CrossMark

The Payne: Self-consistent ab initio Fitting of Stellar Spectra

Yuan-Sen Ting (丁源森)[1,2,3,4,5,7] , Charlie Conroy[5] , Hans-Walter Rix[6] , and Phillip Cargile[5]

[1] Institute for Advanced Study, Princeton, NJ 08540, USA
[2] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA
[3] Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101, USA
[4] Research School of Astronomy and Astrophysics, Australian National University, Cotter Road, ACT 2611, Canberra, Australia
[5] Harvard–Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[6] Max Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany

Abstract

We present *The Payne*, a general method for the precise and simultaneous determination of numerous stellar labels from observed spectra, based on fitting physical spectral models. *The Payne* combines a number of important methodological aspects: it exploits the information from much of the available spectral range; it fits all labels (stellar parameters and elemental abundances) simultaneously; it uses spectral models, where the structure of the atmosphere and the radiative transport are consistently calculated to reflect the stellar labels. At its core *The Payne* has an approach to accurate and precise interpolation and prediction of the spectrum in high-dimensional label space that is flexible and robust, yet based on only a moderate number of ab initio models ($\mathcal{O}(1000)$ for 25 labels). With a simple neural-net-like functional form and a suitable choice of training labels, this interpolation yields a spectral flux prediction good to 10^{-3} rms across a wide range of $T_{\rm eff}$ and $\log g$ (including dwarfs and giants). We illustrate the power of this approach by applying it to the APOGEE DR14 data set, drawing on Kurucz models with recently improved line lists: without recalibration, we obtain physically sensible stellar parameters as well as 15 elemental abundances that appear to be more precise than the published APOGEE DR14 values. In short, *The Payne* is an approach that for the first time combines all these key ingredients, necessary for progress toward optimal modeling of survey spectra; and it leads to both precise and accurate estimates of stellar labels, based on physical models and without "recalibration." Both the codes and catalog are made publicly available online.

Key words: methods: data analysis – stars: abundances – techniques: spectroscopic

Supporting material: machine-readable table

1. Introduction

Large-scale multiplexing spectroscopic surveys are revolutionizing the quality and quantity of spectroscopic data for Galactic archaeology. Surveys such as APOGEE (Majewski et al. 2017), GALAH (De Silva et al. 2015), and *Gaia*–ESO (Smiljanic et al. 2014) are collecting high-quality spectra for 10^5–10^6 stars with a spectral resolution $R \simeq 25,000$, orders of magnitudes more stars than previous samples. Lower-resolution spectroscopic surveys, e.g., RAVE (Steinmetz et al. 2006), *Gaia*–RVS (Recio-Blanco et al. 2016), and LAMOST (Luo et al. 2015), are collecting even larger samples. And upcoming spectroscopic surveys, such as DESI (DESI Collaboration et al. 2016), 4MOST (de Jong et al. 2014), WEAVE (Dalton et al. 2016), MOONS (Cirasuolo et al. 2014), and SDSS-V (Kollmeier et al. 2017), will boost sample sizes at both high and low spectral resolution by another order of magnitude, toward $\sim 10^7$ stars.

However, learning about Galactic archaeology and stellar physics from these spectra depends crucially on our ability to correctly and precisely infer numerous stellar labels from these spectra: stellar parameters and individual elemental abundances. This requires a rigorous method to extract the maximal information from these data, based on physical ab initio spectral models. This is the focus of this study.

A key to rigorous fitting of stellar spectra is the ability to fit all stellar labels (typically >20–50 for stellar spectra) simultaneously (Rix et al. 2016; Ting et al. 2016), principally for two reasons: the spectral features of many elements are

blended in the spectrum, imprinting a covariant signature on the data. And for quite a number of elements, variations in their abundances not only affect the strength of their spectral features, but also alter the stellar atmospheric structure (Ting et al. 2016); this in turn affects the spectral features of other elements, especially in cooler stars. Therefore, spectral modeling should be based on self-consistently calculated models that take into account the dependence of the structure of the atmosphere on various elemental abundances. This dependence is widely implemented for changes in [Fe/H], but not other elements.

In practice, current spectral analyses often fit only small portions of the spectrum to determine any particular elemental abundance, holding the abundances of other elements fixed. And they often require subsequent recalibration of the basic stellar parameters (e.g., $\log g$ and $T_{\rm eff}$) or abundance–$T_{\rm eff}$ trends inferred from the spectral fitting. This motivates the need for the development of a comprehensive approach to study these issues. Here we will present such a method, *The Payne* [8] in this study.

The Payne combines a number of important ingredients: a set of spectral models based on a state-of-the-art line list (P. Cargile et al. 2019, in preparation); models computed that are self-consistently calculated for each set of labels; a robust and flexible "interpolator" in the high-dimensional label space for spectral fitting that can precisely predict spectral model fluxes for arbitrary sets of labels; a well-defined and objective

---

[7] Hubble Fellow.

[8] In appreciation of Cecilia Payne-Gaposchkin's ground-breaking work on physical spectral models.
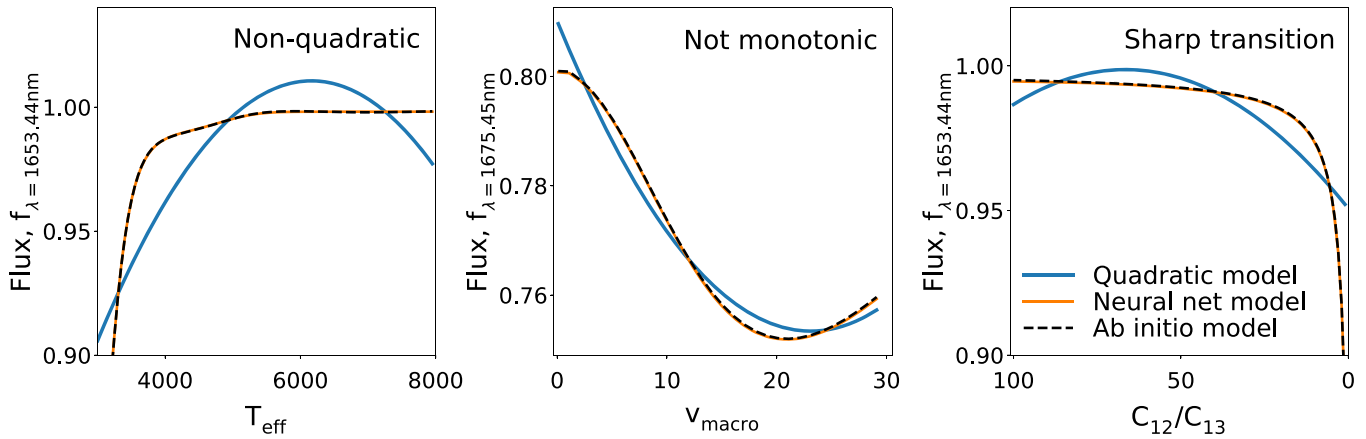
Figure 1. High-fidelity spectral flux interpolation and prediction enabled by *The Payne* compared to a quadratic flux model. In each of the three panels, the dashed line shows the expected flux variation of individual pixels with different label variations. *The Payne* approximates these flux variations through neural networks (orange line), which are more flexible than quadratic models (blue line) in approximating the flux variation across a wide label range, without incurring much additional cost in ab initio model calculation for fitting a spectrum. The three panels show three different scenarios where quadratic models do not approximate the flux well. (a) A spectral region with strong sensitivity to $T_{\rm eff}$. (b) A multiply non-monotonic variation of flux, e.g., with $v_{\rm macro}$. (c) A transition where a label changes abruptly and has little effect below a threshold, e.g., with C_{12}/C_{13}. In contrast, *The Payne* (solid orange line) has no problem approximating these variations.

assessment and mitigation of the wavelength regions where the models have important systematic shortcomings; and a robust estimate of the label estimates from the entire remaining parts of the observed spectra. For modeling stellar spectra, *The Payne* is a fully automated, simple, transparent fitting machinery, given a set of ab initio synthetic spectral models. The codes for running *The Payne* are publicly available on GitHub[9] under an MIT License and version 1.0 is archived in Zenodo (Ting 2019). Moreover, the fitting is very efficient— e.g., fitting 25 labels for an APOGEE spectrum with *The Payne* takes less than one CPU second. *The Payne* differs from *The Cannon* (Ness et al. 2015; Casey et al. 2016) principally in two respects: it is based on physical instead of data-driven models, and it generalizes the "interpolator" beyond the quadratic polynomial implemented in *The Cannon* and Rix et al. (2016). In short, *The Payne* is an approach that for the first time combines all these ingredients, necessary for progress toward optimal modeling of survey spectra; and it leads to both precise and accurate estimates of stellar labels, based on physical models and without "recalibration."

This paper is structured as follows: we introduce *The Payne* and test the interpolator at its core in Section 2. We apply *The Payne* to the APOGEE DR14 data set in Section 3, and present the resulting catalog. We discuss the outlook for stellar spectroscopy in the light of *The Payne* in Section 4 and conclude in Section 5.

2. The Payne

2.1. Motivation

Current approaches to modeling stellar spectra, with either physical or data-driven models, have important limitations that are well documented in the recent literature (Boeche et al. 2011; Adibekyan et al. 2012; Bensby et al. 2014; Blanco-Cuaresma et al. 2014; Nissen et al. 2014; Holtzman et al. 2015; Ness et al. 2015; Boeche & Grebel 2016; Casey et al. 2016; García Pérez et al. 2016; Rix et al. 2016; Ting et al. 2016, 2017a, 2017b; Zhao et al. 2016; El-Badry et al. 2018a, 2018b). In this section we present our approach to addressing

some of these limitations.[10] At the core of *The Payne* is the ability to perform full simultaneous spectral fitting of all stellar labels through an efficient but precise way of "interpolating" a modest set of synthetic model spectra in high-dimensional label space.

The key idea for efficiently interpolating an ensemble of synthetic models is twofold. First, we do not need to create a high-dimensional "grid" of model spectra, which would be computationally prohibitive for, say, 25 labels in this study; with an adaptive approach described below we only create models within the label space spanned by the data and "where needed." Second, we resort to building a generative model for the spectra at arbitrary points (in a portion) of label space, as in Ness et al. (2015) and Rix et al. (2016). If the model for the spectral flux at each pixel is forced to be a quadratic function of the N labels, then only a few times $N \times (N+1)/2$ ab initio spectral models are needed as a basis.

While quadratic models are simple and elegant, they limit the portion of label space over which precise ($\sim 10^{-3}$) flux predictions are possible. For fitting a broad range of stellar labels (e.g., fitting dwarfs and giants or 3000 K $\leqslant T_{\rm eff} \leqslant$ 8000 K simultaneously), quadratic flux models appear too restrictive. Furthermore, for stellar labels such as $v_{\rm macro}$, the variation of flux at any given pixel can be more complicated and often not monotonic. Such complex label dependences of the flux are illustrated by three one-dimensional examples in Figure 1. In this figure we show the continuum-normalized flux as a function of $T_{\rm eff}$, $v_{\rm macro}$, and C_{12}/C_{13}. Here we assume the same Kurucz synthetic models as we will describe in Section 2.4 convolved with the APOGEE averaged line spread function (LSF) to simulate the variations we expect from APOGEE. Clearly a quadratic model cannot capture the behavior of the flux over the entire parameter range, while a more flexible neural network can reproduce the variation in the model very well, as we quantify in greater detail below.

---

[10] For example, to fully harness the information from spectra, a full spectral fitting method can be more advantageous (see the detailed discussion in Ting et al. 2016, 2017a) than methods based on equivalent width, because much of the spectral information is embedded in the subtle blended features.

2.2. Neural Networks for Precise Prediction of a Model Spectrum

The interpolation and approximation of functions with neural networks is based on the idea that most functions can be approximated by a finite composite of simple functions. For *The Payne* we consider an extremely simple neural network architecture, "fully connected" with only two hidden layers. At each wavelength pixel λ, we posit that the flux as a function of stellar label ℓ, can be written as

$$f_\lambda = w \cdot \sigma(\tilde{w}^i_\lambda \sigma(w^k_{\lambda i} \ell_k + b_{\lambda i}) + \tilde{b}) + \bar{f}_\lambda, \qquad (1)$$

where σ is the Sigmoid function $\sigma(x) = 1/(1 + e^{-x})$, and where we have assumed the Einstein convention for index summation. In the training step, we seek the coefficients $(w, \tilde{w}^i_\lambda, \tilde{b}, w^k_{\lambda i}, b_{\lambda i}, \bar{f}_\lambda)$ that best approximate the training spectra as a function of their stellar labels in the least-squares sense. This formalism can be viewed as a straightforward extension of the quadratic flux models (e.g., Ness et al. 2015; Rix et al. 2016), which essentially adopts the flux model $w^i \ell_i + w^{jk} \ell_j \ell_k$. In *The Payne* we consider the nonlinear composite function $w \cdot \sigma(\tilde{w}^i_\lambda \sigma(w^k_{\lambda i} \ell_k + b_{\lambda i}) + \tilde{b})$ to be the more flexible expansion terms. The rationale for adopting a more flexible functional form is similar to taking a higher "expansion" order such that the "Taylor" expansion convergence sphere encompasses a larger region of the parameter space of interest (see Ting et al. 2016, for a more detailed discussion).

The number of "neurons" i in Equation (1) is a free hyperparameter to be optimized. Increasing the number of neurons enables the approximation a more complicated function, but at the risk of overfitting the function. Besides adopting a greater number of neurons, one can also increase the complexity of the neural networks by increasing the number of "layers" by considering the composite of the current composite functions, i.e., $f_\lambda \sim \sigma(\cdots \sigma(\cdots \sigma(\cdots)))$.

Cross-validation experiments described below motivate the following choices. We adopt a model with 2 hidden layers and 10 neurons for each hidden layer. This choice was initially motivated by the fact that the number of free coefficients in this simple neural network model is comparable to that in a quadratic model. At least for stellar spectra, designing the neural networks to have roughly the same number of coefficients of simple polynomials seems to be a robust practical guideline. We checked that adopting a significantly more complex neural network model does not improve the qualitative results of this study, but does lead to overfitting. We train the neural networks by minimizing the L^2 loss, i.e., minimizing the sum of the Euclidean distances between the target (ab initio flux and the model-predicted (or "interpolated") flux at each pixel. We found no need for further imposing explicit L^1 regularization (e.g., Casey et al. 2016) to the networks, as it does not improve the results presented in this study. We limit ourselves to small networks precisely to avoid overfitting, thus regularization is not necessary.

Neural networks are of course not the only flexible model "interpolators;" Gaussian processes or support vector regressions are also employed in related circumstances. For the case at hand, *The Payne* has the advantage of being much faster computationally. While it is more computationally expensive to train neural networks than the quadratic models (each wavelength pixel takes about five CPU minutes), once the neural networks are trained, the speed of inference is about the same as in the quadratic models, and is independent of the size of the training set, because we simply need to evaluate the composite functions. While Gaussian processes are powerful for full Bayesian inferences, predicting a model spectrum at a new label point through Gaussian processes can be extremely slow: it requires the inversion of a matrix, has a complexity of $\mathcal{O}(N^3_{\text{train}})$, and can be very memory-intensive.

Finally, the fundamental idea of *The Payne* is different from some of other previous applications of neural networks in spectral analyses (Fabbro et al. 2018; Leung & Bovy 2019). These studies attempted to map a spectrum to the stellar label through neural networks, but in this study, we are mapping a set of stellar labels to the spectrum. Summarizing the detailed pros and cons of these methods is beyond the scope of this study; here we will only briefly discuss the logic behind our choice. Direct mapping from spectrum to stellar label can be advantageous because the spectral fitting component becomes trivial—evaluating stellar labels in this case only requires evaluating the mapping/function directly, which is extremely fast. On the other hand, mapping f: spectrum $\rightarrow$ label limits the ability to differentiate the function with respect to the label, unlike *The Payne*, which has f: label $\rightarrow$ spectrum. Differentiating the emulating function with respect to label can be useful in many cases—especially at low resolution, comparing $\partial f / \partial$(label) to theoretical line lists can be the key to enforcing that elemental abundances are derived from their corresponding absorption features instead of astrophysical correlations. It also allows us to impose a theoretical prior as was done in Ting et al. (2017b) (but see Leung & Bovy 2019). This reason prompted our choice to map from stellar label to spectrum (see also Dafonte et al. 2016). The downside of this approach, however, is that evaluating the label requires least-squares minimization, which is slower than simply evaluating a function. In short, both types of mapping have their own merits, and the choice of which method to use clearly depends on the applications.

2.3. The Choice of Stellar Training Labels for Building a Spectral Model

Beyond the choice of how to interpolate among a set of model grid points, another essential choice must be made: the size of the training set and the stellar labels at which the ab initio models are to be evaluated to provide training spectra. Formally, we require barely more training spectra than the number of free parameters in the neural networks, which would be 273 training spectra in the case at hand. However, uniformly distributing a few hundred training labels in a high-dimensional ($N_{\text{dim}} = 25$) space would not be optimal because the distribution would be too sparse in the label space, and the interpolation would not be precise. But in generative models such as *The Payne* we need not draw from regular, uniformly spaced training labels.

As discussed in Ting et al. (2015), generating training spectra around the label space that real observed stars are expected to occupy can exponentially reduce the number of models needed. The volume of a hyperellipsoid in a high-dimensional space is exponentially smaller than the volume of a hypercube where the training labels are uniformly distributed. In our illustrative application of *The Payne*, we fit 25 stellar labels, including all elemental abundances with entries in our line list within the APOGEE spectral range. As stellar parameters, we fit T_{eff}, $\log g$, v_{micro}, v_{macro}, and C_{12}/C_{13} along with the 20 elemental abundances (C, N, O, Na, Mg, Al, Si, P,

Table 1
Sampling Scheme for the Model Grid in This Study

Stellar label	Sparse Grid	Refined Grid
n_{grid}	2000	The same
T_{eff}, $\log g$	Draw uniformly from the MIST isochrone convex hull with $T_{eff} \in [3000, 8000]$ K, $\tau_{age} \in [3, 10]$ Gyr	The same
[Fe/H]	Draw uniformly, [Fe/H] $\in [-1.5, 0.5]$	The same
v_{micro}	Draw uniformly, $v_{micro} \in [0.1, 3]$ km s^{-1}	The same
v_{macro}	Draw uniformly, $v_{macro} \in [0, 30]$ km s^{-1}	The same
C_{12}/C_{13}	$f_{MIST}(T_{eff}, \log g, [Fe/H])$ + a scatter of ± 35	The same
[X/Fe]	Draw uniformly, [X/Fe] $\in [-0.5, 0.5]$	$P([X/H]|[Fe/H])$ from the APOGEE fits using the sparse grid, with Δ[Fe/H] $= 0.2$ dex

S, K, Ca, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Ge). We consider a training set of 2000 training spectra. Rix et al. (2016) showed that adopting a larger training set than the free parameters will better constrain the flux variation, especially when the range of the parameter space is large. We found that adopting a 10 times larger training set does not change our results qualitatively in this study. For the 2000 training spectra, we adopt an adaptive refinement technique to decide on the training labels as described below.

We start with a "sparse" set of labels that samples T_{eff} and $\log g$ from the MIST isochrones (Choi et al. 2016) assuming [Z/H] $= -1.5$ to 0.5, $T_{eff} = 3000$–8000 K, and stellar age from 3 to 10 Gyr, covering both dwarfs and giants. We consider states of stellar evolution from the main sequence to the core helium-burning at the red clump. We then use these labels to create two convex hulls for the giants (defined with $\log g < 4$) and the dwarfs ($\log g > 4$) separately in the T_{eff}–$\log g$ space, i.e., minimum polygons that encompass the tracks from the MIST isochrones. Subsequently, we randomly sample T_{eff} and $\log g$ from a uniform distribution within these convex hulls. Analogously, we draw v_{micro} uniformly from 0.1 to 3 km s^{-1} and v_{macro} uniformly from 0 to 30 km s^{-1} with 2000 data points. We have found that this choice spans most of the derived APOGEE label space without requiring extrapolation. We assume a weak prior for C_{12}/C_{13}. We adopt the isochrones' value of C_{12}/C_{13} given the stellar parameters of the training data. But we arbitrarily spread out the C_{12}/C_{13} values on the training set with a uniform distribution of ± 35. Finally, for this sparse grid, we randomly draw all elemental abundances [X/H] from a uniform distribution with the condition $-0.5 <$ [X/Fe] < 0.5. Note that here we train a single spectral model that encompasses both dwarfs and giants.

While the sparse grid is essential to make sure that we capture all cases, spanning a 25-dimensional space with only 2000 training data cannot constrain the flux variation to the necessary precision. Therefore, we need to refine the label space from which we draw our training labels. To do that we train *The Payne* with the sparse grid and fit all APOGEE spectra, which results in an initial distribution of the sample in label space. Then, we resample 2000 training data points with [X/H] drawn from these initial APOGEE label values. We note that APOGEE data do not span the T_{eff}–$\log g$–[Fe/H] space uniformly. Therefore, to avoid only fitting the variation of flux well for the bulk of the data, we do not resample main stellar parameters with the fitted values, but rather we sample T_{eff}, $\log g$, [Fe/H], v_{micro}, v_{macro}, and C_{12}/C_{13} as before. But we adopt [X/H] from the fitted APOGEE values that have consistent [Fe/H]. In other words, we bin the measured (using *The Payne* trained on the sparse grid) [X/H] APOGEE values

according to their fitted [Fe/H] values with a bin size of 0.2 dex. We only sample [X/H] in the corresponding [Fe/H] bin consistent with the newly drawn [Fe/H] training label. And these 2000 resampled training points constitute the final training set. Our sampling scheme is summarized in Table 1.

2.4. The Details: ab initio Models, Line Lists, Spectroscopic Masking, Instrumental Dispersion, and Continuum Normalization

We compute 1D LTE spectral models adopting the state-of-the-art codes ATLAS12 and SYNTHE maintained by R. Kurucz (Kurucz 1970, 1993, 2005, 2013, 2017; Kurucz & Avrett 1981; we reference therein). It is crucial to recalculate the stellar atmospheric structure as we vary the stellar labels to obtain accurate stellar labels from APOGEE, instead of simply running the radiative transfer code. We calculate the stellar atmospheric structure by partitioning the stellar atmosphere into 80 zones of Rosseland optical depth, τ_R, with the maximum Rosseland depth $\tau_R = 1000$. When generating synthetic models, we automate the inspection of numerical convergence for each layer of the stellar atmospheres. We adopt solar abundances from Asplund et al. (2009) and the Arcturus stellar labels from Ramírez & Allende Prieto (2011) throughout this study. We assume a standard mixing length theory with no overshooting for convection. After the stellar atmosphere converges, we produce the synthetic model spectra through the radiative transfer code SYNTHE at the nominal spectral resolution of $R = 300,000$. The synthetic spectra are subsequently convolved to the APOGEE resolution assuming the APOGEE averaged LSF template. We normalize both the synthetic spectra and the APOGEE observed spectra following Ness et al. (2015). In this method, a set of wavelength pixels with the least response to stellar labels, based on the data-driven model *The Cannon*, are selected. A fourth-order polynomial is fitted through the fluxes of these wavelength pixels and is used to approximate the continuum.

A crucial improvement of our ab initio models is the use of an updated line list (P. Cargile et al. 2019, in preparation), which will soon be made publicly available. Improving on the original Kurucz line list, the new line list tweaks three line parameters for every line stronger than 1% at $R = 300,000$ in either the Sun or Arcturus: the central wavelength, the oscillator strength, and the dominant broadening parameter. These line parameters are simultaneously fit to the high-resolution spectral atlas of the Sun and Arcturus in segments of several angstroms in order to capture possible covariance between overlapping lines. We refer readers to the paper for more details. Figure 2 shows a comparative assessment of the new line list. We synthesize spectra at the solar and Arcturus
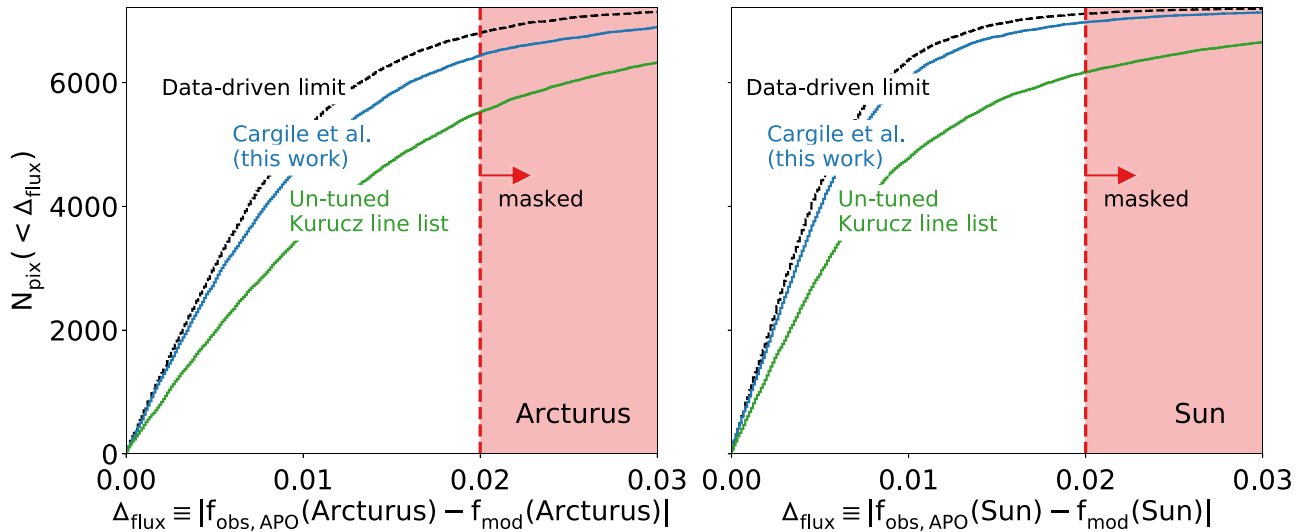
Figure 2. Comparison of the model quality for the improved line list (P. Cargile et al. 2019, in preparation) adopted in this study vs. an untuned Kurucz line list. We generated synthetic models (with those two line lists), adopting fiducial stellar labels for Arcturus and the Sun. These models were convolved with the APOGEE-determined average LSF and compared to the APOGEE spectra of Arcturus and the Sun. The panels show the cumulative distribution of APOGEE wavelength pixels as a function of the absolute deviation of the models and the observations. For comparison, we also convolved the Fourier Transform Spectrometer (FTS) spectra of Arcturus and the Sun, observed at very high resolution ($R = 300{,}000$) and high S/N, with the same LSF and compared those to the corresponding APOGEE spectra (dashed black lines). This comparison reflects the limit of a perfect model. Due to the influence of telluric lines, an imperfect LSF, and perhaps other data-related systematics, the convolved FTS spectra do not exactly match the observed APOGEE spectra. The vertical dashed line shows the model–data mismatch threshold that we adopt in this study for creating a pixel mask for the fitting procedure. Pixels more discrepant than this cut in either the Sun or Arcturus are omitted. The improved line list allows us to discard far less spectral information, and hence improve the precision of our fit.

stellar labels, and convolve and normalize them to the APOGEE resolution with the methods described above. We then compare the models to the observed Arcturus and solar spectra from APOGEE. There is a total of 7214 pixels in an APOGEE spectrum, and Figure 2 shows the cumulative number of wavelength pixels as a function of the absolute deviation of the models from the observations at each pixel.

The model–data match based on the updated line list adopted in this study is shown by the blue line, while the match with models that use the standard untuned line list (available on R. Kurucz's website) are shown by the green line. The shaded regions identify the pixels we mask and eliminate in the subsequent modeling—pixels that have normalized model fluxes deviating by $\geqslant 2\%$ at the APOGEE resolution from the observed spectra, for either Arcturus or the Sun. About 90% of pixels that we mask are due to disagreement with Arcturus, especially in the middle chip of APOGEE, i.e., 15800–16400 Å. The poorer agreement with Arcturus is not surprising because the line list is better calibrated to the Sun than to Arcturus, and because the lower temperature of Arcturus results in more and stronger lines than in the Sun. The 2% cut is chosen to produce a satisfactory balance between the accuracy and the precision of our derived stellar labels. Imposing a more stringent cut will minimize the systematic errors of the spectral models, but at the expense of the precision we can achieve because we are excluding more spectral information. Also note that this binary spectroscopic mask only discards 12% of the APOGEE spectra, and we are still performing full spectral fitting with all stellar labels simultaneously. This should be distinguished from the ASPCAP mask that APOGEE DR14 imposed, where individual abundances are determined with different filters.

Figure 2 shows the comparisons of the APOGEE spectra of the Sun and Arcturus with the convolved version of the FTS spectra of the Sun and Arcturus that have very high signal-to-noise ratio (S/N) and resolution $R = 300{,}000$, which serve as

"perfect model" templates. The convolved high-resolution observed solar and Arcturus spectra do not match their APOGEE counterparts perfectly for several reasons. The APOGEE H-band suffers from severe telluric contamination, which is imperfectly subtracted. Furthermore, the LSF and continuum normalization that we adopt are not perfect and could contribute to this discrepancy. Nonetheless, the convolved FTS spectra set the baseline for the best-case scenario and show that the updated line list is closer to this limit than the original Kurucz line list. We also tested that making a spectroscopic mask at the FTS resolution and subsequently convolving it to the APOGEE resolution does not work. For *The Payne*, it is crucial to make the spectroscopic mask directly in the observable space. The mask is meant to capture both for theoretical imperfections (imperfect line parameters, non-LTE effects, etc.) and for observational problems (LSF, telluric absorption, etc.).

In Figure 3 we further investigate which pixels are masked from the fit. The *y*-axis quantifies how informative each pixel is; it is quantified by the rms of the model variations when sampling the training labels. The *x*-axis shows the absolute deviation of the model from the observed spectrum for both Arcturus and the Sun. The rms is calculated with the refined synthetic model grid used in the final training. The shaded regions show pixels that are excluded from analysis. Figure 3 shows that, overall, there is a weak correlation between the deviation and the strength of spectral features. This trend is expected because stronger lines are harder to model. But as shown, most of the spectral features are included in our fit, and only a minimal number of spectral features are masked.

Finally, we note that our method is completely general and can be applied to other spectroscopic models. We also tried to apply *The Payne* to the untuned Kurucz line list. We showed that, similar to the results using the new line list that we will present in this paper, the fits even with the old line list exhibit
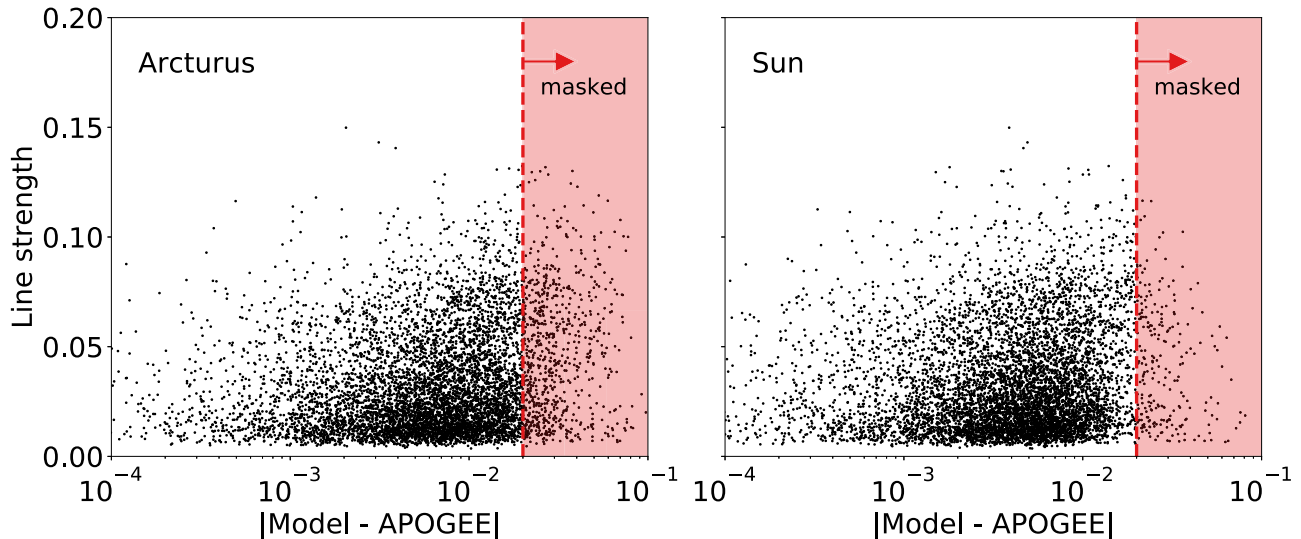
Figure 3. Most of the strongly varying spectral features are retained in *The Payne*'s fitting mask. Analogously to Figure 2, the *x*-axis shows the deviation of the model from the Arcturus and solar spectra observed by APOGEE at the APOGEE resolution. The *y*-axis shows the normalized rms flux for individual pixels, among synthetic training spectra of *The Payne*. Larger *y*-axis values indicate that the pixel contains a strongly varying spectral feature; those are the pixels that can discriminate between models. There is an overall weak correlation between the model deviation and the feature strength because stronger broad features could be harder to model. Nonetheless, with the pixel mask we apply in this study, most strong features remain included in the fit, and we only discard a modest amount of spectral information.

better agreements with the isochrones as well as a flat T_{eff}–abundance trend for open clusters. However, the overall accuracy and precision with the old line list are not as good as with the improved new line list. The worse precision is expected because, with the old line list, we need to mask out significantly more pixels (Figure 2). The slightly worse accuracy (i.e., not as good an agreement with the isochrones) is a bit puzzling. It suggests that the *H*-band spectroscopic models are not consistent throughout all the APOGEE pixels. Checking how the results vary by restricting to different sub-ranges of wavelength could shed light on this issue, but this is beyond the scope of this paper. Moreover, a thorough comparison would also require us to apply the APOGEE's ASPCAP pipeline to the new line list (instead of only applying *The Payne* to the old line list), something that we do not have the tool to perform ourselves. We will defer such detailed explorations to future studies.

2.5. Astrophysical Verification of The Payne

In this section we present two important tests of *The Payne*: first, we compare newly generated ab initio models that were not included in the training step with models predicted from *The Payne*. This step directly tests interpolation errors in the training of the neural networks. Second, we fit noiseless models with *The Payne* to see how well we can recover stellar labels in the case of perfect synthetic models. This step tests how much any interpolation errors in flux space translate into uncertainties in determining accurate stellar labels.

Figure 4 shows how well *The Payne* interpolates synthetic spectra. We trained on 2000 training spectra and test on the additional 850 synthetic validation spectra that are not used in training. The top left panel shows a small range of wavelength, comparing *The Payne* interpolation with the ab initio calculated spectra. The upper case illustrates a spectrum where the interpolation error is small (<0.1%). Most of the validation spectra are in this category. The lower case is one of the few extreme cases where the interpolation is poor (>1%).

The top right panel shows the absolute interpolation errors for different synthetic spectra in different temperature ranges, taking the median over all wavelength pixels. For each synthetic spectrum, the median interpolation error is only about 0.1% with *The Payne*, more accurate than the typical S/N observed by APOGEE. Cooler stars have slightly larger errors because there are more spectral features in cool stars and the imperfectness of continuum normalization becomes more severe. We note, however, that in some cases, the errors can be >1%. We tested that including 10 times more training data and increasing (or decreasing) the number of neurons does not improve these cases. We will leave the fine-tuning of the network architecture and loss function as well as the tailoring of specific regularization to mitigate these extreme cases to future studies. Nonetheless, although not shown, we also tested that with a quadratic model, the interpolation errors are typically an order of magnitude larger, which is not surprising given the large range in T_{eff} and $\log g$ under consideration.

The bottom panels illustrate the pixel-by-pixel interpolation errors, averaging over validation spectra. Plotted on the bottom left panel are the median errors for a randomly selected wavelength segment. Typical pixel-by-pixel errors for *The Payne* are about 0.1%. The results over all wavelength pixels are summarized in the bottom right panel, which illustrates the cumulative number of wavelength pixels as a function of interpolation errors. The solid lines show the median as before, and the dashed lines indicate the 95th percentile (2σ), i.e., pixel-by-pixel, more than half of validation spectra have interpolation errors smaller than the solid line with *The Payne*, and more than 95% of the validation spectra are within the interpolation errors illustrated by the dashed line.

Having established that *The Payne* can interpolate models well, we will now investigate how much the interpolation error in flux space translates into accuracy error in determining stellar labels, i.e., how well *The Payne* can recover the stellar labels in the limit of perfect spectral models with no noise. This will set a lower limit on the accuracy (not precision) with which *The Payne* can recover stellar labels. Figure 5 shows the
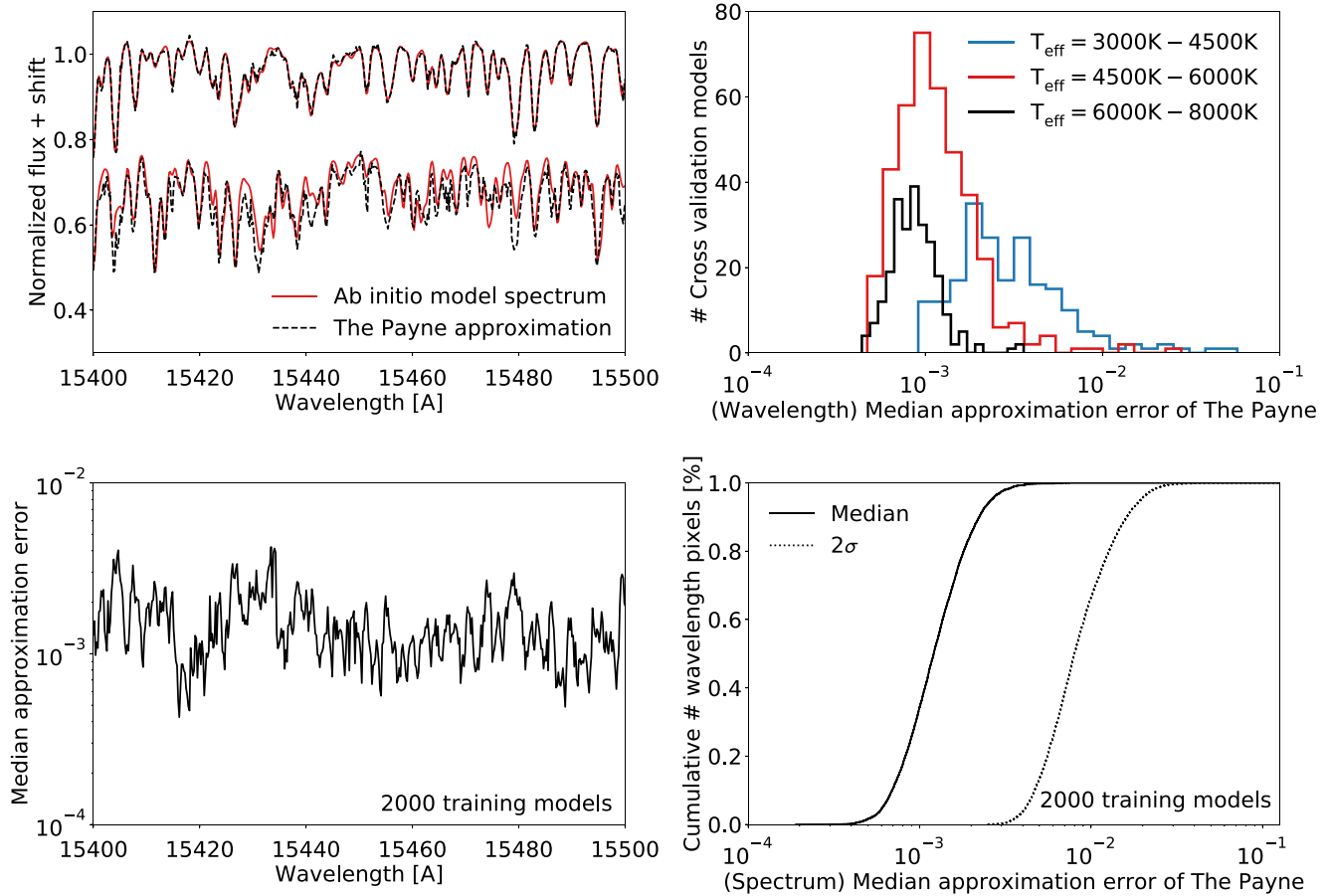
Figure 4. Assessing the quality of flux prediction of *The Payne*. For 25 labels, we have adopted 2000 ab initio models as the training set to build the spectral model, and have used another 850 ab initio models for cross-validation. The top left panel shows two examples of spectral predictions from *The Payne* compared to the directly calculated cross-validation spectra: the upper spectrum show a case where the prediction of *The Payne* and the cross-validation spectrum appear indistinguishable. Most of the validation spectra are in this category. The lower spectrum show an extreme case where the interpolation is poor (errors > 1%). We shift the continuum baseline of the lower spectrum for clarity of the plot. The top right panel shows a more quantitative assessment of how the quality of *The Payne*'s approximate flux prediction varies across label space. We calculate the median of approximation errors of individual testing spectra over all wavelength pixels. Cooler stars have more strong transitions, and hence they are harder to model and interpolate; but even for the cooler case, the median approximation error is less than 1%, which is smaller than the typical S/N of an observed APOGEE spectrum. The lower panels demonstrate the median approximation error of individual wavelength pixels over all testing spectra. The lower left panel illustrates the median approximation error for (across the ensemble of cross-validation spectra) a small segment of the wavelength range, and the lower right panel shows the cumulative distribution for all wavelength pixels. *The Payne* approximates the flux (variation) for each pixel to the level of ∼0.1%.

recovery of stellar labels of the validation spectra by fitting (noiseless) validation spectra with *The Payne*. Throughout this study, we fit spectra by minimizing the χ^2 of the interpolated model to the fitting spectra. The χ^2 minimization is performed using SCIPY.OPTIMIZE.CURVEFIT. When fitting real observed APOGEE spectra, we also take into account the reported uncertainties for individual pixels; pixels masked out by a spectroscopic mask are set to have infinite uncertainties. We have tested that initializing at different initial points for the χ^2 minimization results in the same solutions. This is not surprising because, at the APOGEE's resolution, most spectral features are resolved, and hence the degeneracy of stellar labels is not severe (Ting et al. 2017a). As such, we only run the optimization once for each spectrum. Since generating a spectrum to compare with the fitting spectrum requires only evaluating a function (the trained neural networks), which takes only milliseconds, the optimization typically only consumes one CPU second to fit for an APOGEE spectrum.

The top panel shows 1σ of the label recovery. As shown by the red line, for the bulk of the APOGEE spectra that have

$T_{\rm eff} \simeq$ 4500–5000 K, in the limit of perfect models, *The Payne* can recover labels to an accuracy of $\simeq$0.02–0.1 dex for elemental abundances, 30 K for $T_{\rm eff}$, and 0.05 dex for $\log g$. Some elemental abundances have larger accuracy problems, but these are abundances that have rather weak signatures and/or occur with strong blends. In practice, almost all of the elemental abundances (specifically C, N, O, Mg, Al, Si, S, K, Ca, Ti, Cr, Mn, Fe, Ni,) that we will focus on (14, excluding Cu) in the APOGEE example study have accuracy better than ∼0.05 dex. The blue line shows the accuracy for stars cooler than 4500 K (e.g., Arcturus). Despite having more spectral features, the typical accuracy for cooler stars is two times worse due the larger interpolation errors, as already illustrated in Figure 4. We also note that while there might be biases for individual stars of 0.03–0.1 dex, the bottom panel shows that, if the training sample is a fair representation of the global APOGEE chemical distribution, there is no strong overall bias due to the interpolation error. Plotted is the median deviation of the validation spectra fit to the assumed input. For all abundances, the overall bias is typically less than 0.01 dex.
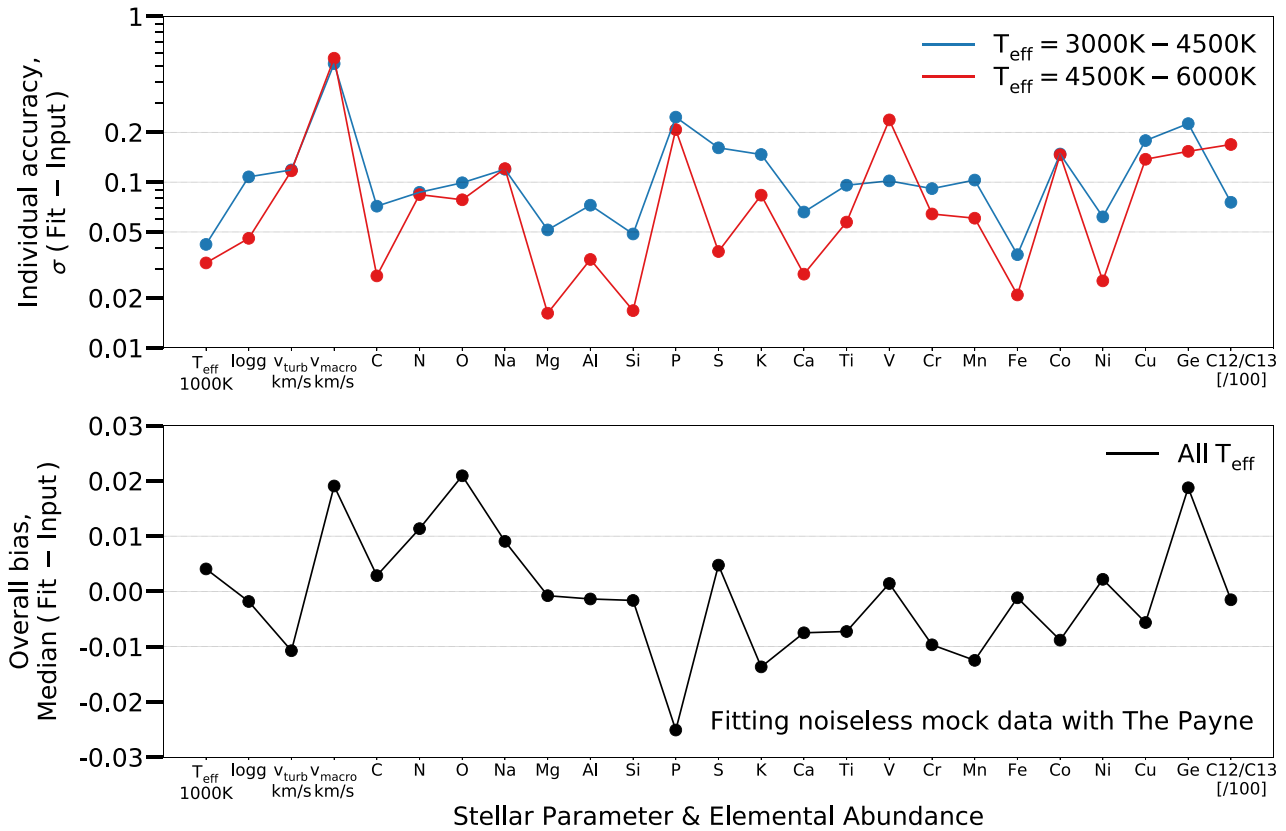
Figure 5. Recovery of stellar labels when applying *The Payne* to fitting noiseless ab initio cross-validation spectra. Shown in the top panel is the deviation of the recovered labels from the input labels, which reflects the systematic flux errors incurred in interpolation with *The Payne*. This sets the floor for the accuracy (not precision, see text for details) of the stellar labels. The red line shows the deviation for the hotter stars and the blue line for the cooler stars. For individual stars, we might incur biases of 0.03 dex in [Fe/H], 0.03–0.1 dex in other elemental abundances (C, N, O, Mg, Al, Si, S, K, Ca, Ti, Cr, Mn, Fe, Ni), 50 K in $T_{\rm eff}$, and 0.05 dex in log g. Hotter stars have errors that are about half of these values because they are less subjected to interpolation error with *The Payne*. In the bottom panel, we show the median deviation of the fit from the input labels. The bottom panel shows that, provided that the validation labels in the mock data are a fair representation of the APOGEE data, there are no strong global biases in the recovery.

Importantly, we emphasize that the results show the accuracy of *The Payne* instead of its precision because at a given stellar label, although *The Payne* could incur a bias, the differential recovery can still be very precise. As we will see in the APOGEE example application below, we achieve an elemental abundance precision of about 0.03 dex for all elemental abundances by fitting the APOGEE spectra.

3. An Illustration of *The Payne*: 25 Stellar Labels from APOGEE Data

As a specific application and illustration of *The Payne*, we fit the entire APOGEE DR14 data set, consisting of ∼270,000 spectra. We only consider the combined APOGEE spectra (instead of individual visits) throughout this study. We train *The Payne* with only 2000 ab initio model spectra, and then fit for 25 stellar labels. We also fit for the radial velocity at the same time during the fit to avoid any radial velocity residual from the APOGEE reduction pipeline. When comparing to APOGEE DR14 values, we will refer to the official APOGEE pipeline, ASPCAP, values, instead of the values from *The Cannon*.

3.1. Fitting the Sun and Arcturus at APOGEE Resolution

We start out by illustrating how well *The Payne* does in fitting Arcturus and the Sun at the resolution of APOGEE (Figure 6). We generated 100 realizations of Arcturus' and the Sun's APOGEE spectra, differing just in Poisson noise of the spectra

(S/N ∼ 400). The violin plots in Figure 6 show the deviations of our fit of all 100 realizations from the Arcturus and solar benchmark values adopted from Ramírez & Allende Prieto (2011) and Asplund et al. (2009). The solid black line shows the corresponding APOGEE DR14 values. Overall, *The Payne* shows comparable deviations from the benchmark values to APOGEE DR14, about 0.1 dex for elemental abundances. Part of the deviations is due to the error in interpolation accuracy described, but they are also partially contributed by the imperfect spectral models. For individual objects, performing full spectral fitting with *The Payne* can be more susceptible to model imperfection due to the covariant spectral features, especially with the lenient cut that we made that keeps almost the full APOGEE spectrum. If we were to make a more stringent cut for the spectroscopic mask, i.e., 0.5% error instead of the fiducial 2% error adopted, as shown by the red dashed lines, the accuracy can get better, with the exception of V, which only has a very weak feature at the solar temperature. But this comes at the expense of the precision of stellar labels for the overall sample because, as illustrated in Figures 2 and 3, with a more stringent cut, we discard a significant portion of the spectra. Therefore, we adopt the fiducial spectroscopic mask throughout this study.

3.2. T$_{eff}$ and log g

Figure 7 shows how well *The Payne* can recover stellar parameters ($T_{\rm eff}$, log g, [Fe/H]) for both giants and dwarfs with
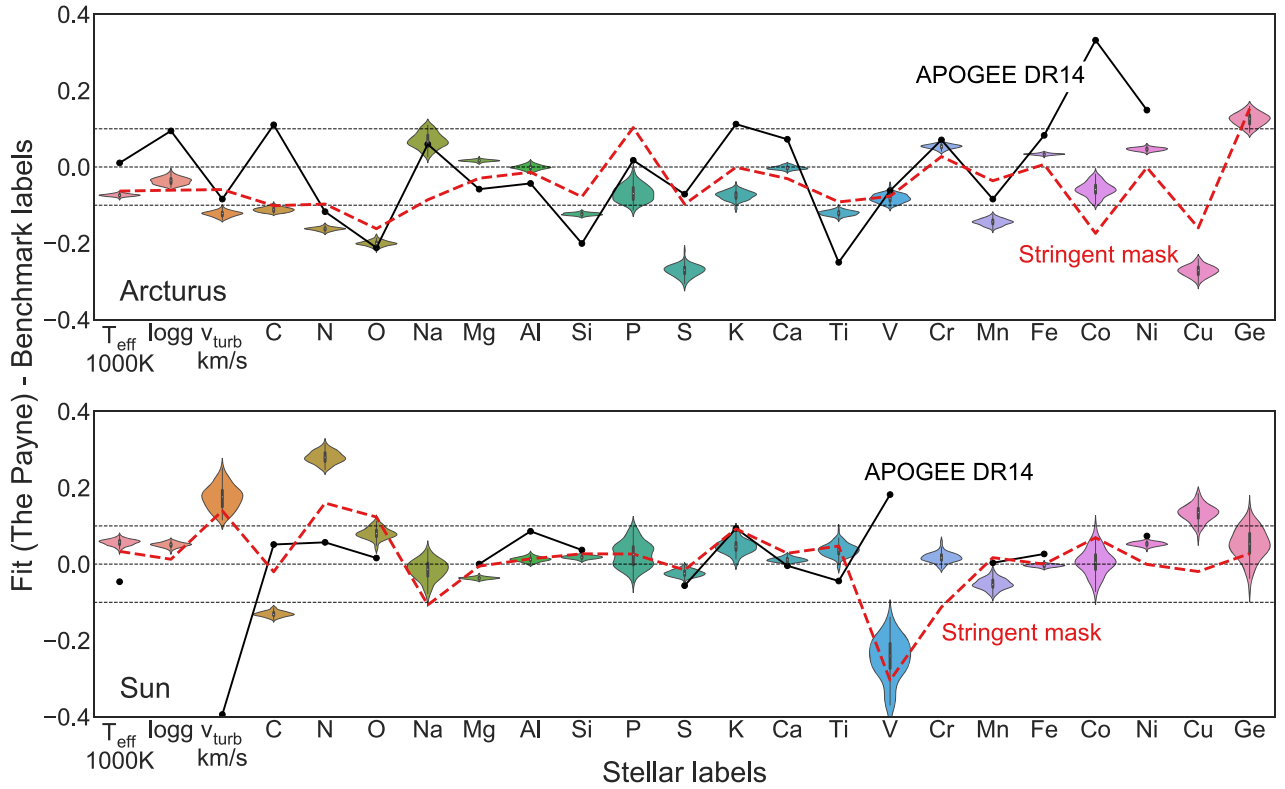
Figure 6. Fitting APOGEE spectra of the Sun and Arcturus with *The Payne*. We adopt the APOGEE Arcturus and solar spectra and generate 100 realizations by sampling the reported uncertainties (S/N ~ 400). We fit these 100 realizations with *The Payne*; for the fiducial case with a spectroscopic mask of 2% error tolerance, we plot the departure of the fitted values from the Arcturus and solar reference stellar labels as the violin plots. The red dashed lines show the fit from *The Payne* where a more stringent spectroscopic mask (0.5% tolerance) is adopted. The APOGEE DR14 fitted values, when available, are overplotted with solid black lines as references. The fitted values are consistent with the reference values to about 0.1 dex in elemental abundances, similar to APOGEE DR14, with the exception of V at solar temperature because V only has an extremely weak feature at the solar temperature. A more stringent spectroscopic mask reduces some biases, but at the expense of the precision for the overall sample.
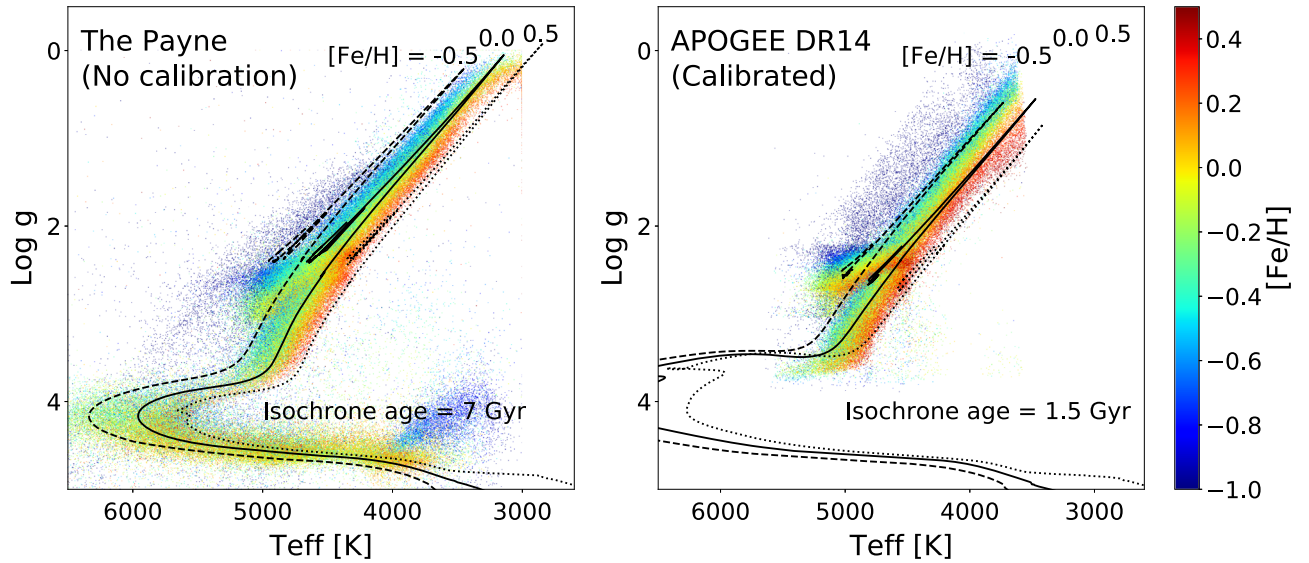

Figure 7. *The Payne* measures physically sensible $T_{\rm eff}$, $\log g$, and [Fe/H] for both giants and dwarfs simultaneously without requiring external calibration. On the left, we show the $T_{\rm eff}$–$\log g$ Kiel diagram from *The Payne* overplotted with MIST isochrones assuming a stellar age of 7 Gyr. On the right, we show the APOGEE DR14 calibrated counterparts and with MIST isochrones at 1.5 Gyr. *The Payne* derives stellar parameters that are consistent with stellar isochrones for both giants and dwarfs with only a single model. For metal-poor dwarfs with $T_{\rm eff} < 4000$ K, the results deviate strongly from the isochrones. This could be because the 1D stellar atmosphere of the Kurucz model is a poor assumption in this regime or simply because the line list is not well calibrated at this temperature and in this metallicity range.

a single self-consistent training model. The left panel shows the values obtained by *The Payne*, and the right panel shows the APOGEE DR14 calibrated counterparts. APOGEE DR14 does

not provide calibrated stellar parameters for dwarfs and sub-giants because the authors found that the current pipeline struggles to provide reliable estimates for non-giants (e.g.,

Holtzman et al. 2015). Overplotted in both panels are the MIST isochrones, but at different stellar ages. *The Payne* derives $T_{\rm eff}$ and $\log g$ that are consistent with the MIST isochrones at 7 Gyr, and the estimates show less scatter at the metal-poor end for the giants compared to APOGEE DR14. The APOGEE team calibrated their values with the photometric $T_{\rm eff}$ and the asteroseismic $\log g$ as we will discuss below, and the calibrated values are more consistent with 1.5 Gyr old MIST isochrones, which might be too young for the bulk of the APOGEE data. It thus suggests that there is a discrepancy between the photometric $T_{\rm eff}$ that the APOGEE values calibrated against, and the spectroscopic $T_{\rm eff}$ from *The Payne* and MIST isochrones at the 100 K level. The figure shows that APOGEE DR14 calibrated values also generally favor more metal-rich estimates than *The Payne*. But this is largely due to their calibration with photometric temperature as we will discuss below.

The Payne does not perform as well for the cooler dwarf stars ($T_{\rm eff} < 4000$ K), especially for metal-poor stars ([Fe/H] < -0.5). This could due to multiple reasons. For example, our adopted line list is only calibrated against hotter and more metal-rich stars— Arcturus ($T_{\rm eff} \simeq 4300$ K) and the Sun ($T_{\rm eff} \simeq 5800$ K). Moving forward, spectral models built from an atomic line list that has been calibrated against a wider array of stars will be very valuable. The failure in the metal-poor dwarf regime could also be due to a breakdown of the assumptions of LTE.

As shown in Figure 7, the $T_{\rm eff} - \log g$ Kiel diagram for dwarfs also exhibits a larger spread than is predicted by the stellar evolution models. Part of this larger spread could be due to the fact a non-negligible fraction of the main-sequence stars could be unresolved binaries. Fitting single-star models to binaries will incur a bias that manifests itself as a thicker sequence in the Kiel diagram (El-Badry et al. 2018a). It is beyond the scope of this paper to fit for binaries, but we caution that the single-star assumption can compromise the abundance precision that we obtain for dwarfs. For giants, the single-star assumption is less of a problem because the giant will outshine its dwarf companion, and giant–giant binaries are rare. We refer readers to El-Badry et al. (2018b), where *The Payne* was adopted to fit for main-sequence binaries by fitting a mixture of (data-driven) stellar models.

In Figure 8 we compare *The Payne* estimates with $T_{\rm eff}$ and $\log g$ derived from other external means. In the left panel we compare the spectroscopic $T_{\rm eff}$ to the $J - K$ color–$T_{\rm eff}$ relation from González Hernández & Bonifacio (2009). For this comparison we only consider giants that have small line-of-sight extinction, i.e., $E(B - V) < 0.02$ from the SFD map (Schlegel et al. 1998), avoid the Galactic disk ($|b| > 30°$), and have color $0.1 < J - K < 0.9$ following González Hernández & Bonifacio (2009). In the right panel we compare spectroscopic $\log g$ for a subset of 3000 stars that have APOKASC v3.6.5 asteroseismic $\log g$ values. Without calibration, the $\log g$ estimates from *The Payne* agree with the asteroseismic $\log g$ values to about 0.07 dex with only a weak metallicity dependence. Overplotted as a red line is the best-fit linear regression. We do not overplot the APOGEE DR14 values because, by definition, APOGEE DR14 $\log g$ are calibrated to match the APOKASC asteroseismic $\log g$ and the photometric temperature. As shown in the left panel, spectroscopic $T_{\rm eff}$ from *The Payne,* however, is typically 100 K cooler than the photometric $T_{\rm eff}$, and shows a dependence on metallicity. It is

hard to speculate what causes this trend, but it could either be inflicted by the inherent differences between *H*-band spectroscopic temperature and photometric temperature, since APOGEE DR14 uncalibrated values also show similar offsets, or it could simply be due to the imperfect spectral model and line list. We found that imposing a more stringent spectroscopic mask does not resolve this issue, indicating that the lower temperature is favored by our spectroscopic model and is not due to interpolation error. But as we will see, even without calibrating this relation, the derived stellar labels from *The Payne* seem to agree well in other plausibility tests that we will present below. So we choose not to calibrate the temperature and will leave the more detailed study of this discrepancy to future studies (e.g., Choi et al. 2018).

One particularly interesting aspect of *The Payne* as shown in Figure 7 is that, besides deriving stellar parameters for the dwarfs, *The Payne* also yields reasonable $T_{\rm eff}$ and $\log g$ for the giants at the cooler end, around 3500 K to 4000 K. In fact, we found that fitting C_{12}/C_{13} is crucial to get $T_{\rm eff}$ and $\log g$ that are consistent with the isochrone at the cooler end for the giants. Since C_{12}/C_{13} spectral features are highly blended with other features, C_{12}/C_{13} can only be reliably derived with a full spectral fitting with all stellar labels simultaneously, an area where *The Payne* excels.

3.3. C_{12}/C_{13} and C/N

The dependence of flux variation on C_{12}/C_{13} is particularly difficult to model. As already shown in Figure 1, the flux variation as a function C_{12}/C_{13} has a sharp transition. Above $C_{12}/C_{13} \simeq 50$, the spectral dependence is very weak, and below $\sim$50 the flux varies strongly with C_{12}/C_{13}. Since carbon molecular features contribute significantly to the *H*-band APOGEE spectra, C_{12}/C_{13} alters the spectra in a significant way. On the one hand, it implies that fitting C_{12}/C_{13} is not only astrophysically interesting, it can also be crucial as part of the spectral fitting, without which the stellar parameters may be biased. But on the other hand, we found that, in the limit of imperfect models, if we do not impose a prior C_{12}/C_{13}, the C_{12}/C_{13} features can be wrongly adopted to adjust the global fit to get a lower χ^2. Therefore, as discussed in Section 2.3, we assume a weak prior for C_{12}/C_{13} from stellar evolution models.

Figure 9 shows the C_{12}/C_{13} values estimated with *The Payne* for all APOGEE stars. On the left, we show the C_{12}/C_{13} values for dwarfs (with $\log g > 4$), and on the right those for giants ($\log g < 4$). Overplotted in black lines are the MIST isochrones for the respective evolutionary states, assuming a stellar age of 7 Gyr and metallicity [Z/H] ranging from -0.5 to 0.5. The C_{12}/C_{13} values for dwarfs are less well constrained and have a larger scatter from the MIST prediction because the spectral response with respect to C_{12}/C_{13} at $C_{12}/C_{13} > 50$ is weak and yields an almost identical spectrum (see Figure 1). As for the giants, the C_{12}/C_{13} values roughly agree with the MIST isochrones, with a sharp transition around 5000 K due to the first convective dredge-up, follow by a second transition as the stars ascend in the H-R diagram in the red giant branch (RGB) and reach a lower temperature. But the transition temperature seems to be smaller than the predictions from stellar evolution models.

We caution readers not to overinterpret the C_{12}/C_{13} results because we have assumed a prior for the C_{12}/C_{13} in the training set. One of the current challenges of full spectral fitting is that, in the limit of imperfect models, one stellar label, such
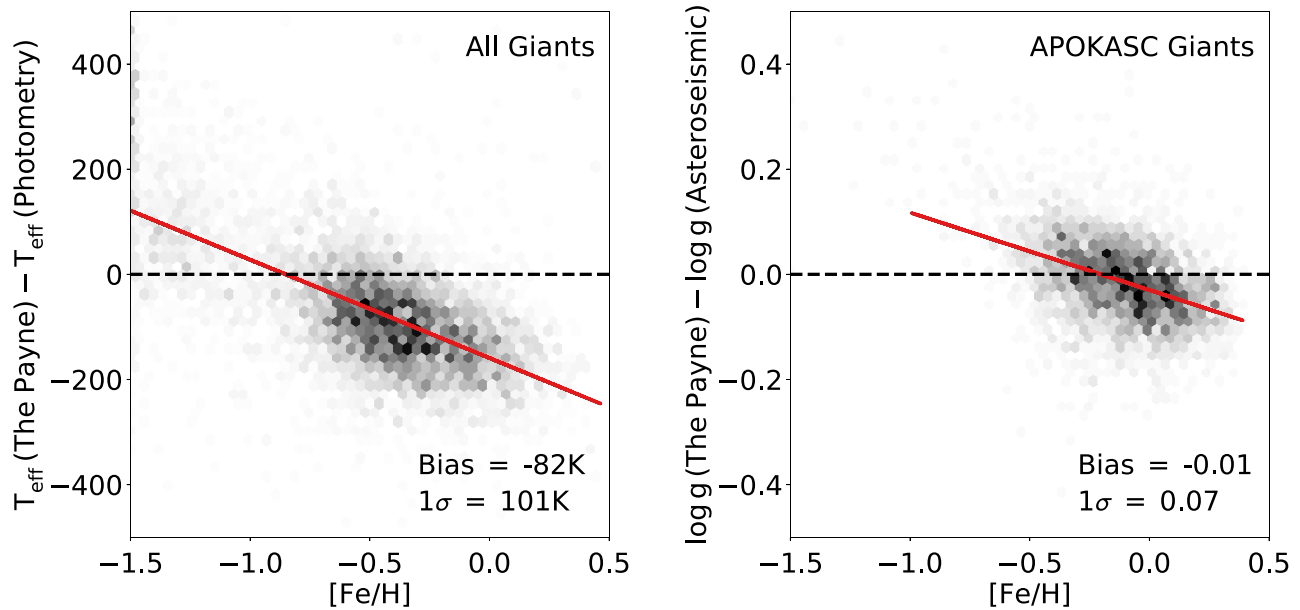
Figure 8. Accuracy of *The Payne* estimates of T_{eff} and $\log g$ compared to independent external constraints. The left panel shows the differences between spectroscopic T_{eff} estimated by *The Payne* and T_{eff} derived from Two Micron All Sky Survey infrared photometry following González Hernández & Bonifacio (2009) (with *The Payne* metallicity as input). We only consider giants that have small extinction $E(B - V) < 0.02$, Galactic latitude $|b| > 30°$, and $0.1 < J - K < 0.9$ to ensure the accuracy of the photometric estimations. The right panel shows the differences between spectroscopic $\log g$ from *The Payne* and APOKASC (v3.6.5) asteroseismic $\log g$. T_{eff} and $\log g$ from *The Payne* agree with these external calibrators to about 200 K and 0.1 dex, and exhibit with a weak metallicity dependence. The red lines illustrate the best-fit linear relation of the metallicity trend.
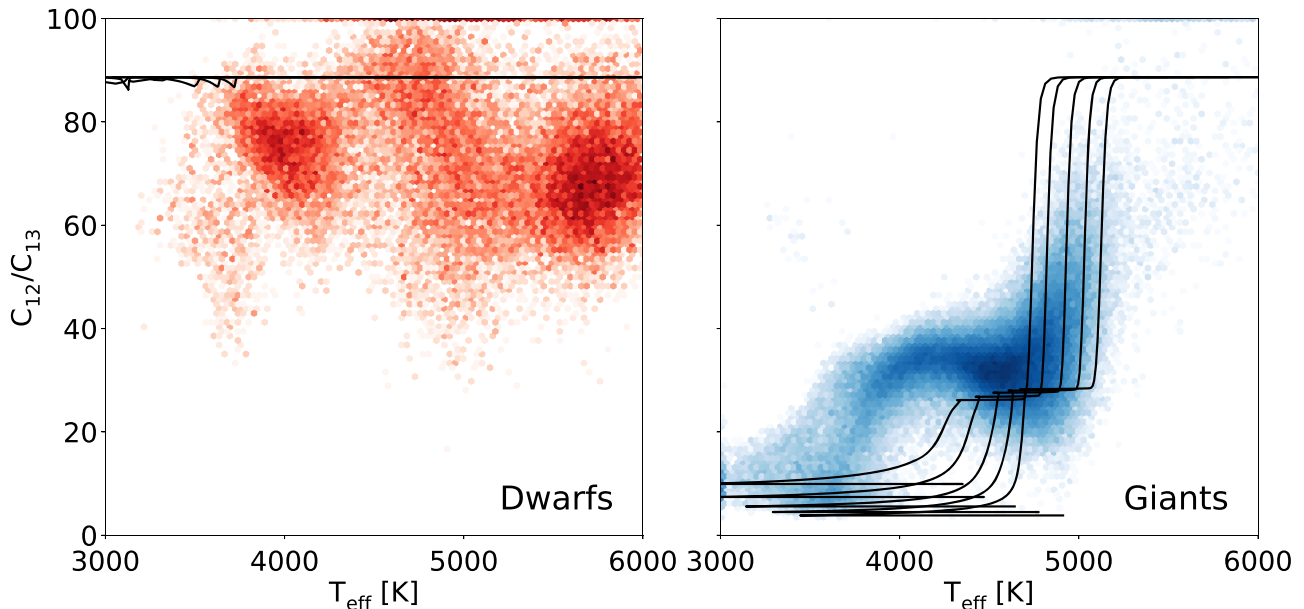
Figure 9. C_{12}/C_{13} estimates from *The Payne*. The C_{12}/C_{13} estimates for dwarf stars ($\log g > 4$) are shown on the left, those for giants ($\log g < 4$) on the right; we only show hexbins that have more than 10 stars. Dwarf stars typically show C_{12}/C_{13} from 60 to 90, but the exact values are not well constrained by *The Payne* in APOGEE because the spectra show little change for $C_{12}/C_{13} > 50$. The black solid lines reflect stellar evolution models from MIST, with a stellar age of 7 Gyr and metallicity $-0.5 < [Fe/H] < 0.5$. On the left panel, we show the isochrones for the main sequence, and on the right the turn-off phase to the helium core-burning red clump phase. For giants, the C_{12}/C_{13} values determined from *The Payne* roughly follow the expected trend with a first drastic transition in C_{12}/C_{13} at 5000 K, caused by the first convective dredge-up, and a second dip at a lower temperature. But we caution that these C_{12}/C_{13} results are partially driven by the prior we impose in the training set (see text for details).

as C_{12}/C_{13}, may in effect "do the work" of another stellar label. As discussed, the reason to include C_{12}/C_{13} is merely to ensure that the stellar parameters are robust at the cooler giant end, as it contributes significantly at this end due to the strong features as well as the second dredge-up. It also shows that C_{12}/C_{13}, in principle, can be fitted simultaneously with all other labels when employing *The Payne*.

Besides C_{12}/C_{13}, the [C/N] ratio of stars will also be modified due to convective dredge-up during the giant phase. In fact, the [C/N] ratio has been shown to be an excellent indicator of stellar mass for giants (Martig et al. 2016; Ness et al. 2016; Ho et al. 2017); how much the dredge-up affects the [C/N] ratio depends crucially on the stellar mass. Since there is a tight correlation between stellar mass and stellar age (given a
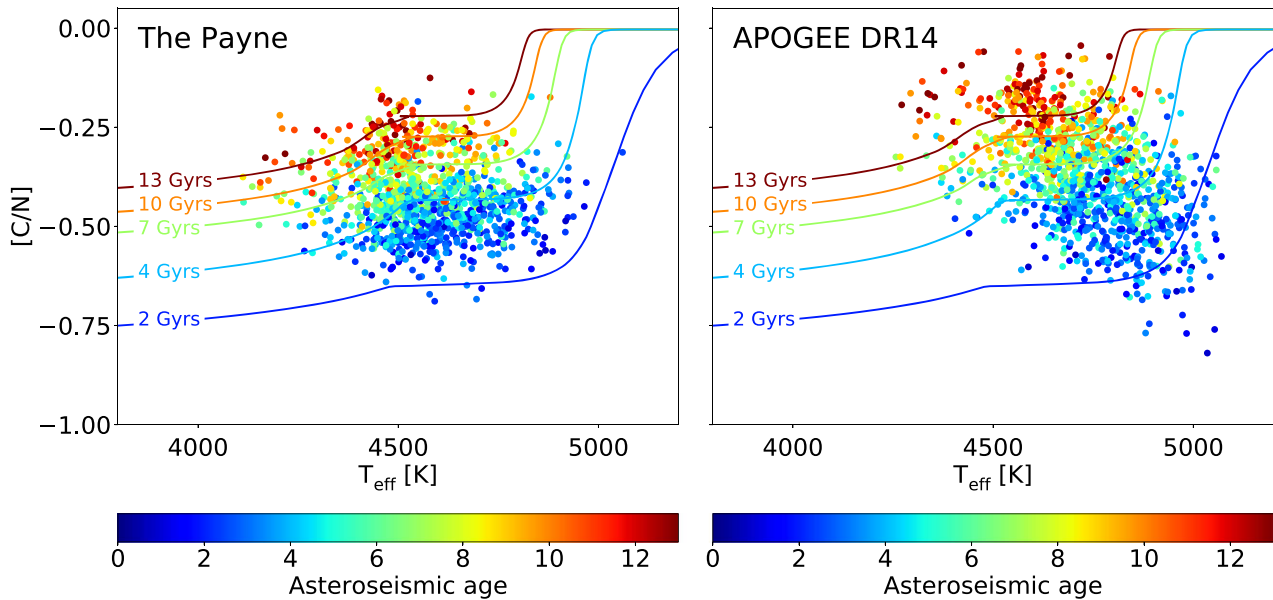
Figure 10. *The Payne* measures accurate and precise [C/N] ratios for field stars. We plot the spectroscopic [C/N] ratios for the subsample of $-0.1 <$ [Fe/H] < 0.1 of the APOKASC sample, color-coded with their corresponding precise asteroseismic ages. We show the estimates from *The Payne* on the left panel, and the APOGEE DR14 values on the right. The solid lines of different colors show the [C/N] predictions from various MIST isochrones of solar metallicity. *The Payne* measurements agree better with the isochrones and show a smaller scatter and bias compared to APOGEE DR14. The excellent agreement with the isochrones affirms the ability to infer stellar ages directly from abundance measurements.

fixed metallicity), determining accurate [C/N] ratios from large spectroscopic surveys is particularly important because they are excellent age indicators for stars. In Figure 10, we overplot the [C/N] ratios of the APOKASC sample, color-coded with their corresponding asteroseismic ages, with the predictions from the MIST isochrones. Since predictions from stellar evolution depend on metallicity, we restrict the APOKASC sample with $-0.1 <$ [Fe/H] < 0.1 and assume solar abundances for the isochrones. We show the results from *The Payne* on the left and APOGEE DR14 on the right. *The Payne* values agree better with the isochrones and show a reduced scatter and bias, especially for the older stars, indicating that our C to N abundances are likely more accurate. The excellent agreement between the stellar evolution models and spectroscopic indices also demonstrates that, by fitting all stellar labels self-consistently and simultaneously, the improved spectral models and stellar evolution models can be accurate enough to allow for a direct inference of stellar ages from spectroscopic indices, going beyond data-driven models.

3.4. Elemental Abundance Patterns

Elemental abundances are often derived from individual spectral lines, one element at a time. A key goal of *The Payne* is to demonstrate that all elemental abundances can be measured from stellar spectra directly from a simple χ^2 fit by fitting all elemental abundances and stellar parameters simultaneously. In this study, we fit for 20 elemental abundances, namely C, N, O, Na, Mg, Al, Si, P, S, K, Ca, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Ge. All elemental abundances show visible absorption lines from our line list in the *H*-band. As already shown in Figure 5, in the limit of perfect models and data, all of these elemental abundances can be extracted with *The Payne*.

However, we found that five elemental abundances (Na, P, V, Co, Ge) cannot be reliably derived with the current

implementation of *The Payne*, an issue also well diagnosed in APOGEE DR14 (e.g., Holtzman et al. 2015). These elements exhibit large scatter in an [X/Fe]–[Fe/H] diagnostic plot or large scatter in the precision test (Section 3.5.3). Elements such as Na, P, V have only weak features ($<1\%$ change in flux for Δ[X/H] $= 0.05$) in the *H*-band, and unfortunately the features are also often blended with the telluric sky lines, an issue compounded by the current interpolation errors from *The Payne*. Although we derived estimates from these elemental abundances, we decided that they are not to be trusted. The reason for a large spread in Co, Ge in an [X/Fe]–[Fe/H] diagnostic is unclear because each of these elements does have a single strong feature in the *H*-band, similar to K, and we have no problem getting reasonable K measurements as shown below. We will defer a more detailed study of the problems to a forthcoming paper. We will focus on the remaining 15 elements, 14 of which (excepting Cu) have been reliably determined in APOGEE DR14 for comparison, and only consider stars with a fitting reduced $\chi^2_R < 50$.

Figure 11 shows the comparison of *The Payne* estimates with the calibrated values from APOGEE DR14, showing a generally good agreement to the level of 100 K in $T_{\rm eff}$, 0.1 dex in $\log g$, and 0.1 dex in [X/H]. *The Payne* favors slightly metal-poor estimates, as already discussed in light of Figure 8. *The Payne* spectroscopic estimates prefer lower temperatures than the APOGEE DR14 values, which are calibrated to photo-metric temperatures. As [Fe/H] and $T_{\rm eff}$ estimates are covariant (e.g., Ting et al. 2017a), this leads to more metal-rich estimates for elemental abundances. Another noticeable deviation is around $\log g \simeq 2.5$. Also shown in Figure 7, the $\log g$ values for red clump stars from *The Payne* are overestimated compared to stellar evolution models. This discrepancy is also consistent with APOGEE uncalibrated values. The reason for this discrepancy is unknown; one possibility is the lack of fitting of the helium abundance. It is conceivable that
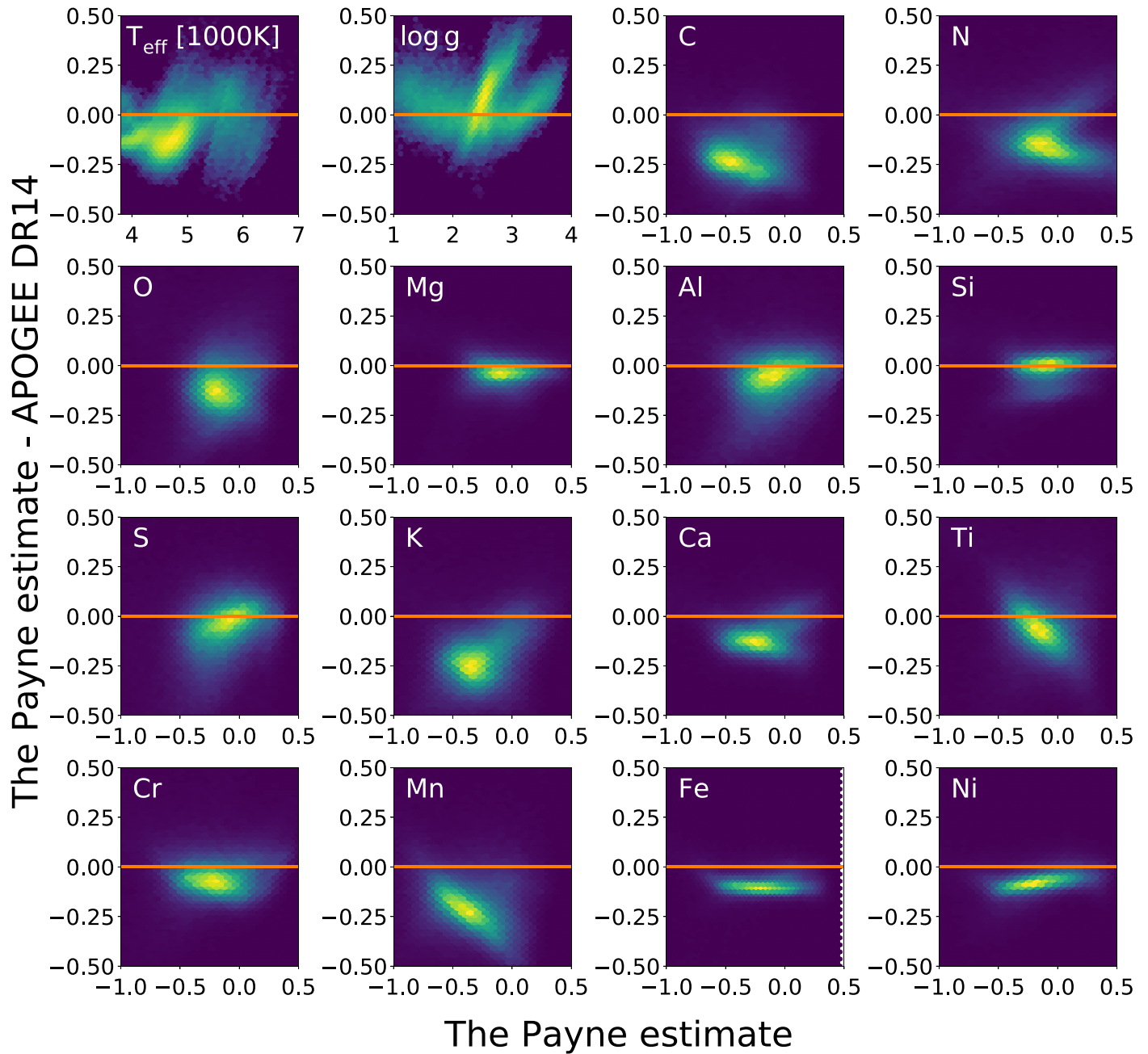
Figure 11. Comparison of *The Payne* estimates with the APOGEE DR14 calibrated stellar labels. We perform a full spectral fitting for all stellar labels as shown in the plot, as well as fitting v_{micro}, v_{macro}, and C_{12}/C_{13} simultaneously. We do not compare Cu since APOGEE does not provide Cu abundances in DR14. Note that, for T_{eff} and $\log g$, we plot the density plot on a log scale to emphasize the contrast since most APOGEE data are concentrated near the red clump T_{eff} and $\log g$. In general, *The Payne* prefers slightly (~0.08 dex) more metal-poor estimates than APOGEE DR14 calibrated values. There is a visible deviation in $\log g$ around $\log g = 2.5$; *The Payne* estimates of $\log g$ for red clumps are slightly higher than the calibrated APOGEE values.

differences in helium abundance between the RGB stars and the red clump stars could explain the $\log g$ discrepancy (e.g., Yu et al. 2018).

Figure 12 shows the [X/Fe]–[Fe/H] derived with *The Payne*. The background demonstrates the elemental abundances of the giant stars ($\log g < 4$) estimated by *The Payne*. Overplotted in white symbols are the literature values. We consider Bensby et al. (2014) to be the main reference, which provides abundances in this plot for O, Na, Mg, Al, Si, Ca, Ti, Cr, and Ni. This main sample is complemented by C abundances from Nissen et al. (2014), K abundances from Zhao et al. (2016), Mn abundances from Battistini & Bensby (2015), and Cu abundances from

Mishenina et al. (2011). For [Fe/H], we adopt [Fe/H] from the same catalog to avoid systematics across different surveys. *The Payne* attains reasonable [X/Fe] –[Fe/H] without any external calibration. The separation of the high-α sequence from the low-α one is clearly visible across all α-elements. Notably, we attain a Ti trend that is consistent with the literature values—resolving one of the persistent problems in APOGEE (e.g., Holtzman et al. 2015). There is a 0.1 dex discrepancy between the literature values and *The Payne* estimates for Si, K, and Ni. But we note that the K abundances from Zhao et al. (2016) adopt non-LTE (NLTE) models. *The Payne* also favors a flat [Mn/Fe] trend, which is at odds with the [Mn/Fe] trend in the literature.
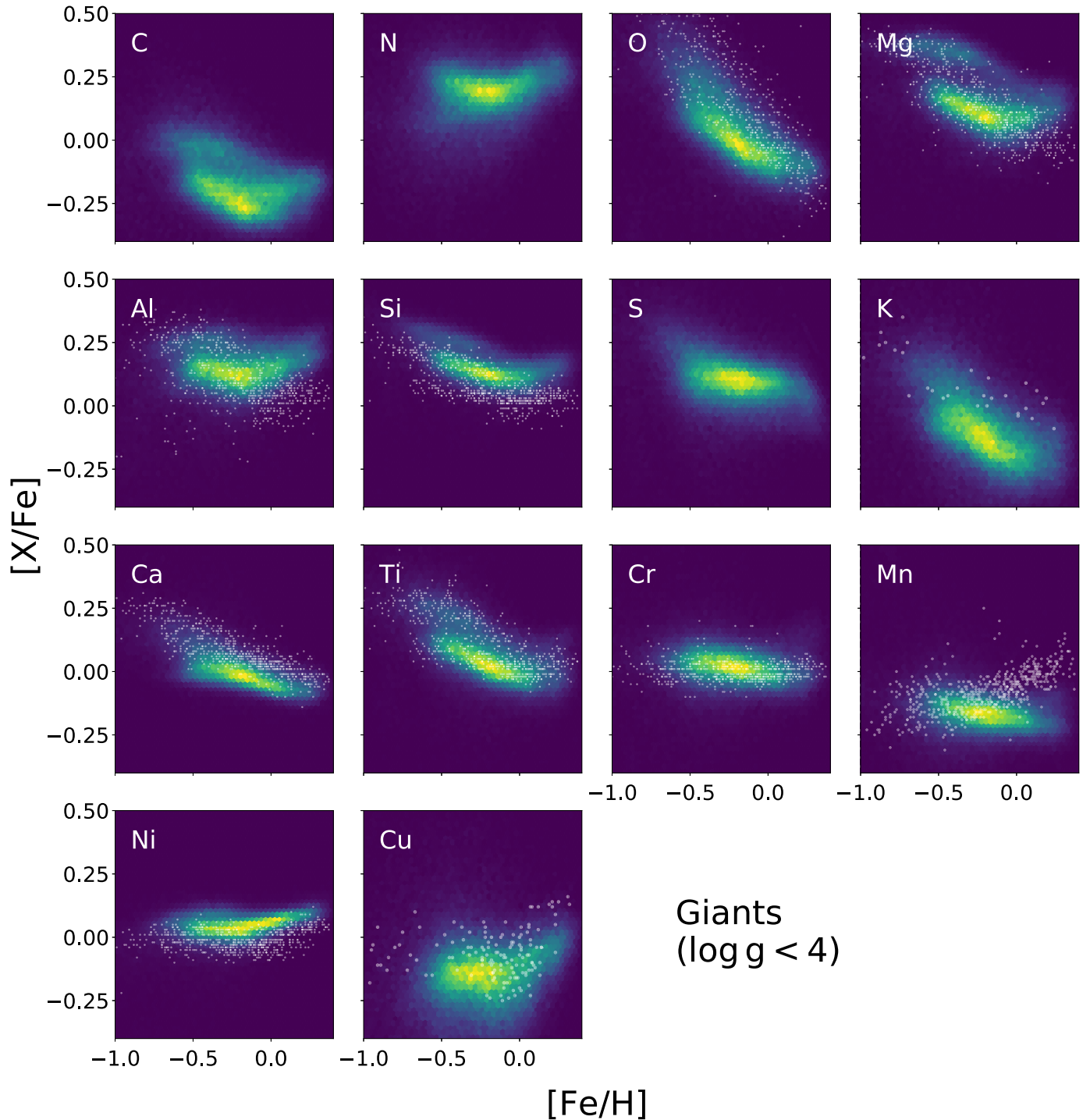
Figure 12. $[X/\mathrm{Fe}]$– $[\mathrm{Fe/H}]$ for 14 elements for the APOGEE DR14 giants ($\log g < 4$), derived with *The Payne*. The background shows the density plot of the label estimates from *The Payne*. Overplotted in white symbols are a compilation of literature values from Mishenina et al. (2011), Bensby et al. (2014), Nissen et al. (2014), Battistini & Bensby (2015), and Zhao et al. (2016). *The Payne* recovers the separation of the high-α and low-α sequences. Elemental abundances from *The Payne* in $[X/\mathrm{Fe}]$ agree well with the literature values without any calibration. Nonetheless, there is a slight offset in Si, K, and Ni compared to the literature values. Also our models prefer a flat $[\mathrm{Mn/Fe}]$ trend, which is in contrast with the literature values. Notably, Ti abundances derived by *The Payne* follow the expected trend, overcoming a persistent issue in APOGEE DR14.

One important improvement coming from *The Payne*, as already demonstrated in Figure 7, is the determination of stellar labels for APOGEE dwarf stars. Figures 12 and 13 demonstrate that *The Payne* recovers consistent abundances for both dwarfs and giants with a few key differences. First, the carbon abundances for the dwarfs are higher than those for the giants, and at the same time, the nitrogen abundances are lower as

expected due to convective dredge-up. Second, since dwarf stars are dimmer, they are dominated by stars that are closer to the Sun, and hence the dwarfs show a more prominent low-α sequence and have relatively fewer high-α stars. The dwarf abundances also seem to agree better with the literature values for Al, Si, K, Mn, and Ni. Since most of the literature values are derived from main-sequence dwarf stars, this agreement is
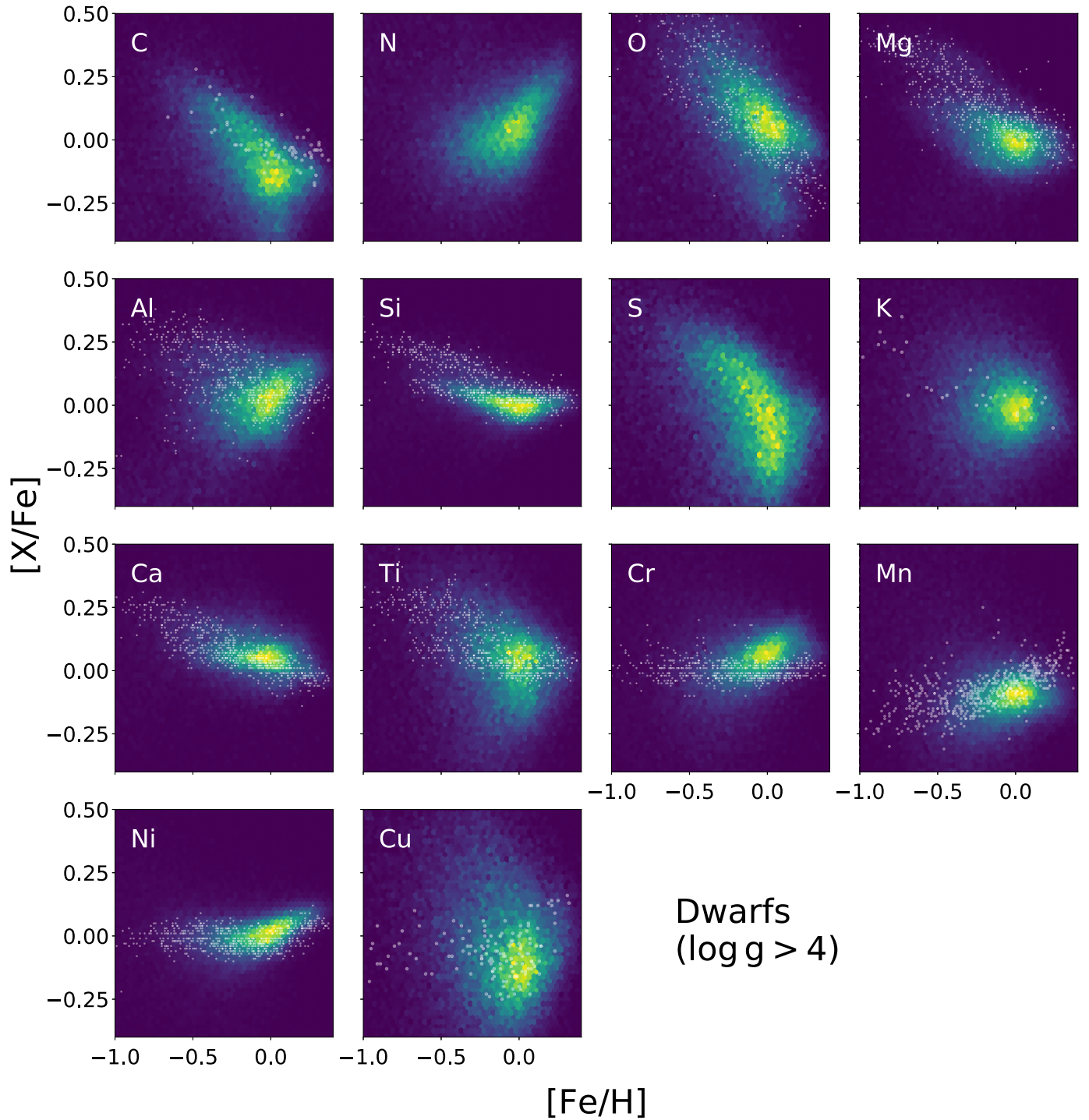
Figure 13. The same as Figure 12, but for dwarf stars. The dwarf abundances estimated by *The Payne* are consistent with the giant abundances with a few exceptions. The C and N abundances for the dwarfs are expected to differ from those for the giants due to convective dredge-up. The Al, Si, K, Mn, and Ni abundances for the dwarfs agree better with the literature values, suggesting that the discrepancies seen in the giants are mostly related to the spectral model. On the other hand, the distribution of Cr abundances among the dwarfs favors an upward trend, and the Ti and S distributions have a larger spread than among the giants. The precision for the dwarfs can also be affected by unresolved binaries, which could constitute a large fraction of the dwarf stars in APOGEE, and might explain the marginally larger abundance spread for the dwarfs. Most dwarf stars are in the low-α sequence because they are, on average, closer than the giants.

encouraging and might suggest that the discrepancies between Figures 12 and 13 are partially due to the imperfect spectral models, or they could also be astrophysically related, such as atomic diffusion in dwarf stars (Dotter et al. 2017). Interestingly, *The Payne* produces upward trends for both Cr and Mn for dwarfs, and thus the dwarf Mn abundances agree with the literature values but the Cr abundances do not. Disentangling

the discrepancies in Cr and Mn requires a careful investigation of the line list, which we will postpone to future studies.

Finally, the dwarf abundances as illustrated in Figure 13 show a marginally larger spread than the giants, suggesting that the precision for the dwarf stars might be inferior to that for the giant stars. This might not be surprising, because a large fraction of main-sequence stars could be unresolved binaries.
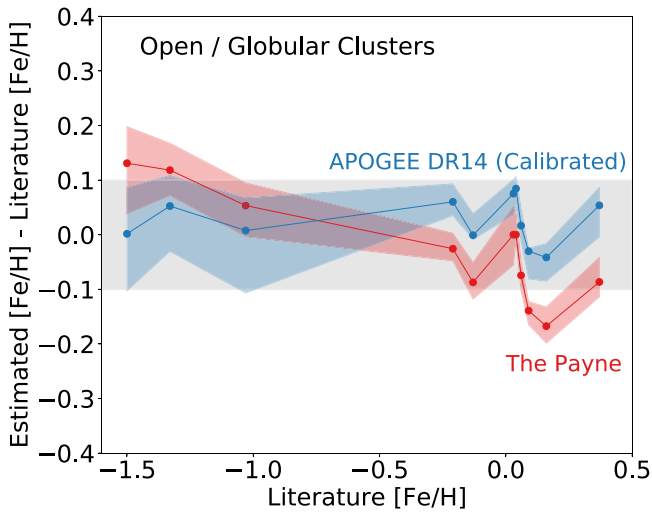
Figure 14. Comparison between literature, *The Payne*, and APOGEE DR14 [Fe/H] abundances for open and globular clusters. We compare *The Payne* [Fe/H] estimates to the literature values of 11 known clusters in APOGEE (in ascending metallicity order, with the numbers of cluster members in APOGEE in square brackets: M3 [73], M5 [103], M107 [18], NGC 2158 [10], NGC 2420 [9], Pleiades [21], NGC 188 [5], M67 [24], NGC 7789 [5], NGC 6819 [30], NGC 6791 [23]) that have [Fe/H] > −1.5, the metallicity where our training set truncates. The solid lines show the median deviation of metallicity, and the shaded areas show the 1σ range from all cluster members. Without any calibration, in the range −1 < [Fe/H] < 0 where most of the APOGEE data reside, *The Payne* derives accurate metallicities that are consistent with the literature values to 0.05 dex. There is a bias of 0.1 dex for the more metal-rich and metal-poor ends. APOGEE DR14 does not show any global trend because the calibrated values from APOGEE plotted here are calibrated against these clusters.

Fitting a single-star model to binaries can affect the precision (El-Badry et al. 2018a). Finally, in a companion paper, El-Badry et al. (2018b) adopted the dwarf abundances in this study to build a data-driven model and successfully fit for the spectra of unresolved binaries, indirectly verifying that the dwarf stellar parameters and metallicities in this study are internally consistent and robust.

3.5. Testing The Payne with Open and Globular Cluster Data

In this section we explore the stellar labels derived with *The Payne* for stars in open and globular clusters with APOGEE spectra. These stars serve as strong tests of *The Payne* owing to extensive literature data and also to the fact that open clusters are believed to be at least approximately chemically homogeneous. The latter fact allows us to empirically test the measurement precision of *The Payne* and also to test for any systematic behaviors in the derived labels as a function of, e.g., T_{eff}.

3.5.1. The Metallicity Scale

In Figure 14, we compare [Fe/H] from *The Payne* with the literature values for 11 known clusters (both open and globular) with more than five identified members in APOGEE and with metallicity [Fe/H] > −1.5, where our training set truncates. The open cluster members in APOGEE are identified in Mészáros et al. (2013). We adopt the median of all members of individual clusters to be the estimate of the cluster metallicity, and the shaded regions show the 1σ metallicity range of all cluster members. Plotted are the differences of *The Payne* and

the APOGEE calibrated metallicity estimates compared to the literature values. By definition, the APOGEE metallicities show no global trend because they are calibrated against these literature values. The deviations of estimates from *The Payne* show a weak dependence on metallicity. The trend is similar to the deviations of APOGEE metallicity before calibration. In fact, this behavior is likely traced back to the T_{eff}–metallicity biases that we see in Figure 8. As the origin of these discrepancies is unclear, we choose not to calibrate our T_{eff} to the APOGEE scale. While we do not conform to the standards, as we have discussed in Section 3.2 and in various accuracy tests throughout the paper, the APOGEE–*Payne* scale seems to be more consistent with the MIST isochrone models.

Interestingly, going beyond the global trend, the APOGEE estimates and *The Payne* estimates show similar relative offsets across various clusters. Since APOGEE and this study adopt very different methods (including different line lists), this suggests that the local correlated deviations from the literature values may be due to the difference between optical spectroscopy (literature values) and *H*-band spectroscopy (APOGEE spectra). Finally, there is a discrepancy in metallicity, but since this is related to temperature, as shown in Figures 12 and 13, it does not affect the study of [X/Fe] since the differences in the two abundances roughly cancel out, as they are both caused by the differences in T_{eff}.

3.5.2. Testing the Abundances

In Figure 15, we show the [X/H] –T_{eff} trend of three largest open clusters in APOGEE. Open clusters are found to be very chemically homogeneous (Bovy 2016; Ness et al. 2018). Therefore, apart from secondary effects such as dredge-up and atomic diffusion (Dotter et al. 2017), their chemical abundances should be independent of their evolutionary state, and hence T_{eff}. This property is usually used to calibrate out any systematic behavior of [X/H] with T_{eff}. As shown in Figure 15, *The Payne* estimates have no significant [X/H]–T_{eff} trend for any of the three clusters, showing that our abundances display no strong systematic error as a function of T_{eff}. However, we caution that 95% of the members of these three clusters are giants. More follow-up studies of dwarf stars in these open clusters are therefore needed to test the stellar labels in the dwarf regime.

Furthermore, as discussed Section 3.3, the C and N abundances of stars are sensitive to stellar ages. Since open clusters have well established ages, they can also be used to check the accuracy of our C to N abundances. In Figure 16, we show the [C/N] ratios of the same three open clusters: NGC 6819 (2.5 Gyr, e.g., Kalirai et al. 2001; Anthony-Twarog et al. 2014), M67 (4 Gyr, e.g., Richer et al. 1998; Sarajedini et al. 2009), and NGC 6791 (8 Gyr, e.g., Grundahl et al. 2008). The top panels in Figure 16 show the measurements from *The Payne*, and the bottom panels show the calibrated abundances from APOGEE DR14. Overplotted are the predictions from the MIST isochrones, taking into account the metallicities of each cluster—[Z/H] = 0 for NGC 6819 and M67; [Z/H] = 0.25 for NGC 6791. The thick black dashed line in each panel shows the MIST prediction for individual clusters given their corresponding stellar ages. As shown, *The Payne* [C/N] ratios agree better with the isochrones, and there is less spread, indicating that our C to N abundances are likely more accurate.
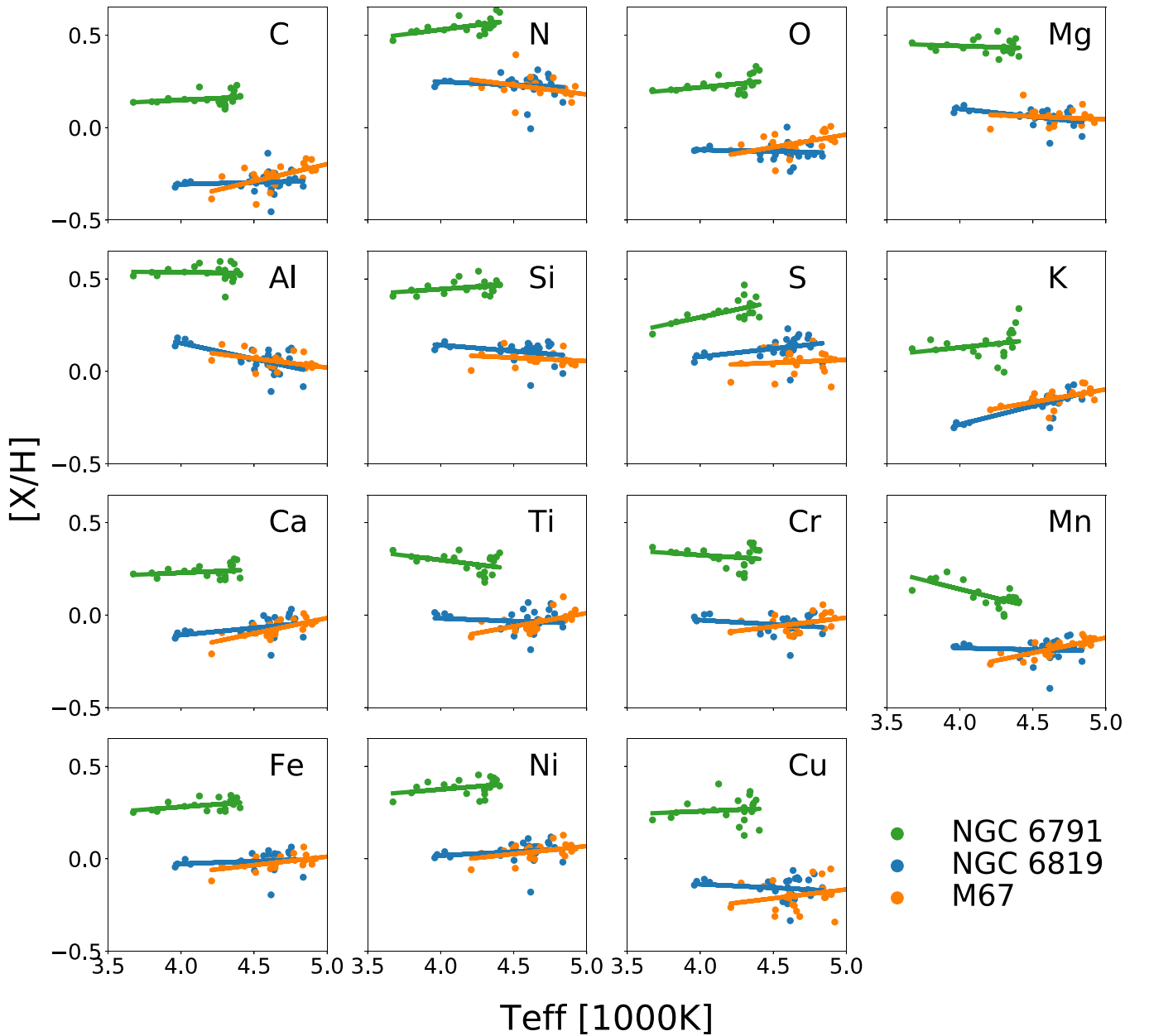
Figure 15. *The Payne* derives elemental abundances without any significant $T_{\rm eff}$ dependence. Elemental abundances of the members of three open clusters—NGC 6819, M67, and NGC 6791—are plotted against their $T_{\rm eff}$. We also overplot the best-fit linear regression in every case, merely to guide the eye. Accurate elemental abundances should show no $T_{\rm eff}$ dependence: for the vast majority of these cluster members, there is no systematic trend of abundances with $T_{\rm eff}$. *The Payne* estimates do not require external abundance–$T_{\rm eff}$ calibrations within this temperature range.

3.5.3. Abundance Precision

Figure 17 shows the elemental abundance dispersion of three open clusters discussed in the previous section. Since open clusters are chemically homogeneous, at least to the level of 0.03 dex (Bovy 2016; Liu et al. 2016), their elemental abundance dispersion gives an independent estimate of the measurement precision. Figure 17 demonstrates that *The Payne* obtains a precision of 0.03 dex for almost all elemental abundances, more precise than APOGEE DR14 calibrated values, especially at the metal-rich end (NGC 6791). We caution, however, that the precision achieved for individual stars clearly depends on their stellar parameters. The open clusters only probe precision at the metal-rich end. Interestingly, we found that fitting C_{12}/C_{13} is the key to getting more

precise abundances at the metal-rich end, presumably also due to a higher contribution from C_{12}/C_{13}, especially for the members of NGC 6791, which are, on average, cooler than stars in the other two clusters (see Figure 15). This might be the reason why APOGEE DR14 is performing somewhat worse in precision at the metal-rich end.

In this test of cluster precision, we only consider cluster members that have median S/N $= 100-300$, the typical S/N of the global APOGEE sample. About 80% of the APOGEE sample has S/N > 100. The black solid line shows the Cramer–Rao bound for a typical APOGEE K-giant with $S/N_{\rm pix} = 200$, i.e., the best precision one could achieve in principle if there is no systematics from spectral models and interpolation (see Ting et al. 2017a, for a detailed discussion on the Cramer–Rao bounds). When calculating the Cramer–Rao
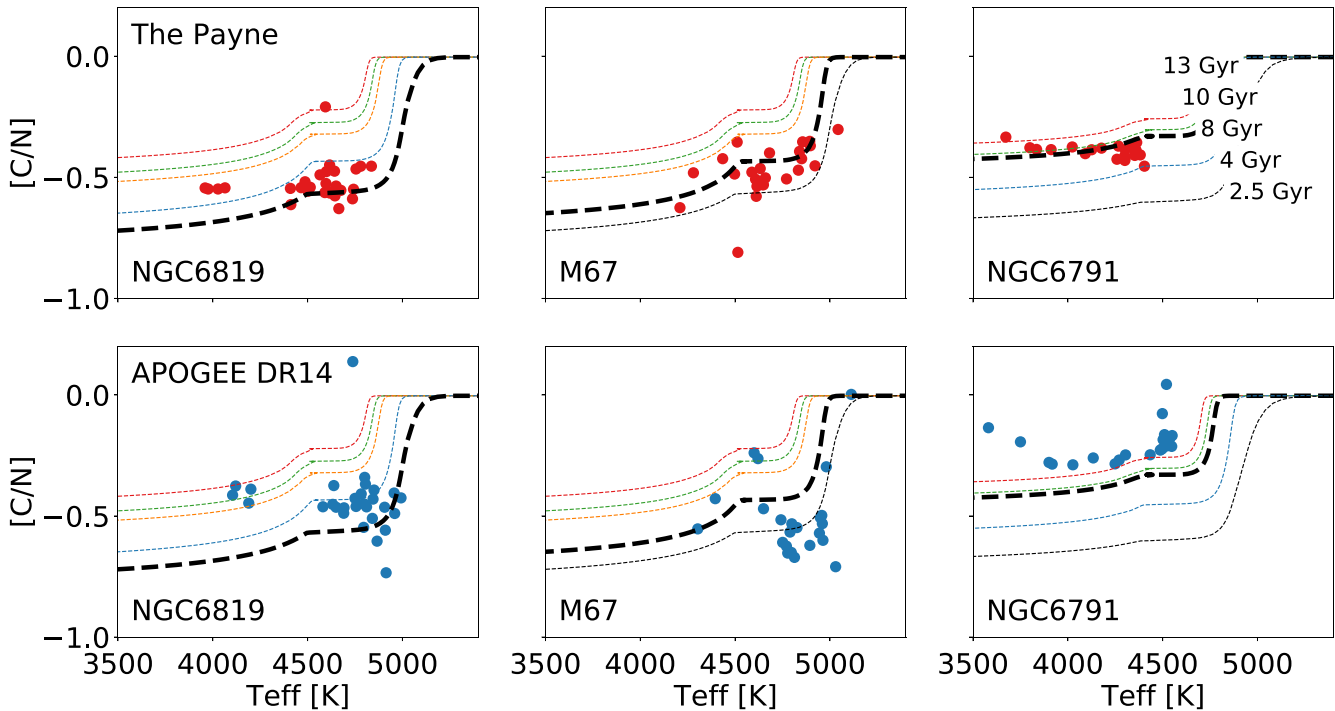
Figure 16. *The Payne* derives [C/N] abundance ratios that are astrophysically consistent with open cluster ages. We consider three open clusters with different stellar ages: NGC 6819 (2.5 Gyr), M67 (4 Gyr), and NGC 6791 (8 Gyr). Overplotted are the MIST stellar evolution models for different stellar ages. When evaluating the stellar evolution models, we also take into account the metallicities of these clusters: [Z/H] = 0 for NGC 6819 and M67, and [Z/H] = 0.25 for NGC 6791. The thick black dashed line in each panel shows the expected C/N ratio as a function of T_{eff} at these cluster ages and metallicities. In the ideal case, the data should follow a tight 1D sequence predicted by the models. The [C/N] ratio changes before and as the stars evolve up the giant branch due to convective dredge-up. The top panels show the [C/N] estimates from *The Payne*, and the lower panels demonstrate the calibrated values from APOGEE DR14. Without any calibration, the [C/N] ratios of *The Payne* agree with the MIST stellar evolution model very well, suggesting that *The Payne* [C/N] estimates are not only precise, they are also accurate.

bounds, we assume the APOGEE LSF as well as the same spectroscopic mask that we impose on the real data. We caution that, while this should mimic the instrumental effect and bad telluric regions, there might be other minor instrumental/ observation effects that are not being accounted for in the Cramer–Rao bound. *The Payne* allows us to get closer to the Cramer–Rao bound, but we are not yet reaching this fundamental limit.

We also tested how our achieved precision varies as a function of S/N by adding noise to the spectra from cluster members, and found that the achieved precision is not very sensitive to S/N. The precision consistently hovers around 0.03–0.05 for S/N > 50, and only grows as $(S/N)^{-1}$ at S/N < 50. Almost all APOGEE spectra have S/N > 50. This is also consistent with both the theoretical expectation and previous empirical studies (Casey et al. 2016; Ness et al. 2016; Ting et al. 2017a), which have demonstrated that spectra are generally rich in information. Even at S/N ∼ 50, precise abundances can be readily achieved through full spectral fitting.

However, why there is a precision ceiling of ∼0.03 dex at higher S/N is unclear. This result is in line with previous studies (Bovy 2016; Liu et al. 2016; Ness et al. 2018), illustrating that open clusters are indeed chemically homogeneous to the level of at least 0.03 dex. The limits derived in Bovy (2016) are plotted as black dashed lines as a reference. These previous studies arrive at this conclusion from either employing a statistical argument (Bovy 2016), a data-driven approach (Ness et al. 2018), or a more careful line-by-line differential analysis (Liu et al. 2016), while our result is based on direct full spectral fitting of physical spectral models to the

data. It is interesting that we are not attaining the Cramer–Rao bound. Some argue that open clusters have intrinsic chemical spreads (Liu et al. 2016) and are inhomogeneous at this level. This might well be the reason we are not reaching the best limit. But we also note that, due to imperfections in the spectral model and interpolation, it is possible that the spread we are measuring is due to systematic errors. A further improvement of *The Payne* will hopefully shed more light on the chemical inhomogeneity of open clusters.

3.6. A Catalog of Stellar Labels for APOGEE DR14 Stars from The Payne

We present all stellar labels (T_{eff}, log g, v_{micro}, v_{macro}, C_{12}/C_{13}, and 15 elemental abundances) in this study in an electronic form with this paper. The catalog is summarized in Table 2. We remove duplicated stars in the APOGEE DR14 catalog and exclude stars that have determined stellar labels that are close to the T_{eff}, log g, or [Fe/H] boundaries of our training set; we only present stars that have 3050 K < T_{eff} < 7950 K, 0 < log g < 5 and −1.45 < [Fe/H] < 0.45. We also further exclude dwarf stars that have T_{eff} < 4000 K because, as shown in Figure 7, our current models cannot determine stellar labels reliably for dwarf stars cooler than this temperature. This leaves a total of 222,707 stars in our catalog.

We caution that in this catalog we keep stars that have large χ_R^2 in the fitting for completeness, but we recommend readers only use stars that show "good" in the "quality_flag" column. This flag excludes all stars with $\chi_R^2 > 50$, a fiducial cut we adopt in this study. It also excludes fast rotators with $v_{macro} > 20$ km s^{-1} (mostly hot stars with $T_{eff} > 6000$ K).
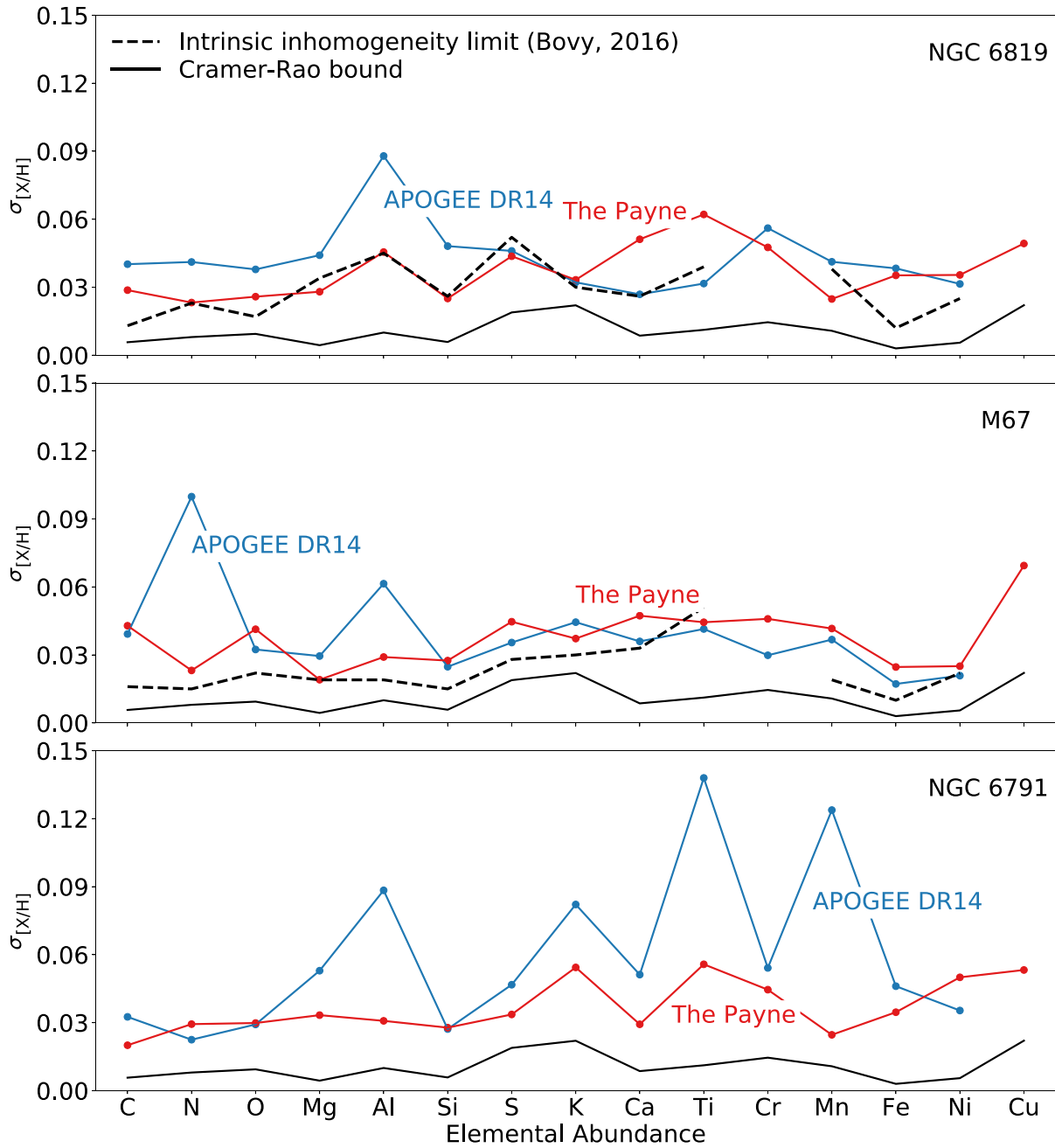
Figure 17. Empirical estimate of elemental abundance precision assuming intrinsically chemically homogeneous star clusters. To have a precision that is more representative of the global sample, we only consider cluster members that have median S/N = 100–300. About 80% of the APOGEE sample has S/N > 100. *The Payne* delivers abundance precision at the ≃0.03 dex level, which is comparable to or in some cases more precise than the APOGEE DR14 precision. The three panels show the variance of various elemental abundance estimates, $\sigma_{[X/H]}$, among members of three open clusters: NGC 6819, M67, and NGC 6791. The black dashed lines indicate the upper limits on the chemical inhomogeneity of these clusters, inferred in a data-driven approach by Bovy ([2016]). The black solid lines show the Cramer–Rao bound of elemental abundances for a typical K-giant, illustrating the theoretical limit on the precision for APOGEE spectra with S/N = 200.

We found that some rapidly rotating stars yield unreliable abundance patterns. But this is expected because here we do not properly account for stellar rotation, $v \sin i$, and our training grid truncates at $v_{\mathrm{macro}} = 30 \, \mathrm{km \, s^{-1}}$, a broadening that is still too small for typical fast rotators. We will explore the inclusion of rapid stellar rotation in the future.

4. Discussion

The Payne provides a straightforward way to perform full spectral fitting with a minimal number of spectral models required; in our case, we only generated 2000 synthetic

ab initio spectra for 25 stellar labels. *The Payne* does not require a boutique spectroscopic mask (e.g., APOGEE/ASP-CAP, García Pérez et al. [2016]), but only a simple spectroscopic mask, constructed algorithmically from the comparison of the synthetic and observed spectra of two standard stars. This appears to be sufficient to attain stellar labels that are more precise and broadly consistent with stellar evolution models. But it is important to emphasize that the main goal of this paper is to lay out this new fitting methodology, using APOGEE merely as a sample application. There are several limitations in the current APOGEE–*Payne* catalog.

Table 2
APOGEE–*Payne* Catalog: APOGEE Stellar Labels Determined with *The Payne*

APOGEE ID	R.A. [deg]	Decl. [deg]	$T_{\rm eff}$ (K)	$\log g$	$v_{\rm micro}$ (km s^{-1})	[C/H]	[N/H]	[O/H]	[Mg/H]	[Al/H]
2M00000233+1452324	0.00975	14.87567	4809.8	4.42	1.62	0.21	0.59	0.33	0.37	0.54
2M00000317+5821383	0.01323	58.36065	3744.9	0.95	1.68	−0.44	−0.05	−0.29	−0.10	−0.12
2M00000662+7528598	0.02762	75.48329	7322.1	4.11	2.05	0.05	0.05	0.04	−0.23	−0.27
2M00011143+6240187	0.29765	62.67188	3839.4	1.08	1.77	−0.48	0.26	−0.25	−0.08	−0.03
...	...	...	...	...	...	...	...	...	...	...

APOGEE ID	[Si/H]	[S/H]	[K/H]	[Ca/H]	[Ti/H]	[Cr/H]	[Mn/H]	[Fe/H]	[Ni/H]	[Cu/H]
2M00000233+1452324	0.36	0.14	0.38	0.39	0.37	0.46	0.25	0.35	0.44	0.28
2M00000317+5821383	−0.12	−0.21	−0.37	−0.34	−0.20	−0.23	−0.39	−0.28	−0.20	−0.38
2M00000662+7528598	0.13	0.02	−0.31	−0.05	−0.31	−0.01	−0.07	−0.13	−0.28	0.91
2M00011143+6240187	0.03	0.03	−0.21	−0.23	−0.08	−0.19	−0.28	−0.15	−0.17	−0.27
...	...	...	...	...	...	...	...	...	...	...

APOGEE ID	C_{12}/C_{13}	$v_{\rm macro}$ (km s^{-1})	$\chi^2_{\rm R}$	Quality Flag
2M00000233+1452324	51.5	1.02	9.9	good
2M00000317+5821383	12.5	0.96	323.9	chi2_R > 50
2M00000662+7528598	79.6	29.99	24.7	vmac > 20 km s^{-1}
2M00011143+6240187	39.0	23.57	280.7	vmac > 20 km s^{-1}; chi2_R > 50
...	...	...	...	...

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

Despite the improvement going beyond the quadratic models and a small median interpolation error of 0.1%, the interpolation error can be larger than 1% in some extreme cases (see Section 2.5), and can still prohibit obtaining absolute abundances to a level of better than 0.05–0.1 dex, especially for the cooler stars. Elements with only very weak and blended features may be more susceptible to the interpolation error, and the absolute abundances for individual stars could be biased up to 0.2 dex. Another limitation of this catalog is that we do not fit for stellar rotation, $v \sin i$, but rather adopt $v_{\rm macro}$ as an approximation. We found that for some hot stars with $T_{\rm eff} \gtrsim 6500$ K, their $v_{\rm macro}$ values reach the boundary ($v_{\rm macro} = 30$ km s^{-1}) of our training set and exhibit seemingly spurious abundance patterns. We create a flag in the catalog for these fast rotators and defer a proper accounting of $v \sin i$ to future studies; ultimately, this can just be another (two) labels to fit. Furthermore, as discussed in Section 3.2, there is a 100 K inconsistency between our spectroscopic $T_{\rm eff}$ and external photometric $T_{\rm eff}$, which appears to favor more metal-poor estimates at the high-metallicity end and more metal-rich estimates at the low-metallicity end, with a discrepancy up to 0.1 dex. The reason for this discrepancy in unknown, but it seems to agree with the APOGEE uncalibrated $T_{\rm eff}$. It thus calls for a more careful analysis of the spectral models adopted in this study and H-band spectral models in general. We also truncate our training set at [Fe/H] = −1.5 and do not analyze metal-poor globular clusters or halo stars. Further, we do not fit for unresolved stellar binaries that might affect the abundances for dwarf stars. Such an analysis was done separately in El-Badry et al. (2018b) using *The Payne*. This illustrates that *The Payne* is not a panacea for stellar spectroscopy—it is only a new methodological framework, and we mention a few areas for future improvement below.

Our analysis of the APOGEE data also made simplifying assumptions about the experimental set-up. First, we convolve all synthetic spectra with a fixed averaged LSF template from APOGEE, assuming that the averaged LSF is an accurate representation for all APOGEE spectra. This is not the case because the instrumental dispersion can vary from fiber to fiber and observation to observation. In this application of *The Payne*, we do not fit for the LSF, but use $v_{\rm macro}$ as a free parameter instead to compensate part of the LSF variation. Second, we normalize synthetic spectra in the same way as we normalize observed spectra; but even with a self-consistent normalization, the normalization scheme can be still problematic at low temperatures. In particular, Ness et al. (2015) derived a set of "continuum pixels" for APOGEE giant spectra with $T_{\rm eff} \simeq 4000$–5000 K. At lower temperatures, this set of pixels that we adopt might no longer be valid reference points, and the systematics between models and observations can still skew the continuum. In the long run, fitting the LSF and continuum along with the stellar labels might mitigate some of the remaining systematics seen in this study.

The success of *The Payne* relies on further key ingredients. One is a robust spectroscopic mask, which we derived here from only the Sun and Arcturus. It will be crucial to have a set of standard stars that all large-scale spectroscopic surveys will observe: due to the subtle combined effect from the instrumental profile, telluric sky lines, and normalization, we found that a robust mask must be made based on observations from the same instrument. It is, for example, not sufficient to make a spectroscopic mask using the highest resolution FTS spectra and convolve the mask to the observable space. Second, *The Payne* must rely on ab initio spectral models that span a broad range of the $T_{\rm eff}-\log g-$[Fe/H] space. Again, a limitation of *The Payne*'s current application derives from the fact that the line list underlying its ab initio models is only calibrated to two stars, the Sun and Arcturus. Models for cooler stars ($T_{\rm eff} < 4000$ K) and more metal-poor stars therefore remain problematic. One future step will be extending the calibration of the line list beyond the Sun and Arcturus as well as constructing a spectroscopic mask beyond using these two stars. It is also essential to explore other more sophisticated options, such as 1D or 3D NLTE models. With *The Payne*,

only a few hundreds or thousands of spectral models are needed to fit for all stellar labels. Therefore, self-consistent NLTE elemental abundances should be possible with *The Payne* (see M. Kovalev et al. 2019, in preparation).

5. Summary

We present *The Payne* as a new framework for fitting stellar spectra with physical models that allows for the simultaneous determination of numerous stellar labels. The key ideas and components of *The Payne* are summarized below:

1. *The Payne* builds generative models that predict spectral fluxes at each pixel for any set of stellar labels, based on a modest discrete set of physical, ab initio model spectra. Only $\mathcal{O}(1000)$ suitable ab initio spectra are needed to create a model for fitting $N_{dim} = 25$ labels. This opens up new avenues for determining stellar labels drawing on computationally expensive synthetic models, such as 3D NLTE models.
2. The generative models for each pixel are built with neural-network-like function, a flexible extension of the quadratic models.
3. At any combination of stellar parameters and abundances, the underlying spectral models for the training step consistently calculate the structure of the atmosphere and the radiative transfer, to yield the ab initio spectra.
4. An auto-refinement technique iteratively determines the points in stellar label space where ab initio spectra are needed to construct a sufficiently high-quality generative model.
5. For fitting of actual survey spectra, we apply a straightforward and algorithmic way to construct a pixel mask that excises wavelengths where observed spectra systematically and prominently differ from the model predictions.
6. The code for *The Payne* is minimal and transparent, and the fitting is very efficient. Fitting 25 labels simultaneously to an APOGEE spectrum, and accounting for radial velocity at the same time, takes no more than a CPU second per spectrum. One significant advantage of this speed is the ability to explore many detailed options and choices in the fitting.

The Payne is a general method that can be applied to any set of synthetic models and survey data. As a sample application, we applied *The Payne* to the full APOGEE DR14 catalog. By directly fitting all stellar labels simultaneously, *The Payne* delivers state-of-the-art results that appear both accurate and precise, without any external calibration. We demonstrate this by comparing the APOGEE labels derived from *The Payne* to APOGEE DR14 results and subjecting them to a variety of physical plausibility tests, illustrating the following points:

1. Fitting all data only with a single generative model, we obtain stellar parameters (T_{eff}, $\log g$, [Fe/H]) for both giants and dwarfs that are broadly consistent with external constraints, although modest [Fe/H]-dependent systematic offsets are evident when comparing to photometric temperatures and asteroseismic $\log g$ values.
2. We obtain detailed elemental abundances for APOGEE dwarf stars, which had not been done within APOGEE before. Broadly, the abundance patterns we see in the dwarfs are consistent with those in the sample of giants.

3. The derived elemental abundances for stars within open clusters show no systematic trends with T_{eff}.
4. We resolve the problem of the "peculiar" [Ti/Fe]–[Fe/H] abundance distribution that has been persistent in APOGEE data releases and obtain a physically plausible abundance pattern.
5. Our derived [C/N] ratios agree well with predictions for changes in surface abundance due to dredge-up in stellar evolution models. The agreement between stellar evolution models and our [C/N] ratios vividly demonstrates that it is possible to infer stellar ages directly from ab initio spectroscopic labels.
6. We achieve an elemental abundance precision of ∼0.03 dex for all elements. On this basis, we confirm that open clusters are chemically homogeneous to a level of at least 0.03 dex.

The Payne offers a very powerful framework to perform accurate, efficient comparison of large data sets to sophisticated spectral models. Work is currently underway to improve the training of *The Payne* among the cool giants, to employ NLTE spectral models, and to apply *The Payne* to a diverse variety of existing spectroscopic and photometric data sets.

ORCID iDs

Yuan-Sen Ting (丁源森) ⓘ https://orcid.org/0000-0001-5082-9536
Charlie Conroy ⓘ https://orcid.org/0000-0002-1590-8551
Hans-Walter Rix ⓘ https://orcid.org/0000-0003-4996-9069

References

Adibekyan, V. Z., Sousa, S. G., Santos, N. C., et al. 2012, A&A, 545, A32
Anthony-Twarog, B. J., Deliyannis, C. P., & Twarog, B. A. 2014, AJ, 148, 51
Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, ARA&A, 47, 481
Battistini, C., & Bensby, T. 2015, A&A, 577, A9
Bensby, T., Feltzing, S., & Oey, M. S. 2014, A&A, 562, A71
Blanco-Cuaresma, S., Soubiran, C., Heiter, U., & Jofré, P. 2014, A&A, 569, A111
Boeche, C., & Grebel, E. K. 2016, A&A, 587, A2
Boeche, C., Siebert, A., Williams, M., et al. 2011, AJ, 142, 193

Bovy, J. 2016, ApJ, 817, 49
Casey, A. R., Hogg, D. W., Ness, M., et al. 2016, arXiv:1603.03040
Choi, J., Dotter, A., Conroy, C., et al. 2016, ApJ, 823, 102
Choi, J., Dotter, A., Conroy, C., & Ting, Y.-S. 2018, ApJ, 860, 131
Cirasuolo, M., Afonso, J., Carollo, M., et al. 2014, Proc. SPIE, 9147, 91470N
Dafonte, C., Fustes, D., Manteiga, M., et al. 2016, A&A, 594, A68
Dalton, G., Trager, S., Abrams, D. C., et al. 2016, Proc. SPIE, 9908, 99081G
de Jong, R. S., Barden, S., Bellido-Tirado, O., et al. 2014, Proc. SPIE, 9147, 91470M
De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, MNRAS, 449, 2604
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
Dotter, A., Conroy, C., Cargile, P., & Asplund, M. 2017, ApJ, 840, 99
El-Badry, K., Rix, H.-W., Ting, Y.-S., et al. 2018a, MNRAS, 473, 5043
El-Badry, K., Ting, Y.-S., Rix, H.-W., et al. 2018b, MNRAS, 476, 528
Fabbro, S., Venn, K. A., O'Briain, T., et al. 2018, MNRAS, 475, 2978
García Pérez, A. E., Allende Prieto, C., Holtzman, J. A., et al. 2016, AJ, 151, 144
González Hernández, J. I., & Bonifacio, P. 2009, A&A, 497, 497
Grundahl, F., Clausen, J. V., Hardis, S., & Frandsen, S. 2008, A&A, 492, 171
Ho, A. Y. Q., Rix, H.-W., Ness, M. K., et al. 2017, ApJ, 841, 40
Holtzman, J. A., Shetrone, M., Johnson, J. A., et al. 2015, AJ, 150, 148
Kalirai, J. S., Richer, H. B., Fahlman, G. G., et al. 2001, AJ, 122, 266
Kollmeier, J. A., Zasowski, G., Rix, H.-W., et al. 2017, arXiv:1711.03234
Kurucz, R. L. 1970, SAOSR, 309, 291
Kurucz, R. L. 1993, in SYNTHE Spectrum Synthesis Programs and Line Data, ed. R. L. Kurucz (Cambridge, MA: Smithsonian Astrophysical Observatory)
Kurucz, R. L. 2005, MSAIS, 8, 14
Kurucz, R. L. 2013, ATLAS12: Opacity sampling model atmosphere program, Astrophysics Source Code Library, ascl:1303.024
Kurucz, R. L. 2017, ATLAS9: Model atmosphere program with opacity distribution functions, Astrophysics Source Code Library, ascl:1710.017
Kurucz, R. L., & Avrett, E. H. 1981, SAOSR, 391, 139

Leung, H. W., & Bovy, J. 2019, MNRAS, 483, 3255
Liu, F., Yong, D., Asplund, M., Ramírez, I., & Meléndez, J. 2016, MNRAS, 457, 3934
Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA, 15, 1095
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Martig, M., Fouesneau, M., Rix, H.-W., et al. 2016, MNRAS, 456, 3655
Mészáros, S., Holtzman, J., García Pérez, A. E., et al. 2013, AJ, 146, 133
Mishenina, T. V., Gorbaneva, T. I., Basak, N. Y., Soubiran, C., & Kovtyukh, V. V. 2011, ARep, 55, 689
Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16
Ness, M., Hogg, D. W., Rix, H.-W., et al. 2016, ApJ, 823, 114
Ness, M., Rix, H., Hogg, D. W., et al. 2018, ApJ, 853, 198
Nissen, P. E., Chen, Y. Q., Carigi, L., Schuster, W. J., & Zhao, G. 2014, A&A, 568, A25
Ramírez, I., & Allende Prieto, C. 2011, ApJ, 743, 135
Recio-Blanco, A., de Laverny, P., Allende Prieto, C., et al. 2016, A&A, 585, A93
Richer, H. B., Fahlman, G. G., Rosvick, J., & Ibata, R. 1998, ApJL, 504, L91
Rix, H.-W., Ting, Y.-S., Conroy, C., & Hogg, D. W. 2016, ApJL, 826, L25
Sarajedini, A., Dotter, A., & Kirkpatrick, A. 2009, ApJ, 698, 1872
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Smiljanic, R., Korn, A. J., Bergemann, M., et al. 2014, A&A, 570, A122
Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645
Ting, Y.-S. 2019, tingyuansen/The_Payne, v1.0, Zenodo, doi:10.5281/zenodo.2949901
Ting, Y.-S., Conroy, C., & Goodman, A. 2015, ApJ, 807, 104
Ting, Y.-S., Conroy, C., & Rix, H.-W. 2016, ApJ, 826, 83
Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2017a, ApJ, 843, 32
Ting, Y.-S., Rix, H.-W., Conroy, C., Ho, A. Y. Q., & Lin, J. 2017b, ApJL, 849, L9
Yu, J., Huber, D., Bedding, T. R., et al. 2018, ApJS, 236, 42
Zhao, G., Mashonkina, L., Yan, H. L., et al. 2016, ApJ, 833, 225