RED GIANT STELLAR PARAMETER ESTIMATION VIA APOGEE SPECTRA

Abhishek Kattuparambil[1]

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

ABSTRACT

The APOGEE survey began in 2011, cataloguing spectra for over 150,000 red giant stars scattered throughout the Milky Way. Since its public release in July 2014, researchers have iterated on a data-driven model (ASPCAP) to predict stellar parameters and abundances from the spectra measurements. We build a generative model trained on the ASPCAP labels, such that a forward-propagated least linear squares will predict a stellar spectra from labels, and a backwards-propagated nonlinear optimizer can derive the stellar labels for a given spectra. Specifically, the model is a linear system comprised of five input labels ($T_{\rm eff}$, log$g$, [Fe/H], [Mg/Fe], and [Si/Fe]), which will determine a five-variable second-order polynomial to individually estimate the flux at each wavelength pixel. We will define a methodology to derive the labels from a raw APOGEE spectrum, demonstrate the results agree with labels from previous ASPCAP iterations, and extrapolate findings to explore relationships between red giant stars. We provide this model as a tool for automated spectra analysis, and its results as an extension of the APOGEE dataset itself.

Keywords: APOGEE Survey — SDSS-III — Stellar Spectra — Kiel Diagram — Generative Models — Metropolis-Hastings Monte Carlo Markov Chain — Nonlinear Optimization

1. INTRODUCTION

The Apache Point Observatory Galactic Evolution Experiment (APOGEE) began in September 2011 at Apache Point Observatory (APO) in New Mexico, aiming to utilize infrared spectroscopy to survey ∼100,000 red giants scattered throughout the Milky Way, as described in Ahumada (2015) and Majewski (2015). APOGEE is one of the four surveys included in the third phase of the Sloan Digital Sky Survey (SDSS-III), and its final spectra measurements on ∼150,000 sources were released to the public in SDSS' Data Release 12 (DR12), however all data is queried from DR16 to ensure consistent data formatting. Barycentric corrections have been applied to each spectra in the APOGEE dataset, reporting measurements with respect to the center of mass of the Solar System (the barycenter), to correct for the relative motion of the Earth with respect to the observed star. Furthermore, the Doppler shift for each visit differs, and the APOGEE dataset counters this issue via visit combination. A unique Doppler shift is derived for each visit depending on the movement of the Earth and the observed star, occassionally requiring the use of synthetic spectra. The wavelengths are accordingly corrected, and the spectra is resampled onto the new basis. Finally, combined spectra is created as a weighted combination of the individual spectra, where the weight assigned to the individual visit is the square of its signal-to-noise ratio (SNR^2). In order to predict these spectra, the generative model must be trained on the determined ASPCAP labels from the allStar catalog, also included in SDSS DR16. The APOGEE Stellar Parameters and Abundances Pipeline (ASPCAP) is a two-phase process for stellar parameter estimation, consisting of an derivation of atmospheric parameters via the entire spectra, and an evaluation of abundances for small wavelength windows where the respective element's spectral features are prominent. The pipeline was initially bolstered by the labels determined by FERRE code from Prieto (2005), which matched observed spectra against synthetic spectra to determine the star's position in label-space. Since then, the methods for APOGEE label derivation have undergone many iterations Holtzman (2015), including complex evaluations of stellar parameters in pre-determined regions which are sensitive to label variation.

2. DATA PREPARATION

We utilize a subset of the entire APOGEE dataset, filtering sources by their `FIELD`, a unique string encoding their location in the sky. After querying stars whose fields are one of `M15, N6791, K2_C4_168-21`, or `060+00`, each of the 3,036 corresponding apStar files are opened to produce the model's dataset. Furthermore, the remaining sources are cross-referenced with the allStar

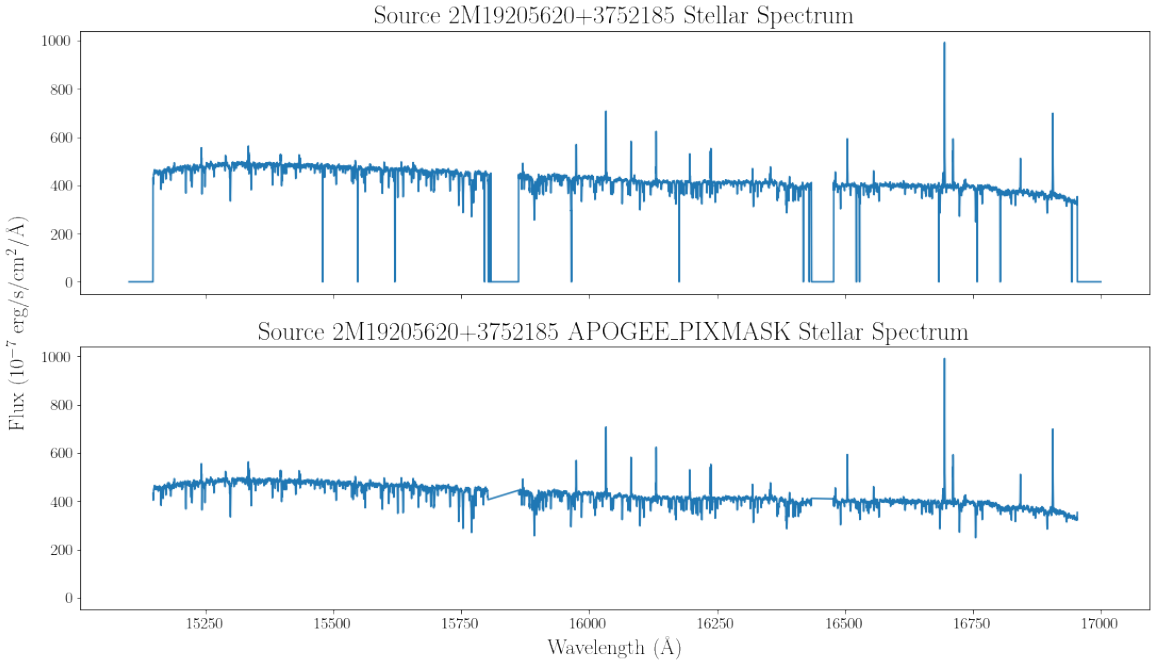

Figure 1. Raw APOGEE spectra (above) and spectra validated via the APOGEE_PIXMASK (below). The gaps between the chips have been removed, alongside a multitude of other erroneous measurements. However, some of the large values remain, and will be filtered out as an outlier in the next section to avoid skewing the normalization.

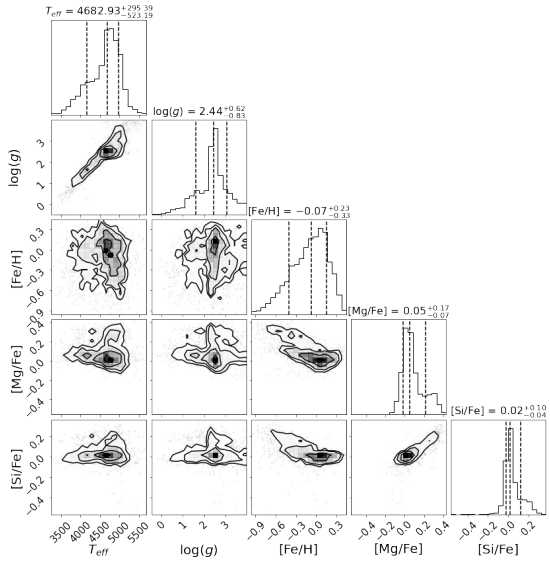

Figure 2. Corner plot showing the distributions and covariances of the labels from our filtered dataset, demonstrating a clear linear relationship between $T_{\rm eff}$ and log$g$. The individual parameter distributions, displayed as kernel density estimations (KDEs), are marked and labeled with the median and the errors, described by the values at the 0.16 and 0.84 quantiles.

catalog, where they must have a large signal-to-noise ratio (SNR > 50). To ensure the dataset only contains red giants, filters are applied on both the effective temperature ($T_{\rm eff}$ < 5700) and surface gravity (log$g$ < 4) to remove dwarfs. The cut on the surface gravity defines an explicit boundary between dwarfs and giants; a late-stage giant has a radius of $100R_\odot$ before the helium flash, resulting in a surface gravity ∼44 times weaker than its radius $15R_\odot$ dwarf counterpart undergoing core helium burning. Finally, we discard metal-poor stars based on their iron abundance ratio ([Fe/H] < -1). The final filtered dataset contains 1855 stars (Figure 1), whose APSCAP labels García-Pérez (2015) will serve as truths for our training and validation sets.

2.1. *Bitmask*

Each spectrum has an assocaited bitmask (APOGEE_PIXMASK), detailing errors and external factors leading to invalid data, including bits marked as saturated, unfixable, cosmic ray, high error, and bad via a multitude of methods. To clean the spectrum to prepare for normalization, pixels marked with error

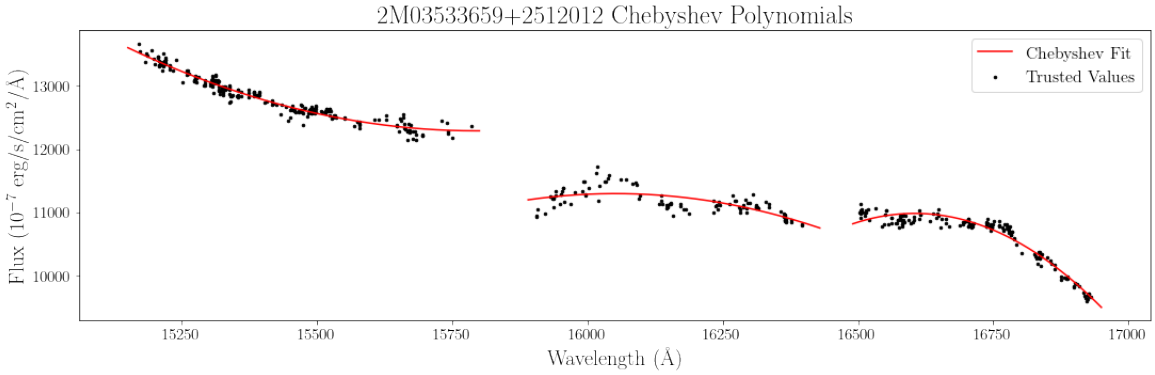

Figure 3. Second-degree Chebyshev polynomials fitted to the final wavelength subset determined by the APOGEE_PIXMASK validations and NPZ trusted wavelengths. An individual polynomial is fit to each chip in the spectra, as they have differing structures. The APOGEE spectra with validated bitmasks is divided by the Chebyshev polynomial to produce the resulting continuum-normalized spectra in Figure 1.

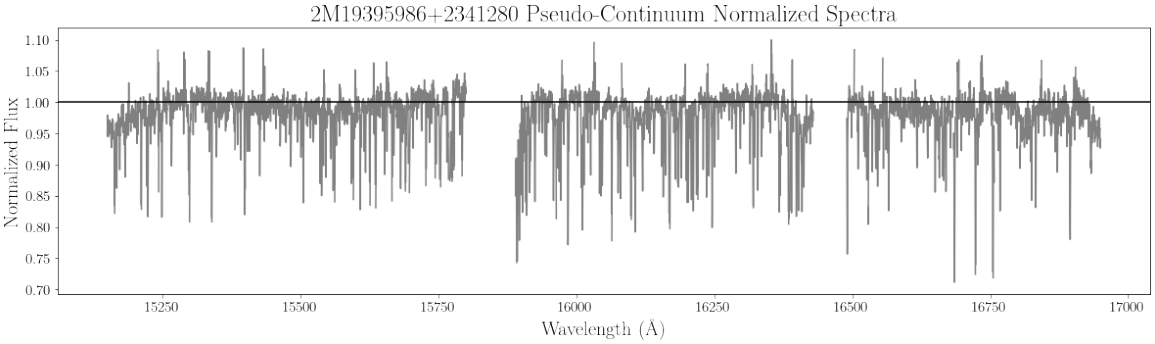

Figure 4. Pseudo-continuum normalized spectrum of source 2M19395986+2341280. The spectrum has been transformed into a range $f_\lambda \in [0.7, 1.1]$, and the absorption lines have been preserved. This model will attempt to fit this normalized spectrum at each pixel wavelength as a function of its stellar labels. As most measurements lie around the normal, the differences between neighboring stars in label-space will be most apparent in the absorption lines.

bits[1] are filtered out. Furthermore, extreme outliers from the flux measurements are discarded, defined as a point further than 5 standard deviations from the mean. This final spectrum is the only flux data used in later sections, invalid bits are propagated with corresponding weights of 0, such that their values are effectively not included in the model.

2.2. *Pseudo-Continuum Normalization*

The observed flux is also sensitive to the absolute magnitude M_G and the distance to the star. The flux is measured in erg/s/cm^2 at each wavelength pixel (Å), which represents the rate of emitted energy traveling through a perpendicular surface area, which is the field of the sky captured by the SDSS telescopes. To create a model that estimates stellar parameters and metal abundances without simultaneously fitting for the magnitude and distance of the star, the spectrum must be normalized to remove these additional dependencies. As displayed in Figure 1, the measurements are split over three chips, whose boundaries are defined at 15800Å-15150Å, 15890Å- 16430Å, and 16490Å- 16950Å. Handling each chip separately, the validated measurements are divided by a second-order Chebyshev polynomial fit to wavelengths that do not contain strong absorption lines; which allows the Chebyshev fit to characterize the sub-spectra solely by their magnitude and distance. To determine which wavelengths do not exhibit strong absorption lines, the model must be iteratively run to ex-

[1] Bits 0-7 and 12 have been found to be most important for filtration

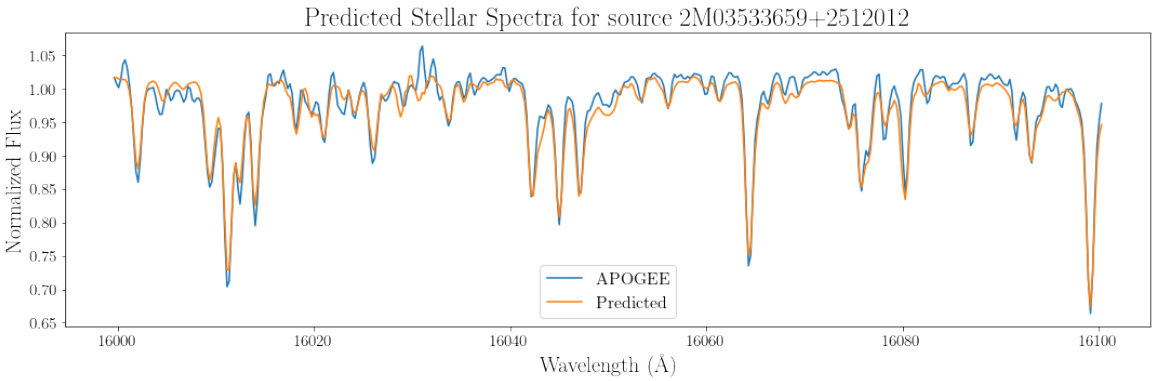

Figure 5. APOGEE stellar spectra for source 2M03533659+2512012 with the model prediction overplotted. The model does well to predict the strength of each absorption line, but struggles to fit the points around and above the normal. This effect is expected, as a small interval around the normal contains a majority of the training points, such that it is difficult to exactly fit for the intrinsic variance in the data. Therefore, the most important predictions are the absorption lines, characterized by large dips where the spectra vary smoothly relative to the stellar labels.

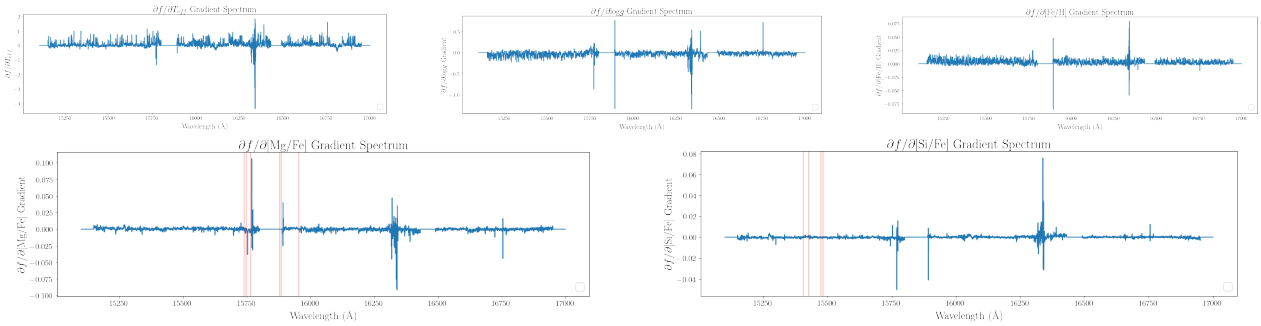

Figure 6. Gradient spectra for each stellar label. The upper three plots are for the stellar labels $T_{\rm eff}$, log$g$, and [Fe/H] respectively. The lower two plots are for the metal abundance ratios [Mg/Fe] and [Si/Fe]. The known absorption lines for Mg and Si are overplotted on the lower graphs, with a few of the lines coinciding with large gradients.

pose wavelengths which are insensitive to variations in the labels. In this case, we have downloaded an NPZ file containing the list of trusted wavelengths from previous ASPCAP iterations.

These wavelengths are ideal for continuum fitting, as they will expose the structure/variance in the data as a function of magnitude and distance. Dividing this structure out of the data places the entire spectrum in the same range such that the training data varies smoothly as a function of the labels. See Figure 1. As shown, our second-order Chebyshev polynomial continuum estimations are only evaluated on insensitive wavelengths with valid bitmasks. Since the continuum is derived from a small subset of trusted wavelengths, this process is a pseudo-continuum normalization, where the continuum is not truly representative of all the data points in our sub-spectrum.

3. THE MODEL

The model's design is based on *The Cannon* (Ness 2015), a data-driven generative model for ASPCAP label derivation. Due to the aforementioned motives of pseudo-continuum normalization, the model requires normalized spectra and associated errors to function. Firstly, we use the five labels ($T_{\rm eff}$, log$g$, [Fe/H], [Mg/Fe], [Si/Fe]) to create a five-parameter second-order polynomial function[2] for each wavelength pixel to predict f_λ, the flux at the given pixel. The final model stitches to-

[2] This function is second-order in the sense that no more than two labels can contribute to a single term; it includes label pairs such as $T_{\rm eff}$*[Fe/H], $T^2_{\rm eff}$, etc.

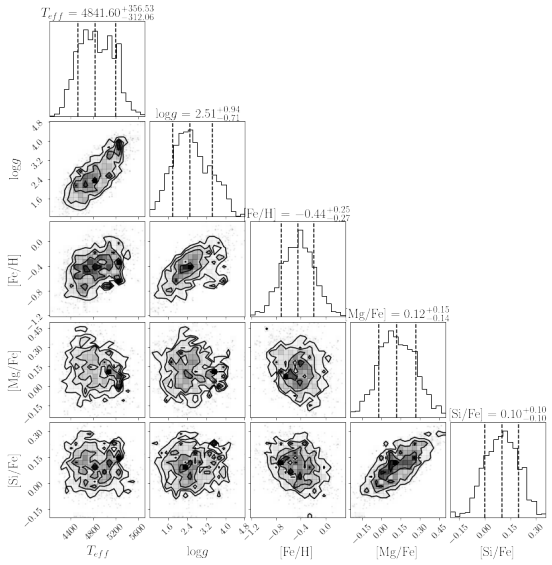

Figure 7. Corner plot showing the distributions and covariances of the predicted labels for the mystery spectra. The individual parameter KDEs are marked and labeled with the median and the errors, described by the values at the 0.16 and 0.84 quantiles. The errors on all of the labels are quite large, and the distributions are flatter than expected when compared to the scatter exhibited in Figure 3.1. The MCMC optimizer imputes the values for bad pixels marked as NaN in the spectra, which will inflate errors as the predicted values will differ at these points.

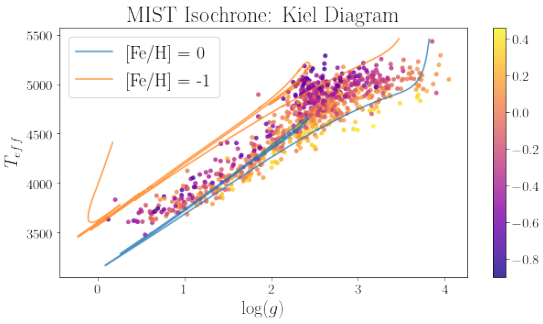

Figure 8. Kiel diagram ($T_{\rm eff}$ v log$g$) of the validation set, constructed with the predicted labels. The points are colored by their iron abundances [Fe/H], with the colorbar shown on the right. There is a clear correlation between $T_{\rm eff}$ and [Fe/H] as hotter stars with the same log$g$ feature lower iron abundances. Furthermore, the graph exhibits a clear linear relationship between the variables of interest, log$g$ and $T_{\rm eff}$. The overplotted isochrones are downloaded from MIST with an age of 6 Gigayears and a metallicity specified in the legend. These isochrones agree with the results of the optimizer, as points with similar predicted metallicities lie along the curve of the isochrones, although the [Fe/H] = -1 isochrone lies above our points, as the smallest predicted metallicities are all above -0.9.

gether the results from each wavelength, such that the entire spectrum can be predicted in one iteration. Additionally, the relationship between the labels and the resulting flux is assumed to be both unique and varying smoothly. Therefore, by iteratively exploring label-space for the best-fitting spectra, an optimizer can predict the ASPCAP labels of a normalized APOGEE spectrum.

3.1. *Training*

Once all spectra and errors in the training set have been normalized, the coefficient determination can be formulated as a weighted least squares problem. Let the number of stars in the training set be **N**, and let the number of pixels be **P**. Therefore, the matrix of normalized spectra **F** is of size (N, P). Since the objective function requires all the second-order terms to derive coefficients, we define a label matrix **A** of size (N, 21) such that each row contains the pre-calculated terms for the fitting polynomial. These terms will include the ASPCAP labels themselves and the second-order terms, as well as a 1 at the end to account for a constant offset. Each label is normalized to order unity since the $T_{\rm eff}$ measurements are of a much larger magnitude than the abundance ratios, making the model disproportionately senesitive to changes in temperature. Finally, the solution of the linear system is defined as Θ, a coefficient matrix of size (21, P), holding the set of 21 coefficients for each pixel. However, solving the system in the current state does not set aside the bad bits previously marked to have erroneous measurements. To propagate the errors through the training stage, we also define a weight matrix **W** of shape (P, N, N), storing a diagonal NxN matrix storing the weight for each star for each pixel wavelength. The weights are simply the inverse of the normalized errors, such that stars with low errors get proportionately large weights. As mentioned in the previous section, invalid bits are assigned a weight of 0, ensuring the model does not factor their erroneous values into its training. See Equation (3.1).

3.2. *Cross-Validation*

To confirm the accuracy of the model, cross-validation can be utilized by comparing the predicted labels of our validation set spectra to their official ASPCAP labels. To estimate the labels, the model must be solved backwards, matching a set of labels to the data based on the similarity (χ^2) between the theoretical spectrum and the observed: this process will be streamlined with `scipy.optimize.curve_fit`, which utilizes the Trust

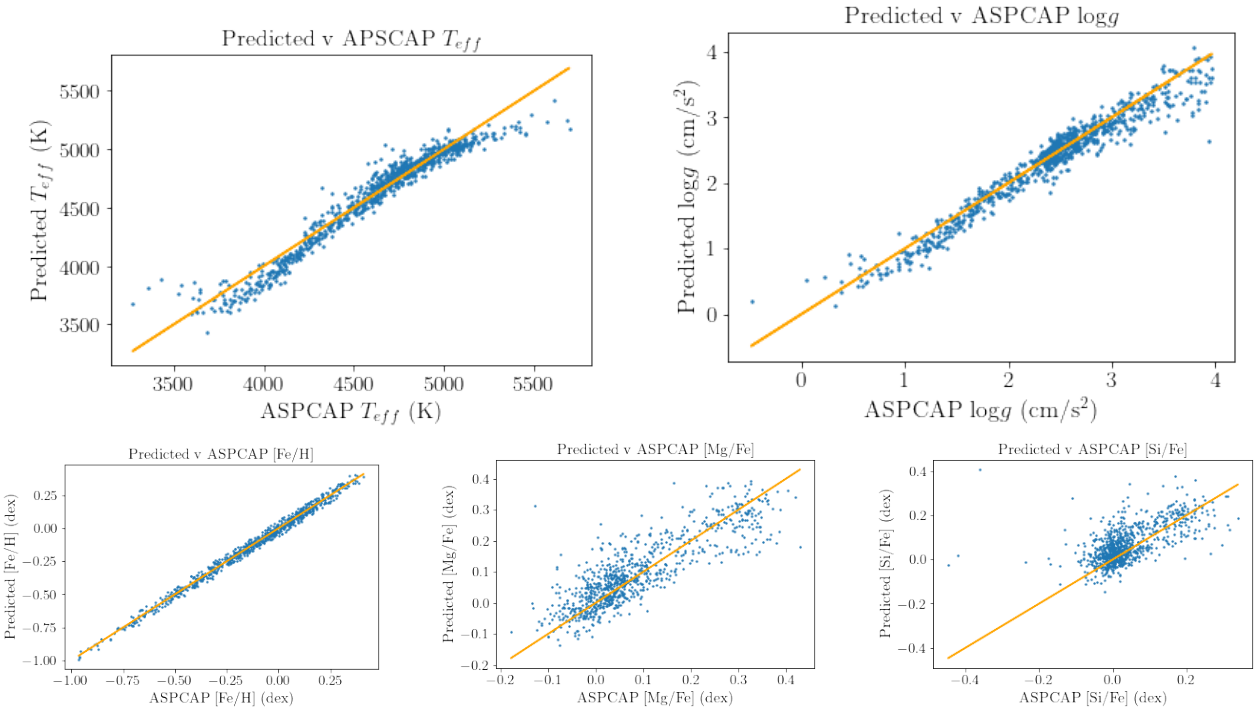

Figure 9. Model predictions vs ASPCAP labels for each stellar label. The upper two plots are the stellar parameters $T_{\rm eff}$ and log$g$. The lower three plots are the metal abundance ratios, specifically [Fe/H], [Mg/Fe], and [Si/Fe] in order. Each graph can be treated as a residual, where the diagonal is the normal. Each parameter demonstrates a linear relationship, showing that the model has truly derived the relation as expected. However, the scatter is quite large for the final two metallicites. Furthermore, the $T_{\rm eff}$ plot exhibits a clear cubic structure, where the model overestimates low temperatures and underestimates large temperatures. This drives an iteration on the model outlined in Figure 3.1.

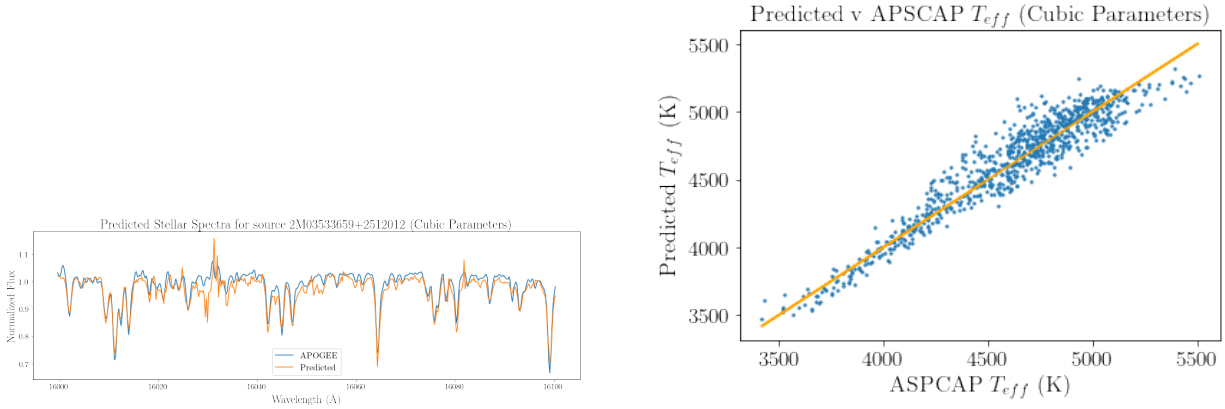

Figure 10. APOGEE and ASPCAP measurements compared to model predictions with cubic $T_{\rm eff}$ terms. This model includes the terms $T_{\rm eff}^3$, log$gT_{\rm eff}^2$, [Fe/H]$T_{\rm eff}^2$, and so on. These additional terms were motivated by the $T_{\rm eff}$ results in Figure 3.1; the model attempts to correct for the cubic residual structure by making the variable $T_{\rm eff}$ vary as a cubic itself. Additionally, the stars with reported temperatures of above 5100 or below 3900 were weighted tenfold, such that the model would be forced to fit them with high accuracy. This resulted in a final pipeline that predicts low and high temperatures more accurately, but produces a larger scatter than its second-order counterpart. As shown on the left, the residuals for predictions around the normal are larger, as the spectra is largely dependent on stars with extreme temperatures, which will have distinctly differing spectra.

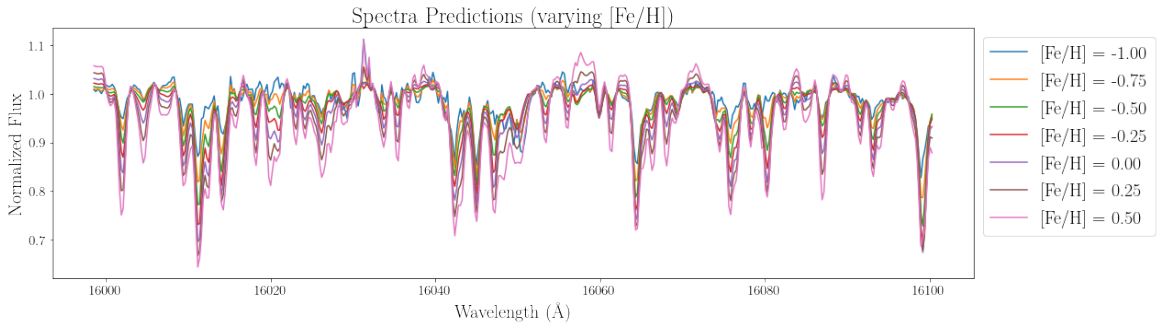

Figure 11. Normalized spectrum predictions for a theoretical star whose metallicity [Fe/H] varies from -1 to 0.5, with stellar parameters set to the average of the dataset: $T_{\text{eff}} = 4682.93$, $\log g = 2.44$. Since the model is trained on normalized labels, these values are functionally set to 1, while the abundance ratios ([Mg/Fe] and [Si/Fe]) are set to 0 to match solar metallicities. These spectra show a clear relationship between the metal abundance ratio [Fe/H] and the flux output, where the larger abundance ratios correspond with larger spectra amplitudes, such that the absorption lines are deeper and the peaks are higher.

$$\begin{bmatrix} T_{\text{eff}_0} & \log g_0 & \dots & [Mg/Fe]_0[Si/Fe]_0 & 1 \\ T_{\text{eff}_1} & \log g_1 & \dots & [Mg/Fe]_1[Si/Fe]_1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ T_{\text{eff}_N} & \log g_N & \dots & [Mg/Fe]_N[Si/Fe]_N & 1 \end{bmatrix} W^{1/2} \begin{bmatrix} \theta_{\lambda_0 0} & \theta_{\lambda_1 0} & \dots & \theta_{\lambda_P 0} \\ \theta_{\lambda_0 1} & \theta_{\lambda_1 1} & \dots & \theta_{\lambda_P 1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{\lambda_0 20} & \theta_{\lambda_1 20} & \dots & \theta_{\lambda_P 20} \end{bmatrix} = \begin{bmatrix} f_{\lambda_0 0} & f_{\lambda_1 0} & \dots & f_{\lambda_P 0} \\ f_{\lambda_0 1} & f_{\lambda_1 1} & \dots & f_{\lambda_P 1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{\lambda_0 20} & f_{\lambda_1 20} & \dots & f_{\lambda_P 20} \end{bmatrix} W^{1/2}$$

Figure 12. The model formulated as a weighted least squares problem. In practice, this will be of the form $AW^{1/2}\Theta = FW^{1/2}$. The weight matrices are set to the reciprocal of the errors from each pixel as reported by the APOGEE dataset. Each weight vector will contain as many values as stars in our training set (N), which must be aligned along the diagonal of an NxN matrix to properly match the dimensions of the other matrices. Each coefficient column θ can be used to predict the flux at a unique wavelength pixel.

Region Reflective (TRF) algorithm. The optimizer will solve the aforementioned linear system as a nonlinear least squares problem, this time solving for the optimal term vector $\overrightarrow{a}$, requiring the per-pixel coefficient array and the flux measurements as inputs. It will iterate on $\overrightarrow{a}$, converging towards an optimal vector which will contain the labels in its first five elements, due to the nature of our **A** matrix construction (3.1). In order to predict labels for the stars in the validation set, the optimizer is applied to each of the normalized spectra. For comparisons between the two sets of labels, see Figure 3.1. Furthermore, the derived coefficients are generalizable to any set of red giant stars; they can be used to estimate stellar labels for spectra not detailed in the APOGEE dataset. In Figure 1, the model (whose label-space is wrapped in a NUTS MCMC sampler) predicts stellar parameters and abundances for a newly introduced spectrum.

4. ASPCAP LABELING

Now that a method has been defined to derive stellar labels from spectra, we will demonstrate relationships between the stars in label space. In order to demonstrate the correlation between T_{eff} and $\log g$, a Kiel diagram is produced to match synthetic MIST isochrones (Figure 1). Once again returning to spectrum prediction via stellar input labels, the spectral variance induced by changes in metallicity is examined in (Figure 3.1).

4.1. *Binaries*

The entire pipeline functions under the assumption that each spectrum is for a single star, such that there are no binary systems recorded in the dataset. As discussed in El-Badry (2017), binary systems produce an amplified spectrum, which should exhibit more variation around the normal due to the stars having differing stellar parameters and compositions. The model fits these spectra under the assumption that both stars in the binary are from the same isochrone, and goes further to fit the mass ratio ($q = m_2/m_1$), v_{macro}, and v_{helio} to recover the stellar parameters of the second star. However, the identification of binary systems is exhaustive, predicted spectra that differ from the normalized APOGEE spectra by some threshold will be evaluated against the binary model, with two superimposed predicted spectra. In these binary models, the stars can contribute a grossly disproportionate amount of flux, such that one is much hotter than the other. Nonetheless, both play a key role in the final prediction, as cooler stars often have naturally stronger absorption lines. The unique features from both of these spectra are superimposed to create the final spectra modeling the binary system. The current model will fit all these binaries as single stars, and will underestimate the surface gravity and overestimate the effective temperature and metal abundance ratios.

5. SUMMARY

We have defined and validated a process for stellar label estimation from continuum normalized spectra, producing results that differ from ASPCAP effective temperatures with a scatter of 50 K, and predicted ASPCAP metallicities with a mean scatter of 0.07 dex, largely due to errors in magnesium abundance ratio estimations. Nonetheless, the accuracy and performance of the model, despite its limited expressiveness, allows for the experimentation and the prediction of both red giant spectra and stellar labels. This is a purely generative model, trained on existing spectra and associated labels, which does not require provision of synthetic spectra to produce a match in label-space. Furthermore, the model has been trained on only 900 sources, with the capability to extrapolate and predict spectra/labels for any plausible red giant star.

REFERENCES

Ahumada, Romina, P. C. A. A. A. 2015, The Sixteenth Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. https://arxiv.org/abs/1912.02905

El-Badry, Kareem, T. Y.-S. R. H.-W. 2017, Discovery and Characterization of 3000+ Main-Sequence Binaries from APOGEE Spectra. https://arxiv.org/abs/1711.08793

García-Pérez, Ana E., P.-C. A. H.-J. A. 2015, ASPCAP: The Apogee Stellar Parameter and Chemical Abundances Pipeline. https://arxiv.org/abs/1510.07635

Holtzman, Jon A., S.-M. J. J. A. 2015, Abundances, Stellar Parameters, and Spectra From the SDSS-III/APOGEE Survey. https://arxiv.org/abs/1501.04110

Majewski, Steven R., S.-R. P. F. P. M. 2015, The Apache Point Observatory Galactic Evolution Experiment (APOGEE). https://arxiv.org/abs/1509.05420

Ness, Melissa, H. D. W. R. H.-W. 2015, The Cannon: A data-driven approach to stellar label determination. https://arxiv.org/abs/1501.07604

Prieto, Carlos A., B. T. C. R.-W. 2005, A Spectroscopic Study of the Ancient Milky Way: F- and G-Type Stars in the Third Data Release of the Sloan Digital Sky Survey. https://arxiv.org/abs/0509812