

# DATA-DRIVEN SPECTROSCOPY OF COOL STARS AT HIGH SPECTRAL RESOLUTION

AIDA BEHMARD<sup>1</sup>

<sup>1</sup>Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA

ERIK A. PETIGURA<sup>2,3</sup>, ANDREW W. HOWARD<sup>2</sup>

<sup>2</sup>Cahill Center for Astrophysics, California Institute of Technology, Pasadena, CA 91125, USA and

<sup>3</sup>Hubble Fellow

(Received; Accepted)

Draft version April 2, 2019

## ABSTRACT

The advent of large-scale spectroscopic surveys underscores the need to develop robust techniques for determining stellar properties ("labels", i.e., physical parameters and elemental abundances). However, traditional spectroscopic methods that utilize stellar models struggle to reproduce cool (<4700 K) stellar atmospheres due to an abundance of unconstrained molecular transitions, making modeling via synthetic spectral libraries difficult. Because small, cool stars such as K and M dwarfs are both common and good targets for finding small, cool planets, establishing precise spectral modeling techniques for these stars is of high priority. To address this, we apply *The Cannon*, a data-driven method of determining stellar labels, to Keck High Resolution Echelle Spectrometer (HIRES) spectra of 141 cool (<5200 K) stars from the California Planet Search. Our implementation is capable of predicting labels for small (<1  $R_{\odot}$ ) stars of spectral types K and later with accuracies of 68 K in effective temperature ( $T_{\text{eff}}$ ), 5% in stellar radius ( $R_{*}$ ), and 0.08 dex in bulk metallicity ([Fe/H]), and maintains this performance at low spectral resolutions ( $R < 5000$ ). As M-dwarfs are the focus of many future planet-detection surveys, this work can aid efforts to better characterize the cool star population and uncover correlations between cool star abundances and planet occurrence for constraining planet formation theories.

*Keywords:* methods: data analysis — methods: statistical — stars: fundamental parameters — surveys — techniques: spectroscopic

## 1. INTRODUCTION

Precise determination of stellar properties (e.g., masses, radii, effective temperatures, elemental abundances) is a challenging, yet essential component of stellar and planetary astrophysics. Accurate measurements of masses ( $M_{*}$ ), radii ( $R_{*}$ ), and temperatures ( $T_{\text{eff}}$ ) are crucial for vetting models of stellar structure and evolution, and the chemical compositions of stellar photospheres reflect formation histories and can link stars to their parent molecular clouds, providing a window into galactic chemical evolution. The burgeoning field of exoplanets also calls for robust methods of determining stellar properties as characterization of planets is predicated on thorough characterization of their stellar hosts.

Stellar spectroscopy has a rich history, beginning with Annie Jump Cannon and her colleagues at Harvard College Observatory who developed the current stellar classification system based upon visual inspection of spectral features. Modern spectroscopic methods involve matching information-rich portions of empirical spectra to benchmark or synthetic spectra generated from model stellar photospheres. Two commonly-used spectral modeling tools are SME and MOOG (Valenti & Piskunov 1996; Sneden 1973), both of which have undergone significant evolution since their inception (e.g., Valenti & Fischer 2005; Valenti et al. 2009; Deen 2013; Brewer et al. 2015; Piskunov & Valenti 2017). However, current model photospheres are limited by an incomplete knowledge of the physics behind stellar attributes; they suffer from poorly constrained atomic and molecular opacities, often assume local thermodynamic equilibrium (LTE), and inaccurately model dynamical effects such as convection or stellar

winds, if at all. While three-dimensional hydrodynamic models have been created that allow for non-LTE conditions, they still suffer from the other aforementioned drawbacks and are computationally expensive. Laboratory studies have refined atomic and molecular data and improved line lists, but departures from solar type atmospheres still present significant modeling challenges.

Stars of spectral types K4 ( $T_{\text{eff}} \lesssim 4700$  K) and later are particularly difficult to model with synthetic spectral techniques as their optical and NIR spectra feature dense clusters of molecular lines that lack reliable opacity data. In the optical regions of K and M dwarf spectra, TiO and VO bands are prominent, as well as hydride bands such as MgH, CaH, and FeH. The NIR regions of M dwarf spectra often feature H<sub>2</sub>O (e.g., Rojas-Ayala et al. 2012). Characterization of late-type stars such as M dwarfs is important because they are common, representing  $\sim 75\%$  of stars in the solar neighborhood (Henry et al. 2006). Small, cool stars are also popular targets for exoplanet surveys as their low  $M_{*}$  and  $R_{*}$  result in deeper transit signals and larger Doppler shifts, increasing the probability of detecting and characterizing small planets.

Empirical methods offer alternative routes for predicting K and M dwarf parameters and abundances. Common proper motion pairs of M dwarfs and F-, G-, K-type stars of known metallicities ([Fe/H]) can be used to calibrate M dwarf metallicities with equivalent widths (EWs) of NIR spectral features (Newton et al. 2014; Mann et al. 2014). Similarly, temperatures ( $T_{\text{eff}}$ ) and stellar radii ( $R_{*}$ ) can be calibrated with EWs of K and M dwarf NIR spectra (Newton et al. 2015), and parallaxes can provide further constraints on stellar properties (Mann et al. 2015, 2017). Empirical as opposed to syn-

thetic spectral libraries composed of touchstone stars with well-measured properties are also capable of predicting accurate parameters for stars of mid-K spectral types and later (Yee et al. 2017).

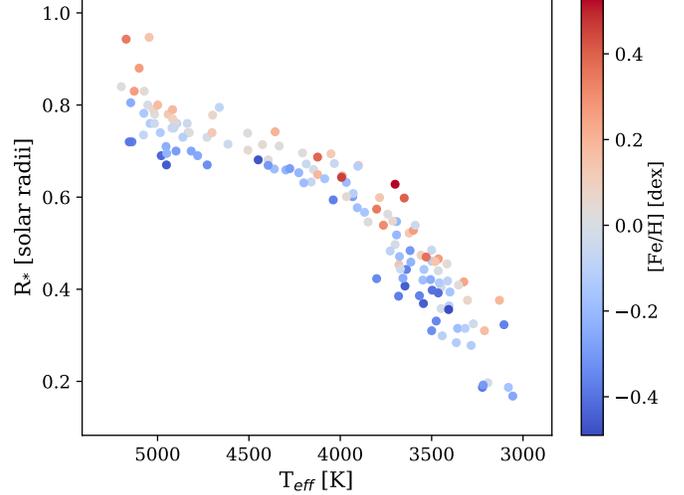
Another promising method for modeling cool stars is offered by *The Cannon*, a data-driven approach to modeling spectroscopic data (Ness et al. 2015). In brief, *The Cannon* predicts stellar parameters and elemental abundances from spectroscopic data via a two-step process: a "training step" where the spectra for a set of reference objects with well-determined parameters and/or abundances are used to construct a predictive model of the flux, and a "test step" where the model is used to infer those of objects given their spectra. Unlike traditional spectroscopic modeling methods, *The Cannon* makes no use of physical stellar models, and does not require an accompanying library of synthetic spectra for reference. Here, we modify *The Cannon* to optimize parameter and elemental abundance predictions for K and M dwarfs with HIRES spectra.

Throughout this work, we refer to stellar parameters and elemental abundances ( $T_{\text{eff}}$ ,  $R_*$ , and  $[\text{Fe}/\text{H}]$ ) as "labels" to be consistent with previous literature on *The Cannon* (e.g., Ness et al. 2015; Casey et al. 2016; Ho et al. 2017) and to adhere to machine learning/supervised methods terminology. We evaluate *The Cannon*'s ability to predict stellar labels in our cool star sample with cross-validation experiments. Cross-validation was carried out by dividing a reference set of cool stars with well-determined labels into training and validation sets. The reference set is pulled from a library compiled by Yee et al. (2017) (see Section 2 for more details). Performance was evaluated by examining how well *Cannon*-predicted labels for the validation set matched those reported in the library. In Section 3, we present *The Cannon*, and outline our implementation and its performance on our cool star sample in Section 4. We find that *The Cannon* can predict labels with precisions of 68 K in  $T_{\text{eff}}$ , 5% in  $R_*$ , and 0.08 in  $[\text{Fe}/\text{H}]$  (dex). Discussion of the results is presented in Section 5.

## 2. COOL STAR SAMPLE

Our spectral library was compiled by Yee et al. (2017) and consists of 404 touchstone stars originating from several source catalogs that span the spectral types  $\sim\text{M5-F1}$  ( $T_{\text{eff}} \approx 3000\text{--}7000$  K,  $R_* \approx 0.1\text{--}16 R_\odot$ ). The stars have spectra obtained from HIRES at the Keck-I 10-m telescope (Vogt et al. 1994) as part of the California Planet Search (CPS). For more details on CPS, see Howard et al. (2010). The HIRES spectra are high-resolution ( $R \approx 60,000$ ) and high signal-to-noise ( $\text{S/N} > 40/\text{pixel}$ , with  $\sim 80\%$  having  $\text{S/N} > 100/\text{pixel}$ ). The spectra originate from the middle HIRES detector CCD chip and contain 16 spectral orders. The HIRES blaze function has been removed and the spectra registered onto a common wavelength scale ( $\lambda = 4990\text{--}6410$  Å) uniform in  $\Delta \log \lambda$  to ensure that linear velocity shifts correspond to linear pixel shifts (Yee et al. 2017). We confined the wavelength range to 13 orders ( $\lambda = 4990\text{--}6095$  Å) to avoid redder portions of the middle HIRES CCD chip that are more affected by tellurics.

To isolate a cool star sample composed of K and M dwarfs, we employed radius and temperature cuts of  $T_{\text{eff}} < 5200$  K and  $R_* < 1 R_\odot$ , leaving 141 stars. These cool stars are primarily drawn from the catalog described in Mann et al. (2015) with  $T_{\text{eff}}$ ,  $R_*$ , and  $[\text{Fe}/\text{H}]$  determined from a combination of spectrophotometry, SED modeling, *Gaia* parallaxes, and EW empirical relations (quoted uncertainties of 60 K,



**Figure 1.** The domain of  $T_{\text{eff}}$ ,  $R_*$ , and  $[\text{Fe}/\text{H}]$  for our reference sample of 141 cool stars pulled from the library outlined in Yee et al. (2017). The cool stars have temperatures and radii that satisfy  $T_{\text{eff}} < 5200$  K and  $R_* < 1 R_\odot$ .

3.8%, and 0.08 dex, respectively). A smaller subset originate from the catalog compiled by von Braun et al. (2014), and have interferometrically-determined  $R_*$  (quoted uncertainties of  $< 5\%$ ). Many of the early K dwarfs in the sample have  $T_{\text{eff}}$  and  $[\text{Fe}/\text{H}]$  determined from LTE spectral synthesis carried out by Brewer et al. (2016) with SME (quoted uncertainties of 60 K and 0.05 dex, respectively), while the sample mid to late K dwarfs have  $T_{\text{eff}}$ ,  $R_*$ , and  $[\text{Fe}/\text{H}]$  determined from a combination of spectrophotometry, SED modeling, parallaxes, and SME analysis carried out by Yee et al. (2017) (quoted uncertainties of 5%, 7.4%, and 0.1 dex, respectively). Because most of these catalogs do not provide a complete set of  $T_{\text{eff}}$ ,  $R_*$ , and  $[\text{Fe}/\text{H}]$  values, Yee et al. (2017) conducted an isochrone analysis using Dartmouth stellar models (Dotter et al. 2008) to obtain a homogeneous label set, and took uncertainties as the 5th and 95th percentiles of the MCMC distributions that resulted from fitting to the stellar model grids. For more details on any of the library catalogs or the isochrone analysis procedure, see Yee et al. (2017).

## 3. THE CANNON

### 3.1. Preparing HIRES spectra for The Cannon

To prepare the spectral library for *The Cannon*, we must ensure that the spectra satisfy certain conditions; the spectra must share a common wavelength grid, be shifted onto the rest wavelength frame, share a common line-spread function, and be continuum-normalized via a method independent of S/N (Ness et al. 2015). The first two conditions are already satisfied for the library spectra, and we can assume that they effectively share a line-spread function, though there may be negligible variation due to variable observation seeing conditions. To carry out normalization, we applied error-weighted, broad Gaussian smoothing with

$$\bar{f}(\lambda_0) = \frac{\sum_j (f_j \sigma_j^{-2} w_j(\lambda_0))}{\sum_j (\sigma_j^{-2} w_j(\lambda_0))}, \quad (1)$$

where  $f_j$  is the flux at pixel  $j$  of the wavelength range,  $\sigma_j$  is the uncertainty at pixel  $j$ , and the weight  $w_j(\lambda_0)$  is drawn from a Gaussian:

$$w_j(\lambda_0) = e^{-\frac{(\lambda_0 - \lambda_j)^2}{L^2}}, \quad (2)$$

where  $L$  was chosen to be  $3 \text{ \AA}$ . If larger  $L$  values are chosen for HIRES spectra, continuum-normalization begins to remove high resolution features. For reference, [Ho et al. \(2017\)](#) used a width of  $L = 50 \text{ \AA}$  to normalize low-resolution Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) spectra ( $R \approx 1800$ ). The Gaussian smoothing procedure is illustrated in [Fig. 2](#).

### 3.2. Training Step

We used *The Cannon 2*, the second implementation of *The Cannon* developed by [Casey et al. \(2016\)](#). Hereafter, we will refer to *The Cannon 2* simply as *The Cannon*. This version builds upon the original with additional features that are designed to aid prediction of a larger label set including elemental abundances that go beyond bulk metallicity ([Fe/H]), such as regularization.

As outlined in [Section 1](#), in the training step, *The Cannon* uses a set of reference objects with well-determined labels to construct a predictive model of the flux at every pixel of the wavelength range that is a function of the stellar labels. Model construction is based on two assumptions: that continuum-normalized spectra with identical labels look identical at every pixel, and that the flux at every pixel in a spectrum changes continuously as a function of the stellar labels. While *The Cannon* can be trained on any set of empirical spectra and their labels, the resultant model will only be capable of predicting labels for spectra with properties that are represented in the training set. In other words, *The Cannon* is not able to accurately extrapolate outside the training set parameter space, so the training set spectra must be representative of the test set spectra in order to predict accurate label values. It is also important to note that the *Cannon*-predicted labels will only be as accurate as those of the training set.

The flux model  $f_{jn}$  for a spectrum  $n$  at pixel  $j$  can be written as

$$f_{jn} = \mathbf{v}(l_n) \cdot \boldsymbol{\theta}_j + e_{jn}, \quad (3)$$

where  $\boldsymbol{\theta}_j$  is the set of spectral model coefficients at each pixel  $j$  and  $\mathbf{v}(l_n)$  is a function of the label list  $l_n$  that is unique for each spectrum  $n$ . The function  $\mathbf{v}(l_n)$  is referred to as the "vectorizer" which can accommodate functions that are linear in the coefficients  $\boldsymbol{\theta}_j$ , but not necessarily simple polynomial expansions of the label list  $l_n$ . The noise term is described by  $e_{jn}$  and can be taken as sampled from a Gaussian with zero mean and variance  $\sigma_{jn}^2 + s_j^2$  where  $\sigma_{jn}^2$  is the uncertainty reported on the input HIRES spectra (flux variance) and  $s_j^2$  is the intrinsic scatter of the model at each pixel  $j$ . This intrinsic scatter can be likened to the expected deviation of the model from the spectrum at  $j$ .

To determine the optimal model labels  $(\boldsymbol{\theta}_j, s_j^2)$ , we can relate the flux model to a single-pixel log-likelihood function:

$$\ln p(f_{jn} | \boldsymbol{\theta}_j, \mathbf{v}(l_n), s_j^2) = -\frac{[f_{jn} - \mathbf{v}(l_n) \cdot \boldsymbol{\theta}_j]^2}{\sigma_{jn}^2 + s_j^2} - \ln(\sigma_{jn}^2 + s_j^2) - \Lambda Q(\boldsymbol{\theta}), \quad (4)$$

where  $\Lambda$  is a regularization parameter and  $Q(\boldsymbol{\theta})$  is a regularizing function that encourages the model coefficients  $\boldsymbol{\theta}_j$  to take on zero values, resulting in a simpler model that is less prone to overfitting. In the case of L1 regularization implemented within *The Cannon*, the regularizing function takes the form

$$Q(\boldsymbol{\theta}) = \sum_{j=0}^{J-1} |\theta_j|. \quad (5)$$

L1 regularization was chosen because *The Cannon* is designed for predicting large sets of elemental abundances, and it is reasonable to assume that only one or a few elemental abundances will affect the flux at a single pixel of the wavelength range. For more details on regularization or the model itself, see [\(Casey et al. 2016\)](#).

In the training step, the log-likelihood is maximized via the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm to derive the best-fit model coefficients  $\boldsymbol{\theta}_j$  and intrinsic scatter  $s_j^2$ :

$$\boldsymbol{\theta}_j, s_j^2 \leftarrow \underset{\boldsymbol{\theta}, s_j}{\text{argmax}} \left[ \sum_{n=0}^{N-1} \ln p(f_{jn} | \boldsymbol{\theta}_j, \mathbf{v}(l_n), s_j^2) \right]. \quad (6)$$

Plugging in the explicit form of the log-likelihood ([Equation 4](#)) leads to

$$\boldsymbol{\theta}_j, s_j^2 \leftarrow \underset{\boldsymbol{\theta}, s_j}{\text{argmax}} \left[ \sum_{n=0}^{N-1} -\frac{[f_{jn} - \mathbf{v}(l_n) \cdot \boldsymbol{\theta}_j]^2}{\sigma_{jn}^2 + s_j^2} - \sum_{n=0}^{N-1} \ln(\sigma_{jn}^2 + s_j^2) - \Lambda Q(\boldsymbol{\theta}) \right]. \quad (7)$$

### 3.3. Test Step

In the test step, we set the model labels  $(\boldsymbol{\theta}_j, s_j^2)$  to the optimized values determined during the training step at every pixel  $j$ , and fit for the label list  $l_n$  for each star  $n$  that minimizes the log-likelihood:

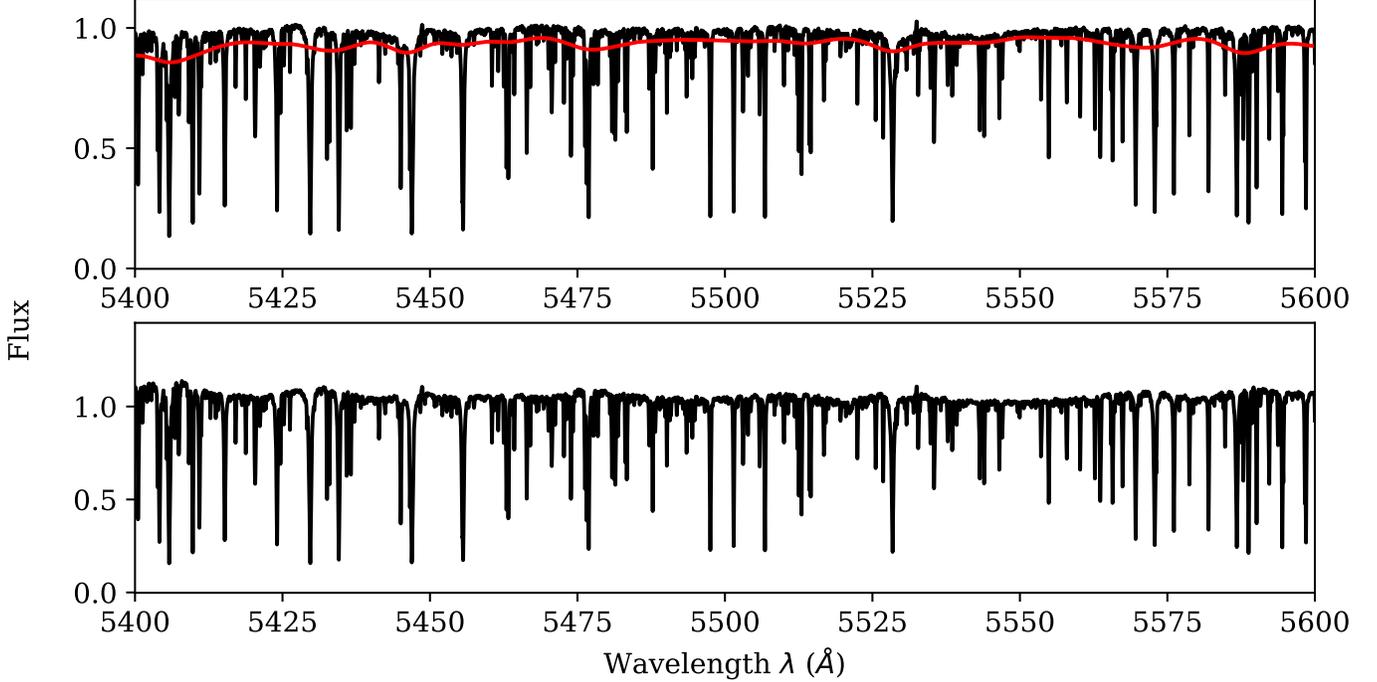
$$l_n \leftarrow \underset{l_n}{\text{argmin}} \left[ \sum_{j=0}^{J-1} -\frac{[f_{jn} - \mathbf{v}(l_n) \cdot \boldsymbol{\theta}_j]^2}{\sigma_{jn}^2 + s_j^2} \right], \quad (8)$$

Optimization of the log-likelihood in the test step is carried out via weighted least squares.

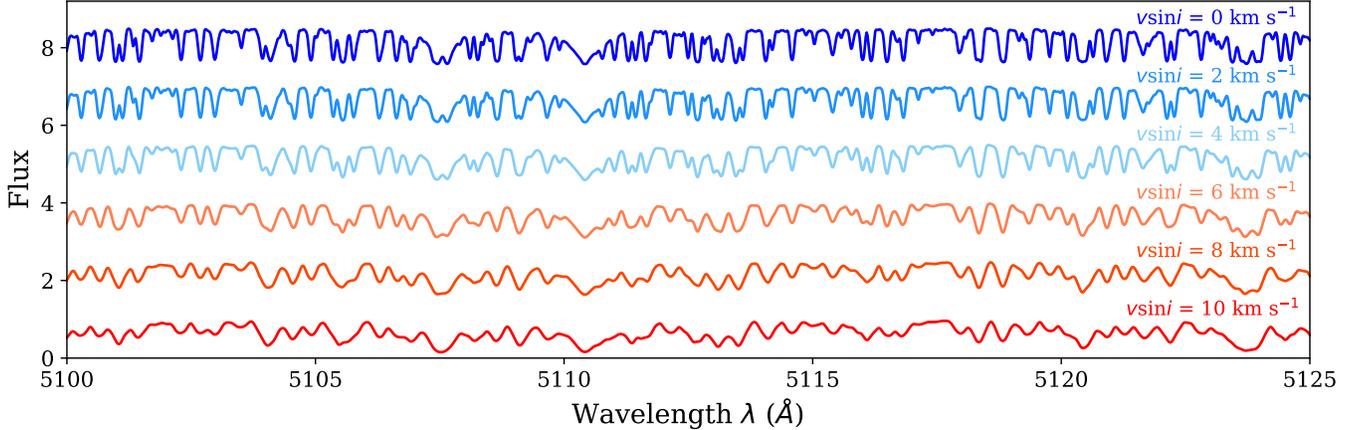
## 4. THE CANNON PERFORMANCE

### 4.1. Building Intuition with Synthetic Spectra

Before running *The Cannon* on our cool star sample, we sought to establish a measure of baseline performance. We did this by constructing a sample of synthetic spectra that mimics the cool star sample: 141 "stars" with the same label values as the true cool sample. Because the labels of the synthetic spectra, by definition, lack uncertainty, they can provide a sense of how well *The Cannon* performs under perfect conditions. The synthetic spectra are generated from the publicly-available code `SpecMatch-Syn` which fits five regions of optical spectra by interpolating within a grid of model spectra from the library described in [Coelho et al. \(2005\)](#). For more details on `SpecMatch-Syn`, see [Petigura \(2015\)](#). See [Fig. 3](#) for examples of synthetic spectra.



**Figure 2.** HIRES spectrum of a reference sample star (HD100623) before and after normalization. The top panel shows the pre-normalized spectrum overlaid with the Gaussian-smoothed version of itself in red, while the bottom panel shows the normalized spectrum after the Gaussian-smoothed signal was divided out. The displayed wavelength region ( $\lambda = 5400\text{-}5600 \text{ \AA}$ ) is a subset of the full wavelength range and was chosen for better visualization of the spectrum and accompanying Gaussian-smoothed curve.



**Figure 3.** Synthetic spectra generated via *SpecMatch-Syn* under the same  $T_{\text{eff}}$ ,  $\log g$ , and  $[\text{Fe}/\text{H}]$  conditions ( $T_{\text{eff}} = 4000 \text{ K}$ ,  $\log g = 4.5 \text{ cm s}^{-2}$ ,  $[\text{Fe}/\text{H}] = 0.2 \text{ dex}$ ) but with varying amounts of additional broadening. From top to bottom, the spectra have  $v \sin i = 0\text{-}10 \text{ km s}^{-1}$  in increments of  $+2 \text{ km s}^{-1}$ .

We tested the validity of *Cannon*-predicted labels through a bootstrap leave-one-out cross-validation scheme where we trained the spectral model on all objects in the synthetic spectral sample but one, and predicted labels for the object that was left out. We carried out this scheme iteratively to pass through the entire sample and predict labels for every object. Following the work of [Ness et al. \(2015\)](#) and [Casey et al. \(2016\)](#), we began with a spectral model in which the label list  $l_n$  was quadratic in the labels, resulting in the following label list:

$$l_n \equiv [1, T_{\text{eff}}, R_*, [\text{Fe}/\text{H}], T_{\text{eff}}^2, T_{\text{eff}} \cdot R_*, T_{\text{eff}} \cdot [\text{Fe}/\text{H}], R_*^2, R_* \cdot [\text{Fe}/\text{H}], [\text{Fe}/\text{H}]^2], \quad (9)$$

Where 1, the first element in the label list, is there to allow for

a linear offset in the fitting ([Ness et al. 2015](#)). We found that modeling the projected rotational velocity  $v \sin i$  as a fitted-for parameter in addition to  $T_{\text{eff}}$ ,  $R_*$ , and  $[\text{Fe}/\text{H}]$  resulted in more accurate labels predictions; a second order model without  $v \sin i$  achieves accuracies of 40 K in  $T_{\text{eff}}$ , 13% in  $R_*$ , and 0.06 dex in  $[\text{Fe}/\text{H}]$ , while a second order model with  $v \sin i$  achieves accuracies of 32 K in  $T_{\text{eff}}$ , 13% in  $R_*$ , and 0.03 dex in  $[\text{Fe}/\text{H}]$ .

Using a third order rather than second order (quadratic-in-label) model with  $v \sin i$  further improves label predictions; a third order model achieves accuracies of 22 K in  $T_{\text{eff}}$ , 8% in  $R_*$ , and 0.03 dex in  $[\text{Fe}/\text{H}]$ . Thus, these tests with synthetic spectra motivate a third order *Cannon* model with  $v \sin i$  included as a label. The third order model results in a label list composed of additional third order cross terms, bringing the

total number of terms up to 20.

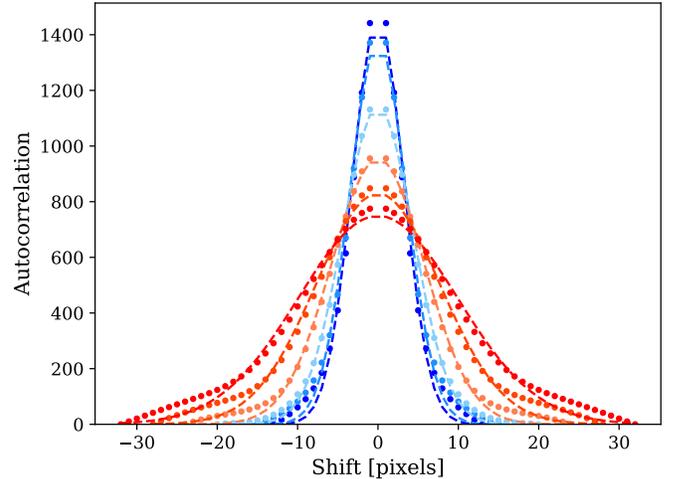
#### 4.2. Cool Star Sample

To run *The Cannon* on the cool star sample, we employed the same bootstrap leave-one-out cross-validation scheme. As in the case of synthetic spectra, the cool star HIRES spectra are best described by a third order model, which is unsurprising given their resolution of  $R \approx 60,000$  ( $\sim 3$  times the resolution of APOGEE spectra). The more flexible model may also better describe our more diverse training set, composed of stars with a wider  $T_{\text{eff}}$  range (APOGEE stars are confined to  $T_{\text{eff}} = 3500\text{--}5500$  K). A third order model fitting for  $T_{\text{eff}}$ ,  $R_*$ , and  $[\text{Fe}/\text{H}]$  achieves precisions of 80 K in  $T_{\text{eff}}$ , 6% in  $R_*$ , and 0.1 dex in  $[\text{Fe}/\text{H}]$ .

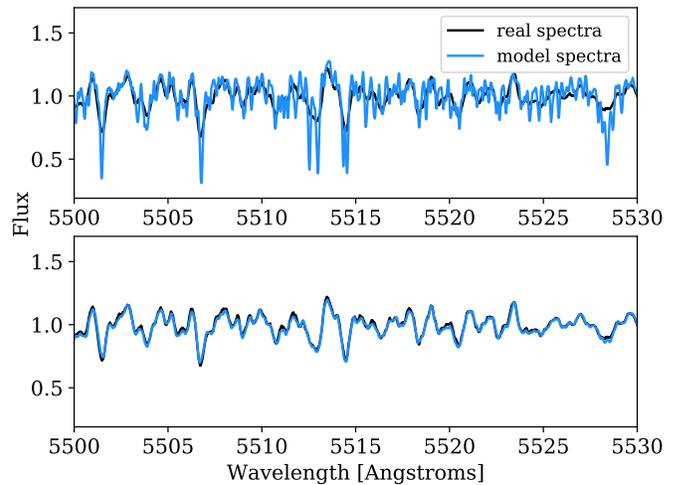
We found that *The Cannon* predicted anomalously poor label values for one source (GL896A). Upon inspection, we found that the spectrum of GL896A exhibits significantly broader features than any other source in our sample. Because GL896A is not well-represented in the training set, *The Cannon* is unable to construct a model that well-describes GL896A (Fig. 5, top panel). While such fast rotators are rare amongst K and M dwarfs, the presence of this target indicates that our implementation of *The Cannon* must still take them into consideration. We modified our implementation by augmenting and diversifying the training set; we created  $x$  copies of each spectrum in the cool star sample (exploring different values of  $x$  to see which resulted in the best performance), and applied differential values of artificial broadening to simulate faster stellar rotation. Artificial broadening was carried out by convolving the spectra with a rotational-macroturbulent broadening kernel described in Hirano et al. (2011).

In order for this scheme to work,  $v \sin i$  must be specified as a fitted-for label as in the tests with synthetic spectra. This is problematic because more than half of the cool stars do not have reported  $v \sin i$  values. We dealt with this by assigning all sources in the augmented sample a new label that describes general broadening, taken to be the FWHM of a Gaussian fitted to the spectral autocorrelation peaks (Fig. 4). This resulted in better flux predictions for the spectrum of GL896A (Fig. 5), and better label predictions overall. The most precise labels are achieved when the cool star sample is augmented by  $x = 5$  (5 copies generated for each spectrum), and the copies are artificially broadened by  $0\text{--}5 \text{ km s}^{-1}$  as the cool star sample does not appear to include any significantly rapid rotators ( $v \sin i > 5 \text{ km s}^{-1}$ ). We ultimately achieved precisions of 68 K in  $T_{\text{eff}}$ , 5% in  $R_*$ , and 0.08 dex in  $[\text{Fe}/\text{H}]$ , and verified that these label predictions vary within the reported precisions for different *Cannon* runs. While it may seem surprising that *The Cannon* achieves better predictions in  $R_*$  for empirical spectra compared to synthetic spectra (5% versus 8% in  $R_*$ ), it should be noted that the synthetic spectra may not accurately reflect the input  $R_*$  values as a conversion to  $\log g$  was required, which in turn required  $M_*$ . We do not have  $M_*$  values for the cool star sample, and instead assumed a linear relationship between  $R_*$  and  $M_*$  ( $M_*/M_\odot = R_*/R_\odot$ ) to obtain mass estimates. This is a valid approximation for the main sequence, but is not perfect.

Because of the large number of terms in our model, we considered overfitting to be a potential issue. That is, overly precise modeling of the training set flux may lead to less accurate label predictions. To address this, we added regularization to our *Cannon* model and assessed whether label prediction improved. We explored a grid of regularization strengths from  $\Lambda = 10^{-6}$  to  $\Lambda = 10^2$  uniform in log space. We found that no



**Figure 4.** The autocorrelation functions of spectra displayed in Fig. 3, marked in dotted lines. The overlaid Gaussian fits are displayed in dashed lines.

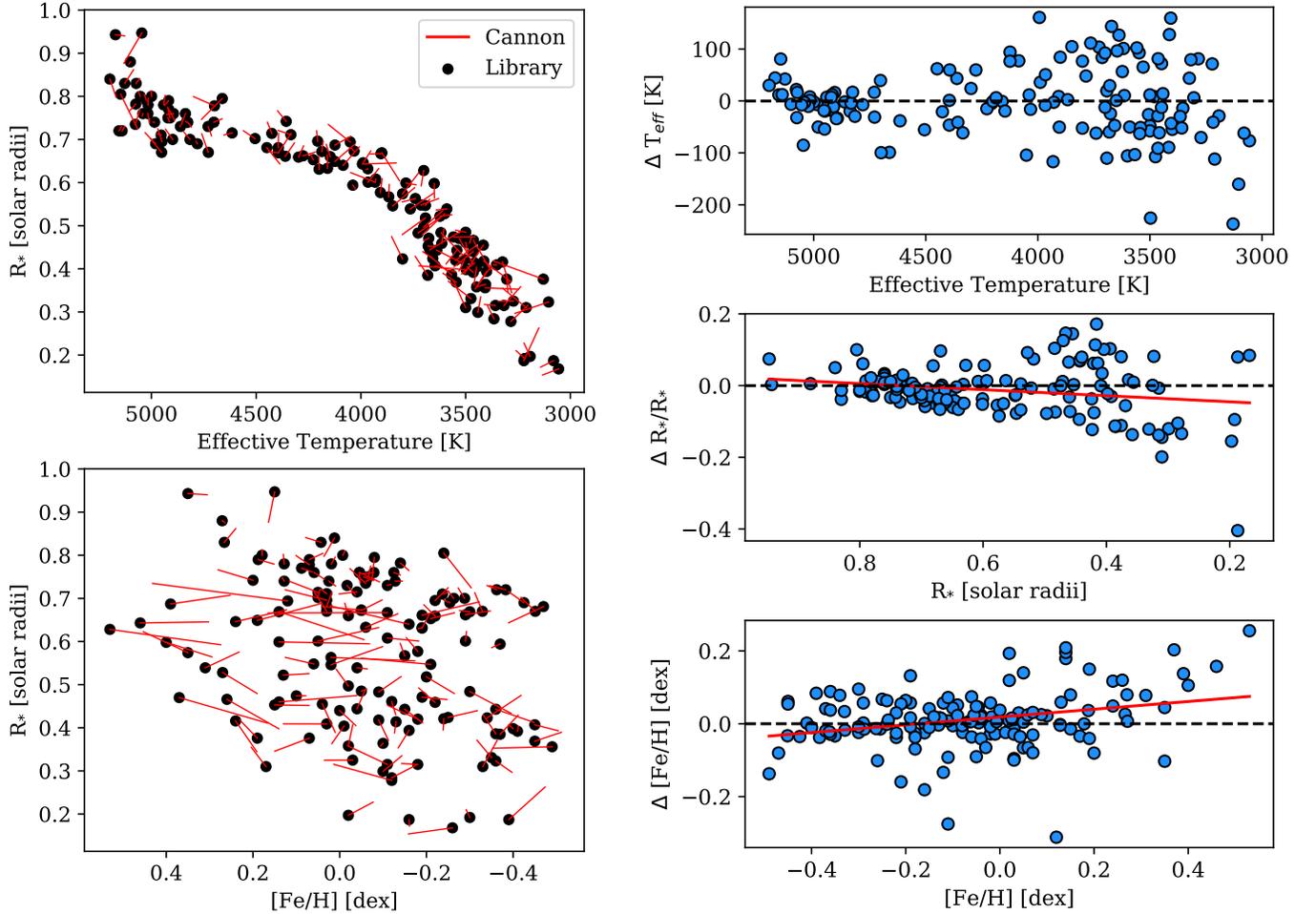


**Figure 5.** The spectrum of GL896A overlaid with the *Cannon* model before augmenting the library with broadened copies of the spectra (top), and after (bottom). The true spectrum is plotted in black while the *Cannon* models are plotted in blue.

matter the regularization strength, adding regularization to the model always resulted in less precise label predictions. It is possible that regularization does not lead to better predictions for the 3 labels ( $T_{\text{eff}}$ ,  $R_*$ ,  $[\text{Fe}/\text{H}]$ ) we are considering because all of these labels affect the flux at each wavelength point. Thus, we do not benefit from regularization that encourages sparsity (L1, encourages the model coefficients to go to zero). L1 regularization may lead to better label predictions if we expand our label set to include elemental abundances, but that is beyond the scope of this study.

#### 4.3. Performance at Low S/N

To investigate the effect of photon shot noise on the precision of label predictions made with *The Cannon*, we carried out the same procedure employed by Yee et al. (2017) for the empirical spectroscopic tool *SpecMatch-Emp*; we isolated a subset of 20 stars from the cool star sample with  $\text{S/N} > 160/\text{pixel}$  and degraded their spectra by injecting Gaussian noise to simulate target S/N values of 120, 100, 80, 60, 40, 20, and 10 per pixel. We generated 20 S/N-degraded spectra for each spectrum in the subset and S/N target value, then com-



**Figure 6.** Comparison of the cool star sample library labels with the *Cannon*-predicted labels ( $T_{\text{eff}}$ ,  $R_*$ ,  $[\text{Fe}/\text{H}]$ ). In the left panel plots, the black points represent the library labels and the red lines represent the *Cannon* labels. The right panel plots display the label residuals, with the red lines denoting possible trends. We note that the slope values of these linear trends are much lower than those of residuals from labels predicted via techniques that make use of empirical spectral libraries (Yee et al. 2017).

pared the precision of the *Cannon* label predictions for the degraded spectra with those of the original  $\text{S/N} > 160/\text{pixel}$  spectra, which we took as ground truth. The results are summarized in Fig. 7.

As expected, lower  $\text{S/N}$  leads to larger median scatter in label predictions made with *The Cannon*. However, the median scatter at  $\text{S/N} = 10/\text{pixel}$  is still low, with 3.5 K in  $T_{\text{eff}}$ , 0.4% in  $R_*$ , and 0.006 dex in  $[\text{Fe}/\text{H}]$ . This demonstrates that *The Cannon* is quite robust, even at low  $\text{S/N}$  values. This performance is better than that achieved by *SpecMatch-Emp*, which has median scatter values at  $\text{S/N} = 10/\text{pixel}$  of 10.4 K in  $T_{\text{eff}}$ , 1.7% in  $R_*$ , and 0.008 dex in  $[\text{Fe}/\text{H}]$  (Yee et al. 2017), though it should be noted that *SpecMatch-Emp* conducted this test with stars spanning the HR diagram while our sample is  $T_{\text{eff}}$ -limited.

Motivated by the small observed scatter in  $[\text{Fe}/\text{H}]$ , we attempted to estimate the minimum change in  $[\text{Fe}/\text{H}]$  that is theoretically detectable. To do so, we considered the difference between two spectra corresponding to stars with slightly different metallicities ( $\Delta[\text{Fe}/\text{H}]$ ). We defined a quantity  $\mathcal{S}$  that relates three quantities:  $\Delta[\text{Fe}/\text{H}]$ ; the derivative of the spectrum with changing metallicity,  $\delta f/\delta[\text{Fe}/\text{H}]$ ; and the flux uncertainty  $\sigma_f$ .  $\mathcal{S}$  can be thought of as analogous to  $\text{S/N}$ . For the  $j$ th pixel, the relation is

$$S_j = \frac{\left(\frac{\delta f}{\delta[\text{Fe}/\text{H}]}\right)_j \Delta[\text{Fe}/\text{H}]}{\sigma_{f,j}}. \quad (10)$$

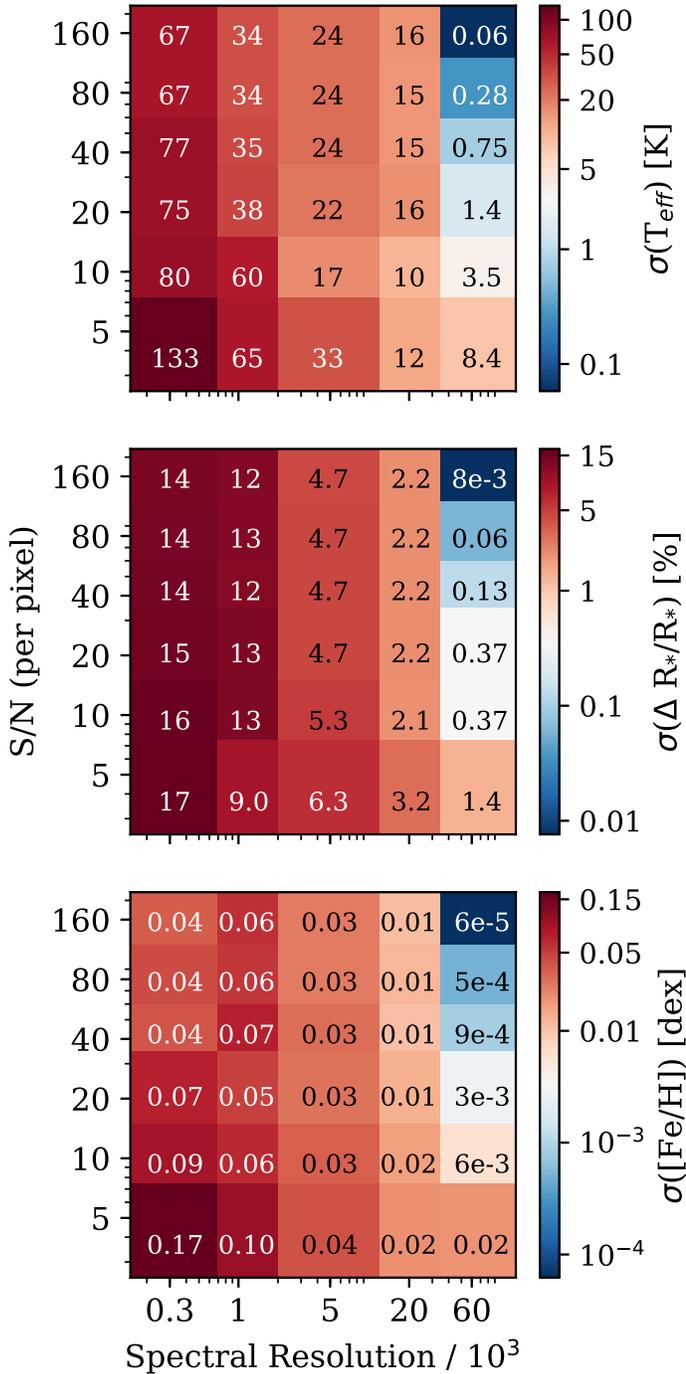
This equation can be rewritten as

$$S_j = \frac{\left(\frac{\delta f}{\delta[\text{Fe}/\text{H}]}\right)_j \Delta[\text{Fe}/\text{H}]}{c_j \langle \sigma_f \rangle}, \quad (11)$$

where  $\langle \sigma_f \rangle$  is the average flux uncertainty and  $c_j$  is directly related to the blaze function. The total  $\mathcal{S}$  (summing over pixels) of the metallicity measurement can be written as

$$\begin{aligned} \mathcal{S} &= \sqrt{\sum_j (\mathcal{S}_j)^2} \\ &= \frac{\Delta[\text{Fe}/\text{H}]}{\langle \sigma_f \rangle} \sqrt{\sum_j \left[ \left(\frac{\delta f}{\delta[\text{Fe}/\text{H}]}\right) / c_j \right]^2}. \end{aligned} \quad (12)$$

Rearranging terms to solve for the minimum theoretically detectable change in metallicity yields



**Figure 7.** Log-log plots showing the median scatter of *Cannon*-derived labels as a function of both S/N and resolution. Each colored block within the subplots represents the median RMS difference in  $T_{\text{eff}}$  (top),  $R_*$  (middle), and  $[\text{Fe}/\text{H}]$  (bottom) predictions from the cool star subset with spectra satisfying  $S/N > 160/\text{pixel}$  when degraded to lower S/N and resolution. The median RMS difference is also explicitly provided within each block in units of K ( $T_{\text{eff}}$ ), solar radii ( $R_*$ ), and dex ( $[\text{Fe}/\text{H}]$ ). The median scatter increases as the S/N and resolution decreases, which is representative of the effect photon shot noise and lower resolution would have on the precision of *Cannon* label predictions for HIRES spectra.

$$\Delta[\text{Fe}/\text{H}] = S \langle \sigma_f \rangle / \sqrt{\sum_j \left[ \left( \frac{\delta f}{\delta[\text{Fe}/\text{H}]} \right) / c_j \right]^2}. \quad (13)$$

A metallicity change is detectable at  $1\sigma$  if  $S = 1$ . For a  $S/N = 10/\text{pixel}$  as considered above, i.e.,  $\langle \sigma_f \rangle = 0.1$ , we find  $\Delta[\text{Fe}/\text{H}] = 0.001$  dex. This is much smaller than the median scatter in  $[\text{Fe}/\text{H}]$  predictions made with *The Cannon* at  $S/N = 10/\text{pixel}$  (0.006 dex). Therefore, the sensitivity of *The Cannon* lies within theoretical bounds.

#### 4.4. Performance at Low Spectral Resolution

While HIRES spectra are observed at  $R \approx 60,000$ , many large spectroscopic surveys are observed at lower spectral resolutions. Thus, it is valuable to quantify how spectral resolution affects the accuracies of label predictions with *The Cannon*. We expected performance to decrease as spectral resolution decreases because lines will blend together, resulting in less spectral information for *The Cannon* to work with.

To investigate spectral resolution dependence, we followed the same procedure used for the S/N degradation test; we used the same subset of 20 stars with  $S/N > 160$  and simulated lower resolution by convolving their spectra with a Gaussian kernel. We again treated the label predictions of the original high resolution ( $R \approx 60,000$ ) spectra as ground truth. We simulated spectra with target resolution values of  $R = 50,000$ , 40,000, 30,000, 20,000, 10,000, 7500, and 5000. The results are summarized in Fig. 7, which also illustrates how the predictions of label precisions are affected when both S/N and resolution are degraded.

As in the case of degraded S/N, the accuracy of *Cannon* label predictions decrease with spectral resolution. At  $R = 30,000$ , median scatter in the labels is 6.7 K in  $T_{\text{eff}}$ , 0.2% in  $R_*$ , and 0.009 dex in  $[\text{Fe}/\text{H}]$ . This performance is better than that of *SpecMatch-Emp*'s at equivalent resolution, with median scatter values of 10.1 K in  $T_{\text{eff}}$ , 1.3% in  $R_*$ , and 0.014 dex in  $[\text{Fe}/\text{H}]$  (Yee et al. 2017).

*The Cannon* also exhibits a much slower reduction in label accuracy as resolution continues to decrease; at  $R = 5000$ , the median scatter in *Cannon* predictions is 24.1 K in  $T_{\text{eff}}$ , 4.7% in  $R_*$ , and 0.026 dex in  $[\text{Fe}/\text{H}]$ , while *SpecMatch-Emp*'s is 962 K in  $T_{\text{eff}}$ , 228% in  $R_*$ , and 0.094 dex in  $[\text{Fe}/\text{H}]$  (Yee et al. 2017). This suggests that *The Cannon* would be a favorable method for predicting labels for spectra from many large, lower resolution spectroscopic surveys (e.g., SEGUE (Beers et al. 2006),  $R \approx 2000$ , RAVE (Steinmetz et al. 2006),  $R \approx 7000$ , LAMOST (Newberg et al. 2012),  $R \approx 1800$ ).

#### 4.5. Performance with Label Errors

To investigate the effect of errors in the library labels on predictions made with *The Cannon*, we followed the same procedure used for the S/N and resolution degradation tests; we used the same subset of 20 stars with  $S/N > 160$  and injected Gaussian noise into the labels to simulate additional uncertainty up to 1x the achievable precisions (68 K in  $T_{\text{eff}}$ , 5% in stellar radius  $R_*$ , and 0.08 dex in  $[\text{Fe}/\text{H}]$ ). We found that the labels are quite robust to realistic random noise in the library labels; adding 1x uncertainty leads to an increase in label prediction uncertainties of 22 K in  $T_{\text{eff}}$ , 4% in  $R_*$ , and 0.06 dex in  $[\text{Fe}/\text{H}]$ . The results are summarized in Table 1.

It is worth noting that the scatter in *Cannon*-predicted values of  $T_{\text{eff}}$  is lower than the original label uncertainty by more than 50%, suggesting that in the limit of a very large library with labels containing a certain amount of random noise, *The Cannon* can derive a model that yields a higher  $T_{\text{eff}}$  precision compared to that of the library spectra. We note that this result is insensitive to zero-point offsets; it is not possible to bootstrap to higher label precisions using *The Cannon*.

**Table 1**

Median RMS scatter in all *Cannon*-derived labels after adding an 1x the amount of uncertainty to all labels

Added uncertainty ( $T_{\text{eff}}$ , $R_*$ , [Fe/H])	$\sigma(T_{\text{eff}})$ K	$\sigma(\Delta R_*/R_*)$ %	$\sigma([\text{Fe}/\text{H}])$ dex
68 K	22	2	0.02
5% $R_*$	17	4	0.02
0.08 dex	10	1	0.06

## 5. DISCUSSION

We evaluated how well *The Cannon*, a data-driven spectroscopic tool, is able to predict stellar labels for cool stars ( $T_{\text{eff}} = 3000\text{--}5200$  K) given high-resolution spectra. With adjustments to the spectral training set, it achieves precisions of 68 K in  $T_{\text{eff}}$ , 5% in  $R_*$ , and 0.08 dex in [Fe/H]. Unlike traditional spectroscopic modeling techniques, *The Cannon* does not rely on stellar models that struggle to reproduce the complexities of cool star spectra. Rather, as a data-driven method, *The Cannon*'s performance improves as the input spectra become more information-rich.

In the case of spectra with perfect labels (no uncertainty) as simulated with synthetic spectra, *The Cannon* achieves label accuracies of 22 K in  $T_{\text{eff}}$ , 8% in  $R_*$ , and 0.03 dex in [Fe/H]. *The Cannon* generally makes better label predictions for synthetic spectra because the labels of real spectra include uncertainties that are endemic to the catalogs from which the cool star sample originates. These catalogs are described in von Braun et al. (2014), Mann et al. (2015), Brewer et al. (2016), and Yee et al. (2017), and present labels derived from a combination of modified SME (Brewer et al. 2015), photometry, parallaxes, interferometry, and empirical relations between the labels and EWs of spectral features. Each of these techniques have associated uncertainties, resulting in less precise label predictions with *The Cannon* when compared to the case of spectra with perfect labels.

Compared to current synthetic spectral techniques (SME, MOOG, etc.), *The Cannon* is better-suited for predicting the labels of cool stars. While the latest iterations of spectral synthesis codes model cool stars more successfully than initial versions with additions such as more accurate radiative transfer algorithms, equations of state, and larger line lists, they still lack complete sets of molecular line opacities and sufficient constraints to fully disentangle the effects of  $T_{\text{eff}}$ ,  $\log g$ , and abundances (e.g., Bean et al. 2006; Piskunov & Valenti 2017).

It is more appropriate to compare *The Cannon* to other data-based techniques such as *SpecMatch-Emp*, a label-predicting spectroscopic tool developed by Yee et al. (2017) that utilizes an empirical spectral library. While *SpecMatch-Emp* achieves accuracies of 70 K in  $T_{\text{eff}}$ , 10% in  $R_*$ , and 0.12 dex in [Fe/H] for stars of spectral types  $\sim$ K4 and later, these label predictions are slightly worse than those achieved by *The Cannon*. In addition, the residuals from label predictions with *SpecMatch-Emp* display linear trends where residuals are more negative for larger values in the label space, and more positive for smaller values in the label space (Yee et al. 2017). These trends are partly explained by considering that the empirical spectral library spans a finite region (convex hull of the label values), and is inclined to pull spectral predictions at the edge of the region towards the center. While the residuals from label predictions with *The Cannon* also display slight linear trends, they are less pronounced and

constitute a smaller source of systematic error (Fig. 6, right panel). This is because the choice of flux model coefficient values allows for some extrapolation outside the finite region spanned by the training set.

While *The Cannon* is a powerful tool for spectroscopic characterization, it has a number of drawbacks. For example, by individually treating each pixel within the spectral wavelength range, it assumes no covariance between flux values of any pixels. However, multiple spectral features can be affected by a single label, such as a particular elemental abundance or ionization state. This motivates converting *The Cannon* into a fully Bayesian framework through the inclusion of priors such as line lists to address covariance of different spectral features, or known correlations between labels such as the Stefan-Boltzman relation.

Although L1 regularization does not improve cool star label predictions for  $T_{\text{eff}}$ ,  $R_*$ , and [Fe/H], L2 regularization may be better suited to such cases where labels do not include large sets of elemental abundances as L2 regularization does not encourage model coefficients to go to zero as rapidly. However, we are also interested in eventually using *The Cannon* to predict elemental abundances, in which case L1 regularization may become a useful feature. For example, we are interested in comparing the C/O ratios of K and M dwarfs to the characteristics of planets they host as such volatile ratios can probe planet formation histories. Ultimately, we will use *The Cannon* to conduct large demographic studies of cool stars with HIRES spectra with the goal of establishing correlations between small, cool stars such as K and M dwarfs and the planets they host. This work has wide potential application given that many future exoplanet surveys are focused on cool stars such as M dwarfs.

We thank Andrew Casey (Monash U.), Anna Ho (Caltech), and Melissa Ness (MPIA) for many useful discussions regarding *The Cannon*. A.B. acknowledges funding from the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745301.

## REFERENCES

- Bean, J. L., Sneden, C., Hauschildt, P. H., Johns-Krull, C. M., & Benedict, G. F. 2006, *ApJ*, 652, 1604
- Beers, T. C., et al. 2006, *Memorie della Societa Astronomica Italiana*, 77, 1171
- Brewer, J. M., Fischer, D. A., Basu, S., Valenti, J. A., & Piskunov, N. 2015, *ApJ*, 805, 126
- Brewer, J. M., Fischer, D. A., Valenti, J. A., & Piskunov, N. 2016, *ApJS*, 225, 32
- Casey, A. R., Hogg, D. W., Ness, M., Rix, H.-W., Ho, A. Q., & Gilmore, G. 2016, *ArXiv e-prints*
- Coelho, P., Barbuy, B., Meléndez, J., Schiavon, R. P., & Castilho, B. V. 2005, *A&A*, 443, 735
- Deen, C. P. 2013, *AJ*, 146, 51
- Dotter, A., Chaboyer, B., Jevremović, D., Kostov, V., Baron, E., & Ferguson, J. W. 2008, *ApJS*, 178, 89
- Henry, T. J., Jao, W.-C., Subasavage, J. P., Beaulieu, T. D., Ianna, P. A., Costa, E., & Méndez, R. A. 2006, *AJ*, 132, 2360
- Hirano, T., Suto, Y., Winn, J. N., Taruya, A., Narita, N., Albrecht, S., & Sato, B. 2011, *ApJ*, 742, 69
- Ho, A. Y. Q., et al. 2017, *ApJ*, 836, 5
- Howard, A. W., et al. 2010, *ApJ*, 721, 1467
- Mann, A. W., Deacon, N. R., Gaidos, E., Ansdell, M., Brewer, J. M., Liu, M. C., Magnier, E. A., & Aller, K. M. 2014, *AJ*, 147, 160
- Mann, A. W., et al. 2017, *AJ*, 153, 267
- Mann, A. W., Feiden, G. A., Gaidos, E., Boyajian, T., & von Braun, K. 2015, *ApJ*, 804, 64
- Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, *ApJ*, 808, 16
- Newberg, H. J., et al. 2012, in *Astronomical Society of the Pacific Conference Series*, Vol. 458, *Galactic Archaeology: Near-Field Cosmology and the Formation of the Milky Way*, ed. W. Aoki, M. Ishigaki, T. Suda, T. Tsujimoto, & N. Arimoto, 405
- Newton, E. R., Charbonneau, D., Irwin, J., Berta-Thompson, Z. K., Rojas-Ayala, B., Covey, K., & Lloyd, J. P. 2014, *AJ*, 147, 20
- Newton, E. R., Charbonneau, D., Irwin, J., & Mann, A. W. 2015, *ApJ*, 800, 85
- Petigura, E. A. 2015, Ph.D. thesis, University of California, Berkeley
- Piskunov, N., & Valenti, J. A. 2017, *A&A*, 597, A16
- Rojas-Ayala, B., Covey, K. R., Muirhead, P. S., & Lloyd, J. P. 2012, *ApJ*, 748, 93
- Sneden, C. A. 1973, Ph.D. thesis, THE UNIVERSITY OF TEXAS AT AUSTIN.
- Steinmetz, M., et al. 2006, *AJ*, 132, 1645
- Valenti, J. A., et al. 2009, *ApJ*, 702, 989
- Valenti, J. A., & Fischer, D. A. 2005, *ApJS*, 159, 141
- Valenti, J. A., & Piskunov, N. 1996, *A&AS*, 118, 595
- Vogt, S. S., et al. 1994, in *Proc. SPIE*, Vol. 2198, *Instrumentation in Astronomy VIII*, ed. D. L. Crawford & E. R. Craine, 362
- von Braun, K., et al. 2014, *MNRAS*, 438, 2413
- Yee, S. W., Petigura, E. A., & von Braun, K. 2017, *ApJ*, 836, 77