weights etc...) Coupled with a hierarchical Bayesian model, which allows for the sharing of parameters between stars, I will apply the method to the first APOKASC sample of Kepler dwarfs and subgiants.

Keywords: n/a

1. INTRODUCTION & AIMS

- Problems with grid based modelling and interpolation
- What science does this limit? Ages, helium, mixing length
- Advantage of HBMs
- Advantage of neural networks and mention Guy's paper in prep?
- Refer to the Serenelli sample
- Goal of the project is to apply our method to a sample of asteroseismic dwarfs and maybe subgiants and compare results.
- Free up mixing length and helium abundance relation.

The scientific goals of this project are as follows:

- Test the use of artificial neural networks to approximate the output of stellar models
- Determine stellar fundamentals for a sample of dwarfs and subgiants using a hierarchical Bayesian model which samples the trained neural network
- Probe the effects of freeing up parameters such as initial helium fraction and the mixing-length theory parameters on the stellar fundamentals and compare to S17

The main collaborators on this project are: Guy Davies (supervisor), Tanda Li (stellar models, MESA) and Lindsey Carboneau (neural networks). I also hope to seek help from other members of the group such as: Josefina Montalban (stellar models, *Cley*) and Warrick Ball (details regarding S17). Some external collaborators who may be able to help include: Victor Silva Aguirre (comparisons with BASTA) and Jamie Tayar (gyrochronology considerations).

2. METHODS

The methods for this project are split into three sections: data, grid and model. These

All code used in this project will be made publicly available upon submission. A private *GitHub* repository has been made which will be made accessible to collaborators[1].

2.1. *Data*

The objects being studied in this project are a subset of the first APOKASC catalog of *Kepler* dwarfs and subgiant stars (Serenelli et al. 2017, hereafter S17). The full dataset comprises asteroseismology for 415 objects with respective fundamental parameters determined through grid based modelling (GBM) using two independent effective temperature scales: Sloan Digital Sky Survey (SDSS) *griz* band photometric temperatures and APOGEE[2] Stellar Parameters and Chemical Abun-

[1] https://github.com/alexlyttle/kepler-dwarfs.git

[2] Apache Point Observatory Galactic Evolution Experiment

Table 1. An extract from Table 3 of **?** comprising measurements of the global asteroseismic parameters from the Sydney pipeline (Huber et al. 2009) with uncertainties from the spread of results from several other pipelines.

Kepler ID	$\nu_{\rm max}$	$\sigma_{\nu_{\rm max}}$	$\Delta\nu$	$\sigma_{\Delta\nu}$
2450729	1053.105	114.904	61.910	2.539
2991448	1111.248	18.148	61.732	0.899
3223000	2573.222	563.234	110.919	1.662
3241581	2807.592	395.565	123.412	2.821
3427720	2726.381	56.767	120.045	0.120

dances Pipeline (ASPCAP) spectroscopic temperatures corresponding to Data Release 13 of SDSS.

I will not consider intermediate-mass stars (those with a convective core) and low-metallicity stars in this project. Therefore, I selected a subset of S17 where stellar mass determined using each temperature scale lie in the range $0.85 < M/{\rm M}_\odot < 1.15$ and metallicity from each scale is in the range $-0.3 <$ [M/H] < 0.3. This subset contains 69 objects which are plot in Figure 1. The values of surface gravity, $\log g$, and [M/H] are determined for each temperature scale by S17. These plots show the subsample contains mostly main sequence stars and subgiants, with a few early red giant branch stars. Five

2.2. *Grid*

The span of the grid is chosen just outside the subset of S17 chosen in 2.1. The grid should be fine enough such that the neural network is able to learn the relationship between fundamentals and observables, but within a computational time of a few weeks. A grid (hereafter `grid1_sun`) was computed by Tanda Li using MESA in late 2019. The models in `grid1_sun`, which may be seen in 1, were terminated at the main sequence turnoff. This figure shows that to include the more evolved stars, $\log(g) \lesssim 4.1$ the grid must be further evolved to at least the early red giant branch stage. I propose an extension to the initial conditions of `grid1_sun` described below, with the bounds of `grid1_sun` given in parentheses where appropriate,

$$
\begin{aligned}
M/{\rm M}_\odot &\in [0.8, 1.2], & \Delta M &= 0.01\,{\rm M}_\odot,\\
[{\rm M/H}]_{\rm ini} &\in [-0.5\,(-0.3), 0.5\,(0.3)], & \Delta[{\rm M/H}]_{\rm ini} &= 0.1,\\
Y_{\rm ini} &\in [0.24, 0.32\,(0.3)], & \Delta Y_{\rm ini} &= 0.01,\\
\alpha_{\rm MLT} &\in [1.6\,(1.7), 2.3\,(2.0)], & \Delta\alpha_{\rm MLT} &= 0.1.
\end{aligned}
$$

The bounds on mass are $0.05\,{\rm M}_\odot$ outside of the range selected in the previous section. The upper bound is conveniently chosen as the point before which stars have a convective core, so there is no need to consider convective core overshooting. Median uncertainties on the masses determined by S17 are $\sim 6\,\%$, so a more conservative cut in mass of the dataset may be needed if we experience sampling issues due to being close to the edge of the grid. The bounds in initial metallicity are 0.2 dex outside of the dataset which comfortably covers the observed range by $\gtrsim 3\sigma$. The grid will be computed with diffusion which affects the surface metallicity, reinforcing the need to extend $[{\rm M/H}]_{\rm ini}$ beyond the data.

Since a primary focus of this project is to test the effects of freeing up $\alpha_{\rm MLT}$ and $Y_{\rm ini}$, their bounds are chosen outside of the expected range of the dataset. For example, taking a helium enrichment ratio of $\Delta Y/\Delta Z = 1.4$ (Brogaard et al. 2012) with upper and lower bounds from the literature of 1.0 and 1.8 respectively, gives approximate values of $Y_{\rm ini}$ between 0.25 and 0.31 for the initial metallicity bounds of the grid selected above. Therefore, the limits are chosen 0.01 outside this range. As for $\alpha_{\rm MLT}$, I refer to hydrodynamical simulations (e.g. Trampedach et al. 2014; Magic et al. 2015) to get a suitable range. For example, figure 2 suggests values between 1.7 and 2.2 to be a good choice of limits on $\alpha_{\rm MLT}$. However, hydrodynamical simulations have been found to disagree with stellar codes using asteroseismic constraints (e.g. see Fig. 9 of Silva Aguirre et al. 2017) so I have increased the limits by 0.1 above to account for these inconsistencies.

The purpose of the grid is to provide inputs and output to an artificial neural network (ANN) which will learn the relationship between the fundamental stellar input parameters (grid inputs) and the observables. These observables are some combination of the following: large frequency separation $\Delta\nu$, either from scaling relations or radial mode frequencies, effective temperature $T_{\rm eff}$, surface metallicity [M/H], radius $R/{\rm R}_\odot$, and luminosity $L/{\rm L}_\odot$.

2.3. *Model*

The model will be split into two components: an ANN trained on the stellar grid, and a hierarchical Bayesian model (HBM) which shares hyper-parameters between stars in the subsample.

2.3.1. *Artificial neural network*

Output from each evolutionary track computed in Section 2.2 will be combined into one master table. A random sample of 20 % will be set aside for validation and the remaining 80 % will be used for training the ANN. The training set will be resampled for mass steps of

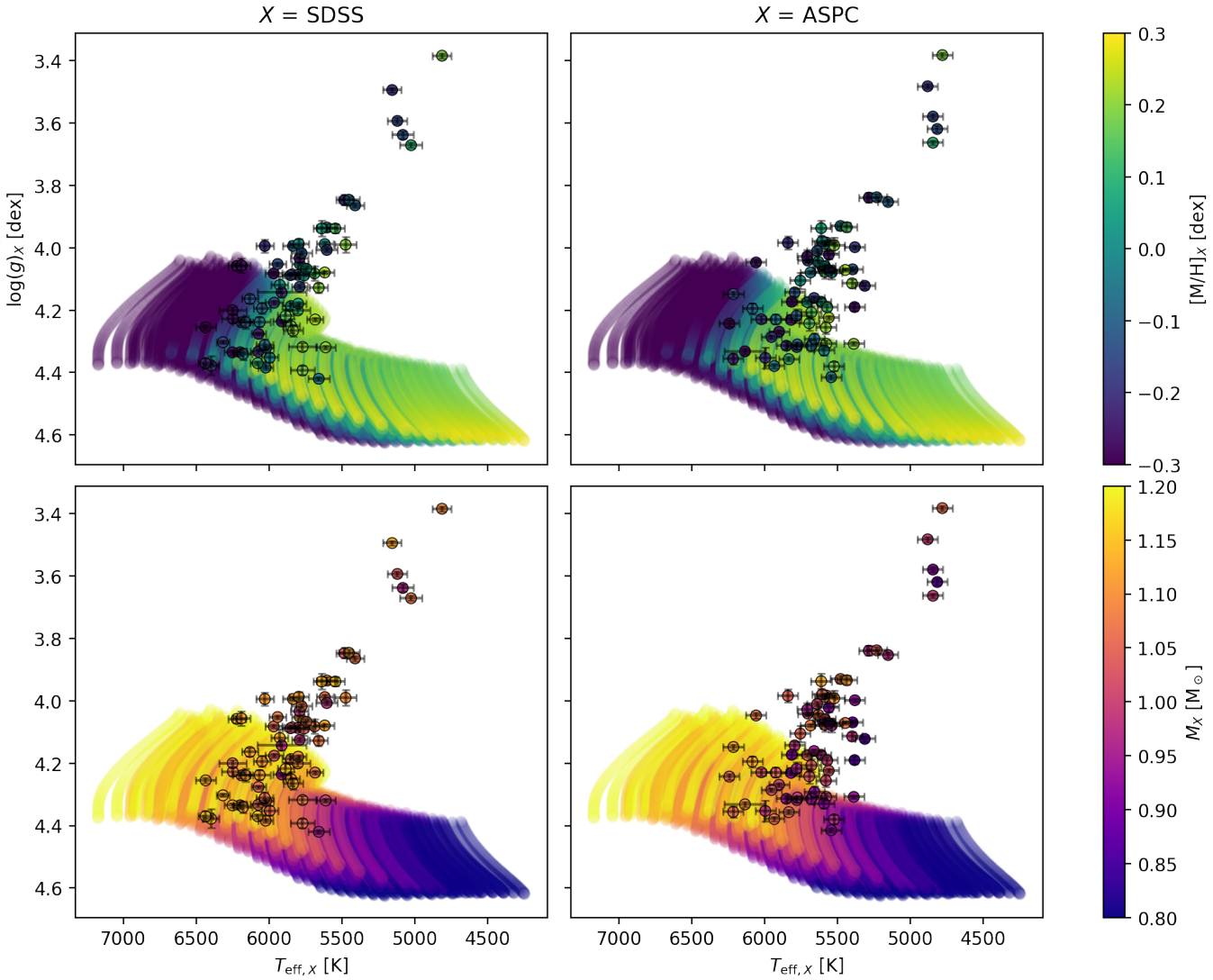

Figure 1. Results from a subsample of S17 containing 69 low-mass objects. The plots comprise results which use SDSS *griz* photometric temperatures (left) and ASPCAP spectroscopic temperatures from DR13 (right). The points in the background are from a grid of main sequence stellar models, `grid1_sun`, described in Section 2.2 with $\alpha_{\rm MLT} = 1.9$ and $\Delta M = 0.04$. The data points are coloured by metallicity (top) and mass (bottom) corresponding to each temperature scale.

$0.04\,\mathrm{M}_\odot$, $0.02\,\mathrm{M}_\odot$ and $0.01\,\mathrm{M}_\odot$. An ANN will be trained on all three datasets separately and evaluated to investigate the approximation ability with sparsity of the grid.

An off-grid validation dataset should also be computed, since the ANN inputs are discrete

The ANN will be a regression neural network with 5 to 10 fully-connected layers of $\sim$ 100 neurons per layer. Finding the optimal network architecture will be achieved with a combination of a grid search and heuristic approach. A grid search will be used to find optimal start points for the learning rate, optimiser, activation function and regularisation. Once these are found, a trial and improvement method will be used to tune the best number of layers and neurons. To save time, tuning will be carried out on random subsamples of the data before being applied to the whole dataset.

Neural networks train better on inputs and outputs between -1 and 1. Positive ANN inputs and outputs may be log-normalised, either by taking the logarithm of the inputs and dividing by the median, or scaling by the log-mean and standard deviation. Similarly, some inputs may be normalised in linear space. To justify the normalisation method, histograms of input and output parameters should be plot. A selection of example ANN inputs and outputs from `grid1_sun` are plot in Figure 3. For instance, these plots suggest parameters such as age and $\Delta\nu$ should be log-normalised whereas mass and metallicity may be scaled linearly.

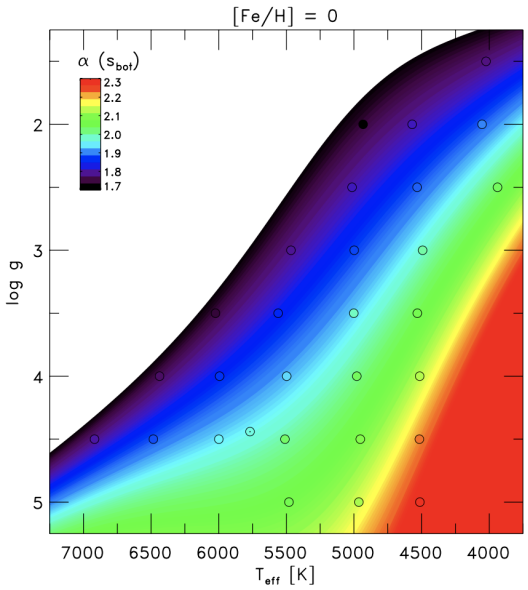

Figure 2. Values of $\alpha_{\rm MLT}$ from the hydrodynamical simulations of solar metallicity stars for position on the $T_{\rm eff} - \log(g)$ plane. Source: Fig. 2 (*left-panel*) of Magic et al. (2015).

2.3.2. *Hierarchical Bayesian model*

The HBM will be used to sample the function generated by the ANN given priors on the individual stellar fundamentals and hyperpriors on shared quantities which we expect to follow a particular distribution. For example, $\alpha_{\rm MLT}$ could be allowed to vary between stars in one model and assume it follows a normal distribution across the sample in another. Additionally, $Y_{\rm ini}$ may be constrained to follow and enrichment law with $\Delta Y/\Delta Z$ as a hyperparameter, or left to vary freely.

3. PLAN

The project plan is summarised in the following sections. I intend to make full use of *GitHub* as a platform for organising this project. The milestones described in Section 3.2 correspond to *GitHub* milestones, which will be achieved via the completion of projects corresponding to those described in 3.1 each comprising one or more branches.

3.1. *Timeline*

The plan is to complete the project by the end of June 2020. A gantt chart is presented in Figure 4 which breaks the project down into 9 sections and 5 milestones. The sections are as follows:

1. Literature review – I plan to read and understand S17, and compile and read papers which cover helium abundance and mixing-length theory, for example
2. Compute grid of stellar models – I will ask Tanda to compute a grid of stellar models as described in 2.2
3. Cross-match dataset with *Gaia* - I will cross-match the dataset with, for example Berger et al. (2018) to get parallax-corrected distances for independent luminosities
4. Preliminary method on MS stars – I will apply the method to a small subset of the data which lie within `grid1_sun`
5. Train ANN on full grid – I will work with Guy and Lindsey on training a neural network on the full grid
6. Construct and run HBM – working with Guy, I will construct and run an HBM using the trained ANN and observables
7. Analyse results and uncertainties – I will analyse the results and consider the effects of systematics (e.g. from the temperature scales)
8. Review method and repeat – I will present my results to colleagues and collaborators and revise the method if appropriate
9. Write paper – I will write a the paper and share with collaborators and co-authors when appropriate

The milestones tie in with the communications strategy described in Section 3.2.

3.2. *Communications strategy*

Updates will be sent to project collaborators at each of the project milestones. The milestones are summarised as follows:

M1. 19th March 2020 – Results from a test of the method will be sent to and discussed with collaborators. Feedback from here may be used to update the method

M2. 10th April 2020 – Results from the HBM on the full dataset will be shared for discussion with collaborators

M3. 1st May 2020 – A report or presentation summarising the results analysis should be made and given to collaborators.

M4. (a) 22nd May 2020 – I will start writing the paper and share with other writers (e.g. Tanda for grid section)

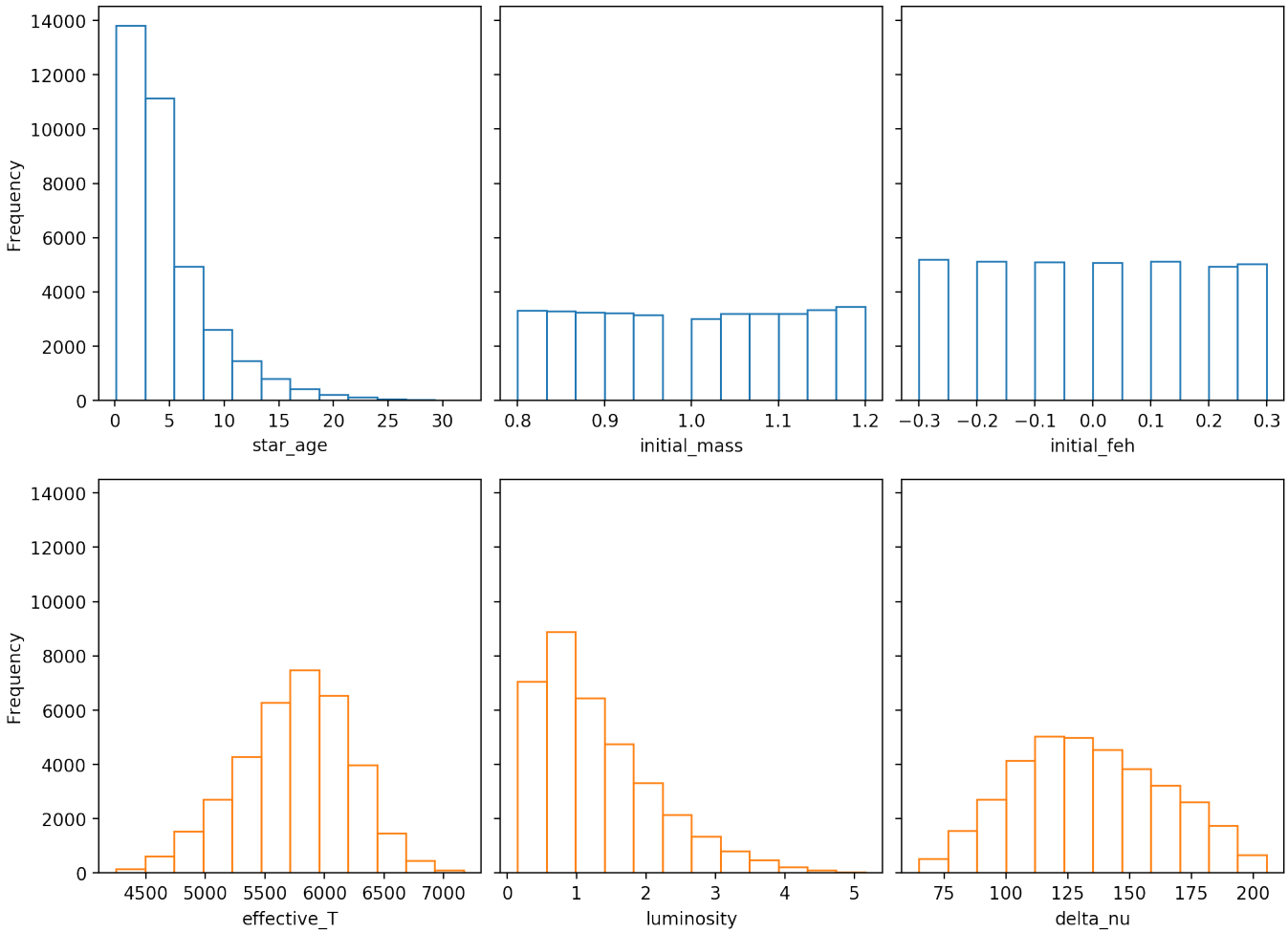

Figure 3. Histograms of selected columns in the output of `grid1_sun` as examples of inputs (*top*) and outputs (*bottom*) to train the ANN.

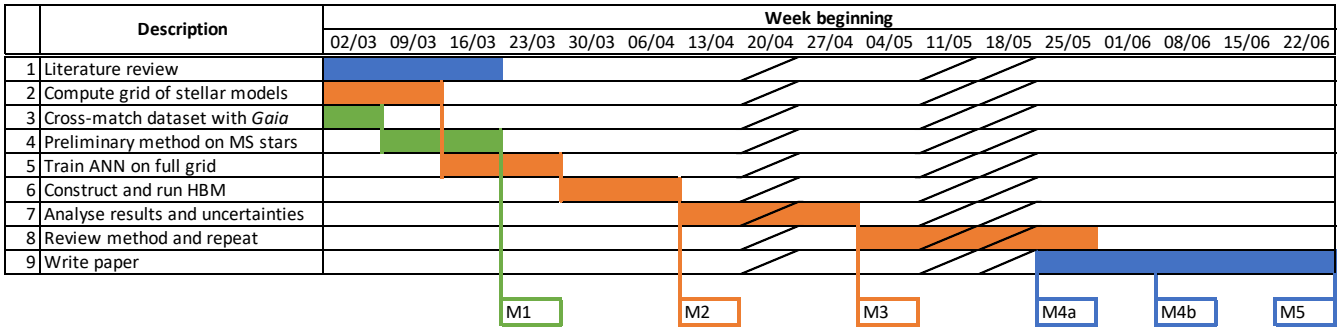

Figure 4. A gantt chart outlining 9 sections of the project and their dependencies. Each colour represents a dependency path through the project. Milestones are labelled by the letter M and described further in Section 3.1. The shaded regions are weeks where I will be away from the University of Birmingham.

(b) 5th June 2020 – Send the paper to co-authors

M5. 27th June 2020 – Submit the paper

These deadlines are marked in Figure 4 and represent key goals which need to be complete before moving forward.

3.3. *Risk and opportunity*

There are a number of risks and opportunities (abbreviated to opp.) associated with this project. For example,

Risk: Grid does not fully cover the data (edge-effects in ANN and HBM)

Solution: Save MESA model files and extend grid further where necessary

Risk: Data loss

Solution: Make full use of version control and store grid output on the research data store (RDS)

Opp.: Complete M2 ahead of schedule

Solution: Consider the effects of stellar model systematics by training on grids from different codes (e.g. *Cley*)

Opp.:

Solution:

4. SUMMARY

REFERENCES

Berger, T. A., Huber, D., Gaidos, E., & van Saders, J. L. 2018, ApJ, 866, 99, doi: 10.3847/1538-4357/aada83

Brogaard, K., VandenBerg, D. A., Bruntt, H., et al. 2012, Astronomy and Astrophysics, 543, A106, doi: 10.1051/0004-6361/201219196

Huber, D., Stello, D., Bedding, T. R., et al. 2009, Communications in Asteroseismology, 160, 74

Magic, Z., Weiss, A., & Asplund, M. 2015, A&A, 573, A89, doi: 10.1051/0004-6361/201423760

Serenelli, A., Johnson, J., Huber, D., et al. 2017, ApJS, 233, 23, doi: 10.3847/1538-4365/aa97df

Silva Aguirre, V., Lund, M. N., Antia, H. M., et al. 2017, ApJ, 835, 173, doi: 10.3847/1538-4357/835/2/173

Trampedach, R., Stein, R. F., Christensen-Dalsgaard, J., Nordlund, Å., & Asplund, M. 2014, MNRAS, 445, 4366, doi: 10.1093/mnras/stu2084