Draft version November 7, 2022
Typeset using LaTeX default style in AASTeX631

Astron 128 - Lab 2: Modelling Stellar Compositions from Spectra

Shrihan Agarwal[1]

[1] University of California, Berkeley, Department of Astronomy, Berkeley, CA 94720

Abstract

The future of our understanding of stellar compositions lies in large-scale spectroscopic surveys. So far, large surveys like Gaia-ESO, APOGEE, LAMOST, have produced hundreds of thousands of spectra. Using APOGEE data as a test set, we develop a method to perform modeling of stellar compositions and properties: temperature, $\log(g)$, [Fe/H], [Si/Fe], and [Mg/Fe]. We continuum-normalize spectra, train a flexible model per APOGEE wavelength pixel on 926 existing ASPCAP labelled spectra, and perform root-mean squared error minimization in order to determine the optimal stellar parameters for the remaining set.

We achieve a strong agreement with ASPCAP labels generated by the "Cannon", with a scatter of 130 K in T_{eff}, 0.22 in $\log(g)$ and 0.08 dex in [Fe/H]. We additionally compare our results with simulated expectations from MIST Isochrones and find reasonable agreement on a Kiel Diagram. We describe the technique by which the model identifies various stellar properties from spectra, and describe an outlook for future uses of this technique and opportunities for improvement.

1. INTRODUCTION

The significant growth of spectroscopic surveys e.g. GAIA-ESO, APOGEE, LAMOST, etc., in the past decade has allowed a renewed understanding of stellar compositions in the Milky Way, stellar clusters, and the Magellanic Clouds. This has made it necessary to develop a technique to perform accurate large-scale modelling of stellar spectra procured.

In this lab, we utilize 1800 unnormalized APOGEE spectra as our spectra. With 900 ASPCAP labels produced using the "Cannon" as a basis to train our model, we predict the stellar properties and compositions of the remaining half of stars: T_{eff}, $\log(g)$, [Fe/H], [Si/Fe], and [Mg/Fe]. We split the data into equally sized halves of training and test data, and use the test data as a proof of concept for the model by comparing its predictions to both ASPCAP and MIST predictions.

We briefly describe our dataset in Section 2. We begin by continuum-normalizing the spectra in order to amplify evidence of an absorption spectrum in Section 3. We train a weighted linear regression model in order to reproduce spectra based on the ASPCAP-derived labels in Section 4. This model is used to search the parameter space of stellar properties and determine the optimal parameters using a RMSE loss fitting routine, as well as an MCMC routine as explained in Section 5. The properties derived from the RMSE fit are presented on standard astrophysical parameter spaces such as the Kiel diagram, and compared with MIST Isochrone expectations, along with other quality of fit parameters in Section 6. We also describe the nature of stellar spectra by varying various stellar parameters in the same section, to gain an intuitive undedrstanding of how spectra can be used to distringuish between stellar properties. Lastly, we present our conclusions in Section 7.

2. DATASET

The Sloan Digital Sky Survey's (SDSS) Apache Point Observatory Galactic Evolution Experiment (APOGEE) is a high resolution, near infrared, spectroscopic survey of 100,000 luminous Milky Way red giant stars (Ahumada et al. 2020). The survey extension, APOGEE-2, is continuing observation on the SDSS 2.5 m telescope, and will also include

Corresponding author: Shrihan Agarwal
shrihan@berkeley.edu

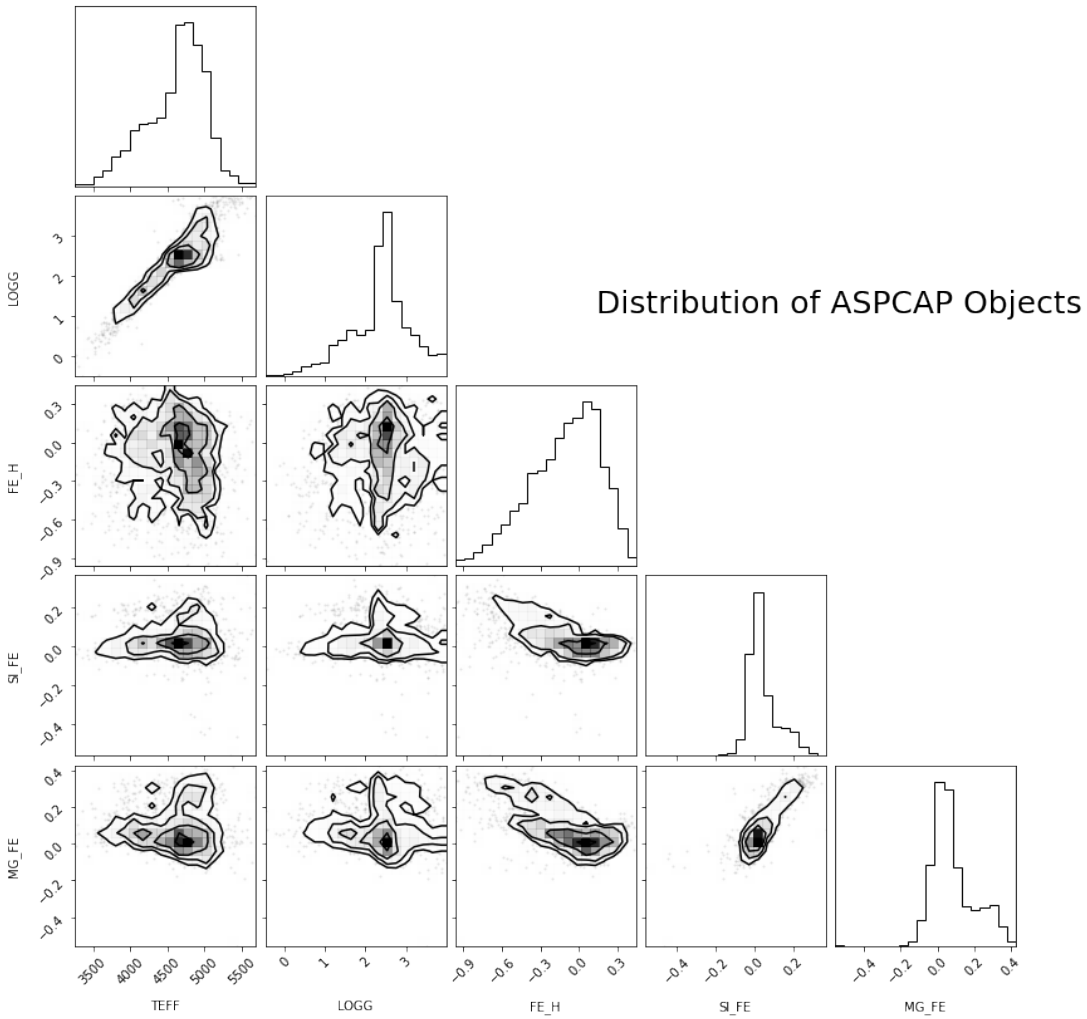


Figure 1: Cornerplot representing the distribution of stellar parameters in the APOGEE dataset. We can make out structure in the data, indicating the existence of a complicated structure in the distribution of stars. We note that stars with a higher $\log(g)$ tend to have a lower T_{eff}, as expected since they are likely larger.

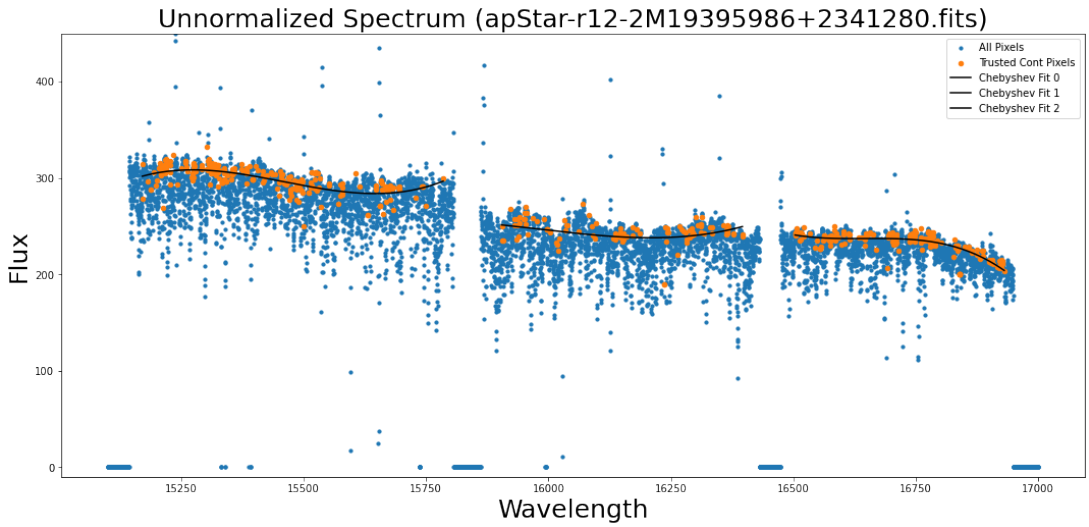
a campaign in the southern hemisphere, providing 400,000 additional stars. We extract all stars unnormalized apStar spectra from the APOGEE database from the following randomly chosen clusters, to produce a diversity of stars in our dataset: M15, 060+00, K2_C4_168-21 and NGC 6791. The spectra are already doppler-shifted to the Earth-Sun barycentric frame, and so it is unnecessary to account for Earth's motion to correct for the shift.

The apStar files consist of 8575 pixel wavelengths, flux, uncertainty, and bitmasks in the 1.5μ to 1.7μ range for each star. We select for only stars with We also discard spectra with low signal-to-noise ratio (SNR 50, as reported in the allStar catalog), the spectra of dwarfs ($\log g$ 4 or T_{eff} 5, 700 K) and stars with low metallicity ([Fe/H] 1). This is in order to remove contaminating stars from the red giant dataset. This cut effectively distinguishes between giants and dwarves, since $\log g$ is a strong function of radius at constant mass, and red giants have low effective surface temperatures. On the HR diagram, this selects for the top right section. This leaves us with 1854 spectra. We begin preprocessing the dataset in the following section.

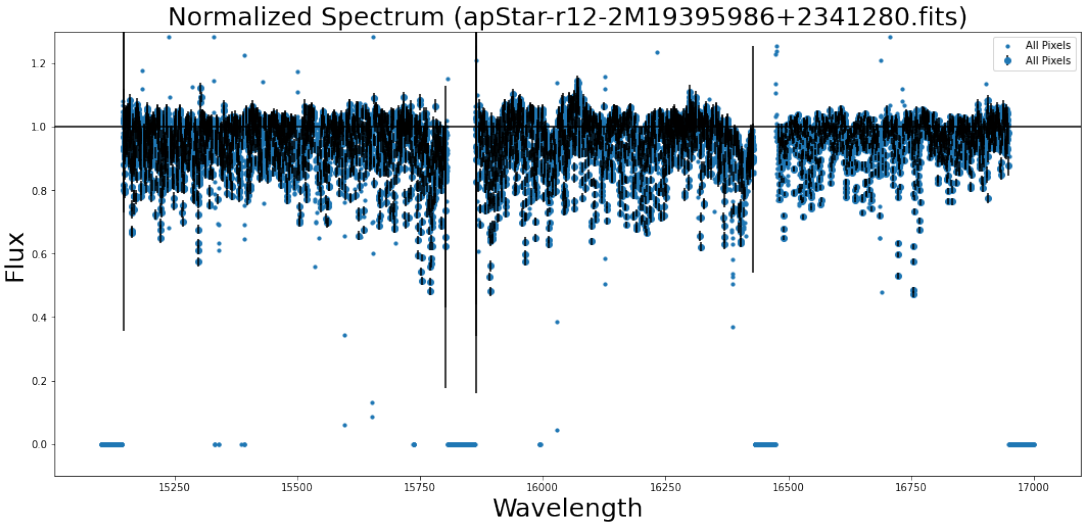
In Figure 1, we show the distribution of APOGEE labels as fit by the "Cannon" (Ness et al. 2015) .

3. CONTINUUM-NORMALIZATION OF APOGEE SPECTRA

3.1. *APOGEE Bitmask*



(a) Unnormalized Spectrum



(b) Normalized Spectrum

Figure 2: For star with 2MASS ID 2M19395986+2341280, we present the Chebyshev fit to the spectrum's pseudo-continuum using the procedure described in 3. The pixels between the APOGEE chips can be seen to have a value of 0, and are avoided by the bitmask. **Top:** Chebyshev Fit is in black, plotted over the orange scatter, or the trusted continuum pixels over which the data was fit, and in blue, we see the remaining part of the dataset, with the absorption line signals. **Bottom:** Post-division, we can see the normalized spectrum and corresponding uncertainties in black. The continuum is presented as a horizontal black line.

We utilize the APOGEE bitmask to filter for cases where the pixel experienced a cosmic ray strike, was saturated, or otherwise determined as "BAD" pixel by APOGEE according to the dark and flat frame subtraction. We flagged bits 0-7 and 12, and set the fractional errors for these values to a large value, 100,000. This ensures these pixels are not used in the fits, while maintaining the length of the spectrum arrays.

3.2. *Continuum Determination*

We use pseudo-continuum normalization for out data, which is better applied to the multiple-chip APOGEE spectra and does not require usage of blackbody models. We determine the pseudo-continuum for each star and each of the three APOGEE chip separately, as they may have systematic errors between them. We select pixels which generally do not have any absorption features, and additionally have no significant error or bitmask issue. In Figure 2 we present

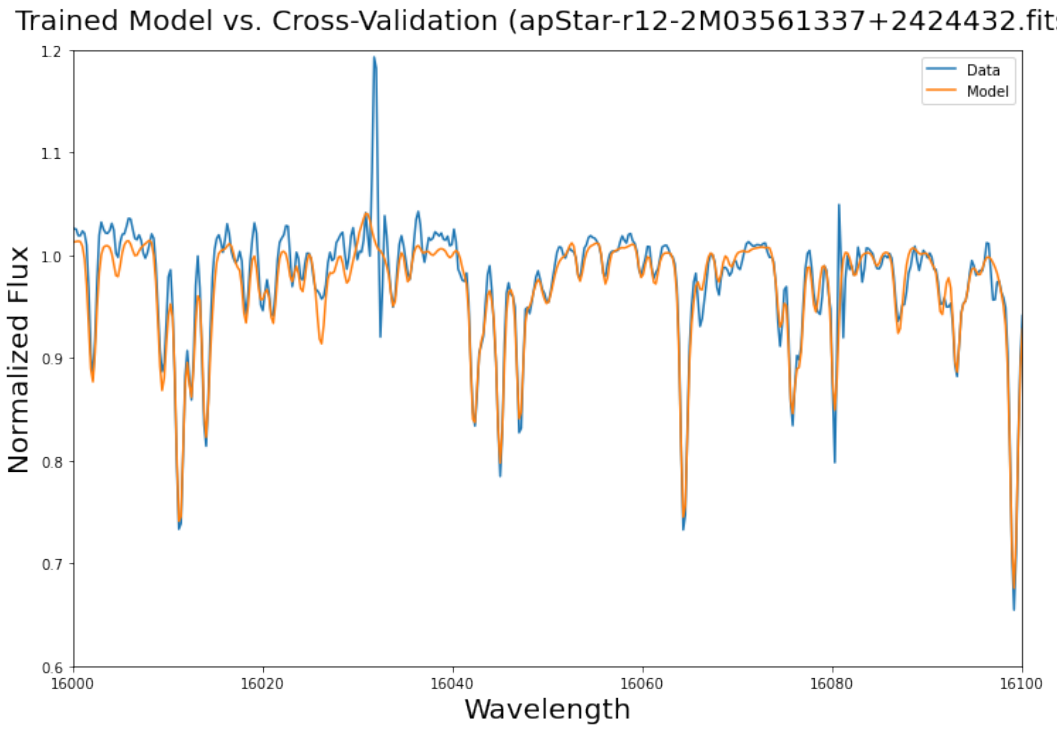


Figure 3: Fitted with the weighted linear regression procedure described in 4. We find remarkable agreement between the model predictions and the true data. Points with significant deviations are likely bad pixels with high error. In the case of the sudden spike at 16030, this is indeed true, and the pixels are assigned high uncertainties for all spectra in the dataset. Our model corrects for this appropriately.

the results of a Chebyshev polynomial fit of degree 3 to our normalized dataset. We find that on using degree 3, we strike a solid balance between bias and variance, as indicated by the chi-squared per degree of freedom.

3.3. *Continuum Normalization*

We then divide both the spectrum and corresponding errors by the normalized value in order to obtain our normalized spectrum. We perform this procedure for all the 1800 spectra in our dataset, and a randomly chosen one is provided in Figure 2. This puts all spectra in a consistent range with selected errors, and is ready to be used for training and prediction.

4. TRAINING A GENERATIVE MODEL WITH NORMALIZED SPECTRA

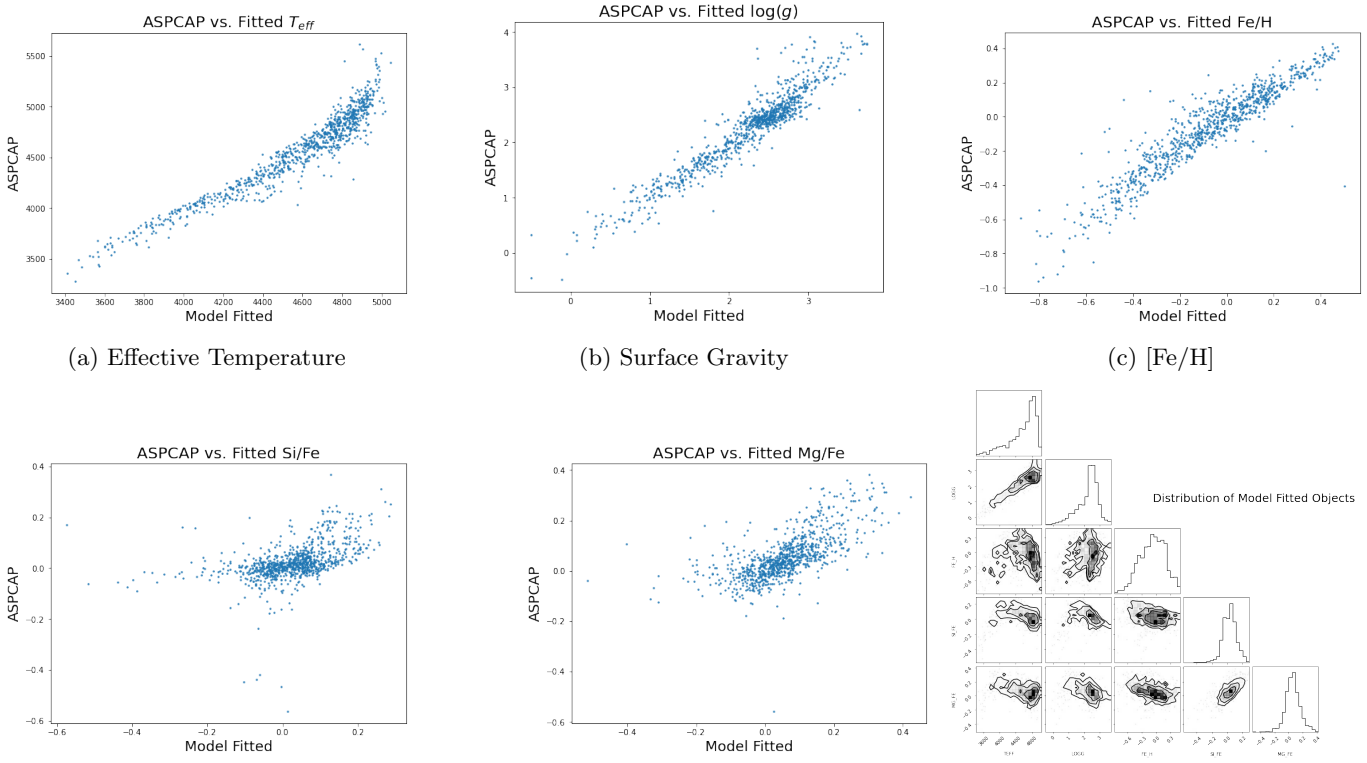
We randomly split the dataset into two equal halves (900 stars) for training and cross-validation respectively. In order to train the model, we use a weighted linear regression approach, with a quadratic feature set, which we have found to be sufficient complexity in order to accurately model the majority of the APOGEE spectra. The benefit of this approach is that the regression can be performed accurately and efficiently, requiring few linear algebra operations.

4.1. *Weighted Linear Least Squares Method*

We generate a quadratic feature set for each star, for the 5 stellar properties, T_{eff}, $\log(g)$, [Fe/H], [Si/Fe], and [Mg/Fe], resulting in a 21 parameter feature set, including the constant coefficient.

$$A_i = [1, T_{eff}, \log(g), [Fe/H], [Si/Fe], [Mg/Fe], T_{eff}^2, T_{eff} \log g, \dots]$$

This can be a row of a matrix A, where the row A_i is for the i'th star in the training set. We perform this on a per-pixel basis, since the principled linear algebra method for weighted linear regression all wavelengths simultaneously requires the generation of a 4D covariance matrix that is computationally intractable to use, without invoking sparse



(a) Effective Temperature (b) Surface Gravity (c) [Fe/H]

(d) [Si/Fe] (e) [Mg/Fe] (f) Model Fitted Cross-Validation.

Figure 4: We compare the model fitted predictions of the various parameters against the ASPCAP labels, ideally seeking a direct correlation between the ASPCAP and model fitted parameters if we assume ASPCAP labels as truth. We find strong agreement and low spread for the first three parameters, with less certainty for [Si/Fe] and [Mg/Fe]. We also present the distribution of cross-validated stars in a cornerplot. This can be compared with the ASPCAP distributions in Figure 1.

matrices. Therefore, for each pixel, using the language of weighted linear regression, we solve for a set of coefficients x such that:

$$Ax = y$$

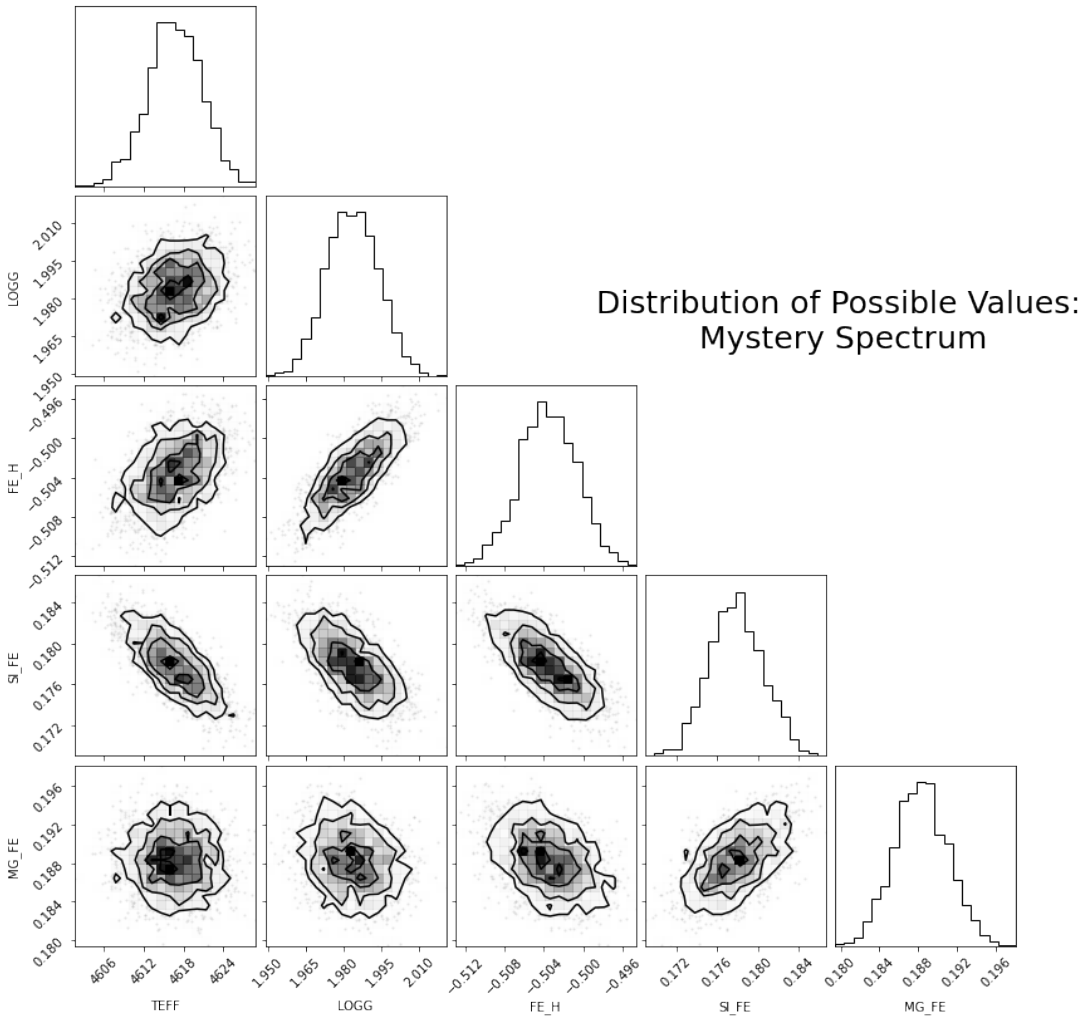
where y is the set of fluxes for all stars in that wavelength. For a weighted linear regression, we weight each of the stars in the training set by the inverse covariance in each flux value. We assume a complete lack of correlated error in all the fluxes. Mathematically, for each pixel, we can represent this as:

$$A^T C^{-1} Ax = A^T C^{-1} y$$

Note: We do not use the SNR value for the fitting, since that is an aggregated value of uncertainty, and will not effectively ignore bitmask pixels. We use the inverse covariance for each pixel as determined by the uncertainty σ_λ for all stars for each wavelength. We also fit for a scatter parameter s_λ, which allows us to correctly weight the uncertainty per pixel, as each pixel may have random uncertainty that differs from any other pixel. We test a grid of 10 values from 0 to 100 and take the best fit s_λ.

An example of a spectral fitting is provided below on the spectrum of a randomly selected spectrum in the cross-validation set 3. It performs similarly well for the majority of stars in the dataset.

With an accurate model that allows us to convert from stellar properties to a spectrum, we allow two different methods, *scipy.curve_fit*, which uses the Trust Region Reflective method, and MCMC, to both fit for the stellar parameters and determine the parameter values. We find that the MCMC is computationally intractable for all the



(a) Posterior Distribution for Mystery Spectrum

Figure 5: The posterior distribution of the mystery spectrum is presented here. We can identify degeneracies between $\log g$ and [FE/H]. We discuss this further in 6. The posteriors seem to underestimate the quanitity of uncertainty, as evidenced by the lower-than expected spread in all the parameters. The posteriors are determined by *scipy.curve_fit*. The mystery spectrum is likely a typical star on the main sequence, with close to average values for all parameters, but extremely low metallicity.

stars in the dataset, but *scipy.curve_fit* can be used for all the stars in the dataset. All determinations of stellar properties beyond this section have been attempted using *scipy.curve_fit*.

5. STELLAR PARAMETER ESTIMATION

5.1. *Root-Mean Squared Error Optimization with* scipy.curve_fit

We use the 'trf' method of *scipy.curve_fit* in order to optimize for the stellar parameter set for each spectrum. For spectra that exceed the maximum number of iterations in order to be fit, throwing a RuntimeError, we ignore the star. In order to improve the runtime, we set a range of reasonable bounds, as well as a starting value that is the mean of the ASPCAP labels for the training data. We obtain good estimates for 924 of the 926 stars, for all 5 parameters. We achieve a strong agreement with ASPCAP labels generated by the "Cannon", with a scatter of 130 K in T_{eff}, 0.22 in $\log(g)$ and 0.08 dex in [Fe/H]. We present results for these fits alongside the comparison with ASPCAP labels and the typical distribution of stars in the entire parameter space in Figure 4.

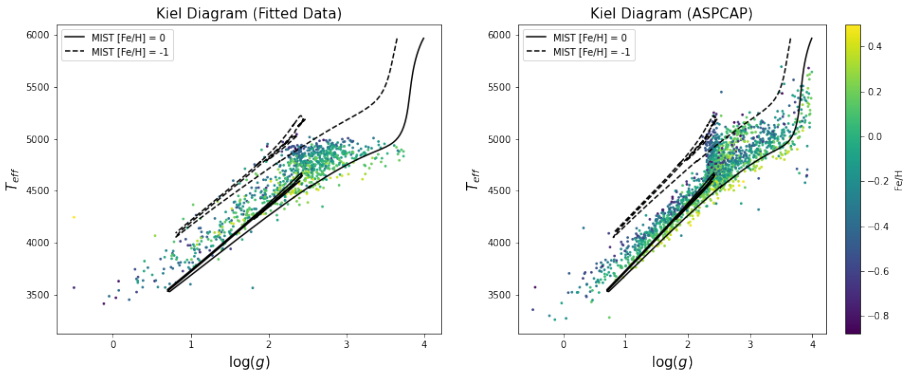


Figure 6: The model trained Kiel diagram (left) alongside the ASPCAP Kiel diagram (right). The black solid line is the MIST Isochrone for [Fe/H] = 0, close to solar metallicity, which seems to agree with the metallicities predicted in that region of the Kiel diagram, according to the colorbar. The black dashed line represents poorer metallicity, confirming both that the drop in metallicity causes a shift in the Kiel diagram to the top right, as we have discovered from our model fitting. The model was unable to fit high T_{eff} spectra, as shown. This is possibly due to the fact that high-temperature spectra have less clear absorption lines.

5.2. *Markov Chain Monte-Carlo*

We utilize MCMC in order to fit for the mystery spectrum provided, set up a log-likelihood function, and minimize it. For our priors, we use uniform distributions over the ranges of values that we used as bounds in the *scipy.curve_fit* routine. The cornerplots for the fitting of the mystery spectra are provided in the Figure 5. The parameter uncertainties are not too large, with a range of 30K for T_{eff}. Considering that the overall scatter of T_{eff} is 130K when compared with ASPCAP labels, it is likely that we are underestimating the uncertainty.

Possible reasons for this are due to the normalization procedure, which may not accurately subtract out the continuum. Additionally, the model may be overfit to the training data. This is a risk in weighted linear regression with no extra regularization term. By overfitting to the training data, the cross-validation data may not by so accurately fit, causing the model on which the least squares minimization is done to be incorrect. Lastly, the data might not be sufficient. We are using 4 randomly chosen sets of clusters, which can have vastly different properties. In Ness et al. (2015) the stars for training are taken from a variety of sources, and so will have more reliable uncertainty estimates, and more diverse training data.

6. STELLAR SPECTRA AS A FUNCTION OF STELLAR AGE AND METALLICITY

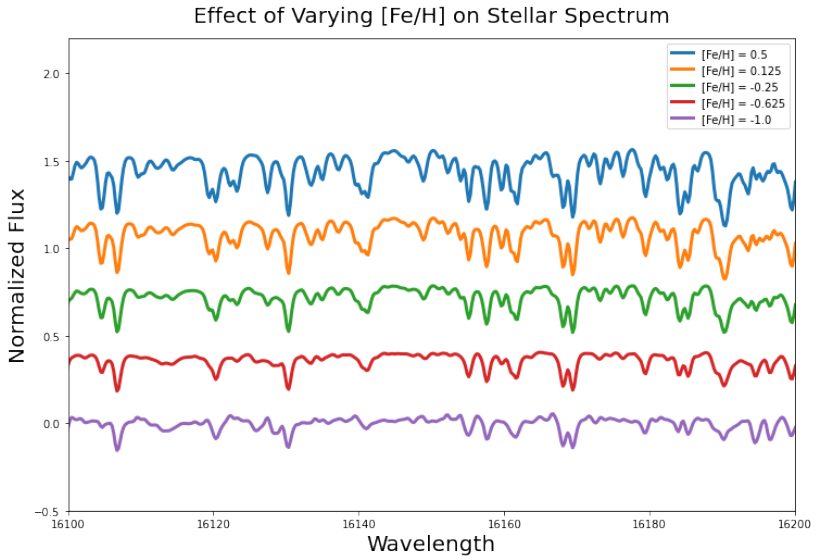
This analysis has allowed us to accurately produce labels for thousands of spectra. This allows us to compare with ASPCAP, as well as theory. We show the Kiel diagram for the model fitted values in Figure 6.

However, it gives further insight into the nature of spectra, and how they can be used to identify particular stellar properties. We use our generative model to predict the spectrum of stars with varying [Fe/H], and with varying log(g) as a proxy for the age of the star on the giant branch.

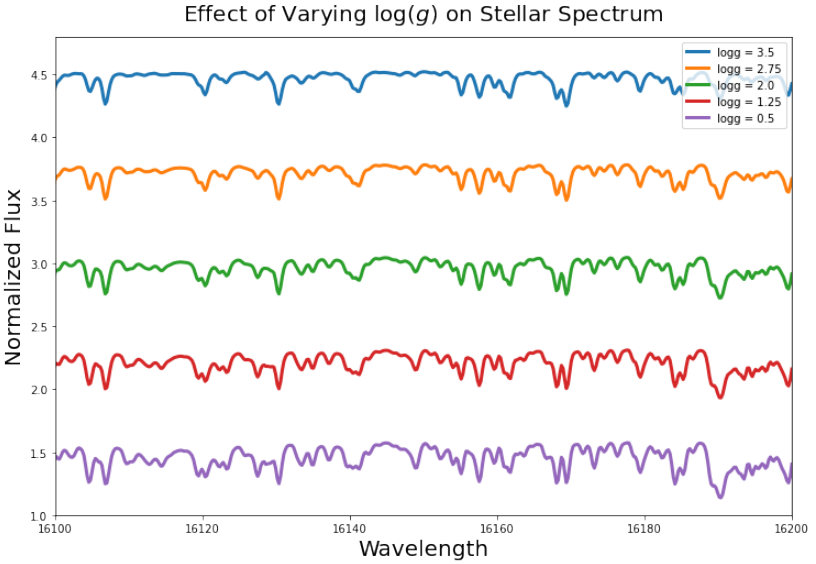
In Figure 7, we both describe and explain the reasoning for the variation. Based on this, it may be difficult to distinguish high-logg, low-metallicity stars from low logg high-metallicity stars, indicating the existence of a degeneracy in the modeling process. Indeed, in Figure 5, where we fit the mystery spectrum, we see that there is a direct uncertainty correlation in the posterior between the [Fe/H] and log(g) parameter, indicating that even the model may find it hard to distinguish between the two. It is still likely possible to distringuish between them. Lower log(g) indicates more absorption lines, since more metals may be present in the star, as it is older. Therefore, it may also be possible to distinguish between the two cases by differentiating between the depth of the absorptionj lines and the number of absorption lines.

7. CONCLUSION AND FURTHER WORK

We achieve a strong agreement with ASPCAP labels generated by the "Cannon", with a scatter of 130 K in T_{eff} , 0.22 in log(g) and 0.08 dex in [Fe/H]. We additionally compare our results with simulated expectations from MIST



(a) Varying [Fe/H]. We find that metal-poor stars tend to have less deep absorption lines, since there are less metals present in order to differentially absorb the wavelengths concerned.



(b) Varying age (with $\log(g)$). We find that aging stars tend to have greater number of absorption lines. This is likely due to the fact that there is an increased optical depth, causing increased selective absorption.

Figure 7: In the figures above we depict the generative model for various stellar parameter values, and find, as expected, that increased metallicity and depth cause an increased absorption - metallicity affecting the depth of the lines, and $\log g$ affecting the number of absorption lines.

Isochrones and find reasonable agreement on a Kiel Diagram. We have developed both a large scale method in order to generate accurate modeling, as well as a smaller scale method to generate accurate posterior distributions using MCMC.

With this technique, datasets from large-scale spectroscopic sky surveys can be accurately and completely evaluated, into the future, advancing our understanding of stellar spectra, and therefore stellar compositions, at all wavelengths, and with relatively small training data.

Further work remains however. The majority of stars in the Milky Way have multiplicities of 2 or greater. If both binaries are unresolvable, only their combined spectra would be detected. This would cause the data to be systematically biased. If the brighter, hotter star has few absorption lines and the dimmer, colder star has many, the brighter, hotter object would wash out the absorption lines, appearing to have a higher T_{eff}. This would hinder our ability to detect more specific [Si/Fe] and [Mg/Fe] compositions, since the absorption lines would be a combination of the two stars. We may be able to more accurately fit for their joint composition by performing a true continuum fit as a sum of two or more stellar continuums. This would still not allow us to distinguish the two stars' stellar compositions however.

Additionally, the modelling process could be significantly improved. Weighted linear regression is prone to overfitting, and may not be the best method to use for a more diverse set of stars, beyond that of the two clusters. In fact, if we look at Figure 6, we notice no predictions are made above a $T_{eff} = 5000$, likely since these were infrequent or even absent in the training dataset. Due to this, the model has overfitted, and not truly learned the variations as a function of temperature and $\log g$. We can improve this by introducing regularization and using a more diverse dataset.

REFERENCES

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3, doi: 10.3847/1538-4365/ab929e

Ness, M., Hogg, D. W., Rix, H. W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16, doi: 10.1088/0004-637X/808/1/16