The Astrophysical Journal, 808:16 (21pp), 2015 July 20 doi:10.1088/0004-637X/808/1/16

THE CANNON: A DATA-DRIVEN APPROACH TO STELLAR LABEL DETERMINATION

M. Ness[1], David W. Hogg[1,2,3], H.-W. Rix[1], Anna. Y. Q. Ho[1], and G. Zasowski[4,5]
[1] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany; ness@mpia.de
[2] Center for Cosmology and Particle Physics, Department of Phyics, New York University, 4 Washington Pl., Room 424, New York, NY 10003, USA
[3] Center for Data Science, New York University, 726 Broadway, 7th Floor, New York, NY 10003, USA
[4] Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
Received 2015 January 16; accepted 2015 June 3; published 2015 July 14

ABSTRACT

New spectroscopic surveys offer the promise of stellar parameters and abundances ("stellar labels") for hundreds of thousands of stars; this poses a formidable spectral modeling challenge. In many cases, there is a subset of *reference objects* for which the stellar labels are known with high(er) fidelity. We take advantage of this with *The Cannon*, a new data-driven approach for determining stellar labels from spectroscopic data. *The Cannon* learns from the "known" labels of reference stars how the continuum-normalized spectra depend on these labels by fitting a flexible model at each wavelength; then, *The Cannon* uses this model to derive labels for the remaining survey stars. We illustrate *The Cannon* by training the model on only 542 stars in 19 clusters as reference objects, with $T_{\rm eff}$, log g, and [Fe/H] as the labels, and then applying it to the spectra of 55,000 stars from *APOGEE* DR10. *The Cannon* is very accurate. Its stellar labels compare well to the stars for which *APOGEE* pipeline (*ASPCAP*) labels are provided in DR10, with rms differences that are basically identical to the stated *ASPCAP* uncertainties. Beyond the reference labels, *The Cannon* makes no use of stellar models nor any line-list, but needs a set of reference objects that span label-space. *The Cannon* performs well at lower signal-to-noise, as it delivers comparably good labels even at one-ninth the *APOGEE* observing time. We discuss the limitations of *The Cannon* and its future potential, particularly, to bring different spectroscopic surveys onto a consistent scale of stellar labels.

Key words: methods: data analysis – methods: statistical – stars: abundances – stars: fundamental parameters – surveys – techniques: spectroscopic

Supporting material: machine-readable table

1. INTRODUCTION

The vast spectroscopic stellar surveys of recent years (e.g., *SEGUE*, Beers et al. 2006, *RAVE*, Steinmetz et al. 2006, *LAMOST*, Newberg et al. 2012, *APOGEE*, Majewski 2012, *Gaia-ESO*, Gilmore et al. 2012, *GALAH*, Freeman 2012) hold tremendous astrophysical promise, but at the same time present formidable data analysis and modeling challenges. One of these challenges lies in consistently and accurately determining what we call "stellar labels," that is, stellar parameters and element abundances, from survey spectra. These labels are usually determined from comparison of the data with synthetic model spectra, with approaches often customized specifically to the particular wavelength region of a given survey (e.g., Lee et al. 2006; Boeche et al. 2011; Bailer-Jones et al. 2013; Mészáros et al. 2013; Liu et al. 2014; Smiljanic et al. 2014).

The stellar photosphere models that are relied upon for stellar label determination have physical ingredients that are incomplete and simplified. For computational feasibility, almost always 1D stellar photosphere models are adopted for large surveys, often assumed to be in local thermal equilibrium; these approximations are both severe. In many cases, the model spectra do not account for all relevant molecular opacities, for convection, stellar winds, and the chromosphere. As a consequence, it happens that different research groups obtain discrepant results for same stars, resulting from analysis across different wavelength regions and different input assumptions and methods used (e.g., Allende Prieto et al. 1999; Hinkel et al. 2014; Jofré et al. 2014). Even when the input assumptions are held fixed, differences in the employed analysis methods lead to substantial differences in assigned labels (e.g., Smiljanic et al. 2014).

Stellar labels are commonly determined by fitting a grid of model spectra (with known labels) to the data using some minimization technique, often restricted to a masked portion of the spectrum that is focused on the absorption line (regions) deemed to be most reliable or relevant. Stated minimal signal-to-noise ratio (hereafter S/N) requirements to obtain robust labels in this way are S/N $\sim$ 100 per resolution element, especially if the labels are to include individual element abundances. Often, a post-calibration procedure is applied to bring the stellar labels derived by such a fitting pipeline in accord with external information of higher fidelity: for example with stellar labels from benchmark stars studied at high resolution or well characterized open and globular cluster stars (e.g., Kordopatis et al. 2013; Mészáros et al. 2013; Jofré et al. 2014). These calibration stars are also used by surveys to provide a reasonable estimate of their label accuracy. In practice, different surveys or different pipelines end up delivering labels with different calibrations, causing their stellar parameters or their abundances to be on slightly different scales. This complicates inter-survey comparisons and constitutes a major challenge of the era of such large data sets.

In this paper we propose and lay out a data-driven approach to deriving stellar labels from stellar spectra in the context of large spectroscopic surveys, which we dub "*The Cannon*."[6] The main practical strengths of *The Cannon* are that it requires

[5] NSF Astronomy and Astrophysics Postdoctoral Fellow.

[6] The name *The Cannon* is inspired by the astronomer Annie Jump Cannon, who was the pioneer in producing stellar classifications without any input of physical models!

no physical model of the spectra, it is enormously fast, it can obtain labels of comparable accuracy to that quoted in current physics-based approaches but at far lower S/N, and it offers a consistent way to cross-calibrate surveys. To achieve this, *The Cannon* relies on the existence of a subset of objects within a survey (*reference objects*) for which the stellar labels are known and cover label space sufficiently.

In this context, the term "*labels*" refers to the pieces of information that characterize and determine a stellar spectrum; these labels are commonly and sensibly split into *stellar parameters* and *element abundances*, although in the context of *The Cannon* it makes sense to treat them on a par. In most cases, it suffices to think of the labels as $T_{\rm eff}$, log g, and the element abundances [X/Fe], athough stellar rotation, microturbulence, age, and so forth can also be thought of as labels. It is central to the approach we lay out here that objects with the same labels have (nearly) identical spectra and that spectra vary smoothly with label changes. This must be true, if the set of labels is comprehensive enough so that it fully specifies the star; but if the labels are (for example) only $T_{\rm eff}$, log g, and [Fe/H] then this is an approximation. These three labels are typically described as stellar parameters and are by far the most important to describe the overall behavior of the spectral flux of red giant stars.

There are fundamentally two steps in *The Cannon*. The first step, or *training step*, is to create from the spectra of the reference objects a very flexible generative model (with ~80,000 parameters) that describes a probability density function (pdf) for the flux at every pixel in the continuum-normalized spectrum as a function of the labels. The second step, or *test step*, assumes that this same generative model holds for all the other objects in the survey (dubbed *survey objects*). Then, the spectra of the survey objects and the generative model from the reference objects allow us to solve for—or infer—the labels of the survey objects. Taken together the training step and the test step effect a *label transfer*, transferring the known labels in the reference objects to the survey objects.

To make such an approach straightforward, we must assume that the reference objects and the survey objects were observed with an identical instrumental setup, a condition well-satisfied with the large surveys listed above. We take the generative model for the continuum-normalized flux at each of $N_{\rm pix}$ pixels to be a polynomial function of all the labels, and hence the model is defined by its $N_{\rm pix}$ sets of polynomial coefficients. In practice, there may be different circumstances that make stars suitable reference objects. They may be members of star clusters: there, external data and the fact that clusters are in good approximation single stellar populations (which have to fall onto an isochrone) lend credibility to their stellar labels. Alternatively, reference objects could be stars for which labels have been derived separately from spectra of particularly high S/N, or at other "easier" or more extensive wavelength regimes (for example, in the optical as opposed to the infrared). Finally, they may be subsets of stars for which other approaches to get stellar parameters (for example, astroseismology) provide accurate stellar labels.

The Cannon is a *generative model* of the observed spectra; that is, it constructs, as a function of labels, a pdf for the observed flux as a function of wavelength. In many real cases the training data will be much higher in S/N than the test data (standard stars tend to be bright and well observed) and with *The Cannon* it is possible to transfer the labels from high S/N reference objects to lower S/N survey objects with high fidelity. In what follows, we show that *The Cannon* behaves very well as the S/N (or observing time) is decreased.

In this paper we use the *APOGEE* survey as the sole example. However, *The Cannon* can be applied to any stellar survey.

Our most basic implementation of *The Cannon* that we present includes only three labels, but this can easily be extended to additional labels (for example, [α/Fe], [X/Fe]) and also more comprehensive models (for example, Gaussian processes). Additionally, as we are using the information in every pixel, this methodology is effective at determining labels at lower S/N than minimization techniques.

The Cannon is similar to the MATrix Inversion for Spectral SythEsis (*MATISSE*) and University of Lyon Spectroscopic analysis Software (*ULySS*) procedure for derivation of stellar parameters (Recio-Blanco et al. 2006; Koleva et al. 2009) in that it uses the full spectrum (and not just a line list) for label determination. However, *MATISSE* employs a large grid of synthetic spectra and is thus limited in all the ways that physics-based methods are limited. The method outlined in Re Fiorentin et al. (2007) proposes the possibility of using real data as a model but implements a different, principle component analysis technique to estimate stellar labels. Empirical stellar libraries have been used previously as a reference set of spectra, including with *ULySS* (e.g., Wu et al. 1998; Prugniel et al. 2011) and also by Soubiran et al. (2011), in order to determine stellar labels directly from observed spectra. A big part of why *The Cannon* is successful at lower S/N is that it uses all of the pixels in the data.

We have adopted a bottom-up approach for *The Cannon*, starting with the most basic implementation and successively adding complexity to the generative model to determine the least complex implementation that works. The aim of this paper is not to explore all possible models that may work for this approach, nor to converge on an optimal model, but to use the simplest model that validates the underlying methodology as successful. With additional complexity, for example partial labels on the reference objects and adding errors to the labels of these objects, it may be preferential to adopt a different form of model entirely, such as a Gaussian Process, considered in the discussion section. One key advantage of the simple model we adopt is in its relative simplicity, which makes *The Cannon* computationally trivial to run to return stellar labels for large data sets.

In laying out the methodology of this approach we first describe the *APOGEE* data set and the way we process the data for both reference objects (542 stars) and survey objects (~55,000 stars from DR10). We then describe perhaps the simplest implementation of label-transfer possible, using a first-order linear model. We found this first-order model to be insufficiently flexible to describe the labels of the stars and extended our model to quadratic form, which satisfactorily describes the label-space of the training data. The success of this model is demonstrated by running *The Cannon* through the DR10 data available through the SDSS-3 data server, the results for which we provide in an online machine-readable table.

2. DATA

The Cannon expects (in its simplest form, presented here) all spectra—for reference and survey objects—to be continuum-

Table 1
Partial Column Excerpt from the Online Table of Stellar Labels ($T_{\rm eff}$, log g, and [Fe/H]) Determined by *The Cannon* for the 55,000 Stars Released in 170 Fields from *APOGEE*'s Data Release DR10

Star ID (2MASS)	$T_{\rm eff}$ (K)	log g (dex)	[Fe/H] (dex)	$\sigma(T_{\rm eff})$ (K)	σ(log g) (dex)	σ([Fe/H]) (dex)	χ^2	$d_{\rm ref}$	EFLAG
21353892+4229507	4160.4	1.62	0.05	3.24	0.008	0.004	2.59	0.03	0
21354474+4250256	4824.4	4.41	0.15	8.8	0.013	0.005	0.83	0.13	0
21354775+4233120	4704.1	2.50	0.05	9.2	0.022	0.011	0.83	0.03	0
21355458+4222326	4837.2	2.52	−0.33	6.8	0.015	0.007	1.02	0.11	0
21360285+4231145	4620.0	2.09	−0.43	9.6	0.023	0.011	0.94	0.15	0
21360822+4225525	4809.6	2.75	−0.03	6.9	0.016	0.008	1.15	0.017	0

(This table is available in its entirety in machine-readable form.)

normalized in a consistent way, and sampled on a consistent rest-frame wavelength grid, with the same line-spread function (LSF). It also assumes that the flux variance, from photon noise and other sources, is known at each spectral pixel of each spectrum. In principle, *The Cannon*, as described below, is applicable to any large, homogeneous spectroscopic data set meeting these criteria. Here, we use the *APOGEE* DR10 data (S. R. Majewski et al. 2015, in preparation) to illustrate and showcase *The Cannon*. Because all of the exposition of the method underlying *The Cannon* involves specificities of the data, we spell out the characteristics of the *APOGEE* data and our adjustments to it in Section 2.1. However, we stress that the approach is more widely applicable.

2.1. Specifics of the APOGEE Data Set

The data set used for this functional demonstration of *The Cannon* is that of the *APOGEE* survey (Majewski et al. 2012, S. R. Majewski et al. 2015, in preparation). *APOGEE*, part of the SDSS-III[7] (Eisenstein et al. 2011), is a high resolution ($R\sim$ 22,500), high signal to noise (S/N ~100), H-band (15200–16900 Å)[8] spectroscopic survey of primarily red giant stars spanning the bulge, disk, and halo of the Milky Way (Zasowski et al. 2013). *APOGEE*'s *ASPCAP* pipeline provides the stellar labels for these stars, which include stellar parameters and multiple elemental abundances, in addition to numerous flags that warn of problems with the spectra or problems with the label determination for the spectra (or both). This pipeline is based on χ^2 fitting of the data to 1D LTE models for seven labels ($T_{\rm eff}$, log g, [Fe/H], [α/Fe], [C/M], [N/M], and micro-turbulence; A. E. García Pérez et al. 2015, in preparation).

We use here spectra from the set of 55,000 stars that were released as part of the SDSS Data Release 10 (DR10) (Ahn et al. 2014), focusing on data from the *apStar* and *aspcapStar* FITS files. The *apStar* files include single-visit and combined spectra for a given star that are fully reduced, resampled, and shifted to the stellar rest frame. The *aspcapStar* files contain the combined spectrum for a given star that has also been pseudo-continuum normalized by the ASPCAP pipeline, along with the best-fitting synthetic spectrum and stellar labels. The *apStar* data, which are not pseudo-continuum-normalized by *APOGEE*, enables us to evaluate the performance of *The Cannon* at lower S/N, by testing it on the individual visit spectra provided in these files.

[7] www.sdss.org

[8] Due to gaps between the instrument's three detectors, the spectra are divided into three pieces: ~151500–15800 Å, 15890–16430 Å, and 16490–16900 Å.

The pixel-by-pixel inverse variances are critical for all steps of *The Cannon*: continuum normalization, training step, and test step. The error arrays of the uncertainty at each pixel in the spectra are provided by *APOGEE* in their fits files. We adopt these vectors directly and additionally set any anomalous values in the spectra, with 0 flux or very high error values to a very large error value, for computational stability.

Aside from photon noise, a number of other factors can contribute to the errors of any pixel in *APOGEE* spectra: poor sky subtraction, cosmic rays, reduction induced errors, high persistence, and other noise sources. In addition to the variance arrays, one can also use any bad pixel masks, where the inverse variance and weighting of that pixel becomes ~0. We find that adopting additional masking from the bad pixel masks degrades our results when using the combined *aspcapStar* spectra. However, our results are improved for individual visit spectra in the *apStar* files when pixels flagged in the bad pixel mask array provided by *APOGEE* are rejected, by assigning the large weighing in the error on those pixels, for the individual visit spectra. We therefore only implement the bad pixel masks from *APOGEE* for our tests on single visit spectra.

The resampled, reduced, and combined spectra are available for about 47,000 survey stars in 150 DR10 fields in the *aspcapStar* files. There are a further 9000 star observed in commissioning mode only, which are available in the radial velocity combined but not continuum-normalized data format in the *apStar* files.

We also apply *The Cannon* to the commissioning data and caution the reader about the fidelity of these results, as the LSF of the commissioning data are different from the main survey and consequently different from the reference data set of stars in the training step. Those objects are flagged as commissioning data (see Table 1). Typically, the LSF for reference and all survey objects is the same within a given survey. As *The Cannon* calculates a separate spectral model for each survey it is applied to, survey homogeneity is sufficient; we do not need to know the actual LSF. This assumption breaks down if the survey stars are observed under a different instrumental setup to the reference objects, as is the case with commissioning data from *APOGEE*. In this case, the LSF would have to be adjusted in a separate step, not introduced as a separate label; labels in the current context are strictly properties of the stars, not the experimental setup.

2.2. Choosing Reference Objects for the Training Step

For the training step in *The Cannon* we must choose a set of reference (or training) objects for which we have spectra from the survey under consideration and *also* high-fidelity labels

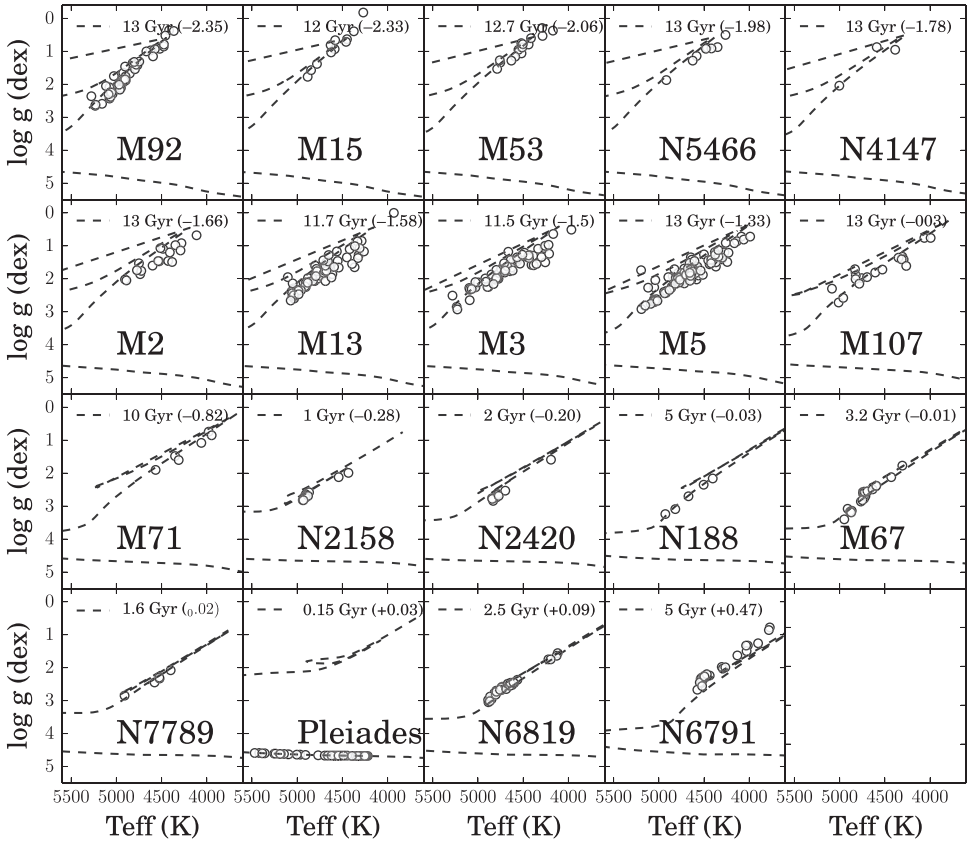

Figure 1. *ASPCAP*-corrected DR10 labels for the training step in $T_{\rm eff}$–log g plane for 542 stars in the 19 clusters for which parameters are provided by *APOGEE* (Mészáros et al. 2013). The age and [Fe/H] of the isochrones (in parentheses) is shown in each sub-panel. All labels adopted from the *ASPCAP*-corrected values of DR10 except for the Pleiades.

(that is, stellar parameters and element abundances that are deemed both accurate and precise). The set of reference objects is critical, as the label transfer to the survey objects can only be as good as the quality of the reference label set. Also, as *The Cannon* may have to interpolate and extrapolate to new parts of label space as it encounters new kinds of spectra among the survey objects, the quality of the label transfer depends on the extent to which the reference objects cover label space and the density with which they cover it. The performance of a data-driven model like *The Cannon* will depend strongly on the size and quality of its training set of reference objects.

In practice, also for the *APOGEE* data, one (but not only) good option for a set of reference objects can be built from members of well-studied open and globular clusters that have been observed in the context of the survey (Mészáros et al. 2013; Zasowski et al. 2013). There is a variety of reasons why the stellar labels for cluster stars may be particularly accurate and robust. For one, they could have their labels derived from independent, high resolution spectral analysis of these stars, for example, from observations in a well understood portion of the optical wavelength region. The labels are of course a property of the star, and hence do not have to arise from the survey data at hand. They may have been derived from different data.

For the case of *APOGEE*, we will use as reference objects 542 members of 19 globular and open clusters (Mészáros et al. 2013). These are the very same objects as used by the *APOGEE* survey for their own calibration of the DR10 data release and represent the documented reference objects that are available. Some objects were removed from the full list available as their cluster memberships were incorrect. In their stellar labels they span the range of $3500 < T_{\rm eff} < 5300$ K, $0 < \log\ g < 5$, and $-2.5 <$ [Fe/H] < 0.45.

Exactly which stellar labels we adopt for these reference objects is critical to the subsequent output, and hence we discuss it in detail in Section 2.4.

Another reason why cluster members make for good reference objects is because we can expect their stellar parameters to fall onto a single isochrone and to have near-identical abundances (at least for open clusters). This provides additional constraints on the labels. We exploit that expectation in the case of *APOGEE* and define "Isochrone-corrected labels," where we use Padova isochrones at the literature age and [Fe/H] of each cluster (see Figures 1 and 2, and Section 2.4).

2.3. Consistent Continuum-normalization

The Cannon operates on continuum-normalized spectra. Continuum-normalization that is based on quantiles of the data (medians or 90th percentiles or the like) are very S/N dependent, for example, because pixels that are clearly *not* continuum in high S/N spectra are completely consistent with being continuum at lower S/N. Therefore, to make *The Cannon* as independent of S/N as possible, we base the continuum estimation on a pre-tabulated set of wavelength locations that we know (iteratively, from running *The Cannon* itself, see Section 5.3) are not strongly affected by absorption lines.

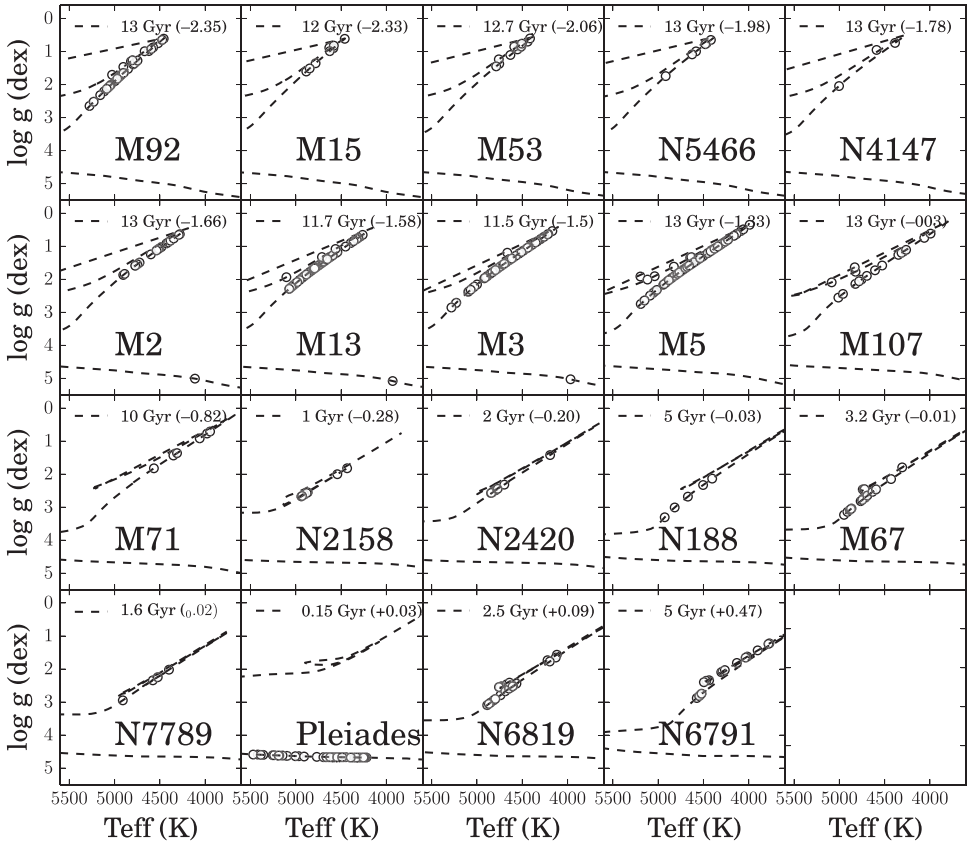

Figure 2. Stellar labels for all reference objects as is Figure 1, except that the log g values have been adjusted from the *ASPCAP*-corrected value to exactly match the isochrone, we refer to this set of labels as "isochrone-corrected" labels to differentiate them from the correction in Figure 1.

To initialize the continuum-pixel determination, we define a preliminary pseudo-continuum-normalization by using polynomial fit to an upper quantile (for example, 80% or 90%) of the spectra, determined, for example, from a running median. For this pseudo continuum-normalized *APOGEE* spectra we use a running quantile across 50 Å of the spectra, taking the 90th percentile. This is effective, but S/N-dependent.

After a training step using spectra of reference objects that have been normalized by this pseudo-continuum, *The Cannon* can provide an improved identification of continuum regions in the spectrum: we take those pixels to be continuum that show nearly unity flux in the spectral model's baseline spectrum (see Section 3), and at the same time show almost no dependence in their normalized flux on the stellar labels. That is, for the *APOGEE* data, we can determine with *The Cannon* the "true" continuum, using the model derived from the pseudo-continuum-normalized spectra for the reference objects provided by *APOGEE*, as described in Section 5. This constitutes a data-driven method for finding continuum pixels, and we find it to have only a very small systematic dependence of the spectra on S/N (see Section 5.3).

Given the continuum pixels, we implement a least-squares fitting to the *APOGEE* spectra of a low-order Chebyshev polynomial, fitting only to the determined continuum pixels outlined in Section 5.3. We treat each of the three chips separately, and find a second-order Chebyshev polynomial to be sufficient to apply to the data provided by *APOGEE*. We apply this normalization to both *aspcapStar* and *apStar* files. Treating the three chips separately, we fit the polynomials over the wavelength regions of (i) 15150–15800 Å, (ii) 15890–16430 Å, and (iii) 16490–16950 Å. Fitting a polynomial has the disadvantage that they are poorly constrained at the edges of the data. An alternative implementation could use a more sophisticated sine or cosine function in place of a polynomial fit.

Figure 3 shows an example of this iterated normalization applied to survey spectra with different labels. To illustrate the result, Figure 3 shows typical *APOGEE* spectra and demonstrates how the spectra vary as a function of metallicity at a given temperature, and as a function of temperature at a given metallicity. For a clearer view of individual absorption line features, we use narrower regions marked in this figure, (A) and (B), for all subsequent examination of the spectral data.

2.4. Labels for the Reference Objects in APOGEE

Which values to adopt for the labels of the reference objects used in *The Cannon*'s training step is a critical issue in any survey. We discuss two options here for *APOGEE*. First, we adopt the DR10 *"ASPCAP-corrected"* stellar parameters (Mészáros et al. 2013) that are available for each of the reference objects as their labels, in order to place the output of *The Cannon*'s test step for the survey objects on the *APOGEE ASPCAP* scale (Figure 1). "*ASPCAP*-corrected" labels were not available for the cluster comprised of main sequence stars, the Pleiades cluster and for this cluster we made our own corrections in $T_{\rm eff}$ and log g and assumed a single literature value for the [Fe/H] label, as described below. The reference set of stars we use is the very same stars used by *APOGEE* to post-calibrate the output of *ASPCAP* to a physical stellar parameter scale (Mészáros et al. 2013). Adopting the

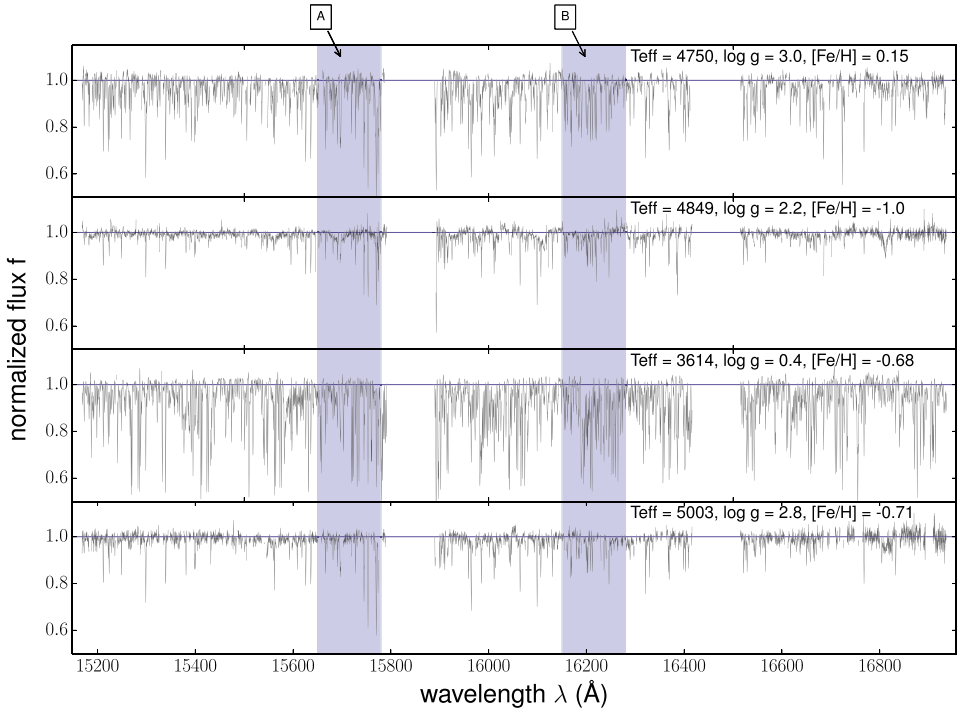

Figure 3. Continuum-normalized spectra for stars across a range of stellar labels; at top, two stars of similar temperatures at different metallicities and at bottom, two stars of similar metallicities and different temperatures. The gray shaded regions A and B indicate *APOGEE* sample wavelength regions used for subsequent figures in the paper.

ASPCAP-corrected labels provided and documented by *APOGEE* has the important advantage that we can test exactly how well we can reproduce the results from *APOGEE* for the survey stars via label-transfer from only 542 stars. A limitation of this reference set of objects is that main sequence stars are not well sampled, which, as we discuss in Section 5, limits our ability to determine labels for these stars at the test step.

The corrections to the labels made by *APOGEE* described in Mészáros et al. (2013) are based on the cluster data and applied to the immediate output of the *ASPCAP* pipeline that arose from comparisons to a library of stellar models. Temperature corrections are determined by comparing the infrared flux temperatures of the stars (Gonzalez et al. 2009), log g corrections are from the offset between *ASPCAP* results and *Kepler* astroseismic results for stars in common and [Fe/H] corrections are from the difference between the *ASPCAP* and the literature value of each cluster. The *APOGEE* corrections determined in Mészáros et al. (2013) are valid only for stars with log $g < 3.5$ and are not implemented for the dwarfs. We adopt the *ASPCAP*-corrected [M/H] values for these clusters and these are corrected to the [Fe/H] of the clusters and so we adopt this label as an [Fe/H] (that is, this label from *APOGEE* therefore, does not explicitly use [Fe/H] lines, but is derived from an [Fe/H] correction). The analysis in Mészáros et al. (2013) is restricted not only to giants but also stars with S/N > 70, determined to be the minimum S/N for reliable stellar parameters by *APOGEE*.

These corrections implemented by *APOGEE* in $T_{\rm eff}$, log g, and [Fe/H] place the giants in the cluster stars on or near the iscohrones (see Figures 7 and 8 in Mészáros et al. 2013). As there are no *ASPCAP* corrections implemented for the 65 main sequence stars among the reference objects that we use, we instead determine temperatures for these dwarfs, which are all in the Pleiades, using the same correction method as in Mészáros et al. (2013). We determine the infrared flux temperature for the stars from Gonzalez et al. (2009) and apply a correction to the *ASPCAP* output based on the offset in the temperature scales. For the dwarf stars in the Pleiades, we find the following relation: $T_{\rm corrected} = 0.855{*}T_{ASPCAP} + 1206.7$.

We do not attempt an individual metallicity correction for each dwarf star in the Pleiades but rather set all [Fe/H] of the dwarf spectra to [Fe/H] = 0.03 (Barrado y Navascués et al. 2001). Tests on the input labels to *The Cannon* demonstrate that there is only a small degradation of the results caused by adopting a single [Fe/H] for every cluster star for the literature value of the cluster, instead of individual *ASPCAP*-corrected [Fe/H] values for the stars. To determine the log g for these Pleiades main sequence stars, we shift the stars vertically to their nearest positions on an appropriate age–metallicity Padova isochrone of 150 Myr at [Fe/H] = 0.03 (Girardi et al. 2000). Due to the high differential reddening to the Pleiades, and the subsequent large temperature errors using the IR flux method that result from this, we only selected the 65 from a total of 72 Pleiades dwarfs, eliminating those with high extinction of SFD (corrected) $E(J - K) > 0.30$ (Schlafly & Finkbeiner 2011).

Adopting the input labels from the *ASPCAP*-corrected parameters determined from calibrations to literature cluster values also transfers the errors from the *ASPCAP* pipeline: of <150 K in $T_{\rm eff}$, <0.2 dex in log g, and <0.1 dex in [Fe/H]. The uncertainties on the input labels will be included as an input parameter of the labels in a future development stage of *The Cannon*. Inclusion of uncertainties may be particularly relevant when introducing multiple labels of individual elements.

For a comparative analysis to the "*ASPCAP*-corrected" labels (Figure 2), we adopt a log g label for all of the training stars not from the *Kepler* scale, but rather from the best vertical fits to the isochrone for the ages and metallicities for the

clusters from the literature (with the temperatures fixed). We call these the *"Isochrone-corrected"* labels, where we use Padova isochrones at the age and [Fe/H] of each cluster.

3. *THE CANNON*'S TRAINING STEP: MAKING A GENERATIVE MODEL

We now lay out the spectral model, whose parameters are determined from the spectra and stellar labels of the reference objects in the training step. Such a generative model is based on two basic notions: first, that the continuum-normalized spectra of stars with identical labels look near-identical at every pixel, save for the observational errors and some intrinsic scatter. This must be true if the set of labels is exhaustive. In practice, that is an approximation, as for example, the spectra of stars with identical $T_{\rm eff}$, log g, and [Fe/H] may differ, as these stars have different [α/Fe], age, or rotation. Second, we presume that the expected flux at every pixel changes continuously with changes in the labels. Importantly, the model is a probabilistic generative model that produces, for every object spectrum at every wavelength, a pdf for the flux, with an expectation value (mean) and a variance.

We presume there are $N_{\rm ref}$ reference objects n, each of which has a continuum-normalized flux measurement $f_{n\lambda}$ at wavelength λ. Each of the training spectra (of index) n has K labels ℓ_{nk}, each of which is (for now) presumed to have negligible uncertainty and contained (possibly with transformations; given below) within a label vector $\boldsymbol{\ell}_n$.

We then presume that for any star, n at any pixel, λ the flux $f_{n\lambda}$ can be described as some smooth function of the star's labels ℓ_{nk} ($T_{\rm eff}$, log g, [Fe/H], ...). The observations $f_{n\lambda}$ will differ from such a model by the observational noise (from all relevant sources), $\sigma_{n\lambda}$. But even for perfect measurements we presume that there will be deviations from the above approximate model for the true flux, characterized by a scatter s_λ, which is a property of any particular pixel; we will subsume s_λ under the noise.

Generally, we take a spectral model to be characterized by a coefficient vector $\boldsymbol{\theta}_\lambda$ that allows to predict the flux at every pixel $f_{n\lambda}$ for a given label vector $\boldsymbol{\ell}_n$:

$$f_{n\lambda} = g\left(\boldsymbol{\ell}_n \middle| \boldsymbol{\theta}_\lambda\right) + \text{noise}. \tag{1}$$

As a specific, but still flexible functional form for the spectral model we presume that it can be written as a linear function of some vector $\boldsymbol{\ell}_n$ built from the labels:

$$f_{n\lambda} = \boldsymbol{\theta}_\lambda^T \cdot \boldsymbol{\ell}_n + \text{noise} \tag{2}$$

where $\boldsymbol{\theta}_\lambda$ is the set of spectral model coefficients at each λ. Each element of $\boldsymbol{\ell}_n$ can be some (possibly complicated) function of the full set of K labels, $\boldsymbol{\ell}_n$, which results in the flexibility of this model. The noise is an rms combination of the associated uncertainty variance $\sigma_{n\lambda}^2$ of each of the pixels of the flux from finite photon counts and instrumental effects and the intrinsic variance or scatter of the model at each wavelength of the fit, s_λ^2. This model assumes that the noise model is noise $= [s_\lambda^2 + \sigma_{n\lambda}^2]\ \xi_{n\lambda}$, where each $\xi_{n\lambda}$ is a Gaussian random number with zero mean and unit variance.

The simplest spectral model is that in which the label vector $\boldsymbol{\ell}_n$ is linear in the labels, that is, in the vector of the individual labels themselves:

$$\boldsymbol{\ell}_n \equiv \left[1, \ell_{n1} - \overline{\ell_1}, \ell_{n2} - \overline{\ell_2}, \cdots, \ell_{nK} - \overline{\ell_K}\right], \tag{3}$$

where the first element "1" will permit a linear offset in the fitting. The $\overline{\ell_k}$ are offsets (possibly means of the training data) to keep the model "pivoting" around a reasonable point in label space. This model leads to the single-pixel log-likelihood function

$$\ln p\left(f_{n\lambda} \middle| \boldsymbol{\theta}_\lambda^T, \boldsymbol{\ell}_n, s_\lambda^2\right) = -\frac{1}{2}\frac{\left[f_{n\lambda} - \boldsymbol{\theta}_\lambda^T \cdot \boldsymbol{\ell}_n\right]^2}{s_\lambda^2 + \sigma_{n\lambda}^2} - \frac{1}{2}\ln\left(s_\lambda^2 + \sigma_{n\lambda}^2\right). \tag{4}$$

The vector $\boldsymbol{f}_\lambda$ is the set of spectral flux values for all N objects at the one wavelength λ. This is a likelihood function for the labels and parameters: it provides a pdf evaluation at the data given settings of all the labels and parameters. We can set the coefficients $[\boldsymbol{\theta}_\lambda, s_\lambda^2]$ either by optimizing the likelihood (4) over all reference objects or by applying priors and performing some form of probabilistic inference (with, say, Markov Chain Monte Carlo techniques). Here we will optimize for now, which can be done separately for each pixel λ, where we are treating the spectral model coefficients $\boldsymbol{\theta}_\lambda$ and the scatter s_λ^2 as free parameters, and the labels in the label vector $\boldsymbol{\ell}_n$, ℓ_{nk} as fixed:

Then, in the training step of *The Cannon* we exploit the fact that we know the $f_{n\lambda}$ and the $\boldsymbol{\ell}_n$, which permits to solve for the coefficients and the scatter of the spectral model:

$$\boldsymbol{\theta}_\lambda, s_\lambda \leftarrow \overset{\text{argmax}}{\boldsymbol{\theta}_\lambda, s_\lambda} \sum_{n=1}^{N} \ln p\left(f_{n\lambda} \middle| \boldsymbol{\theta}_\lambda^T, \boldsymbol{\ell}_n, s_\lambda^2\right). \tag{5}$$

The linear-in-labels form (3) has a number of useful properties. The coefficient vector $\theta_{\lambda 0}$ has a simple interpretation; it is the "baseline spectrum" of the spectral model. The next coefficient vectors, $\theta_{\lambda k}$, linear in $T_{\rm eff}$, log g, and [Fe/H], describe the lowest-order dependence of the spectrum on these labels. In practical terms, the optimization of the model parameters $\theta_{\lambda k}$, at fixed scatter s_λ^2 is a pure linear-algebra operation (weighted least squares); simultaneous optimization of all the parameters $[\boldsymbol{\theta}_\lambda, s_\lambda^2]$ is only nonlinear in the s_λ^2 parameter.

The (perhaps) second-simplest spectral model is that in which the vector $\boldsymbol{\ell}_n$ is quadratic in the labels: so this label vector is described as

$$\boldsymbol{\ell}_n \equiv \left[1, \ell_{n1} - \bar{\ell}_1, \ell_{n2} - \bar{\ell}_2, \ldots, \ell_{nK} - \bar{\ell}_K, \left(\ell_{n1} - \bar{\ell}_1\right)\left(\ell_{n1} - \bar{\ell}_1\right), \left(\ell_{n1} - \bar{\ell}_1\right)\left(\ell_{n2} - \bar{\ell}_2\right), \cdots, \left(\ell_{nK} - \bar{\ell}_K\right)\left(\ell_{nK} - \bar{\ell}_K\right)\right], \tag{6}$$

where the quadratic terms contain all possible products exactly once.

For the training step of *The Cannon*, this quadratic-in-labels form of the spectral model (6) is similar to a the linear-in-labels form (3) in a number of ways. It is still the case that optimization of the model, at fixed scatter s_λ^2 is a pure linear-algebra operation (weighted least squares), except that $\boldsymbol{\ell}_n$ has become longer for a given number of labels. However, the test step on the survey (described in the next section) of the

quadratic-in-labels form will no longer be simple; it will require nonlinear optimization to estimate the labels.

The coefficients $\theta_{\lambda 0}$ can still be seen as an estimate of the *baseline spectrum* (provided that the offsets $\overline{\ell_k}$ are the mean tag values); the first-order coefficients $\theta_{\lambda k}$ can still be seen as first derivatives of the expected spectrum with respect to each of the k labels, but now evaluated at the baseline spectrum; the second-order coefficients $\theta_{\lambda k k'}$ can now be seen as mean second derivatives of the expected spectrum with respect to pairs of labels k and k'.

4. *THE CANNON*'S TEST STEP: LABELING SURVEY SPECTRA

In the previous section, we trained or fit the parameters of a data-driven probabilistic generative model for stellar spectra from the reference objects serving as training data. This model has the property that, given labels (and noise variance estimates), it produces a pdf for the continuum-normalized flux, that includes both observational and intrinsic scatter. In this section, we are going to solve the inverse problem: we have spectra, but we do not have labels for them. In this case, we will use inference and the just determined spectral model to obtain labels for the untagged survey spectra, which we also refer to as the "test data" in what follows.

In the test data there will be M spectra m, each of which—as in the training data—has a continuum-normalized flux measurement $f_{m\lambda}$ at each wavelength λ, and an associated observational uncertainty variance $\sigma^2_{m\lambda}$. Just as in the training step, we consider the same likelihood function given in Equation (4). But now we view it as a function of the *labels*, instead of the function parameters $\boldsymbol{\theta}_\lambda$ and scatter s^2_λ.

In the test step of *The Cannon* we use the spectral model coefficients and scatter, $(\boldsymbol{\theta}_\lambda, s^2_\lambda)$, to be exactly those that were determined in the training step. We then take the entire $N_{\rm pix}$ spectrum of survey star m, $f_{m\lambda}$ and optimize for the labels of that star:

$$\{\ell_{mk}\} \leftarrow \mathop{\rm argmax}_{\{\ell_{mk}\}} \sum_{\lambda=1}^{N_{\rm pix}} \ln p\left(f_{m\lambda} \left| \boldsymbol{\theta}_\lambda^T, \boldsymbol{\ell}_m, s^2_\lambda\right.\right). \quad (7)$$

The labels ℓ_{mk} for each survey star m can be obtained either by maximizing the likelihood function, or else by applying priors and performing probabilistic inference. Again, we will optimize here. Our optimization is not convex in general, but in practice it is insensitive to initialization. The right-hand sides of the training step (5) and test step (7) look formally quite analogous. But in the test step we optimize over the labels, considering all pixels of one survey object at a time. In contrast, in the training step, we optimize over the spectral model coefficients and scatter, considering all reference objects at one pixel at a time.

When we use the simple linear-in-labels form (3) for the mean model, the optimization to obtain maximum-likelihood labels (given parameters $[\boldsymbol{\theta}_\lambda, s^2_\lambda]$) is simple linear least-square fitting. This optimization is obtained by straightforward linear algebra on the spectral pixels $f_{m\lambda}$, and standard frequentist confidence intervals can be obtained similarly. When we use the quadratic-in-labels form (6) for the spectral, there is no simple linear-algebra operation that optimizes the likelihood. Instead an optimization function is used, the python curve_fit routine, which uses a nonlinear least squares fit to fit the function to the data.

We have described how we construct a spectral model from the reference objects in the training step and then estimate stellar labels for survey stars with that model in the test step. We now present in Section 5 the results of implementing our model for all *APOGEE* data, where we applied a quadratic model: linear in the coefficients and nonlinear in the label-inference. For the quadratic model we then show this applied to the DR10 data, including at lower S/N, and investigate different input training labels.

5. RESULTS WITH *APOGEE* DATA

We now present the results for applying *The Cannon* to *APOGEE* data.

To apply *The Cannon* to *APOGEE* data, we first train the quadratic model in Equation (6) using the reference data and three labels chosen as described in Section 2.4. We then apply this model to all of the DR10 continuum-normalized data, using continuum-normalized *aspcapStar* spectra described in Section 5.3. We use a leave-one-out cross-validation test to explore which complexity the spectral model must have and how comprehensive the set of reference objects should be. We then proceed with the same set of reference objects in the label transfer to the entire DR10 in the test step. In particular, we also apply the test step to spectra from individual *APOGEE* visits that have far shorter exposure times, and hence lower S/N than the co-added spectra in DR10, in order to explore and illustrate how well *The Cannon* does at modest S/N (with appropriate continuum fitting).

5.1. The Choice of the Spectral Model Complexity

To evaluate *The Cannon*'s label-transfer we have to settle on a suitable functional form for the spectral model (1). To start, one could consider picking the simplest—namely linear-in-label—spectral model, comprised of only four coefficients at every pixel (3). However, through take-one-out tests on the set of reference objects (see Section 5.2), we found that this simple linear model was too inflexible to describe the spectral flux dependence on the labels. As a consequence, the labels that emerged from the test step applied to the reference objects showed large and systematic deviations compared to "known" input label values, especially at the extremes of the labels' ranges.

This is perhaps not surprising, as absorption features, particularly strong lines, are known to vary nonlinearly as a function of stellar labels. If one were to insist on a label transfer with a first-order spectral model, the systematic discrepancies could presumably be reduced by selecting only weak-line regions, but at the severe price of leaving much of the spectral range unexploited. Therefore, we have not pursued the linear-in-labels ansatz for the spectral model.

The next simplest spectral model, the quadratic-in-labels case, presumes that the continuum-normalized flux is a general second-order polynomial of the stellar labels, $f_{n\lambda} = \boldsymbol{\theta}_\lambda^T \cdot \boldsymbol{\ell}_n + \text{noise}$ (2), but where $\boldsymbol{\theta}_\lambda$ now contains 10 elements at every pixel. For the case of the three labels ($T_{\rm eff}$,

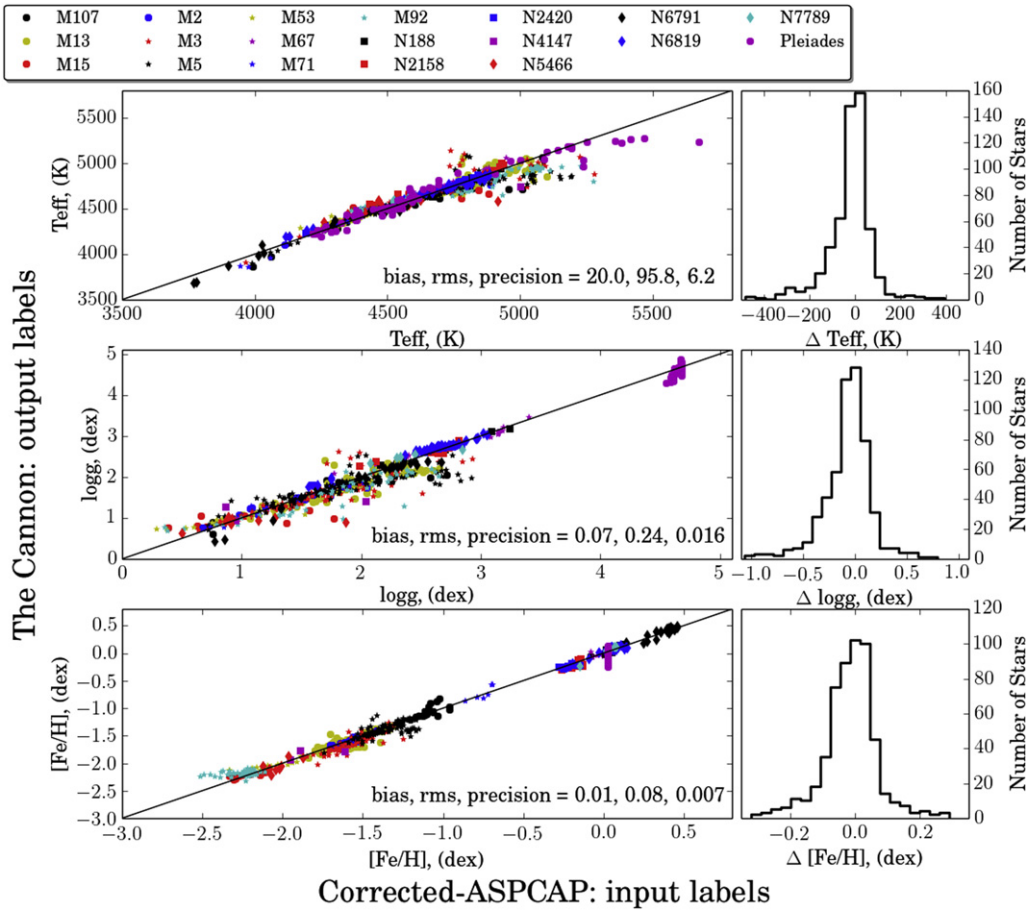

Figure 4. The take-one-star-out cross-validation of the 542 stars in the training data set using the quadratic model in Equation (8) and corresponding histograms at right, showing *The Cannon* output—*APOGEE* input labels.

log g, [Fe/H]) the label vector ℓ_n becomes

$$\ell_n \equiv \Big[1,\, T_{\rm eff},\, \log\, g,\, [{\rm Fe/H}],\, T_{\rm eff}^2,\, T_{\rm eff}\cdot \log\, g,\, T_{\rm eff}\cdot [{\rm Fe/H}],\, \log\, g^2,\, \log\, g\cdot [{\rm Fe/H}],\, [{\rm Fe/H}]^2\Big]. \quad (8)$$

We will use this quadratic-in-labels spectral model throughout the rest of the paper. The exploration of higher-order polynomials for the spectral model at every pixel, or even a Gaussian process at every pixel, is beyond the scope of this paper. For any other application the complexity of the spectral model (for example, is a quadratic model good enough?) should be examined.

5.2. Validation on Take-one-out Stars from the Reference Objects

As a first illustration of how well *The Cannon* works in practice, we perform a take-one-star out test on the set of reference objects. For the take-one-star out test we train the spectral model iteratively on the spectra of all but one of the $N_{\rm ref}$ (=542) reference objects, and then apply *The Cannon*'s test step to the spectrum of that remaining object. If we repeat this procedure $N_{\rm ref}$ times, we have a first powerful test of how the result of this parameter transfer compares to the (known) labels for the reference objects. Here we only consider three labels ($T_{\rm eff}$, log g, [Fe/H]), and the results are shown in Figure 4.

This figure clearly shows how well *The Cannon* works, at least in the circumstance at hand. *The Cannon*'s purely mathematical approach of label transfer estimates the stellar labels (at least) as well as the astrophysical *ASPCAP* pipeline, over the full label range of our the reference data. The rms of the difference between the *ASPCAP* and *The Cannon* values for the three labels are 95 K in $T_{\rm eff}$, 0.24 in log g, 0.08 in [Fe/H], with biases Δ that are 3–7 times smaller. These variances inherently include some portion of the uncertainties on the input labels (from ASCPAP corrected values, of $T_{\rm eff} < 150$ K, log $g < 0.2$ dex, and [Fe/H] < 0.1 dex Mészáros et al. 2013). The precision values stated in Figure 4 are the formal uncertainties in the labels arising in the test step's optimization; for the S/N of the spectra in this take-one-out test, these errors are very small. It is important to remember that the one left-out object and its spectrum are completely detached from the training step, except that they have the same experimental set-up and are likely drawn from a part of label space well-represented by the remaining reference objects.

There are a few outliers in Figure 4, cluster members of M3 in particular, that are offset in $T_{\rm eff}$ and log g space. The Pleiades cluster, which has only spectra for main sequence stars, shows the poorest determination in the [Fe/H] label. We assigned all its members a single [Fe/H] as reference labels, unlike the other reference objects, where we used their *ASPCAP*-corrected labels from DR10. The rms is comparable

to the estimated *APOGEE* errors. The log g label has the largest relative rms in the *ASPCAP–The Cannon* comparison, larger than the *APOGEE* uncertainty, suggesting an internal uncertainty of <0.1 dex in log g determined by *The Cannon*. If we adopt instead of *ASPCAP*-corrected log g labels the isochrone-corrected log gs (see Section 2.4), the rms improves by 10% in $T_{\rm eff}$ and log g.

The outlying stars in Figure 4 may be due to an anomalous scale of the input labels of these stars compared to the other training data, or it may be a consequence of the model being too inflexible to properly describe how flux changes with labels across the parameter space of the training data set. The temperature of the dwarfs is offset low at increasing temperature, compared to the input labels, so the model may be limited in describing the difference between dwarf and giant spectra. There is a flattening at the low metallicity end of the model in [Fe/H] in the output labels at [Fe/H] < -2.2. However, this value of [Fe/H] $= -2.2$ also corresponds to the literature value of this cluster, M5 (Mészáros et al. 2013). The lower metallicity of the *ASPCAP* label may represent internal scatter in the *ASPCAP* results. The fact that Figure 4 shows only very small systematic offsets and such tight scatter leads us to conclude that for the current context the quadratic-in-labels spectral model is sufficient in the label transfer.

Interestingly, an analogous take-one-cluster-out test significantly increases the scatter in the label transfer, increasing the rms differences to <150 K in $T_{\rm eff}$, <0.4 dex in log g, and <0.12 dex in [Fe/H]. This indicates that our training set is sufficiently small that each cluster matters for a good label transfer. One particular case in the *APOGEE* context is the Pleiades cluster: it is the *only* cluster for which dwarf stars have been observed and hence we can draw reference labels for main sequence stars.

We now turn to illustrating where the information that led to the accurate label transfer (Figure 4) came from in the spectra. Figure 5 shows—across the narrow regions (A) and (B) of the spectra, marked in Figure 3—the first coefficient vectors $\theta_{0,1,2,3}$ of the spectral model (those linear in the three labels), which were fit for in the training step for the quadratic-in-labels model in Equation (6).

The top panel of Figure 5 shows the zeroth order-coefficient vector θ_0, or the baseline spectrum, of the model. The mid panel shows the coefficients that are simply linear in $T_{\rm eff}$, log g, and [Fe/H]. In the top panel of Figure 5, the red, blue and green shaded wavelength regions with the 5% highest coefficient values $|\theta_{1,2,3}|$ in the [Fe/H], log g and $T_{\rm eff}$ labels, respectively. These regions indicate where the spectra's flux levels strongly vary with these labels. This also highlights that different parts of the spectrum depend differently on the labels. Note there are many regions where the [Fe/H] label dominates in contribution to the flux. For the first label vector for example in the middle panel of Figure 5, there is typically asymmetry for a given absorption feature, in the flux and the labels. There are very few regions where the flux is a function of only one of the labels, and pixels are typically co-variant. (that is, the same pixel will have a higher flux at both lower $T_{\rm eff}$ and higher [Fe/H]). This simply reflects well-known co-variances between, for example, temperature and [Fe/H]. The strongest log g dependence is typically associated with weak lines including the wings of the feature and the [Fe/H] label, with strong lines, particularly the depth of the line.

The bottom panel of Figure 5 shows the scatter vector of the spectral model, indicating the dispersion of the flux of the training data around the best-fit spectral model at each pixel. The scatter is small and this indicates that our model is a good representation of the data. However, the scatter is highest where the most information in the spectra are contained. This implies that either our quadratic-in-labels spectral model is still somewhat too restricted, or that the labels of our training data set are imperfect or incomplete (for example, lacking [α/Fe] as a label), or a combination of these effects. From the coefficients of an initial fit of this spectral models (see, for example, the middle panel of Figure 5), the continuum pixels have been determined following Section 2.3. These are marked in the cyan dots in the top panel of the figure, and are used for an iterated, consistent continuum-normalization for all spectra, both of the reference and of the survey objects.

To demonstrate the fit of the model to the *APOGEE* spectra at test time, using the continuum normalized test stars shown in Figure 3, we plot the best fit model and the corresponding spectra for the regions A and B highlighted in this figure. This is shown in Figure 6, which demonstrates that the model is an excellent fit to the data: the three labels used to describe the flux are sufficient.

5.3. Identification of APOGEE Continuum Pixels

The continuum pixels shown in Figure 5 for wavelength regions A and B have been determined from the training step with a quadratic-in-labels model operating on spectra normalized by their preliminary pseudo-continuum, using the coefficients returned (see Section 2.3). About 35% of the pixels in the resulting baseline spectrum (the vector θ_λ^0) have flux levels within 1% of unity. However, not all these pixels are suitable continuum pixels, as many of them have significant dependencies, $\theta_\lambda^{1,2,3}$, on the three labels. In practice, a good set of continuum pixels can be identified from the *APOGEE* spectra using a flux cut in the baseline spectra of the model, 1 ± 0.15 (0.985–1.015), combined with the smallest 20–30 percentile of the first order coefficients, $\theta_\lambda^{1,2,3}$, which retains between 5% and 9% of pixels. We found empirically that changing the latter percentiles to $(\theta_\lambda^1, \theta_\lambda^2, \theta_\lambda^3) < (1e^{-5}, 0.0045, 0.0085)$ returns only 6.5% of the pixels, but ultimately makes for an even better match to the *ASPCAP* label scale; we adopt this procedure. We use the inverse variance weighting of these pixels for the corresponding second order Chebyshev polynomial fit, adding an additional error term that is set to 0 for continuum pixels and a large error value for all other pixels so that the new error term σ_λ' for each pixel becomes: $\sigma_\lambda' = \sigma_\lambda + \sigma_{0|\rm LARGE}$.

As we show explicitly in Section 5.6, we find this to provide a robust continuum-normalization across the stars that are within the parameter range of the training set, across all S/N.

5.4. The Cannon's Label Transfer for APOGEE DR10

Going beyond leave-one-out tests on the set of reference objects, we now apply *The Cannon* to effect a label transfer to the entire *APOGEE* DR10. We take the spectral model built in the *APOGEE* cluster stars in Sections 5.1–5.3, and apply the test step with this model to all DR10 spectra. Remarkably, we are able to reproduce well the *ASPCAP* labels for DR10 spectra. We have run *The Cannon* through all 47,000 stars in 150 fields in DR10 contained in the available *aspcapstar* files

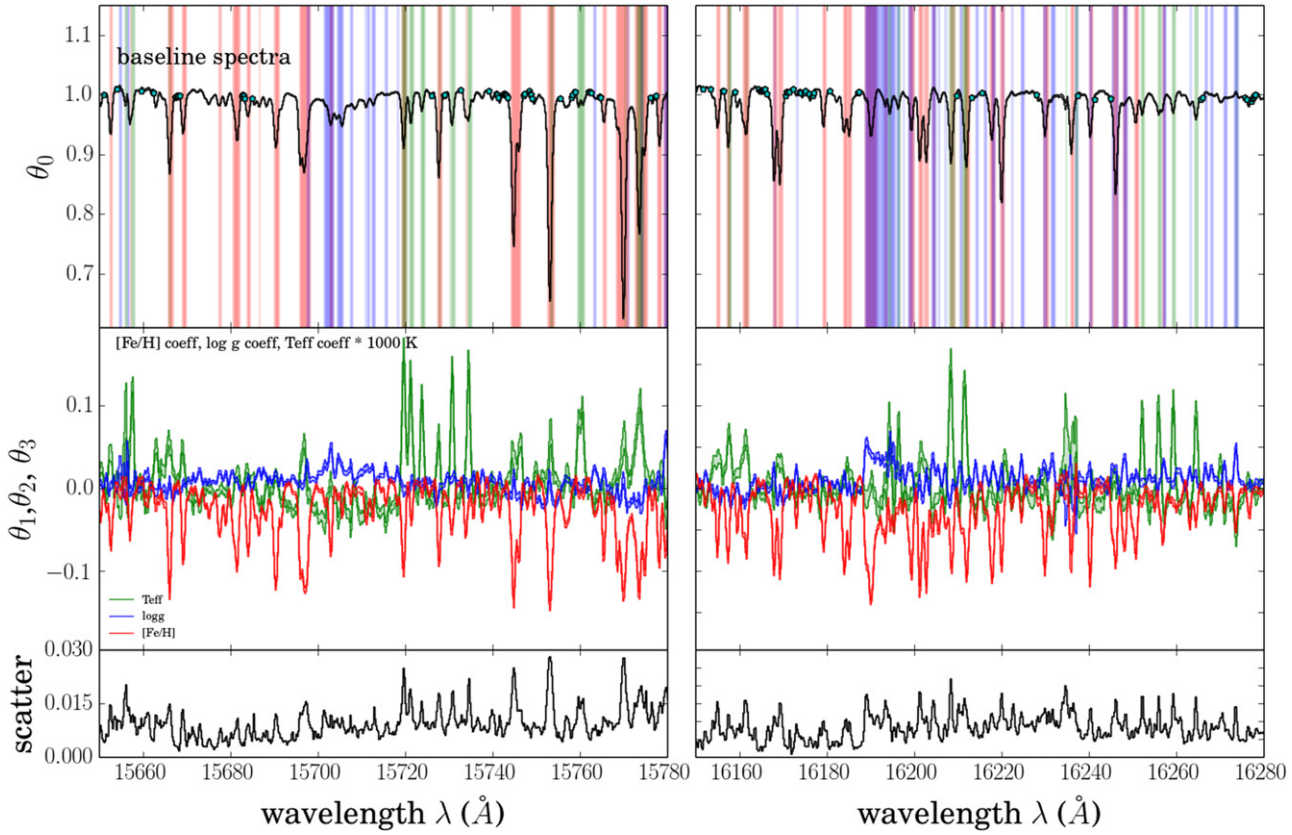

Figure 5. First-order coefficients and scatter across the sample regions of the spectra from Figure 3, (A) and (B). Top panel: the baseline spectra representing the first coefficient from the set of reference spectra; middle panel: the next three coefficients (θ_1, θ_2, θ_3), which correspond to the labels ($T_{\rm eff}$, log g, [Fe/H]); bottom panel: the scatter of the fit with a tenfold expanded vertical scale. The red, blue, and green areas in the top panel encompass the wavelength regions with the 5% highest (absolute value) coefficients for the [Fe/H], log g and $T_{\rm eff}$ labels, respectively. The $T_{\rm eff}$ coefficient has been multiplied by a factor of 1000 simply to show this coefficient on a similar scale to the other coefficients. This indicates where the flux in these spectrum is particularly sensitive to the labels. Note that the [Fe/H] label is dominant in the contribution level and from the top panel it is clear that there is significant covariance between the labels and there are only a few regions of log g sensitivity. The filled dots in the baseline spectrum in the top penal indicate the wavelengths at which the dependencies on all labels are weak, which we operatively identify as continuum pixels (see Section 5.3).

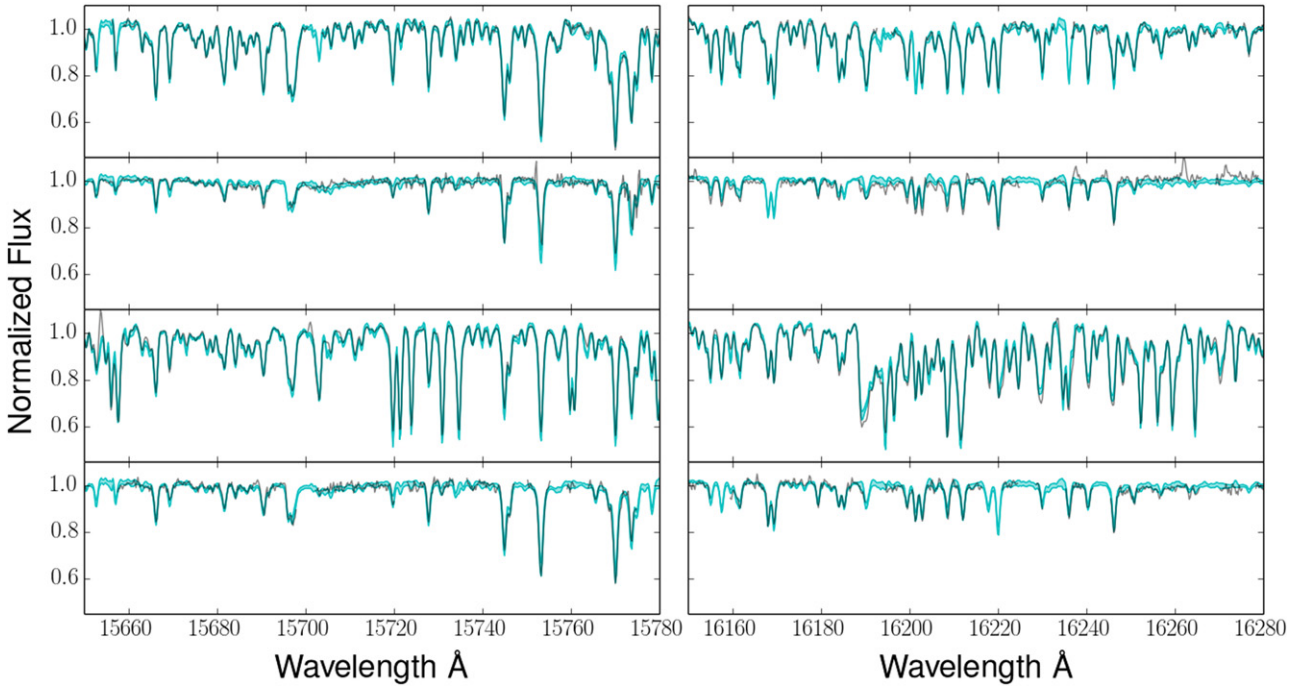

Figure 6. Four stars in Figure 3 across narrow wavelength intervals A and B, as described in Section 2.3. The spectra of the stars are plotted in black with the models in cyan, including the span of the scatter of the fit, generated by *The Cannon*.

as well as an additional 4800 stars in 20 commissioning fields for which no *ASPCAP* parameters were provided in DR10, made available in the (non pseudo-continuum-normalized) *apStar* files. We also have run *The Cannon* through the additional commissioning stars available in the *apStar* files across the fields and in total this comprises 55,000 DR10 stars in 170 fields.

These results of using *The Cannon* to return labels for all DR10 stars is provided online in Table 1. We provide two columns in this table which indicate the label-space returned for each test object, with respect to the reference objects in the training set. We provide an extrapolation flag, EFLAG, in the table, which indicates if the test star lies outside of the label-space of the reference objects (set to 1 if so). We also determine a distance measurement, $d_{\rm ref}$ defined as the distance between the test star (with labels $T_{\rm eff\,(test)}$, log $g_{\rm (test)}$, $[{\rm Fe/H}]_{\rm (test)}$) and the nearest reference object (with labels $T_{\rm eff\,(ref)}$, log $g_{\rm (ref)}$, $[{\rm Fe/H}]_{\rm (ref)}$), normalized to the maximum distance, so values lie between 0 and 1 (see Equation (9)). Additionally, in the online version of this table, we include a number of important *ASPCAP* flags in the online table. Stars with STAR BAD flag set (in ASPCAPFLAG) may have unphysical stellar parameters and commissioning stars are marked with "C." The fidelity of the commissioning stars is uncertain given their different LSF from survey test and training data. The velocity scatter (σ_v) from the *ASPCAP* results as well as the *APOGEE TARGET2* flag (TARG2) are also included

$$d_{\rm ref} = \{\overset{\rm argmin}{\text{over all } n}\} \sum_{1}^{k} \left(\frac{\Delta \ell_{kn}}{{\rm var}_k} \right)^2, \quad (9)$$

where $\Delta \ell_{kn}$ for the three labels is $= (T_{\rm eff\,(test)} - T_{\rm eff\,(ref,}{}_{n)})$, $(\log\ g_{\rm (test)} - \log\ g_{\rm (ref,}{}_{n)})$, $([{\rm Fe/H}]_{\rm (test)} - [{\rm Fe/H}]_{\rm (ref,}{}_{n)})$, and ${\rm var}_k$ is the variance of the reference object ensemble distribution in label l_k.

For the 28,700 stars with parameters (removing all stars flagged as STAR BAD) provided in DR10, we find we reproduce the *APOGEE* labels as follows: $T_{\rm eff} = +12 \pm 85$ K, $\log\ g = -0.04 \pm 0.18$ dex, and $[{\rm Fe/H}] = +0.01 \pm 0.10$ dex in [Fe/H]. The rms errors are comparable to the error estimates for *APOGEE* parameters in Mészáros et al. (2013) of $\delta(T_{\rm eff})$ <150 K, $\delta(\log\ g) < 0.2$ dex, and $\delta([{\rm Fe/H}]) < 0.1$ dex in. The typical internal precision on the measured parameters from *The Cannon* is $\delta(T_{\rm eff}) < 5.6$ K, $\delta(\log\ g) < 0.01$ dex, and $\delta([{\rm Fe/H}]) < 0.006$ dex.

The comparison of *The Cannon* with *ASPCAP* showing the bias, rms and formal precision for the labels ($T_{\rm eff}$, log g, [Fe/H]) in six sample fields, with bulge, disk, and halo targeting, is illustrated in Figure 7. As for all stars in the survey with *ASPCAP* labels, these fields show that we reproduce the *ASPCAP*-corrected stellar parameters with typical rms uncertainties of $T_{\rm eff} < 100$ K, log $g < 0.20$ dex, and $[{\rm Fe/H}] < 0.10$. These variances are slightly smaller than expected from our cross-validation leave-one-star-out test. This may be because the median of the stellar labels for the DR10 survey object are near the median labels of the reference objects from the training step. They are not concentrated to the extreme ends of the range, which have a higher weighting in evaluating the test data with cross validation. *The Cannon*'s label transfer also returns values for those ≈15% of stars in DR10 are that must be main-sequence dwarf stars. These are not shown in Figure 7, as *APOGEE* does not report *ASPCAP*-corrected dwarf parameters for DR10. We exclude the stars flagged as bad overall using the STAR BAD flag from *ASPCAP* (this includes stars flagged by *ASPCAP* as having bad $T_{\rm eff}$, bad log g, high χ^2, an effective temperature more than 1000 K from photometric temperature for dereddened color, rotation, S/N low (<50), or if the parameter is near the *ASPCAP* grid edge). Note we can not return parameters for spectral types not included in our training set (see Section 5.5).

In Figure 7 we show the label differences of *The Cannon*—*ASPCAP* for the 1400 stars from the six sample fields as a function of *ASPCAP* $T_{\rm eff}$, log g, and [Fe/H]. There are weak trends; at low $T_{\rm eff} \sim 3700$ K, we find temperatures about 100 K cooler than *APOGEE* and at low log g we find ∼0.15 dex larger log g than *APOGEE*. At the lowest metallicities $[{\rm Fe/H}] < -2.0$, we typically report higher metallicities on the order of 0.05 to 0.3 dex. Figure 8 emphasizes (compared to Figure 7) the slight systematic deviations at the lowest temperatures. These could be addressed by increasing the complexity of the spectral model; however, they occur beyond the temperature range covered by the reference sets, and hence are inherently less robust (and are flagged as such).

We show *The Cannon*'s resulting label distribution in the $T_{\rm eff}$–log g plane from *The Cannon* for the stars in DR10 in Figure 9. This figure shows the result when *ASPCAP*-corrected labels are used for the reference objects in the training step; Figure 10 shows the analogous results but for isochrone-corrected reference labels. There are 35,000 stars in these Figures that remain after excluding stars with the STAR BAD flag set, with velocity scatter >10 km s^{-1} and telluric calibration target set. These figures also show the labels for the 15% stars with log $g > 4$ dex that must be main sequence stars.

In short, for all stars with good *ASPCAP* labels, we find excellent agreement between *The Cannon* and *ASPCAP* by adopting ASCPAP corrected labels in the training step. In addition, we are able to derive plausible parameters for dwarf stars in DR10. However, the $T_{\rm eff}$–log g plane for the stars shows a deviation from the giant branch of the isochrone at low log g (see the right panel of Figure 9). This is a consequence of the input labels of the training spectra.

If instead we use the isochrone-corrected log g labels to fix *The Cannon*'s spectral model (Section 2.2), the results of the label transfer deviate slightly from the *ASPCAP* scale in each of the parameters. However, with these new log g labels, we find a broad giant branch width that is consistent with expectations in $T_{\rm eff}$–log g space given the metallicity of these stars (see the right-hand panel of Figure 10).

This comparison again illustrates both the power of *The Cannon* to transfer labels, but also its dependence on the choice of suitable reference labels.

Currently, no priors are incorporated in *The Cannon* to place the resulting label estimates near physically plausible isochrones. Nonetheless, almost all stars lie in physical spaces on the isochrones as shown in Figures 9 and 10 validates the labels. The labels for the main sequence stars are presumably much more poorly determined, given the limitations of the reference objects in the training step. Remarkably, though, only a handful of stars at low [Fe/H] and low log g do not lie near conceivable isochrones.

At metallicities $[{\rm Fe/H}] < -0.25$ the red clump is offset too high in the log g label. This is noted in Bovy et al. (2014) who estimate that this offset shifts the red clump and red giant branches

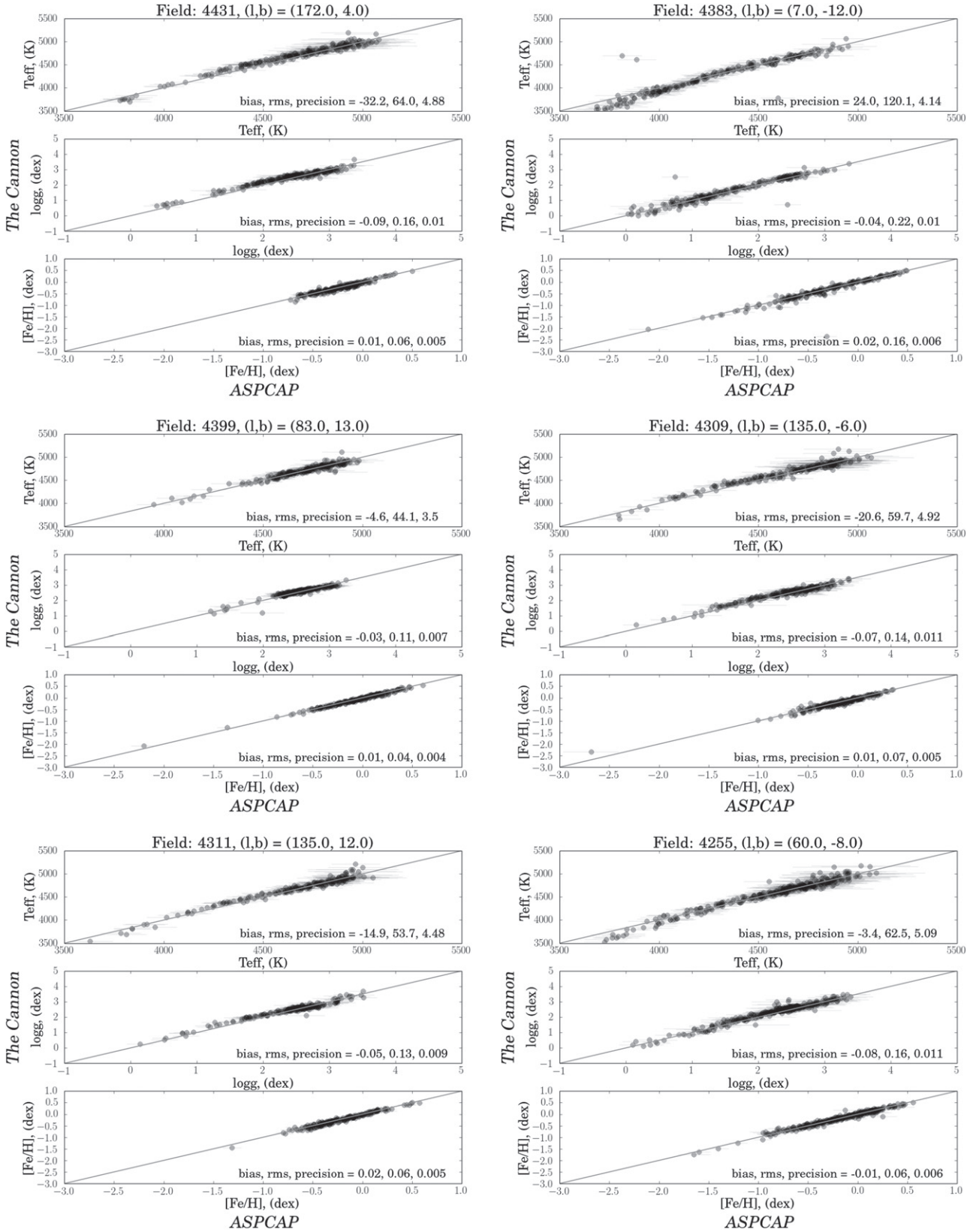

Figure 7. *ASPCAP* DR10 vs. *The Cannon* for six different fields including in the disk, bulge, and halo. The number of stars for each subfigure is 211 (4431), 207 (4384), 217 (4399), 210 (4309), 198 (4311), 319 (4255). Each panel lists the mean difference between the labels (bias), the scatter between the labels (rms), and the formal uncertainly returned by *The Cannon* (precision).

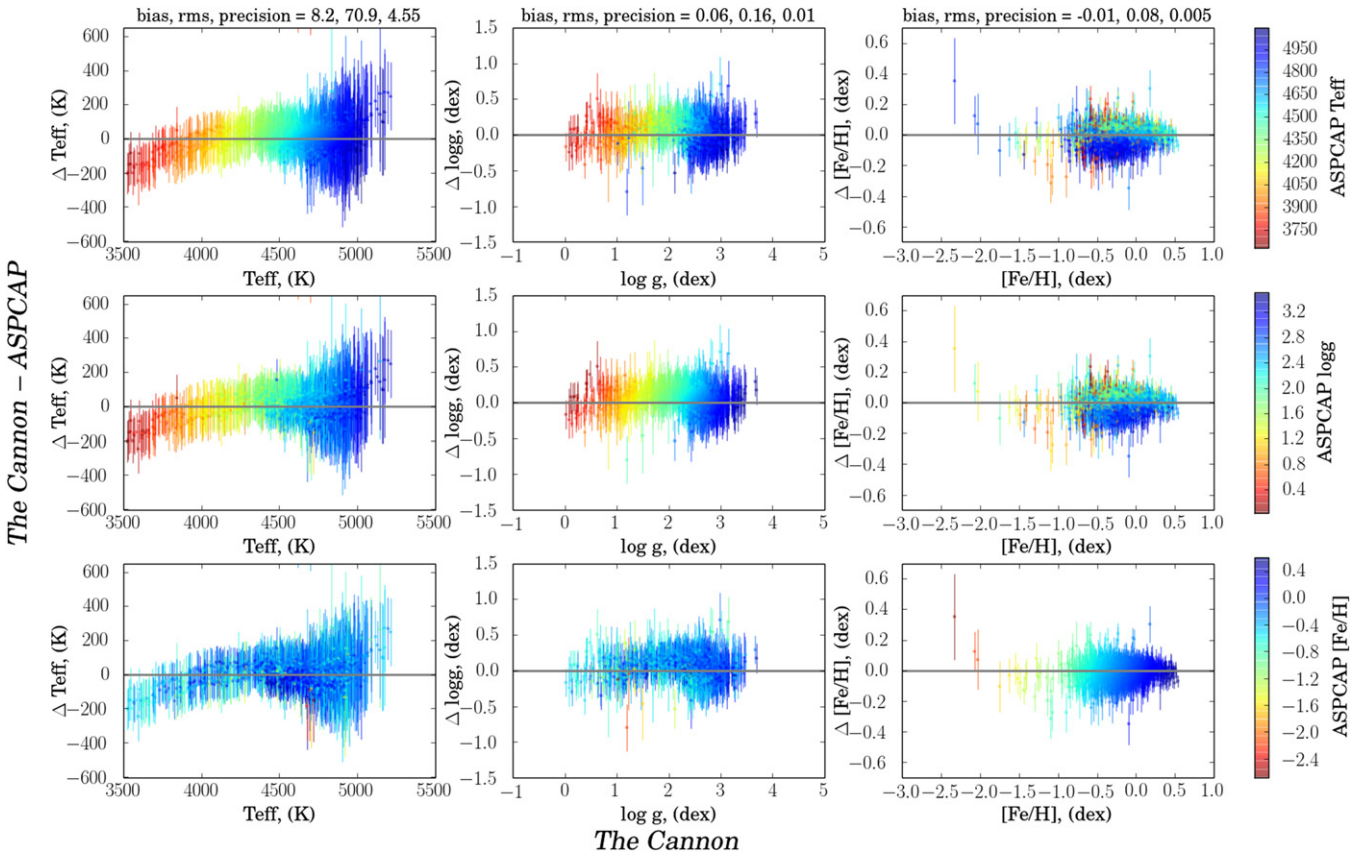

Figure 8. Difference between the labels ($T_{\rm eff}$, log g, and [Fe/H]) derived by *The Cannon* and their *ASPCAP* DR10 values for all the 1400 stars shown in Figure 7. The error bars are dominated by those quoted by *ASPCAP*. There are systematic offsets at the coolest temperatures.

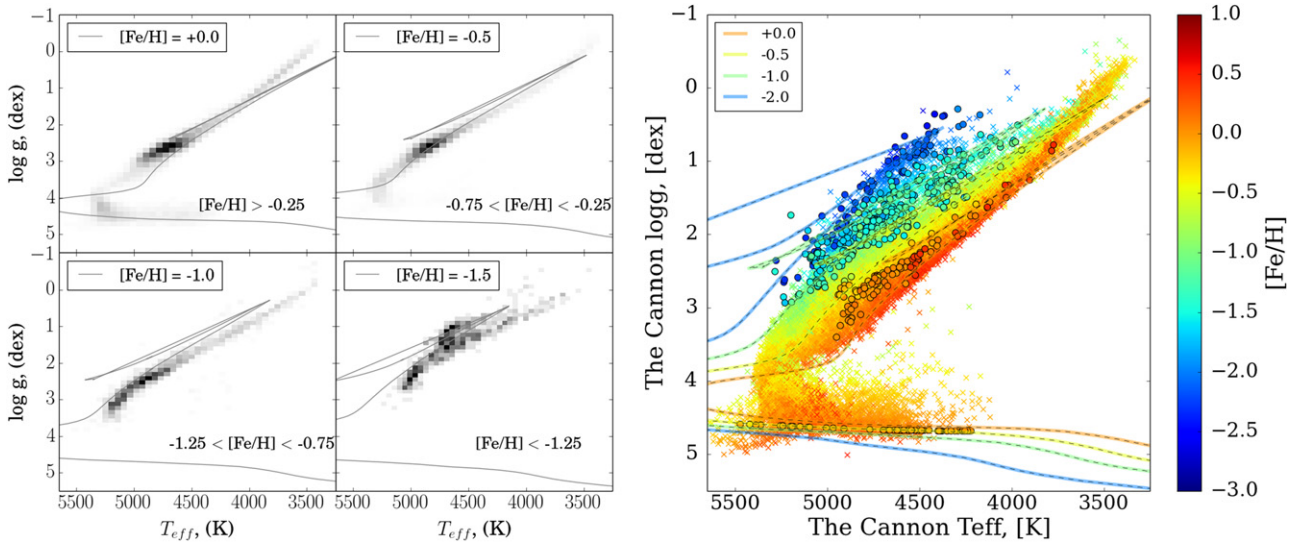

Figure 9. Labels for the ∼35,000 stars from DR10 derived by *The Cannon* based on *ASPCAP*-corrected labels for the set of reference objects. The set of panels on the left shows $T_{\rm eff}$–log g in four metallicity bins. There are ∼19,000, 13,000, 1600, and 1000 stars in the most metal-rich to metal-poor metallicity bins, respectively. The isochrones plotted are 10 Gyr Padova isochrones at the metallicities marked in the upper left hand corners of each sub-panel. The panel on the right shows all stars colored in [Fe/H] on the four isochrones. Note that the log g distribution at low log g is narrow and offset from the giant branch. Reference objects are shown as open circles.

(RGBs) 0.2 dex closer together. Our log g labels in Figure 9 are essentially identical to *APOGEE ASPCAP* labels (offset −0.04 dex in log g for DR10) and the left panels show that the red clump stars which should be seen as a density maxima of stars around log $g \sim 2.5$, $T_{\rm eff} \sim$ are offset to higher log g than the red clump branch of the Padova isochrone, for stars [Fe/H] < −0.25. This offset is on the order of 0.2 dex at [Fe/H] = −0.5 and is present for both *ASPCAP*-corrected labels and isochrone-corrected labels. This may indicate that the *ASPCAP* temperature scale is offset too cool in DR10 (but as a function of [Fe/H]).

So far, we have not yet explored the (possibly systematic) uncertainties of the label estimates, arising from the

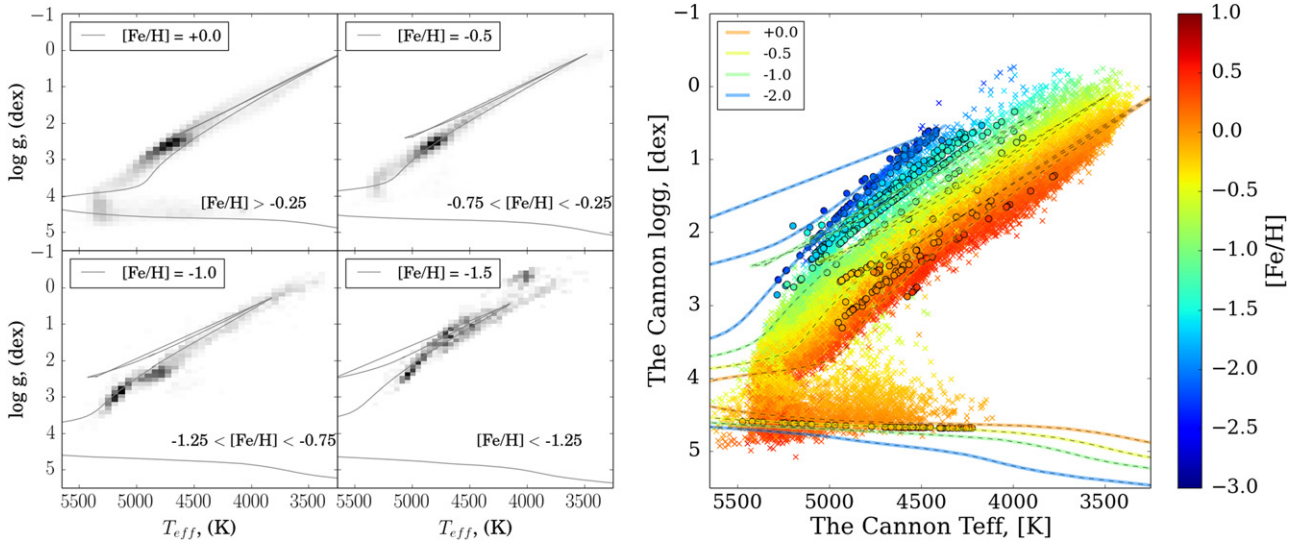

Figure 10. Same as Figure 9 but based on the "isochrone-corrected" labels for the reference objects. In this case, the labels follow the red giant branch on the isochrones. Note that there is nothing in the mathematics of *The Cannon* (Equation (1)) that forces resulting labels to lie in physically plausible locations in label space. This is illustrated by the tiny fraction of objects that lie between the main sequence and the giant branch. That most labels lie in physically sensible portions of the $T_{\rm eff}$–log g plane is a testament to both the quality of the label coverage in the set of reference objects and to the power of *The Cannon* approach. This is all the more remarkable given that there are basically no main sequence stars among the reference objects.Reference objects are shown as open circles.

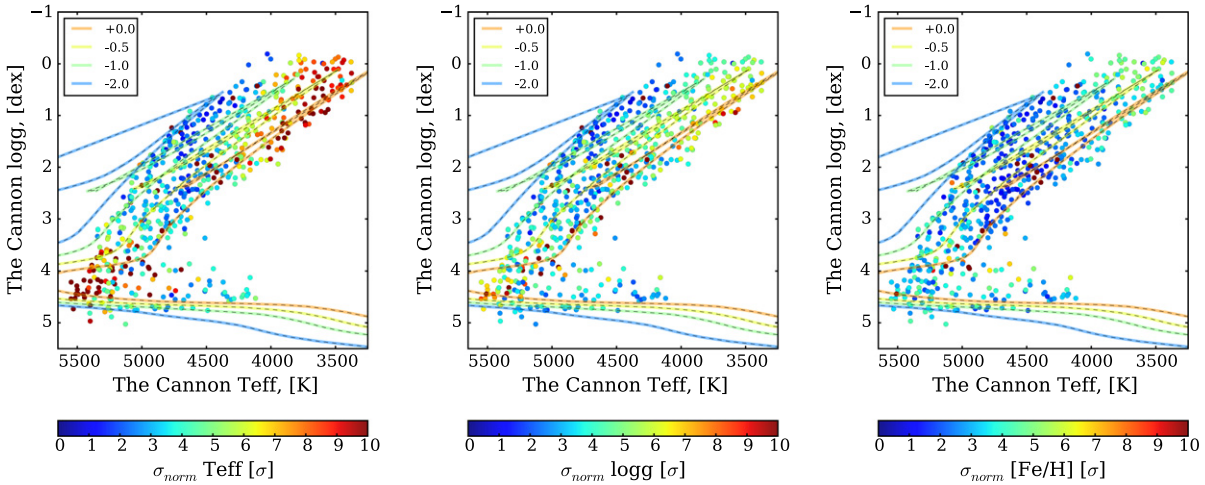

Figure 11. Standard deviation in the labels returned in $T_{\rm eff}$, log g and [Fe/H], shown in the $T_{\rm eff}$–log g plane, normalized by the optimization error on each measurement, for 20 bootstrapping tests of the training set. The representative sample of ∼670 stars shown here has been drawn from an equal sampling of a grid spaced by 100 K in $T_{\rm eff}$, 0.25 dex in log g and 0.25 dex in [Fe/H] from the labels returned using the model trained on the isochrone-corrected reference objects. The location of the reference objects is shown in the gray shaded regions in the panel. Note the narrow region of reference objects also on the main sequence. The highest scatter in the labels is seen for regions where the labels are extrapolated. These figures are shown for the isochrone-corrected labels discussed in Section 2.4.

incomplete, and in part sparse coverage of label-space by the reference objects. Though the labels in Figure 10 lie mostly in physically plausible locations, the fidelity of the labels in particular in the extrapolated part of label space must be scrutinized. To do this, we created twenty different spectral models, by bootstrap-sampling from the set of reference objects and ran the training step on each of them. Using these twenty different spectral models, we derived twenty different label estimates for a sub-set of the survey objects; we picked one survey object in each cell of a three-dimensional grid in $T_{\rm eff}$, log g and [Fe/H], of 100 K, 0.25 dex, and 0.25 dex, respectively. We then took the dispersion among these label estimates as a diagnostic of the uncertainties arising from the sparseness of the training set; this is shown in Figure 11, normalized by the formal error on each label. The location of the reference objects is indicated in the gray shaded regions. These figures show that in the parts of label space well-covered by reference objects, these two uncertainties are comparable. In the extrapolated regions of label space (here, the main sequence turn-off and the metal-rich tip of the RGB), the dispersion among the bootstrap-returned labels is considerably higher than the formal uncertainties; here the incomplete coverage of label space by reference objects becomes the dominant uncertainty.

Both methods rely on extrapolation of labels, *The Cannon* uses the extrapolation of the spectral model, *ASPCAP* performs the extrapolation at the label-inference stage (see Mészáros et al. 2013). Figure 12 compares the stellar labels from *The Cannon* to those from *ASPCAP*. It is only in these extrapolated regions that the labels from *The Cannon* and *ASPCAP* deviate beyond the estimated errors of the *ASPCAP* pipeline. Again, in these regions, neither survey is calibrated to empirical ground truth. These figures also show

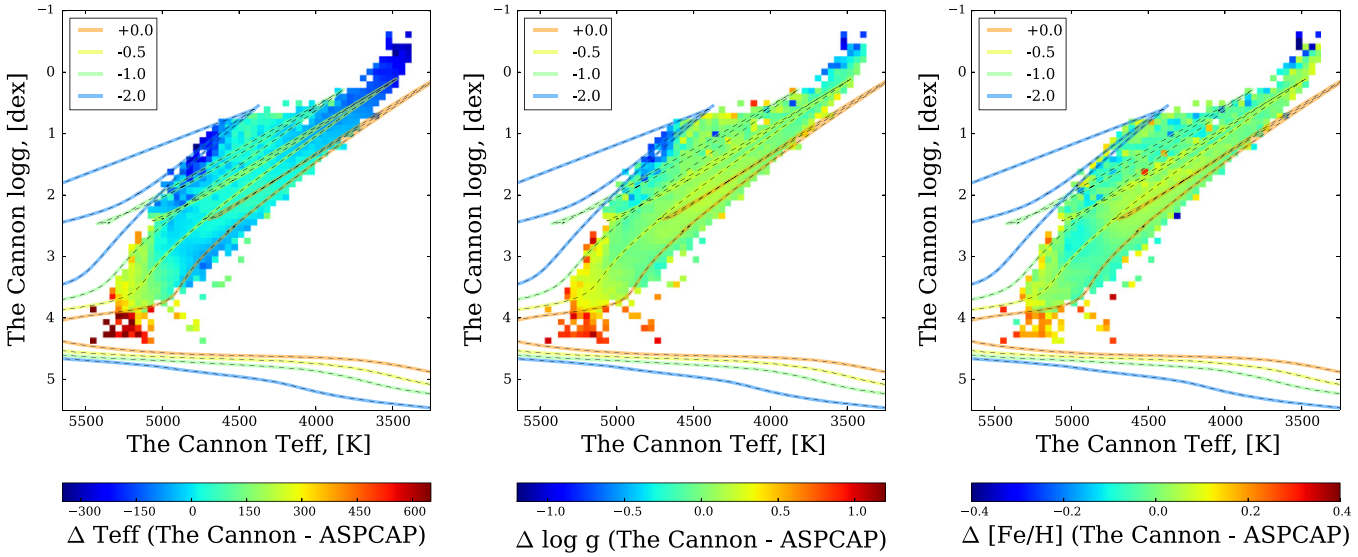

Figure 12. Difference in labels between *The Cannon* and *ASPCAP* indicating the regions of extrapolation where the difference in the labels extends beyond the estimated errors of the methods, due to the limited sampling of the reference objects which does not fully cover the label-space of the survey. The *ASPCAP*-corrected training labels were used to generate the model applied at the test step on the DR10 data.

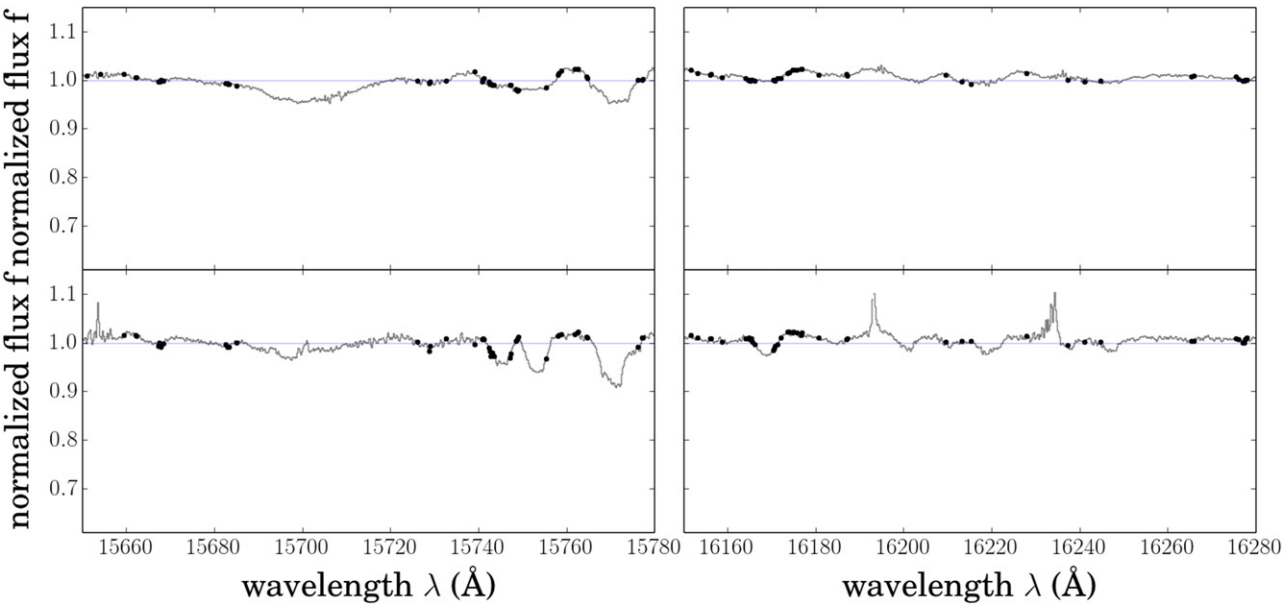

Figure 13. Examples of hot rotating dwarfs in the *APOGEE* DR10 data across regions A and B, comparable to Figure 5. These types of stars are *not* included in our set of reference objects. Therefore, the label transfer by *The Cannon* leads to grossly unphysical label estimates.

that the deviations between *The Cannon* and *ASPCAP* are systematic and not random. In general, these regions of extrapolation will be directly dependent on the survey and the sets of reference objects. Figure 12 highlights the regions of missing label space where stronger constraints on the labels are needed; that is reference objects. In general however, this approach allows the propagation of the uncertainties from imperfect sets of reference objects to the resulting label estimates of the test objects.

5.5. Failures: Types of Spectra not Represented Among the Reference Objects

The dwarf spectra in our reference set only come from the Pleiades cluster, at a single metallicity. This restricted sample limits our ability to determine the stellar parameters for dwarf stars. Given these training data, our model *can* differentiate dwarfs from giants, as long as their spectra are comparable to that of the Pleiades. However, none of the dwarfs in our training set are hot rotating objects with broad line features. Three examples of stars with broad line features that are in the test data but not included in our training data set are shown in Figure 13.

It is possible to differentiate these stars with *The Cannon* because they are output in non-physical space in $T_{\rm eff}$–log g, and present as a group of very metal-poor, [Fe/H] ~ -2.0, low log g stars $\sim$0, with cool temperatures $\sim$4000 K. The metal-poor solution determined by *The Cannon* reflects the dearth of lines in the spectra for these hot stars, given the training model. This group of stars is flagged in *ASPCAP* with a `ROTATION WARNING` flag set. We therefore

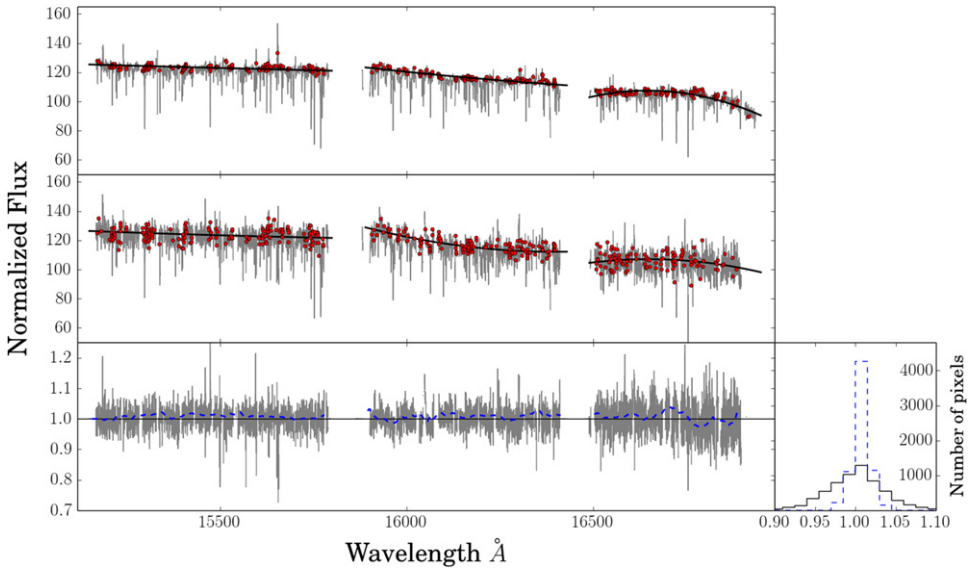

Figure 14. Comparison of the continuum normalization of the same star at high and modest S/N. The *APOGEE apStar* combined visit spectra is shown in the top panel (S/N = 120) and the *apStar* spectra for the fourth visit (S/N = 25) is shown in the second panel. The bottom panel is the ratio of the continuum-normalized spectra of the high and medium S/N spectra and the blue dashed line is a running median of this ratio over 20 Å, showing a small bias. The histogram of this ratios and of its median are given in at the right of the bottom panel.

are able to exclude these stars from our analysis using this condition.

5.6. Performance at modest S/N

By identifying "true" continuum pixels we have been able to implement a simple continuum-normalization that is robust across low and high S/N and that is valid across the parameter range of our training set. To examine how *The Cannon* performs at lower S/N, we have taken individual visits from the *apStar* fits files, when there are ⩾4 visits, and run *The Cannon* on a single visit spectra, when consistently continuum-normalized (Section 5.3). Note, that we have not simply added noise to the combined DR10 spectra for our low S/N tests, which would bypass the question of how consistently the continuum can be defined at different S/N levels. Instead, we have treated single-visit spectra as (formally) independent survey objects. Figure 14 shows a comparison of a sample star for a single visit and combined visits (>4 total visits). Figure 15 presents the results of *The Cannon* compared to *APOGEE* for these stars, showing *only* the *APOGEE* stars with errors of <150 K in $T_{\rm eff}$ and <0.25 dex in log g, across four S/N intervals, from 20 < S/N < 30 to 100 < S/N < 200.

These figures illustrate that our approach to continuum normalization works well for both of these S/N regimes and is S/N independent, which is not true for a weighted-quantile normalization. At the highest S/N (and *APOGEE* estimates a upper noise floor of 200 although stars do measure above this), the rms difference between *The Cannon* and *ASPCAP* is comparable to the *ASPCAP* measurement errors, at 73 K in $T_{\rm eff}$ 0.18 dex in log g and 0.11 dex in [Fe/H]. At a S/N of 30–50, the rms error increases to 100 K, 0.2 dex, and 0.10 dex in $T_{\rm eff}$, log g, and [Fe/H], respectively. At an S/N of 20–30 the rms error is significantly higher and here the internal errors of *The Cannon* become comparable to typical minimization methods and at S/N < 20 exceed them. With this method we can return stellar parameters of $T_{\rm eff}$, log g, [Fe/H] to as good a precisions as minimization techniques ($T_{\rm eff}$ < 100 K, log g <0.2 dex, [Fe/H] <0.1 dex) with an S/N of ⩾25.

6. DISCUSSION

We have demonstrated with *The Cannon* that it is possible to label stellar spectra from extensive homogeneous surveys with stellar parameters and abundances (collectively "stellar labels"), using not physical stellar models but rather a *training set* of reference objects. These reference objects must have trustworthy labels and spectra with the same resolution, line-spread function, and wavelength coverage (though not necessarily S/N) as the data on the survey objects that require labeling. Except for the fact that the reference objects must have been assigned labels themselves somehow, presumably on the basis of physical models for stellar structure and photospheres, we do not rely on explicit stellar photosphere models for the spectra. *The Cannon* is based on the premise that (continuum-normalized) spectra of stars with the same labels look the same, and that spectra vary smoothly with changing labels. This makes it possible to propose a simple mathematical model for the spectrum as a function of the labels, and fix this model in the *training step*, operating on the spectra of the reference objects. In the subsequent *test step* that same model can assign labels (and their uncertainties) to all other objects in the survey.

In a first application of *The Cannon*, on the *APOGEE* DR10 data, we focused on the three most important labels, $T_{\rm eff}$, log g, [Fe/H], and derived them for essentially all of the 55,000 DR10 survey stars, based on a training step that involved only 542 reference objects, i.e., 1% of the survey. Remarkably, *The Cannon*'s label transfer results in stellar parameters and metallicites that are as precise and accurate as those derived from *APOGEE*'s pipeline *ASPCAP*. In addition, *The Cannon* appears to produce—at least in this present implementation—plausible labels for 6000 main sequence stars in *APOGEE*, even though only ∼60 main sequence stars were used in the training step (all of which are members of the

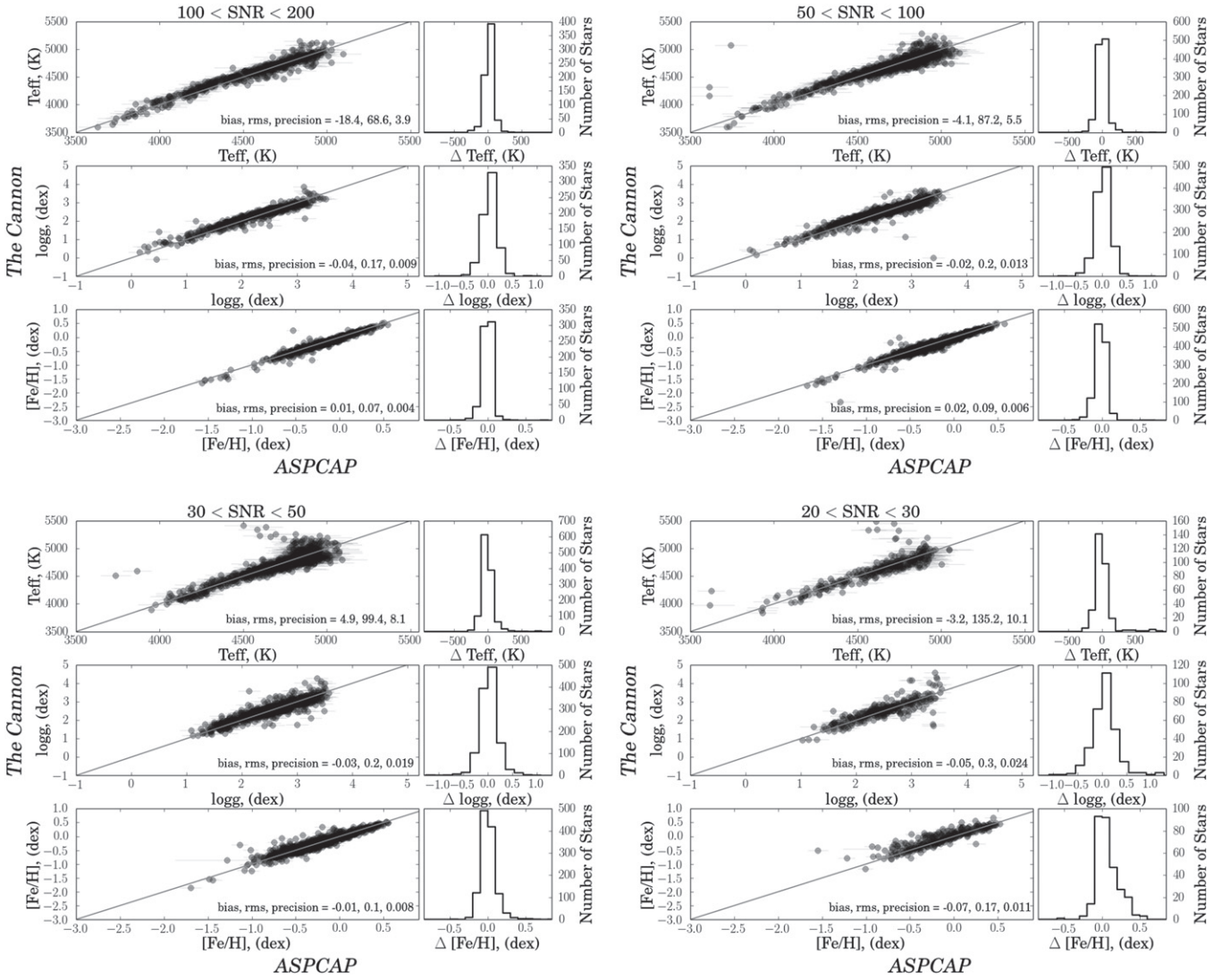

Figure 15. Illustration of *The Cannon*'s ability to estimate labels for spectra of modest S/N. Shown is the comparison of *The Cannon* labels derived for some single visit spectra, compared to the *ASPCAP* label values derived from the co-added high S/N spectra. The single vista spectra are grouped in four different regimes of S/N. There are 60 stars in the 20 < S/N < 30 bin, 1200 stars in the 30 < S/N < 50 bin, 1100 stars in the 50 < S/N < 100 bin and 670 stars in the 100 < S/N < 200 bin. Note that the rms difference between those two label estimates increases more slowly than expected from the S/N of the single visit spectra: label transfer with *The Cannon* therefore enables label estimates at modest S/N. Each S/N regime shows the corresponding histograms of *The Cannon—ASPCAP* for each label, at right.

Hyades cluster). It is also remarkable that the log g–$T_{\rm eff}$ diagram of Figure 10 shows basically no stars outside the physically plausible regime, although *The Cannon* knows nothing here about stellar evolution save the training step.

Our application to *APOGEE* illustrates a number of strengths and practical advantages of such an approach. First, *The Cannon* is computationally very fast. It trains fast and then delivers the labels for 55,000 stars of the *APOGEE* DR10 sample in reasonable time on a single laptop: it takes <0.1 s on a 2.6 GHz intel core i7 to determine three labels for each survey star, without any attempt at code optimization. This is because *The Cannon* only involves linear algebra and (in the test step) the well-behaved optimization of a few parameters with an analytic model for the spectrum.

Second, we have explicitly demonstrated that *The Cannon* can deliver labels, at least these three labels, with nearly the same precision at much lower S/N than commonly deemed necessary. The rms difference between *ASPCAP* labels from spectra with S/N $\geqslant$ 150 and *The Cannon* labels for the same stars from S/N $\sim$ 50 survey spectra is only 30% larger than the *ASPCAP* error bars. *The Cannon* exploits the information at all pixels and certainly the labels $T_{\rm eff}$, log g, [Fe/H] effect many different parts of each spectrum. How this S/N behavior scales to label sets of higher dimension, encompassing, for example, individual abundances, remains to be seen. Part of the reason for this good behavior at low S/N is presumably that *The Cannon* contains a generative model of the intensity or flux density. Given labels, the model provides a Gaussian pdf for the flux density at every wavelength. This pdf is convolved (trivially) with the Gaussian uncertainty assigned to each pixel measurement in the data when the comparison is made between the observed data in the survey object spectra with the generative model, straightforwardly accommodating heteroscedastic uncertainties from spectrum to spectrum.

Third, *The Cannon* requires and provides a continuum estimate that remains unbiased among spectra of different S/N. The training step of *The Cannon* itself identifies the pixels that have near-unity flux in preliminarily normalized

spectra, *and* that show little flux variation with label changes. Those pixels are, conceptually and practically, good approximations to pixels to which to fit a smooth continuum. Our initial application to *APOGEE* spectra indicated that biased continuum fits to spectra would be the main source of poor label estimates from lower S/N spectra using *The Cannon*, and may well also be for label estimates based on physical models.

Our initial application of *The Cannon* to *APOGEE* DR10 data involved a number of important approximations and illustrated several important limitations. We discuss some of these now, along with the benefits and costs of relaxing them. Some of these limitations are attributable to the particular implementation of *The Cannon*, which is just the tip of a large iceberg of potential methods for transferring labels from a set of reference objects to a set of unlabelled survey objects. Other limitations are inherent to the overall approach. The current limitations mainly revolve around the (reference) labels on the one hand, and the choice of the spectral model on the other hand.

Three important issues arise around labels. First, the reference labels are so far assumed to be perfectly known, but in reality are both noisy and potentially biased. In turn, we presume that we have simply no additional information about the labels of the survey objects. Yet, we know *something* about the unlabelled stars (for example, from photometry, and stellar evolution models). Second, any choice for the dimensionality of the label space, 3D in our sample application, will be incomplete in an astrophysical sense. Clearly, stars with identical $T_{\rm eff}$, log g, [Fe/H] may have different spectra, for example, because they differ in [α/Fe] or $v_{\rm rot}$. Third, no set of reference objects will cover the label space comprehensively, especially if one considers high-dimensional label spaces (*APOGEE*'s DR12 published 16 labels per star!).

The general approach to the first and second issues is to expand the scope of the model, which warrants substantive discussion. The model currently only generates spectra by providing a pdf over spectral pixel intensities given a set of labels. Symbolically we could write that *The Cannon* in its current implementation learns or provides a conditional pdf $p(f_\lambda|\ell, \boldsymbol{\theta}_\lambda, s_\lambda^2)$ (see Equation (4)). Given a prior on the label space $p(\ell)$, *The Cannon* could straightforwardly become a generative model of both the spectral pixel intensities *and* the labels.

This would also make it possible to learn the spectral model $\boldsymbol{\theta}$ from reference objects with noisy labels: at the moment, we effectively assume for the reference objects that $p(\boldsymbol{\ell}_n)$ is a delta-function at the known labels. For noisy labels of the reference objects we would set $p(\boldsymbol{\ell}_n)$ instead to reflect the label uncertainties. One would then, however, have to optimize simultaneously $\boldsymbol{\theta}_\lambda$ for *all* λ pixels and the labels $\boldsymbol{\ell}_n$ for *all* n reference objects. Missing labels among some of the reference objects could be treated pragmatically as simply having very large uncertainties.

Thinking of *The Cannon* as a model for both the spectral intensities and the labels also shows how any (much more limited) external information on the labels of the survey objects could be incorporated. One learns from the reference and survey objects simultaneously (effectively, lifting the separation of training and test step), by optimizing $\boldsymbol{\theta}_\lambda$ and $\boldsymbol{\ell}_n$, where the index n now encompasses both the reference and survey sample. The difference between reference and survey objects now simply consists of how tightly constrained their $p(\boldsymbol{\ell}_n)$ is. For survey objects, $p(\boldsymbol{\ell}_n)$ will likely be broad, for example, constraining label-combinations to physically plausible isochrones. This would combine aspects of *The Cannon* with the approach taken by Schönrich & Bergemann (2014).

A generative model of both the spectra *and* the labels would in principle be much more powerful than the current generative model of spectra alone. That said, these joint optimizations of parameters and labels would be expensive and multi-modal, so it might not be computationally tractable.

In this initial implementation of *The Cannon* we restricted ourselves to producing only the maximum-likelihood estimates for both the $\boldsymbol{\theta}_\lambda$ in the training step, and the $\boldsymbol{\ell}_n$ in the test step, with label errors only coming from the inverse covariance matrix at that point. A full inference would be expensive, especially in the test step (labeling the survey objects); the test step model is nonlinear and inference would require sampling or harsh approximations. The correct way to proceed in the full interference case would be in a fully Bayesian framework, where test and training happen at the same time and are not separated out, as is the case now. This is far more computationally expensive than the current approach. Indeed, all straightforward implementations of a joint inference of parameters and labels given noisily labeled training data and unlabeled test data are computationally intractable at present. There are promising approaches that involve either variational inference or Gibbs sampling, but developing either into a practical approach is a significant research project, not just in astrophysics but also in inference.

In the application of *The Cannon* to *APOGEE* we dealt with systematic errors in the reference labels by adjusting them, given the unphysically narrow giant branch returned for DR10 data at low log g. We empirically found by adopting a very naive calibration that shifts the stars to the nearest position on the isochrone from the *ASPCAP* value described in Section 2.2, the stars were returned in a $T_{\rm eff}$–log g space, across metallicity, in line with expectations of the physical label-space of stars. This suggests that there is some problem with the input labels in either the $T_{\rm eff}$ or the log g dimension adjusted from *Kepler* results in DR10.

The issues raised above on the dimensionality and on the coverage of label space by the reference objects, are linked: the basic implementation of *The Cannon* presented here considers only three labels ($T_{\rm eff}$, [Fe/H], and log g), and we know that the label-space has many more dimensions. Conceptually, it is trivial to extend *The Cannon* to even much larger numbers of labels per star. For example, a next generation could include [α/Fe] or [X/Fe] labels for elements X; but also stellar rotation, or photospheric turbulence could be labels, provided suitable sets of reference values exist. The only limitation—and it is a *substantial* limitation—is that as the label-space grows, we presume that the training set must grow to fill it. After all, *The Cannon* can only be as good as its training set. We have shown that the set of 542 reference objects (1% of the survey) does well for three labels. However in general, the training set needs may scale up as badly as exponentially with the dimensionality of the label space. Therefore, it is at this point an open issue, to how many label-dimensions *The Cannon* continues to be useful and practicable. In expanding the list of labeled stars, one option is to identify critical targets for careful labelling, using new (possibly expensive) data and (definitely expensive) human time to obtain good labels, on the same abundance and

stellar-parameter scale as the labels we already have in our limited training set. Heuristically, we want new labeled targets to be in parts of the label space not covered (or poorly covered) by the existing training set. More quantitatively, we could use ideas from experimental design or active learning (Settles 2012) to make optimized choices. Good technology here could permit us to expand the dimensionality of the label space while growing the size of the training set as minimally and as objectively as possible.

The Cannon is related in a number of ways to supervised methods within the domain of machine learning. On the one hand, *The Cannon* is a pure supervised classification method: it is trained on data and labels from a population that is assumed to have perfect labels and it is applied to data for which labels are assumed to be unknown. However, *The Cannon* is also very unlike standard machine-learning methods in a critical respect: It makes no assumption that the test data and the training data are statistically similar, or drawn from the same noise distribution. Indeed, the main reason that the method is written as a generative probabilistic model is precisely so that it can account for the changing noise model (changing noise variances) from object to object and pixel to pixel. Standard supervised methods from the machine-learning literature (such as Random Forest, Breiman 2001, Deep Learning (e.g., LeCun et al. 1989; Bengio 2009; Schmidhuber 2015), and Kernel Support Vector Machines (Smola & Schlkopf 2004), do not have the property that they can account for variable noise models. These traditional machine-learning methods perform very badly as the training data become different from the test data (as they do in our S/N experiments in Section 5.6).

In this sense, *The Cannon* is less like a standard machine-learning method and more like one of the new methods being developed to account for differences between the training set and the test set. In some sense, *The Cannon* is really a Transfer Learning method (e.g., Pan & Yang 2010) because it learns on data with one noise model (or many noise models) and then is employed on data drawn from a new set of noise models. In the future, as *The Cannon* is understood and developed further, we expect there to be enhancements that benefit from new developments in machine learning. For example, the fact that the test set might (or does) span a different part of label space than the training data might be accommodated by ideas from Concept Drift (e.g., Widmer & Kubat 1996) or Model Adaptation (e.g., Duan et al. 2012), both of which are being developed precisely to account for the problem that in many real-world applications of machine learning.

An important aspect of *The Cannon* as presented here, concerns the spectral model itself. With the pixel-by-pixel polynomial ansatz for the spectral model $\boldsymbol{\theta}_\lambda$ we engender two important consequences. First, we need to pick a functional form for the spectral model (Equation (1)), which we took empirically to be a quadratic-in-labels form of Equation (6). We arrived at that choice by empirical experimentation with this particular data set, but this choice can be generalized. Indeed, the polynomial family is probably not the best family of functions to be exploring, since they extrapolate badly (edge effects) and require explicit, qualitative choices about order and cross-terms. It is probably better to eventually move to a non-parametric form for the functions, such as Gaussian Processes. In this case, model complexity would be controlled by continuous parameters and the functional form could become very complex at the pixels where the data in the training step warrant it. This would be a natural extension of what has been implemented here.

Second, our current ansatz for the spectral model treats all pixels independently, which they plausibly are only in their noise properties. This approximation was made to make the system fast; training (learning) can take place at each wavelength independently and (in principle) in parallel. However, it is not a good approximation for many reasons. One of these is that the finite resolution of the spectrograph correlates nearby pixels; the generative spectral model cannot vary substantially over wavelength differences that are far smaller than the spectrograph resolution. This point of prior information is not used at all in the model.

A much more complex imperfection of the independent-pixel assumption is that there are multiple lines from the same element and same ionization state. These are expected to be co-variant in any sensible model. We do not make any use of such information; indeed no line list enters *The Cannon* at any stage. These decisions were made for good, pragmatic computational reasons. A better model would permit itself to know about the spectrograph resolution and either know about or discover sets of lines that vary together. However, any such generalization will come at substantial computational cost.

Both the application to *APOGEE* data and the possibilities to apply *The Cannon* in a broader context, bring the question of suitable sets of reference objects into focus. Indeed, in the long run the biggest practical problem in applying *The Cannon* may not linked to the mathematics of the method *per se*, but to the actual availability of sufficiently many and sufficiently diverse reference objects in the survey to cover label space in the training step. The most glaring issue in the *APOGEE* DR10 case, even with only three labels per star, is that fact that all main sequence stars in the reference set of objects come from only one cluster, without any range in metallicity. With, for example, the DR12 of *APOGEE*, training sets of much higher dimensionality are becoming available (especially [X/H] of individual elements). While this prospect is exciting, it will exacerbate both the question of how to make labels space coverage sufficient for the training step, and how to assert the accuracy of the training labels in the first place.

Fortunately, there are in principle quite a number of options for picking reference objects. One could pick a subset of survey objects (sensibly covering label space) where the spectra have exceptionally high S/N, lending particular credence to the labels derived from physics-based models. On these, one would train the spectral model and then transfer labels to the remaining survey objects, effectively deriving most of the survey labels from the observations of highest S/N. Alternatively, as we did here, one can choose reference objects where special circumstance (cluster membership, astroseismological information) lend particular credence to their labels.

As labels are a property of the star, not of the data set at hand, they can come from completely independent sources of information. Even if we understood absolutely nothing about near-IR spectra, but had labels for the 542 reference objects from optical spectroscopy, we could have derived the labels for the *APOGEE* DR10 stars as well as *ASPCAP*. This leads to perhaps the most exciting long-term prospect of *The Cannon*: bringing qualitatively different stellar surveys—surveys that use different instruments, working in different wavelength regions at different resolutions and S/Ns—onto a consistent stellar parameter and chemical abundance scale. So long as

different surveys can agree on benchmark stars and best values for the stellar labels, and so long as those training sets are large enough and span enough of the label space, *The Cannon* (or a future upgrade that implements some of the ideas in this section) can be used to ensure that all of the surveys are delivering stellar parameters on the same system. *The Cannon* will not make the data coming from any survey more *accurate*, but it might serve to make the whole industry of stellar parameter estimation and element abundance tagging more *precise* and consistent.

This prospect of survey self-labeling (for example, from high S/N to low S/N) and the prospect of cross-survey calibration brings even more urgency to assuring that sufficient calibration observations are in place and that the different major spectroscopic surveys have sufficient sample overlap.

We would like to thank Daniel Foreman-Mackey (NYU), Morgan Fouesneau (MPIA), Jon Holtzman (NMSU), Keivan Stassun (Vanderbilt University), Jennifer Johnson (OSU), and David Sontag (NYU) for valuable discussions. D.W.H. was partially supported by the NSF (grant IIS-1124794), NASA (grant NNX08AJ48G), and the Moore–Sloan Data Science Environment at NYU. The research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP 7) ERC Grant Agreement No. [321035]. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

REFERENCES

Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2014, ApJS, 211, 17
Allende Prieto, C., García López, R. J., Lambert, D. L., & Gustafsson, B. 1999, ApJ, 527, 879
Bailer-Jones, C. A. L., Andrae, R., Arcay, B., et al. 2013, A&A, 559, A74
Barrado y Navascués, D., Deliyannis, C. P., & Stauffer, J. R. 2001, ApJ, 549, 452
Beers, T. C., Lee, Y., Sivarani, T., et al. 2006, MmSAI, 77, 1171
Bengio, Y. 2009, in Foundations and Trends in Machine Learning Vol. 2, (Now Publishers), 1
Boeche, C., Siebert, A., Williams, M., et al. 2011, AJ, 142, 193
Bovy, J., Nidever, D. L., Rix, H.-W., et al. 2014, ApJ, 790, 127
Breiman, L. 2001, Mach. Learn., 45, 5
Duan, L., Tsang, I. W., & Xu, D. 2012, ITPAM, 34, 465
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Freeman, K. C. 2012, in ASP Conf. Ser. 458, Galactic Archaeology: Near-Field Cosmology and the Formation of the Milky Way, ed. W. Aoki et al. (San Francisco, CA: ASP), 393
Gilmore, G., Randich, S., Asplund, M., et al. 2012, Msngr, 147, 25
Girardi, L., Bressan, A., Bertelli, G., & Chiosi, C. 2000, A&AS, 141, 371
Gonzalez, O. A., Zoccali, M., Monaco, L., et al. 2009, A&A, 508, 289
Hinkel, N. R., Timmes, F. X., Young, P. A., Pagano, M. D., & Turnbull, M. C. 2014, AJ, 148, 54
Jofré, P., Heiter, U., Soubiran, C., et al. 2014, A&A, 564, A133
Koleva, M., Prugniel, P., Bouchard, A., & Wu, Y. 2009, A&A, 501, 1269
Kordopatis, G., Gilmore, G., Steinmetz, M., et al. 2013, AJ, 146, 134
LeCun, Y., Boser, B., Denker, J. S., et al. 1989, Neural Computation, 1, 541
Lee, Y. S., Beers, T. C., Sivarani, T., et al. 2006, BAAS, 38, 168.15
Liu, C., Deng, L.-C., Carlin, J. L., et al. 2014, ApJ, 790, 110
Majewski, S. R. 2012, BAAS Abstracts, 219, 205.06
Mészáros, S., Holtzman, J., García Pérez, A. E., et al. 2013, AJ, 146, 133
Newberg, H. J., Carlin, J. L., Chen, L., et al. 2012, in ASP Conf. Ser. 458, Galactic Archaeology: Near-Field Cosmology and the Formation of the Milky Way, ed. W. Aoki et al. (San Francisco, CA: ASP), 405
Pan, S. J., & Yang, Q. 2010, IEEE Transactions on Knowledge and Data Engineering, 22, 1345
Prugniel, P., Vauglin, I., & Koleva, M. 2011, A&A, 531, A165
Re Fiorentin, P., Bailer-Jones, C. A. L., Lee, Y. S., et al. 2007, A&A, 467, 1373
Recio-Blanco, A., Bijaoui, A., & de Laverny, P. 2006, MNRAS, 370, 141
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Schmidhuber, J. 2015, NN, 61, 85
Schönrich, R., & Bergemann, M. 2014, MNRAS, 443, 698
Settles, B. 2012, Synthesis Lectures on Artificial Intelligence and Machine Learning (San Rafael, CA: Morgan & Claypool Publishers), 1
Smiljanic, R., Korn, A. J., Bergemann, M., et al. 2014, A&A, 570, A122
Smola, A. J., & Schlkopf, B. 2004, Statistics and Computing, 3, 199
Soubiran, C., Katz, D., & Cayrel, R. 2011, A&A, 525, A71
Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645
Widmer, G., & Kubat, M. 1996, Machine Learning, 23, 69
Wu, Y., Singh, H. P., Prugniel, P., Gupta, R., & Koleva, M. 1998, A&AS, 133, 221
Zasowski, G., Johnson, J. A., Frinchaboy, P. M., et al. 2013, AJ, 146, 81