CrossMark

APOGEE Net: Improving the Derived Spectral Parameters for Young Stars through Deep Learning

Richard Olney[1], Marina Kounkel[2] , Chad Schillinger[1], Matthew T. Scoggins[2] , Yichuan Yin[1], Erin Howard[2], K. R. Covey[2] ,
Brian Hutchinson[1,3], and Keivan G. Stassun[4]

[1] Dept. of Computer Science, Western Washington University, 516 High St., Bellingham, WA 98225-9165, USA
[2] Dept. of Physics and Astronomy, Western Washington University, 516 High St., Bellingham, WA 98225-9164, USA
[3] Computing and Analytics Division, Pacific Northwest National Laboratory, 902 Battelle Blvd., Richland, WA 99354-1793, USA
[4] Department of Physics and Astronomy, Vanderbilt University, VU Station 1807, Nashville, TN 37235, USA

Abstract

Machine learning allows for efficient extraction of physical properties from stellar spectra that have been obtained by large surveys. The viability of machine-learning approaches has been demonstrated for spectra covering a variety of wavelengths and spectral resolutions, but most often for main-sequence (MS) or evolved stars, where reliable synthetic spectra provide labels and data for training. Spectral models of young stellar objects (YSOs) and low-mass MS stars are less well-matched to their empirical counterparts, however, posing barriers to previous approaches to classify spectra of such stars. In this work, we generate labels for YSOs and low-mass MS stars through their photometry. We then use these labels to train a deep convolutional neural network to predict $\log g$, T_{eff}, and Fe/H for stars with Apache Point Observatory Galactic Evolution Experiment (APOGEE) spectra in the DR14 data set. This "APOGEE Net" has produced reliable predictions of $\log g$ for YSOs, with uncertainties of within 0.1 dex and a good agreement with the structure indicated by pre-MS evolutionary tracks, and it correlates well with independently derived stellar radii. These values will be useful for studying pre-MS stellar populations to accurately diagnose membership and ages.

Unified Astronomy Thesaurus concepts: Astroinformatics (78); Computational methods (1965); Young stellar objects (1834); Low mass stars (2050); Stellar classification (1589)

Supporting material: machine-readable table

1. Introduction

Spectroscopy is a powerful technique for measuring stellar properties, and in recent years, large surveys such as Sloan Digital Sky Survey (SDSS) Apache Point Observatory Galactic Evolution Experiment (APOGEE; Abolfathi et al. 2018), RAVE (Kunder et al. 2017), and GALAH (Buder et al. 2018) have observed 10^{5-6} stars each. This necessitates an effective method to uniformly and efficiently process these spectra to extract the stellar properties (e.g., effective temperature (T_{eff}), surface gravity ($\log g$), and metallicity (Fe/H)).

A common approach to spectral analysis relies on comparisons between the target spectrum and a grid of spectral standards (that may be difficult to come by for specific source types or wavelength ranges) or synthetic templates (that may systematically differ from the real data). Synthetic templates typically offer more regular coverage of parameter space, enabling an individual targets' parameters to be inferred precisely by using a higher-order function to interpolate the goodness of fit parameters between the points for which the grid is defined. Many surveys adopt this approach in constructing their stellar parameter pipelines (e.g., García Pérez et al. 2016), but the process is computationally intensive, particularly when fitting multiple parameters simultaneously, as well as determining the corresponding uncertainties.

Computational efficiency aside, the reliability of parameters determined via direct model fitting can also vary strongly as a function of target type, with young stellar objects (YSOs) representing a particularly challenging target class (e.g., Doppmann et al. 2005). Spectral fits can return reasonably accurate estimates of T_{eff} for YSOs, but $\log g$ has proven more difficult to accurately

constrain. This parameter is particularly valuable for YSOs, as it serves as a proxy for stellar age and is, therefore, of great value for calibrating pre-MS evolutionary models or inferring star-formation histories within a given star-forming complex. The APOGEE survey has conducted extensive surveys of several nearby star-forming regions, providing a valuable opportunity to infer $\log g$ and age constraints for large samples of YSOs, but those constraints have been difficult to achieve in practice. For APOGEE data in particular, obtaining reliable $\log g$ values for dwarf or pre-MS stars have been challenging: neither APOGEE's primary stellar parameter pipeline, ASPCAP (Holtzman et al. 2015; García Pérez et al. 2016), nor the community-provided Payne model-fitting framework (Ting et al. 2019) released $\log g$ estimates for dwarf stars, due to the presence of clear systematic errors in uncalibrated values and the lack of a densely sample comparison sample for deriving calibration relations. The IN-SYNC pipeline (Cottaar et al. 2014; Kounkel et al. 2018), developed and optimized for YSO spectra, provided $\log g$ values whose age dependence agrees well with physical models (i.e., older populations have higher $\log g$ than the younger ones; see Section 4 for discussion), but the precise values also show unphysical systematics, likely due to mismatches between the empirical and theoretical spectra.

Data-driven analysis pipelines eliminate errors due to model mismatches, by training prediction systems with empirical spectra for stars with well-determined stellar parameters. One data-driven method that has demonstrated considerable success in assigning labels to APOGEE spectra is 'The Cannon' (Ness et al. 2015), which uses a reference sample of APOGEE spectra to train a generative model that can then be used to infer stellar parameters for any object with an APOGEE spectrum. Using

the full wavelength information within the spectrum, and training the parameters of the generative model on empirical standards (i.e., the high-quality APOGEE spectra with known stellar labels that eliminate the potential for model-data mis-fitting), the Cannon is able to provide parameters of comparable quality to ASPCAP's for APOGEE spectra with a signal-to-noise ratio (S/N) $\geqslant 25$. As informed by a set of 60 dwarf calibrators in the Hyades, the 2015 Cannon results also included realistic $\log g$ values for the upper main sequence (MS), but became incomplete for dwarfs with $T_{eff} < 4500$ K, similar to the limit reached by the Payne results. While data-driven models offer important performance and calibration advantages, they are unable to overcome the limits of their training sets.

Neural networks offer a promising data-driven method for inferring accurate stellar parameters from spectra and with a potentially greater flexibility in inference methods than is offered by the polynomial formalism adopted in the Cannon. Neural networks are a common machine-learning model, in which multiple nonlinear transformations of the input features are performed before assigning an output classification. A number of studies have demonstrated the ability of neural networks to classify stellar spectra and derive stellar parameters and abundances, (e.g., Bailer-Jones et al. 1997; Bailer-Jones 2000; Bazarghan & Gupta 2008; Fabbro et al. 2018; Leung & Bovy 2019; Sharma et al. 2020). Neural networks also offer important efficiencies for processing large data sets. Direct fitting can only consider a single spectrum at a time, redoing the same operation regardless of how similar two target spectra may be. On the other hand, neural networks can process in excess of 10^6 observations in under an hour. As with all data-driven methods, however, the network must be trained on a reliable reference sample whose parameters span the full range of interest, making the construction of a label-set as important as the construction of the network itself.

In this paper, we aim to train a deep neural network (DNN) to accurately classify APOGEE DR14 spectra of dwarfs, giants, and pre-MS stars. To realize this goal, we first supplement the stellar labels provided by the Payne with a set of labels inferred from *Gaia*, Two Micron All Sky Survey (2MASS), and Pan-STARRS photometry and astrometry for YSOs and M dwarfs with APOGEE spectra. In Section 2, we describe the data and the procedure used to generate the labels. In Section 3, we then construct a convolutional neural network (CNN) that predicts parameters from the APOGEE spectra, which we refer to as the APOGEE Net. In Section 4, we highlight analysis that could be derived from these spectral properties. Finally, we summarize our results and discuss the implications in Section 5.

2. Data

2.1. APOGEE

APOGEE is a high-resolution ($R \sim 22{,}500$) near-infrared (1.51–1.7 μm spectrograph mounted on the Sloan Foundation 2.5 m telescope (Gunn et al. 2006; Wilson et al. 2010; Blanton et al. 2017; Majewski et al. 2017). APOGEE is capable of observing up to 300 targets in a field of view with radius of 1°5. Over the years, the survey and its targeting priorities has evolved. The primary objective of both APOGEE-1 and APOGEE-2 has been to observe red giants to trace the dynamical and the chemical patterns of the Galaxy (e.g., Hayden et al. 2015; Bovy et al. 2016; Anders et al. 2017; Zasowski et al. 2017). However, among other

programs, it has also observed a number of star-forming regions, including Orion Complex (Da Rio et al. 2016, 2017; Kounkel et al. 2018), NGC 1333 (Foster et al. 2015), IC 348 (Cottaar et al. 2015), NGC 2264, as well as several more evolved clusters, and some of the nearby MS stars.

As of the public Data Release 14 (Abolfathi et al. 2018), over 263,000 stars in the bulge, disk, and halo have been observed. We restrict the current analysis only to these sources.

The sources in the catalog typically have been observed for multiple epochs. Therefore, the data are stored in two formats: "apVisit," which contains the raw spectrum at a particular epoch, and "apStar," in which the Doppler shift has been removed from the epochs, placing them all in a common rest frame with identical wavelength solution across all sources, and multiple visits for the same source are combined into one, increasing the resulting signal-to-noise (Nidever et al. 2015).

2.2. Payne Labels

Ting et al. (2019) used their newly developed spectral interpolator, along with a new grid of Kurucz spectral models calculated with an improved line list (Cargile et al. 2019), to identify best-fit models and infer revised stellar parameters for 222,707 spectra within the APOGEE DR14 data set. The labels inferred from this interpolator, dubbed "The Payne" in honor of Cecilia Payne-Gaposchkin's seminal work in physically-based stellar models, are comparable to the calibrated parameters provided in DR14 for stars along the Red Giant Branch. Moreover, the Payne provides realistic T_{eff} and $\log g$ labels for warmer ($T_{eff} > 4250$ K) MS stars, for which calibrated parameters are not available in the DR14 data set. Typical uncertainties, both random and systematic, are $\sim$100 K in T_{eff}, $\sim$0.1 dex in $\log g$, and $\sim$0.03 dex in abundance space (Ting et al. 2019).

However, the Payne has nonphysical correlations between $\log g$ and Fe/H toward cooler dwarfs. Indeed, these systematics dominated the information content of the Payne labels to such a degree that dwarfs stars with $T_{eff} < 4000$ K were intentionally removed from the Payne outputs to avoid potential misinterpretation of the spurious correlation between the inferred $\log g$ and [Fe/H] values.

In training APOGEE Net, we adopt labels for T_{eff}, $\log g$, and Fe/H from the Payne's outputs for stars with $T_{eff} > 4000$ K or $\log g < 3.5$. As the Payne does not yet produce reliable labels for stars with lower T_{eff} and higher $\log g$ than these limits, in the next section, we use empirically calibrated photometric relationships to generate new labels for 4480 low-mass MS stars and 2446 YSOs.

2.3. Deriving Alternate Labels for Pre-MS and Low-mass MS Stars

2.3.1. MS Stars

To derive labels of the low-mass MS stars, we rely on various empirical photometric relationships. We begin by identifying bona fide lower MS stars with potentially erroneous Payne labels, using 2MASS and *Gaia* DR2 photometry, as well as *Gaia*'s parallax measurements. We selected low-mass MS stars by requiring 2MASS photometry of $0.7 < J - K < 1.05$, $3.37 < g - K < 8.46$, and $5 < M_K < 10$. To ensure accurate M_K, we further restricted the sample to only those sources in which the *Gaia* DR2 measured $\pi/\sigma_\pi > 3$; this removed any giants with erroneous M_K values due to spurious, low-quality parallaxes. Since the empirical relations used in this section

were calibrated for MS stars, we flag all spectra within APOGEE fields covering known star-forming regions and young clusters —namely the Orion Complex, Perseus clusters, NGC 2264, and the Pleiades—for a separate label generation procedure, which is described in the next section.

We infer metallicities for these stars using the relation by Hejazi et al. (2015)

$$[\mathrm{Fe/H}] = A_1 + A_2(g-K) + A_3(J-K) + A_4(g-K)^2 \\ + A_5(J-K)^2 + A_6(g-K)(J-K) \tag{1}$$

in which

$$A = [-14.259, 0.0519, 29.5926, \\ -0.0529, -17.6762, 0.7032].$$

This calibration is derived from a sample of 73 M dwarfs with high-quality SDSS+2MASS photometry and robust metallicity estimates inferred from high-resolution spectroscopy of a common proper motion companion (18) or from moderate-resolution spectra of the M dwarf itself (53). The derived calibration is applicable for stars between K6 and M6.5, with $-0.73 \leqslant$ [Fe/H] $\leqslant +0.3$ dex, $3.3 \leqslant g-K \leqslant 8.46$ and $0.71 \leqslant J-K \leqslant$ 1.01. By design, these limits nearly exactly match the color–magnitude cuts used to select candidates for our alternate, pre-MS focused label generation procedure. We do allow a modestly broader range of $g-K$ colors, by a few hundredths of a magnitude in each direction, as the extrapolated metallicities will nonetheless likely be more accurate than the Payne parameters, which clearly suffer from systematic errors in this space.

To estimate T_{eff}, we use the relation derived from a sample of 183 M dwarfs with accurate ($\sigma_\pi < 5\%$) pre-*Gaia* parallaxes and reliable spectrophtometric data by Mann et al. (2015)

$$T_{\mathrm{eff}} = B_0 + B_1(r-J) + B_2(r-J)^2 + B_3(r-J)^3 \\ + B_4(r-J)^4 + B_5(J-H) + B_6(J-H)^2 \tag{2}$$

with

$$B = [2.151, -1.092, 0.3767, -0.06292, \\ 0.003950, 0.1697, -0.03106]$$

which hold for $4.6 < M_K < 9.8$, $2700 < T_{\mathrm{eff}} < 4100$ K, and $-0.6 < [\mathrm{Fe/H}] < 0.5$.

Finally, to estimate $\log g$, we use the relation calibrated from the analysis of *Y*-band spectra of 29 M dwarfs (Veyette et al. 2017)

$$\log g = 7.912 - 0.1880 \times [\mathrm{Fe/H}] - 1.334 \times 10^{-3} \times T_{\mathrm{eff}} \\ + 1.313 \times 10^{-7} \times T_{\mathrm{eff}}^2, \tag{3}$$

which hold for $3200 < T_{\mathrm{eff}} < 4100$ K and $-0.7 < [\mathrm{Fe/H}] < 0.3$.

2.4. YSOs

The APOGEE fields designed to target star-forming regions nonetheless include sources other than YSOs. To calculate parameters for a pure YSO sample, we restricted the sample to likely cluster members tabulated by Kounkel et al. (2019), with spectra publicly released in DR14 (Abolfathi et al. 2018).

We attempted multiple approaches to interpolate each YSO's photometry onto a grid of isochrones to generate initial T_{eff} and

$\log g$ labels (Figure 1). We considered various isochrones for this purpose, including those from Baraffe et al. (2015), PARSEC (Marigo et al. 2017), and MIST (Choi et al. 2016). We assessed combinations of various photometric bands from 2MASS, *Wide-field Infrared Survey Explorer*, and *Gaia*, and explored the ability to assign reliable stellar labels using standard isochrone fits (as implemented via the "isochrones" python package, developed by Timothy Morton).

Ultimately, we were unable to achieve astrophysically realistic parameters for the majority of YSOs in our sample using standard interpolation methods. This is due to several factors that reinforce one another:

1. *Extinction:* Many of the YSOs in our sample possess nontrivial extinctions due to the interstellar or circumstellar dust that is typically found in star-forming regions. A YSO's photometry can be plausibly explained by a degenerate set of parameters that lie along an extinction vector in a single color–magnitude space. This degeneracy can typically be broken, however, by simultaneously fitting multiband photometry to leverage differences in the slopes of the isochrones and extinction vectors in different projections of color–magnitude space. The isochrones package provides a multiband fitting capability, which we utilized to infer a maximum likelihood A_V value along with each YSO's other stellar parameters (i.e., mass and age). The quality of these estimates, and the resolution of degeneracies between extinction and stellar properties, depend on the underlying agreement between the YSO's observed colors and those expected from reddened stellar models, the factor we consider next.

2. *Systematic offsets between isochrones and empirical data:* As shown in Figure 1, many YSOs lie outside the bounds of an unreddened grid of isochrones. This offset is most prominent in *J–K*, and larger for members of the youngest clusters (i.e., $\Delta_{(J-K)} \sim 0.3$ mag. in NGC 1333, IC348, and NGC 2264), but still present in $\sim$120 Myr old stars (i.e., $\Delta_{(J-K)} \sim 0.05$ mag for Pleiades members).

   In principle, extinction may contribute to this offset but cannot explain it entirely. This is best illustrated by the *J–K* offset between the isochrones and the low-mass stars in the Pleiades, as the extinction measured toward the Pleiades is far too low to produce the necessary $E(J–K)$ excess (Stauffer et al. 1998; Schlafly & Finkbeiner 2011, $A_V = 0.12$; $E(J–K) = 0.178$). This previous work has shown that these stars have offsets with respect to synthetic colors that are inconsistent with standard reddening vectors (Bell 2012; Covey et al. 2016).

   Even worse, the youngest YSOs possess intrinsic infrared excesses due to emission by warm circumstellar dust, which are not included in standard isochrones and are inconsistent with color offsets predicted by standard extinction vectors. These astrophysically meaningful systematic offsets between isochrones and empirical photometry are not gracefully handled by standard isochrone fitting methods, which attempt to reproduce the offset with secondary model parameters (i.e., extinction), but self-consistent solutions are difficult to achieve when the offset is inconsistent with standard extinction curves and significantly larger than the YSOs photometric errors.

3. *Error Normalization and Grid Edge Effects:* The isochrones package fundamentally performs a chi-squared minimization
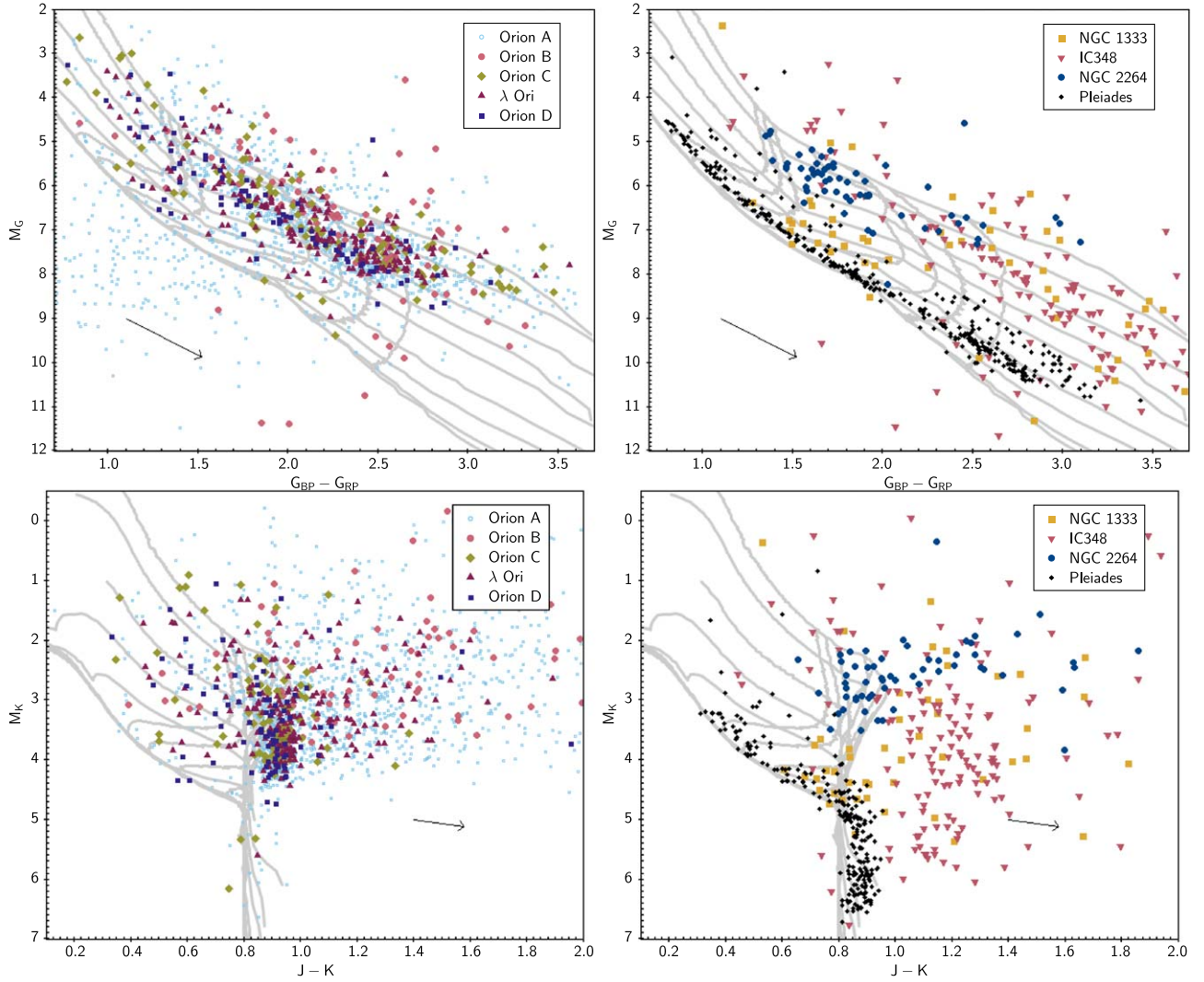
Figure 1. Color–magnitude diagrams constructed using *Gaia* and 2MASS photometry for the YSOs across the various regions. The gray lines show the isochrones at ages of 1, 2, 5, 10, 20, 50, 100, 200, and 300 Myr, and 8.5 dex, as well as the evolutionary tracks for 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1 $M_\odot$ stars from the PARSEC isochrones (Marigo et al. 2017). The black arrow shows the extinction vector corresponding to 1 A_V.

of the residuals between each YSO's photometry and that predicted for synthetically reddened stellar models across a range of physical parameters. In this process, the residuals in each band are normalized by the associated photometric errors to identify the global best-fit solution. This process works well for sources whose photometry is consistent with standard reddened stellar spectra, and where the model-data agreement is primarily disrupted by random photometric noise. As outlined above, however, many of the YSOs in our sample do not meet this criteria: the quality of their fits is fundamentally limited by the systematic offsets with respect to the data, a factor that is decoupled from the precision of their photometry. In this case, the figure of merit drives the fit to areas of parameter space that minimize systematic model offsets, particularly in the photometric bands with the highest average precision. This often drives the fit to select models with unrealistic combinations of $T_{\rm eff}$ log g and A_V, and at least one parameter (most often A_V) hitting the limits of the parameter space included in the search.

Ultimately, we were unable to overcome these limitations with the standard isochrone fitting approach. Our output labels,

for example, typically featured a strong, but nonphysical, correlation between $T_{\rm eff}$ and A_V, and these effects did not substantially diminish when we eliminated the bands with the strongest model-data systematic offsets (i.e., the 2MASS photometry) or restricted the sample to sources with lower intrinsic reddening (i.e., Class III/Weak T Tauri stars versus Class II/Classical T Tauri Stars).

As an alternative to traditional isochrone fitting, and to more gracefully fit YSOs whose photometry lies beyond the edges of the standard isochronal grid, we used a neural network trained on synthetic photometry to infer labels for YSOs across a broad range of parameter space. It proved to be less biased to systematic effects of the individual bands, rather, through assigning different weight to the inputs and treating them as a whole, it resulted in more physically realistic solutions. Specifically, we constructed a CNN with three convolutional layers using max pooling and two fully connected layers. This network was trained on parameters from synthetic stars drawn from the PARSEC isochrones (they provided a better convergence compared to other isochrones we tested). The synthetic stars were generated using a uniform distribution of
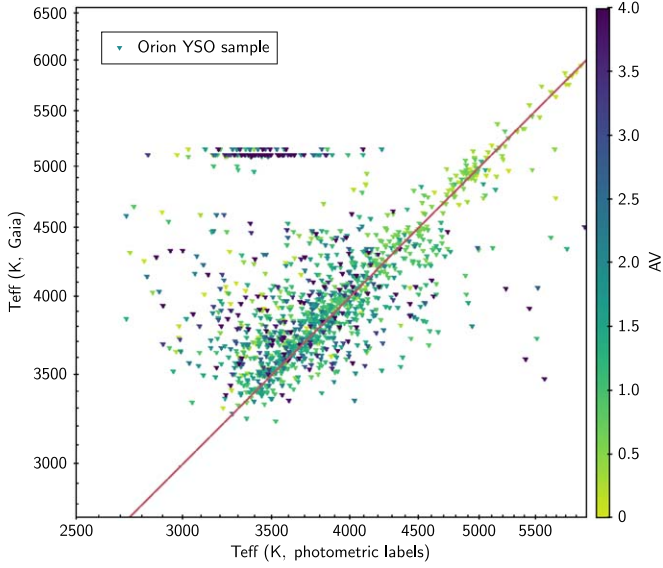
Figure 2. Comparison of the T_{eff} derived from the photometric CNN vs. T_{eff} from Gaia Collaboration et al. (2018) for the YSOs. Each point is color coded with the A_V from Kounkel et al. (2018).

stellar masses from 0.08 to 3 $M_\odot$, ages from 1 to 100 Myr, extinction from 0 to 20 A_V, and distance from 50 to 1000 pc. Only isochrones with Fe/H = 0 were used, which is consistent with the nearby star-forming regions (e.g., D'Orazi et al. 2009, 2011). The empirical parameters the network used to evaluate the labels included nine photometric bands, stellar radii r_*, stellar luminosities $\log L$, and the distance. In cases where the photometry in a particular band was too faint to be reliably detected in the real data, it was set to the limiting magnitude ($G < 19$, $G_{BP} < 20.5$, $G_{RP} < 17.5$, $J < 17$, $H < 16$, $K < 16$, $W1 < 16$, $W2 < 16$, $W3 < 13.5$), with only G band being required. The additional two parameters r_* and $\log L$, were drawn from the isochrone but modeled after those reported by *Gaia* DR2, in that they were only reported if $r_* > 0.5 r_\odot$, $\log L > -1.54$, and $M_G < 10$. Similarly, with the photometric bands, they were set to the limiting cases if they were not detectable. No scatter due to uncertainties was applied to the synthetic parameters.

The CNN was trained on ~42,000 synthetic stars to predict ages, masses, A_V, T_{eff}, and $\log g$ based on these input parameters. Applied on a separate synthetic sample that was generated similarly to the one on which it has been trained, not accounting for any uncertainties or systematic offsets, the neural network could recover $\log g$ with a precision of 0.01 dex, and $\log T_{eff}$ with a precision of 0.003 dex.

When applied to the real data, the CNN produced $\log g$ that are consistent with the isochrones corresponding to the typically accepted ages of the individual clusters covered by APOGEE (Section 4). Additionally, it could generally recover T_{eff} estimated by *Gaia* DR2 with no evidence for bias in the quality of fit as a function of independently inferred extinction values (Figure 2).

3. APOGEE Net

The APOGEE Net was designed to take in the raw spectra and to return the predictions on T_{eff}, $\log g$, and Fe/H (Figure 3). The sources, labels of which were determined in Section 2, were split into three different subsets: a training set on which the model is trained, a held out development set that is used to evaluate the
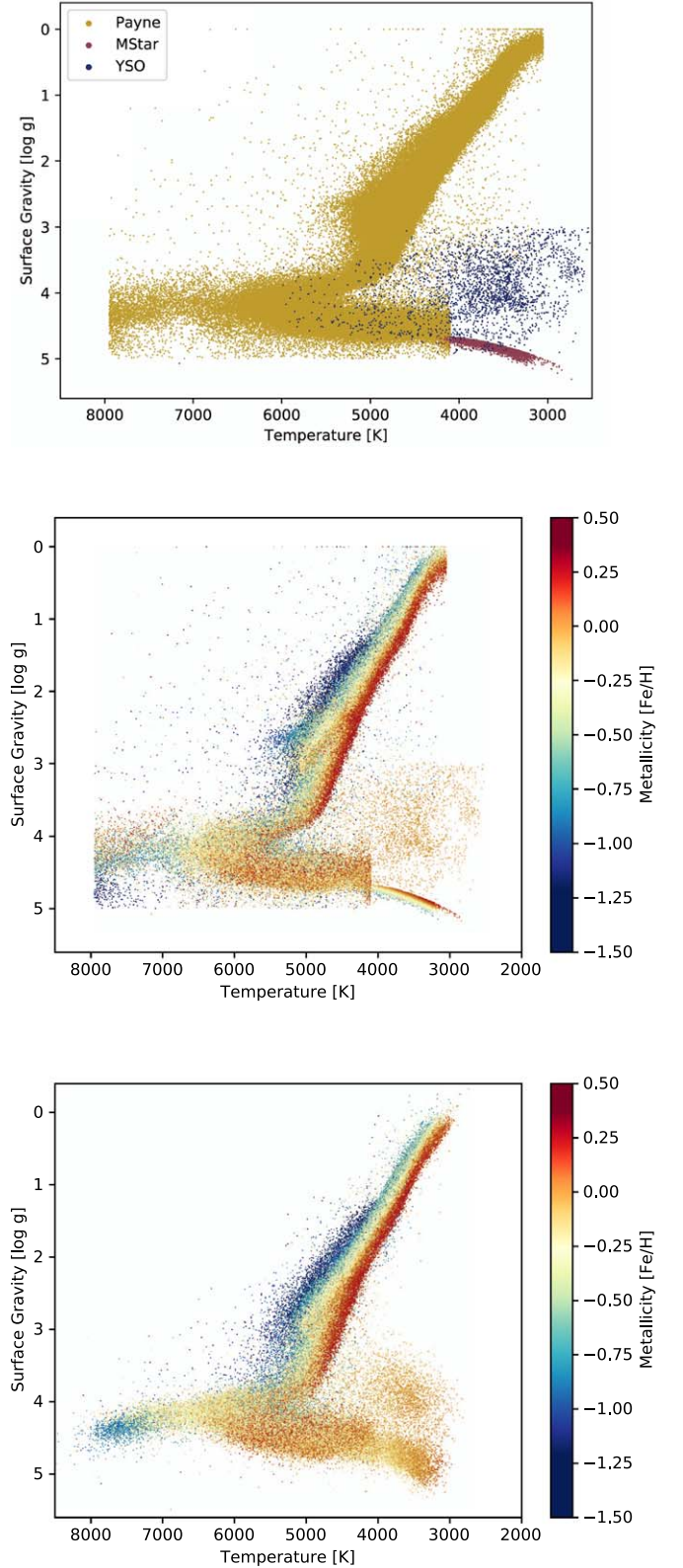
Figure 3. T_{eff} and $\log g$ distribution of the sample. Top panel: categorization of the input parameters based on the origin of the measurements. Middle panel: the input parameters used in training. Bottom panel: the resulting predictions from the APOGEE Net.

model's generalization performance during training and to tune hyperparameters, and finally a held out test set that is used to evaluate the model's performance once it has completed training.

Table 1
Normalization values for T_{eff}, $\log g$, and Fe/H

$\mu_{\log g}$	2.88
$\sigma_{\log g}$	1.16
$\mu_{T_{eff}}$	4716.92
$\sigma_{T_{eff}}$	733.01
$\mu_{Fe/H}$	−0.22
$\sigma_{Fe/H}$	0.30

Table 2
Combined Standardized MSE Loss

	MSE loss		
	Train	Dev.	Test
Full	0.044	0.049	0.063
MStar	0.08	0.098	0.196
YSO	0.153	0.168	0.217

For the Payne catalog, the split between the train, dev, and test sets was 80/10/10%. Due to a smaller number of sources in other categories, to have a sufficient number of sources in the test set, M dwarf and YSO catalogs were split 60–20–20%.

3.1. Feature and Target Preprocessing

While we experimented with using various lossless normalization techniques for the input flux (i.e., not altering the underlying shape of the spectrum, merely scaling it) these standardized fluxes failed to converge in training, and we found that training on the raw flux from the "apStar" readily converged to good results. We did not investigate the performance of the "apStar" spectrum normalized in a way that removes the underlying shape of the spectral energy distribution (SED), as, depending on the spectral type, such normalization may be uncertain and result in additional noise in the line profile. The main benefit of the normalization would be the removal of the extinction signature from the spectrum; however, it should be possible for a neural network to learn to ignore reddening from the raw flux as well. Nonetheless, comparison in the performance between normalized and raw spectra could be a fruitful avenue for further investigation.

In contrast, in order to predict T_{eff}, $\log g$, and Fe/H simultaneously, it was necessary to normalize these target values; normalizing the targets put the losses (and gradients during training) onto a comparable scale. To normalize, we calculated the mean (μ) and standard deviation (σ) of each target variable using the training set and then standardized all prediction targets across all sources and all sets (train, development, and test). Specifically, we normalized as follows:

$$\frac{x_i - \mu_x}{\sigma_x}, \qquad (4)$$

where x denotes a target variables (T_{eff}, $\log g$, or Fe/H) and i denotes a specific data point. Normalization values can be found in Table 1.

For evaluation purposes, the model's predictions are converted back to their physical units using the inverse relations from the above.

3.2. Convolutional Spectral Model

Our model is a one-dimensional CNN, inspired by the VGG16 CNN architecture (Simonyan & Zisserman 2014). It consists of 12 convolutional layers separated every two layers by a max-pooling layer, followed by two fully connected layers (See Appendix A for term definitions and details). The model was implemented in PyTorch (Paszke et al. 2017). The architecture of the network is defined precisely in Appendix B.

3.3. Training and Tuning

We trained the APOGEE Net using stochastic gradient descent to minimize mean squared error (MSE) loss. To improve the model's ability to generalize to new data, we employed early stopping; i.e., we stopped training when performance on the development set begins to decrease, which is indicative of overfitting to the training set at the expense of ability to be generalized. After each full pass through the training data, the model's performance is evaluated on the development set. If the development set performance has improved, as measured by a decrease in loss, then the model is saved. If, however, the loss on the development set does not improve after five consecutive evaluations, training is stopped. If the loss improves before the fifth evaluation, the model is saved, the counter resets, and training resumes.

After a modest amount of hyperparameter tuning, we settled upon the following hyperparameter configuration: learning rate of 0.001, dropout rate of 0.1, and a training batch size of 128.

3.4. Model Adaptation

In order to obtain high model performance on our set of interest (YSOs and M-Stars), we explored various strategies for adapting a model trained on a larger set of data to our smaller set of YSOs and M-Stars. Training on the full set of stars allowed us to achieve a good performance on the Payne subsample, but because it is much larger than the other two, APOGEE Net struggled to achieve acceptable loss for YSOs and M-stars. Another strategy was to first train on all stars, and then further train exclusively on the subset of interest. Unfortunately, this approach also proved suboptimal: while the model performance did improve significantly for M-Stars and YSO stars, it resulted in a dramatic degradation in performance for the red giants in the Payne catalog.

Instead, we used a stratified sampling strategy to balance the sizes of these three subsets in the training sample. After initially training APOGEE Net to convergence on the entire data set, we continued training on the YSO and M-Star samples plus a random 5% of the Payne catalog. Each time the model stopped from early stopping, the last best performing model was reloaded, and another random 5% of the Payne catalog was selected to train on. Through this method, we were able to focus training on the M-Star and YSO data without losing performance on the broader Payne data set. This process continued until the performance on M dwarfs and YSOs did not show continuing improvement. Final normalized MSE loss performances for training, development, and testing after tuning are reported in Table 2, and the resulting performance for each parameter in the native units is shown it Table 3.

	Train	Dev.	Test
		Full	
$\log g$	0.189	0.203	0.216
$T_{\rm eff}$ (K)	144.82	158.72	183.74
Fe/H	0.0769	0.0792	0.0914
		MStar	
$\log g$	0.186	0.216	0.346
$T_{\rm eff}$ (K)	129.69	167.15	273.7
Fe/H	0.129	0.137	0.18
		YSO	
$\log g$	0.366	0.364	0.4
$T_{\rm eff}$ (K)	413.92	436.87	490.99
Fe/H	0.0616	0.0669	0.0879

3.5. Uncertainties

Fundamentally, the predictions of a CNN are deterministic: after a model is trained, passing the same set of inputs always results in the same outputs. As is, the CNN is unable to realize the uncertainties in either the original data or in its predictions (other than through a difference relative to the input labels).

However, given that the data themselves are uncertain, it is possible to vary inputs within the errors and retain the same underlying information. Each one of the realizations of the same spectrum would be perceived by the CNN as a distinct input, and would produce a slightly different prediction from the original. Measuring the scatter in these predictions can give an estimate of the uncertainties on the per source basis, akin to a Markov chain Monte Carlo. Although, unlike MCMC, this analysis is not particularly costly in terms of the computational time.

APOGEE has measured per pixel errors in flux. Therefore, at every pixel, we generated a random value drawn from a normal distribution, multiplied it by the corresponding uncertainties, and added this noise profile to the flux. Some pixels (such as those near the chip gaps, or those that correspond to the telluric lines) had abnormally high uncertainties, so to prevent them from skewing the model, we capped the maximum allowed error at five times the mean in the spectrum.

This procedure was repeated to generate 100 different realizations for each spectrum, and all of them were passed through the APOGEE Net. The mean and standard deviation values were then measured for each parameter for each source.

3.6. Validation of the Stellar Parameters

We report on the resulting predictions with the corresponding uncertainties in Table 4.

The typical agreement between the input labels and the resulting predictions is 100 K in $T_{\rm eff}$, 0.15 dex in $\log g$, and 0.07 dex in Fe/H (Table 3, Figure 4, left). The scatter in $T_{\rm eff}$ and $\log g$ is slightly higher for the YSOs, as it improved on some of the systematic issues the photometric labels had, which were originally derived somewhat crudely, fine-tuning them based on the overall grid. For the M-stars, comparison between the labels and the predictions for Fe/H is slightly offset from the line of unity, with predictions somewhat compressing the range of Fe/H offered by the labels, having fewer sources as metal-rich, and fewer as metal-poor, but showing a good linear agreement overall.

The typical reported uncertainties are 25 K in $T_{\rm eff}$, 0.04 dex in $\log g$, and 0.015 dex in Fe/H; thus, they underestimate the scatter between the labels and the predictions by approximately a factor of four. In comparison to other pipelines, the uncertainties for the same sources are not strongly correlated, but they are generally comparable to the errors reported by the IN-SYNC pipeline, and approximately a factor of two smaller than those reported by ASPCAP pipeline. The reported uncertainties strongly depend on the S/N of the spectrum, as well as on the spectral parameters (Figure 4, right). As expected, the parameters of hotter or more metal-poor stars would be more uncertain due to a fewer number of lines that can be used to determine stellar properties. Similarly, sources with higher $\log g$ would have shallower lines, resulting in more uncertain predictions. Although the APOGEE Net had no information on the uncertainties in the labels, and the errors were generated from slightly perturbing the input spectral fluxes and taking an rms of the resulted predictions, it was able to reproduce physically expected trends.

At low S/N (<10–20), the uncertainties in all parameters reach a ceiling of $\sim$200 K in $T_{\rm eff}$, $\sim$0.2 dex in $\log g$, and $\sim$0.1 dex in Fe/H. This ceiling suggests that the uncertainties for these sources are underestimated and, furthermore, that the CNN does not derive any meaningful information in the low-S/N spectra. This brings into question the reliability of other parameters (e.g., $v \sin i$, radial velocity) derived by other means in these low-S/N spectra. Some of the sources with S/N ~ 0 have unphysical parameters, i.e., located far outside of the bounds of the Figure 3. These sources were removed from the catalog.

A similar approach to determining spectroscopic parameters from the APOGEE spectra for the M dwarfs was recently undertaken by Birky et al. (2020), based on the Cannon (Ness et al. 2015) data-driven spectral modeling code (Figure 5). While some YSOs are included in their sample, they tend to have higher χ^2 for the fit compared to the rest of the sample. In particular, because their code did not train to distinguish between metallicity and $\log g$, it considered YSOs to be more metal-rich than they are likely to be due to deeper spectral lines. In the M-dwarf sample, for the metallicity, there is a good agreement at Fe/H < 0, but at Fe/H > 0, they tend to systematically differ by a factor of two; it is not clear why. The predicted $T_{\rm eff}$ appears to be comparable between the two works, with a scatter of $\sim$100 K. However, because they have not included any sources hotter than $\sim$4100 K, they tend to run into the edge effects at $\sim$4000 K, with sources piling at the boundary. We note that a similar effect occurs in this work as well at $T_{\rm eff} \sim 8000$ K—any stars intrinsically hotter than this temperature are interpolated toward this value. This potentially explains the excess of metal-poor stars in Figure 3. For this reason, we do not include any stars hotter than 6700 K in Table 4. In the future, however, by generating more reliable labels for massive stars and including them in the training, it would be possible to minimize these edge effects.

4. Results

4.1. Overall Properties

As can be seen in Figure 6, the quality of the derived $T_{\rm eff}$ and $\log g$ for YSOs improves with each method for extracting them. The original parameters from the IN-SYNC pipeline are systematically offset from the isochrones, have an odd shape of

Table 4
Results for Combined Evaluation

APOGEE ID	α (J2000)	δ (J2000)	$\log g$ (label, dex)	$\log g$ (prediction, dex)	$\sigma \log g$ (dex)	T_{eff} (label, K)	T_{eff} (prediction, K)	σT_{eff} (K)
2M00003379+7940362	0.140803	79.676727	4.94	4.47	0.06	3307	3357	28
2M00013219+0016012	0.384140	0.267008	4.74	4.64	0.08	3997	3998	47
2M00024474+6158060	0.686431	61.968346	4.74	4.50	0.10	3787	4062	50
2M00025988+0148410	0.749506	1.811404	4.72	4.65	0.07	3959	3919	32
2M00030930+0110025	0.788757	1.167374	4.73	4.66	0.07	3947	3974	33

Fe/H (label, dex)	Fe/H (prediction, dex)	σ Fe/H (dex)	S/N	Data Set	Data Type
−0.03	−0.05	0.01	137.0	Train	Mstar
−0.34	−0.25	0.02	57.2	Train	Mstar
0.02	−0.15	0.02	57.7	Train	Mstar
−0.20	−0.19	0.02	82.1	Train	Mstar
−0.21	−0.24	0.02	88.1	Train	Mstar

Note.

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

the MS, and have various unphysical gaps, the most notable of which is at $T_{eff} = 3600$ K. The labels derived from the photometry show similar agreement to the isochrones in the sequences of the ages of the individual regions, and it renormalizes the derived parameters to the appropriate range, but it does show a somewhat peculiar behavior especially at low T_{eff} as it attempts to reconcile the differences between various bands and the isochrones.

Finally, the APOGEE Net connects the derived parameters for the YSOs to those for the M dwarfs and the red giants, making it possible to interpret the resulting values and removing all of the systematic gaps that persisted in the previous iterations. However, there may still be a weak systematic offset in the shape of the isochrones and the YSOs at $T_{eff} < 3500$ K, as at a particular age, the isochrone tends to result in somewhat lower $\log g$ at lower T_{eff}. On the other hand, $\log g$ traced by the individual stellar populations either remains flat throughout or slightly increased toward higher values at low T_{eff}, oriented in parallel to the MS. Considering that the photometric labels showed the opposite trend at lower T_{eff}, it is unclear how such a potential discrepancy could be better rectified in the future, or if the cause of the discrepancy is necessarily in the predicted parameters as opposed to the isochrones.

The spectroscopic parameter space that the low-mass YSOs occupy can be rather cleanly separated from the other stellar objects, as they have $\log g$ typically lower than the MS stars, and T_{eff} typically cooler than the red giants. Thus, by selecting the sources in this parameter space, we can robustly identify young stars in the catalog and look at their spatial distribution (Figure 7). With a simple cut, restricting the selection bound by (T_{eff}, $\log g$) of (2200, 4.4) (4500, 3.9), and (3200, 1.7), it is possible to recover all of the star-forming regions in the sample: the Orion Complex, Perseus clusters NGC 1333 and IC348, as well as NGC 2264. Some of the sources from Pleiades end up in this selection as well. Surprisingly, some fields (K2_C4_172-20 and K2_C4_177-21) also appear to include sources from the Taurus Molecular Clouds. While Taurus has been observed with APOGEE, the dedicated observations of young stars in this region have not begun until after the release of DR14. Thus, the sources we see have been observed serendipitously.

The aforementioned selection does not recover all of the young stars that have been observed, particularly those that are hotter or those that are somewhat more evolved. Additionally, approximately 6% of the sources from this simple selection are scattered all across the sky, many at high galactic latitudes, which do not appear to be associated with any particular star-forming region, suggesting that they are impostors, i.e., contamination.

The distribution of the selected sources in the spectroscopic space is comparable to the distribution of sources in the HR diagram. Even the impostors appear to be bona fide sub-giants (i.e., appearing above the MS and below the red giant branch. It is unclear what mechanism may drive them, although these sources appear to be fast rotators). Furthermore, this comparison demonstrates that it is possible to use the spectroscopically derived parameters to effectively derive stellar properties, even for sources that are located in regions of high extinction, those have very uncertain parallaxes, or no parallax measurements at all.

4.2. Ages

It is possible to use the derived $\log g$ values as a proxy for age. We have not explicitly interpolated the T_{eff} and $\log g$ across the isochrones; however, we divided the sample into five bins: $\log g < 3.6$ dex ($\lesssim 1$ Myr), $3.6 < \log g < 3.8$ dex (1 ~ 2 Myr), $3.8 < \log g < 4.0$ dex (2 ~ 3 Myr), $4.0 < \log g < 4.2$ dex (3 ~ 5 Myr), and $4.2 < \log g$ dex ($\gtrsim 5$ Myr). We then constructed a density map of sources in each bin (Figure 8). This is done to further test the $\log g$ on the more granular scale, in order to compare these age estimates to the known distribution of ages in various populations in the sample.

In Orion, the distribution of ages is largely consistent what has been previously measured for each individual region. For example, in λ Ori, the central cluster is ~5 Myr, and the outer shell that has been triggered by a supernova is consistent with what has been previously measured by Kounkel et al. (2018). In the vicinity of Orion B, there is a clear separation between ~1 Myr old clusters, NGC 2068 and NGC 2024, a somewhat older (2–3 Myr) cluster σ Ori, and 3–5 Myr extended population associated with it. A similar agreement with previously measured ages can also be observed in ψ^2 Ori.
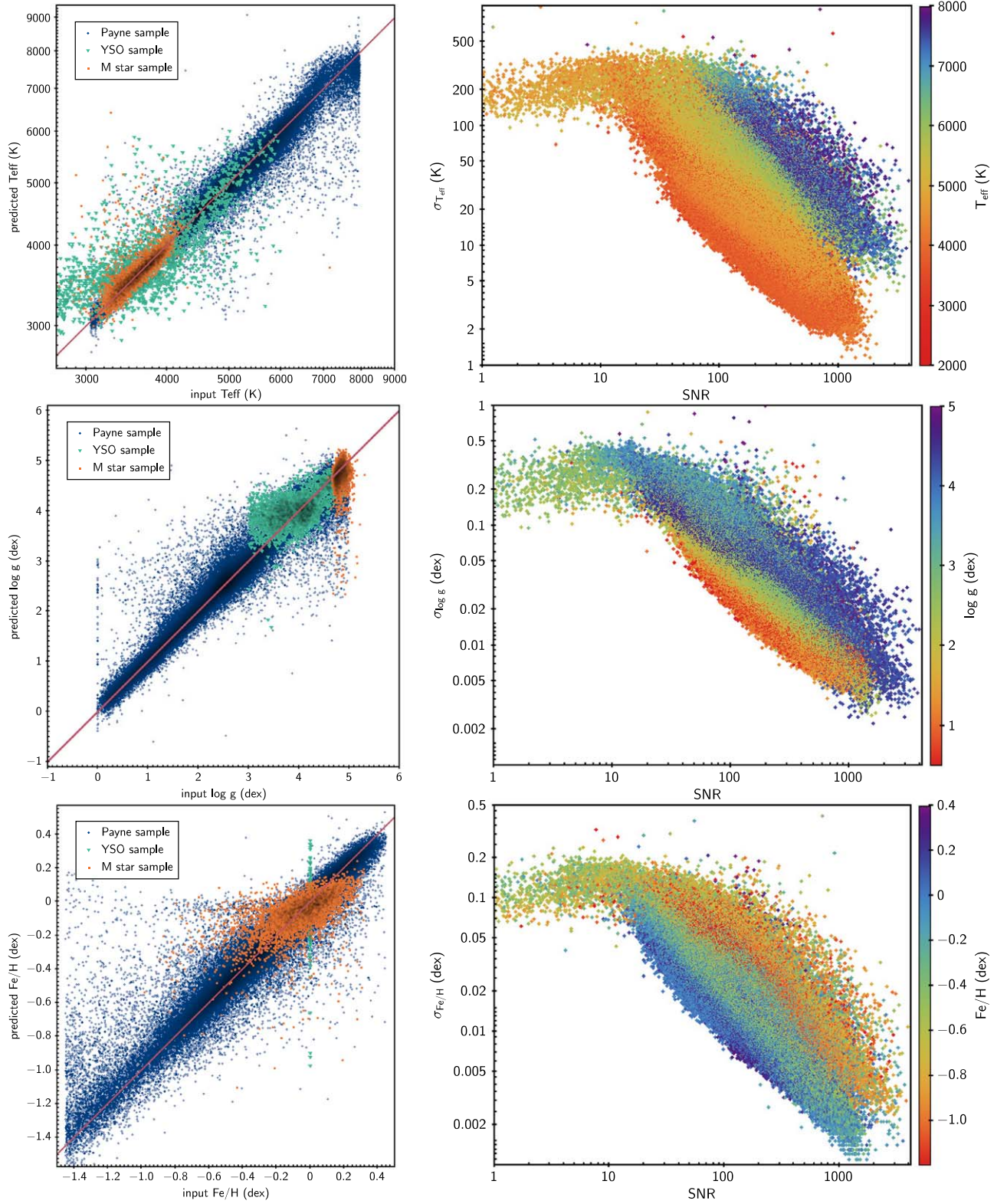
Figure 4. Left panel: comparison between the input labels and the predicted spectroscopic parameters. Right panel: uncertainties as a function of S/N.

In the ONC, Beccari et al. (2017) have observed three distinct populations with mean ages of 1.2, 1.9, and 2.9 Myr, and they observed that these populations, while overlapping, cover different volumes in the sky, with the youngest one being the most centrally concentrated and the oldest one being the most diffuse. However, they could not rule out that these

populations are not due to an unresolved binary and tertiary sequences. Cross-matching against their catalog of sources does indeed show that the these populations are separated in the log g space, independently confirming different ages (Figure 9, top). While there is some crossover between the sources, the "Very Young" sources peak at log g = 3.7 dex, corresponding
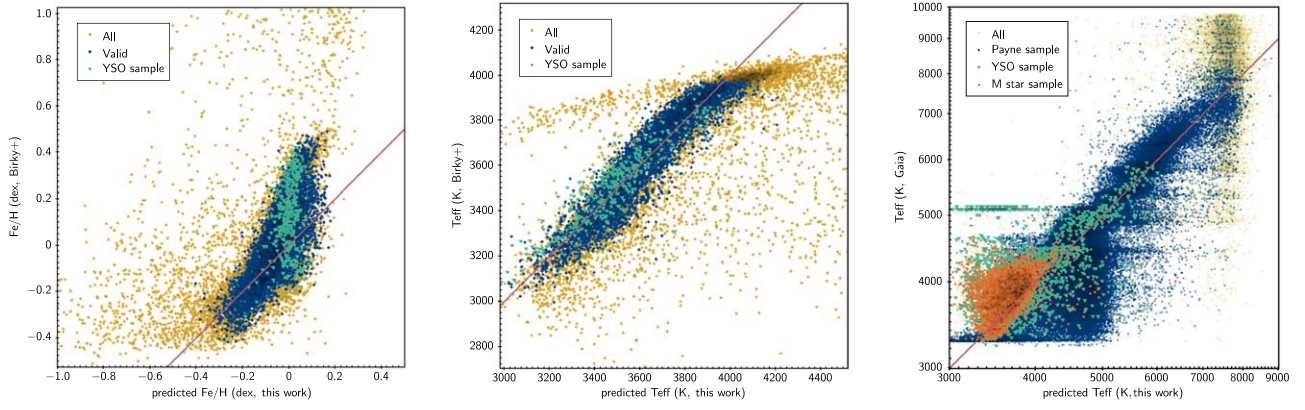
Figure 5. Left and middle panels: comparison of Fe/H and T_{eff} derived in this work compared to those in Birky et al. (2020). The sources in the Valid sample satisfy the quoted ranges of the parameter space from Birky+ over which their extra extrapolation is applicable, combined with having uncertainties in the predicted parameters in our work of $\sigma_{T_{eff}} < 100$ K and $\sigma_{\log g} < 0.1$ dex (typically corresponding to S/N > 20), restricting the comparison to 5893 stars. The YSO sample is a subset of the Valid sample, with the overlapping sources from Kounkel et al. (2019). Right panel: comparison of the derived T_{eff} in this work and T_{eff} from *Gaia* DR2. Note the edge effect at $T_{eff} \sim 4000$ K for the sample from Birky et al., and 8000 K for the sample in this work.

to ~1 Myr. The "Young" sources peak at $\log g = 3.9$ dex, or at the age of ~2 Myr, and the "Old" sources peak at $\log g = 4$ dex, or at ~3 Myr, consistent with the original age estimates.

It should be noted, though, that in the full spectroscopic sample, there do not appear to be distinct age sequences; rather, the distribution appears to be more continuous. This does not appear to be explained by the smearing of the sequences from the uncertainties. But, examining the density of sources at given age slices in Figure 8, similarly to in Beccari et al. (2017), we do indeed find that the youngest sources are located primarily close to the center of the cluster, while the older stars are more distributed throughout the cluster. A similar trend can also be seen in Perseus, in IC348.

Along L1641, there are curious chains of coeval stars running in parallel to the cloud. Hacar et al. (2018) have found a large network of gas fibers toward the ONC, any one of which is in the process of forming stars along its length, and the stars that would form from the same gas fiber are approximately coeval. Hacar et al. (2016) have previously searched for chains of stars in the APOGEE data in Orion A that are comoving in the position–position–velocity diagram, although the stellar ages have not been considered. Only a few such chains have been found in L1641, and they did not exhibit any preferential orientation relative to the cloud. It is unclear whether the chains we see in this work exhibit just a chance alignment or if their arrangement could be thought of as significant.

The derived parameters are not strongly confused by multiplicity—the spectroscopic binaries identified by Kounkel et al. (2019) do not appear to occupy a systematically different T_{eff} or $\log g$ space from other single stars in their corresponding clusters (Figure 9, middle panel). The uncertainties in the parameters also do not appear to be affected. Metallicity is the only parameter where there might be a slight, barely significant systematic shift, with spectroscopic binaries being, on average, more metal-poor by 0.02 dex. There still may also be nonsystematic offsets that might affect the parameters of binary stars. A more advanced fitting of multiple templates to blended spectra would be needed to properly disentangle the companions in order to understand the magnitude of these potential offsets.

However, there is a strong difference in $\log g$ as a function of whether a YSO has a protoplanetary disk or not (Figure 9, bottom), with the disk-bearing Class II sources having $\log g$ 0.2 dex lower than their diskless Class III counterparts, although there is much overlap between the two populations. In a given

cluster, this effect translates to 1–2 Myr systematic difference in age. A similar separation in $\log g$ has previously been observed by Yao et al. (2018) using the values derived from the IN-SYNC pipeline. It is notable that despite the fact that the aforementioned pipeline does not produce reliable absolute calibration, it does discriminate between the effects of veiling from disks and other spectral parameters. Thus, it is likely that the systematic difference we observe is real and not due to Class II sources being deviating from the model produced by APOGEE Net.

This difference is also apparent when examining the distribution of A_V (Figure 10), as YSOs with a higher degree of extinction from their parental cloud and protoplanetary disk are more likely to be somewhat younger compared to those whose extinction is driven only by the external foreground dust. It should be noted that values of A_V for the individual stars are available only for the Orion regions and that they were computed using the previously measured T_{eff} from the IN-SYNC pipeline (Kounkel et al. 2018).

One way to independently validate our spectroscopically determined $\log g$ is with the stellar radii that we have previously measured empirically via the bolometric flux from the SED and parallax in Kounkel et al. (2018) using methods described by Stassun et al. (2017). Since g is proportional to M/R^2, we expect that the spectroscopic $\log g$ will correlate strongly with the SED-based radius, with some secondary dependence on the stellar mass. And because the young, low-mass stars in our sample are expected to still be contracting along roughly vertical Hayashi tracks in the HR diagram, we can take T_{eff} as a proxy for the stellar mass. As expected, Figure 11 shows that our spectroscopic $\log g$ are well correlated with the SED-based radii overall, especially when controlled for T_{eff}. Indeed, stars of a given T_{eff} (i.e., mass) show a very strong correlation between $\log g$ and radius, and cooler (lower-mass) stars show smaller radii for a given $\log g$.

4.3. Metallicity

Stars that form from the same parental cloud form with the same chemical composition; therefore, they should have comparable Fe/H. Indeed, the GALAH survey did not reveal any chemical inhomogeneity in any of the elements inside the Orion Complex (J. Kos et al. 2020, in preparation), to within
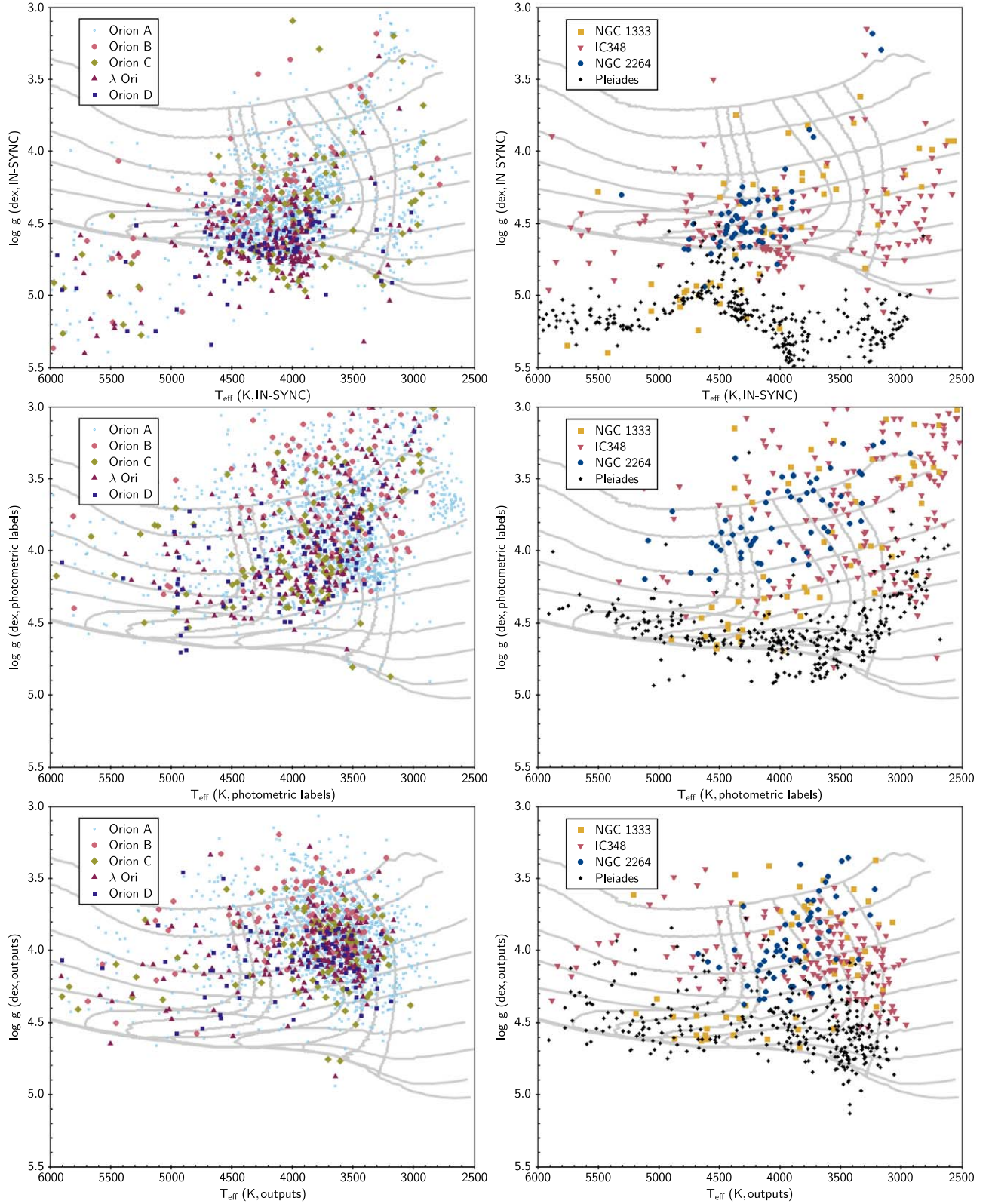
Figure 6. Distribution of T_{eff} and $\log g$ values for the YSOs across the various regions. The compilation of the measurements from the IN-SYNC pipeline (top panel), photometrically derived input labels (middle panel), and the resulting output from the APOGEE Net (bottom panel) are shown. The gray lines show the isochrones at ages of 1, 2, 5, 10, 20, 50, 100, 200, and 300 Myr, and 8.5 dex, as well as the evolutionary tracks for 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1 $M_\odot$ stars from the PARSEC isochrones (Marigo et al. 2017).

0.03 dex. Furthermore, the young populations in the solar neighborhood all have near-solar metallicity (D'Orazi et al. 2009, 2011).

Abundances derived by the APOGEE Net do show similar agreement—the dispersion in Fe/H abundances is within 0.04–0.05 dex in the overall sample of young stars, as well
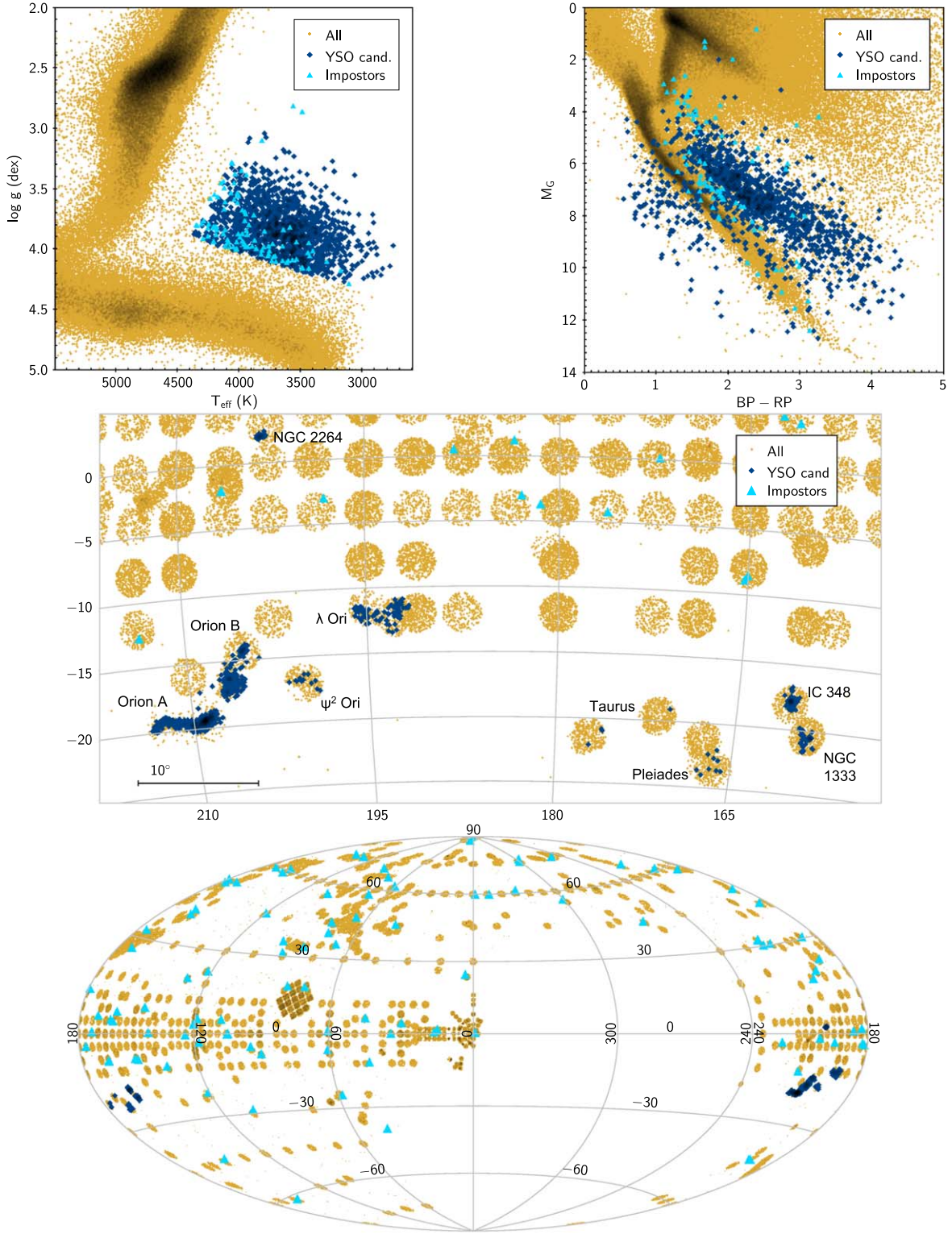
Figure 7. Top panels: selection of the YSOs based on the predicted stellar parameters. Objects are considered YSO candidates if they fall into the parameter space bounded by (T_{eff}, $\log g$) of (2200, 4.4) (4500, 3.9), and (3200, 1.7). The top left panel shows the spectroscopic parameters, and the top right panel shows the HR diagram constructed using *Gaia* photometry for the same sources. The sources in blue that fall bellow the photometric MS are known YSOs, most likely with poor *Gaia* parallaxes and/or fluxes. Bottom panel: spatial distribution of the identified candidates in galactic coordinates. Sources that do not coincide with nearby star-forming regions are flagged as the potential impostors (6% of the sources in the selected area).

as inside any given population (Figure 12). There are slight systematic inhomogeneities, however: hotter stars, as well as stars that have lower $\log g$ tend to be marginally more metal-poor, whereas cooler dwarfs tend to be marginally more metal-rich. The effect is larger than the typical estimated errors, but it is not greater than than a few 0.01 dex.

These systematics slightly affect the average Fe/H between different populations. While most of them have an average
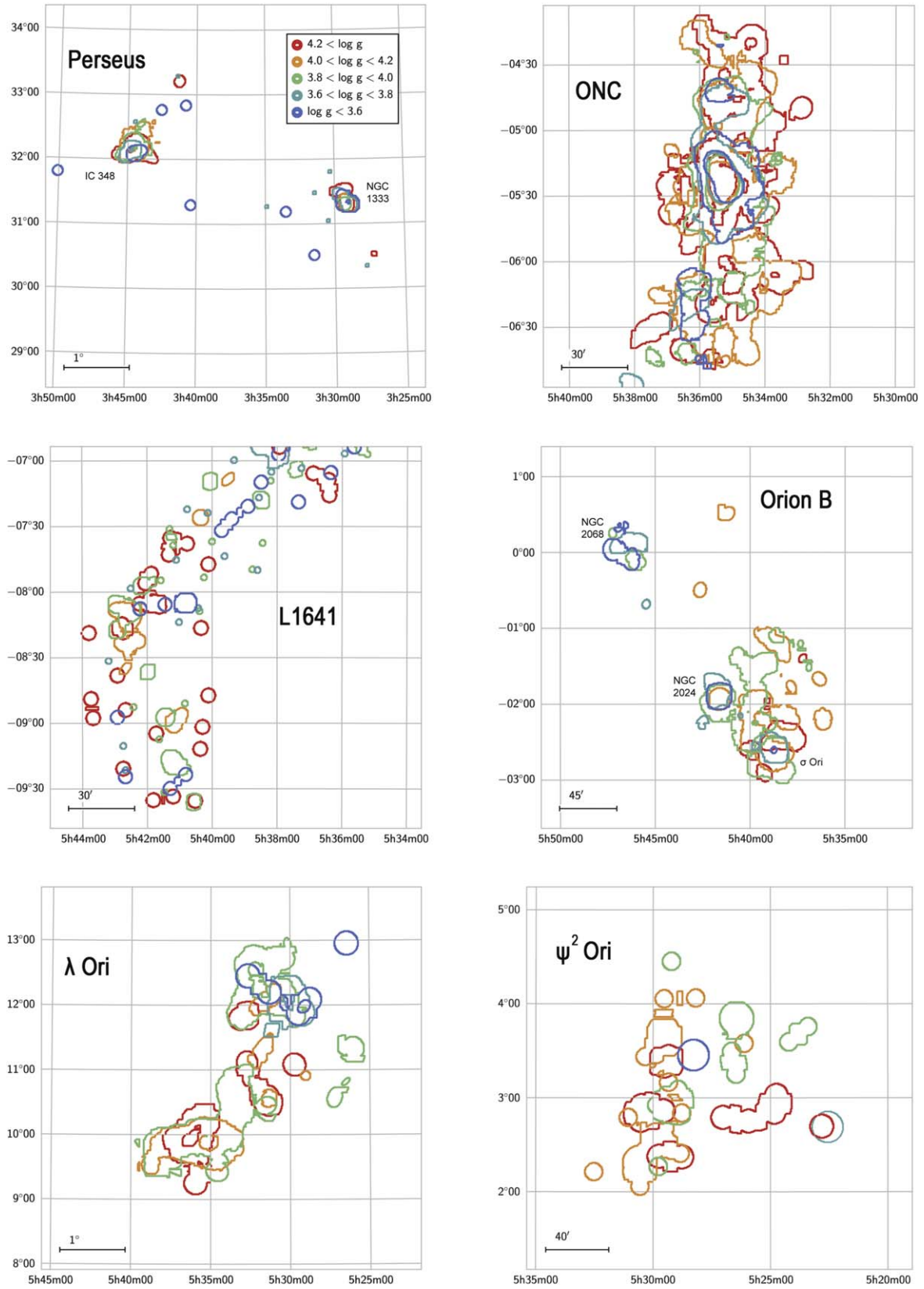
Figure 8. Contour plots showing the density distribution of sources with different log g values, as a proxy for age, color coded from the youngest ($\lesssim 1$ Myr, blue), to the oldest ($\gtrsim 5$ Myr, red).

Fe/H ~ 0 dex, two of the youngest populations (NGC 1333 and Orion B), appear to be somewhat offset, with average Fe/H ~ -0.02 dex. Furthermore, the furthest cluster, NGC 2264, appears to be somewhat metal-poor at Fe/H ~ -0.06 dex—notably in part because low-mass stars cooler than ~ 3900 K were not targeted to be included in the sample.
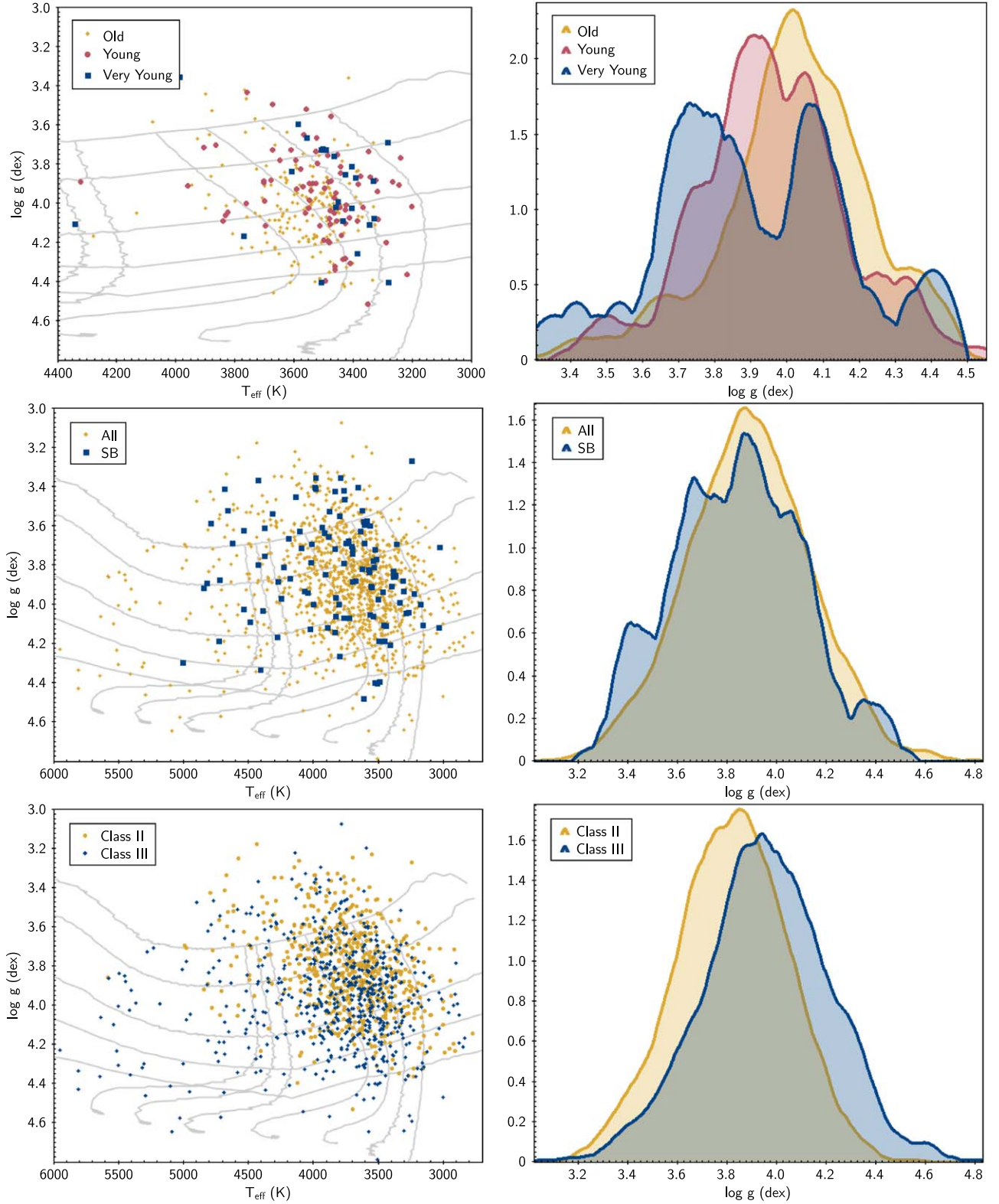
Figure 9. Left panels: distribution of T_{eff} and $\log g$ for the sources in Orion A shown against 1, 2, 5, 10, and 20 Myr isochrones. Right panels: Kernel Density Estimation of the $\log g$ with the kernel size of 0.1 dex. The top panels show the three populations identified by Beccari et al. (2017). The middle panels highlight spectroscopic binaries from Kounkel et al. (2019). The bottom panels show the difference in the distribution of disk-bearing Class II YSOs and diskless Class III photospheres.

5. Discussion

Machine learning is an effective approach for classifying stellar spectra, and for the first time we have applied it to

spectra of young stars to derive meaningful T_{eff} and $\log g$ that are interpolatable over the isochrones to determine ages. The performance of these parameters can rival the usage of
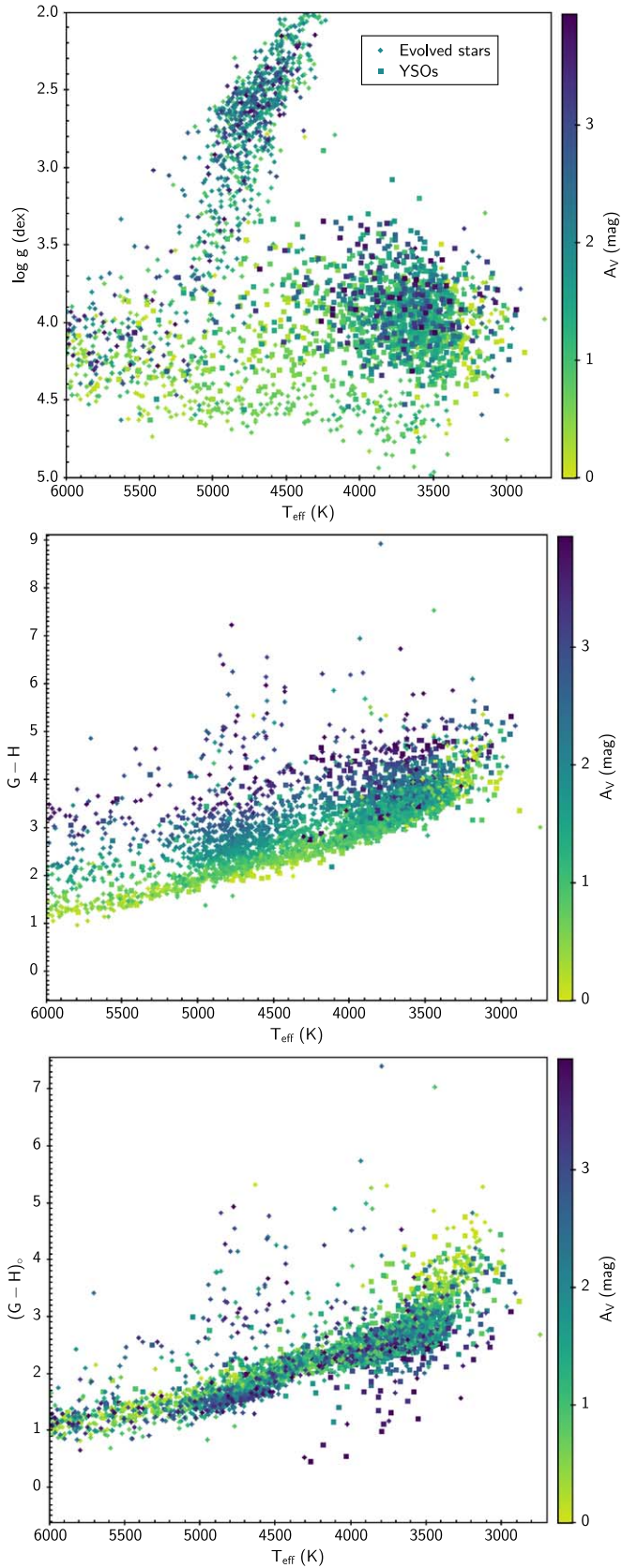
Figure 10. Top panel: distribution of T_{eff} and $\log g$ of sources toward Orion, color coded by the measured A_V from Kounkel et al. (2018). Extinction tends to be higher in the youngest stars, decreasing incrementally as they age. The dwarfs in the field are all nearby, without much dust along the line of sight. Red giants tend to be more extinct the further away they are. Middle panel: T_{eff} vs. $G - H$ color. Bottom panel: T_{eff} vs. the unreddened $G - H$ color.
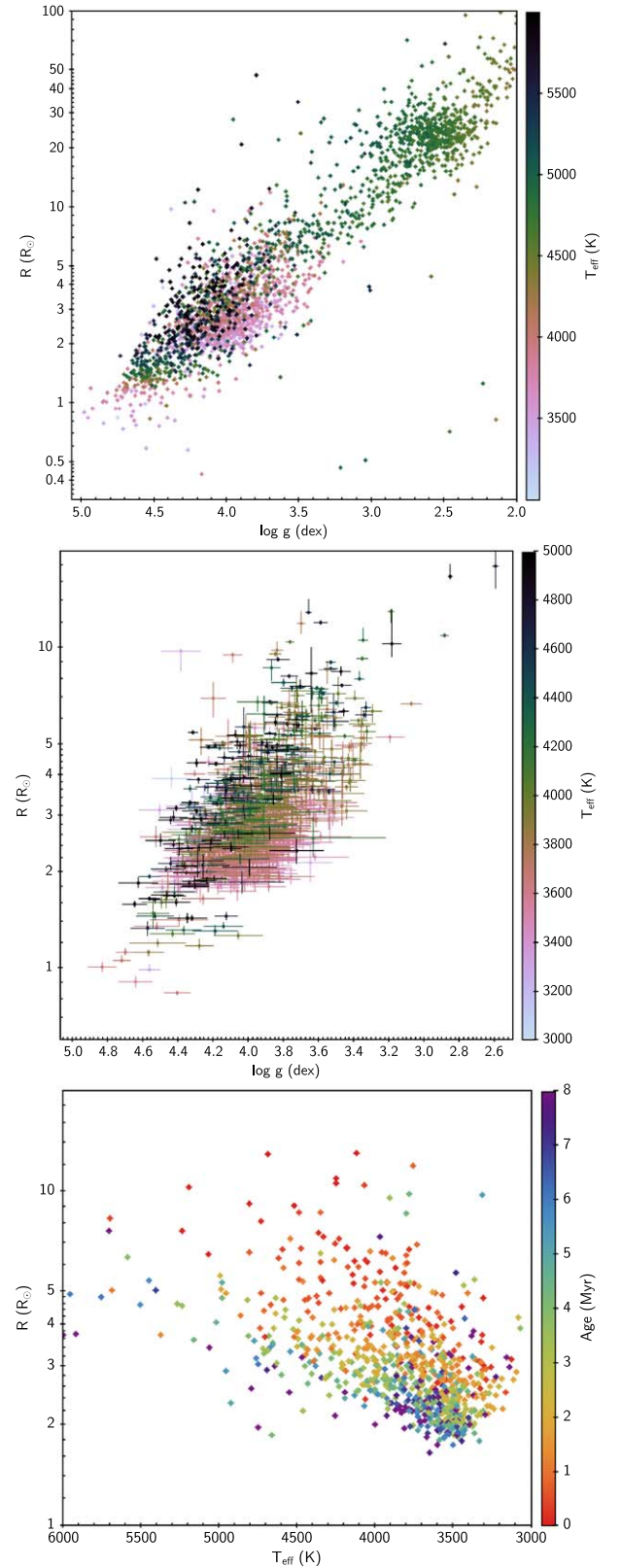
Figure 11. Comparison of spectroscopically derived $\log g$ and stellar radii measured from the SED fitting from Kounkel et al. (2018), color coded by T_{eff}. The top panel shows all sources toward Orion, and the middle panel shows only the YSOs. The bottom panel shows the relation between the stellar radii and T_{eff}, color coded by the photometrically derived ages of YSOs.
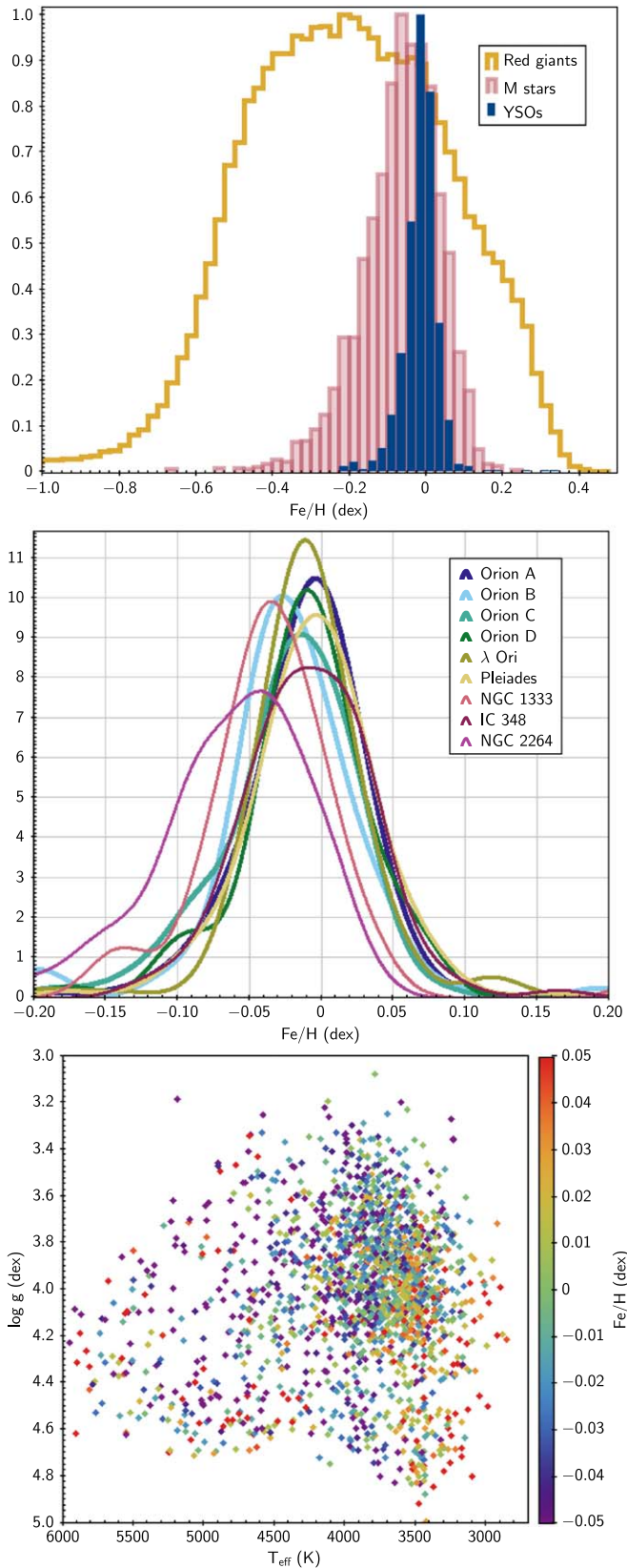
Figure 12. Top panel: distribution of Fe/H in the different samples, with $T_{eff} < 5000$ K. Middle panel: distribution of measured Fe/H for each young population. Bottom panel: Fe/H of YSOs as function of T_{eff} and $\log g$.

photometric color–magnitude diagrams, offering two major advantages over them: spectroscopic parameters are unaffected by extinction or by the binary sequence. Using both of these

approaches in conjunction with one another could allow for a more detailed analysis of the star-forming history inside young populations.

The main limiting factor for machine learning is the existence of reliable input labels over which it would be possible to generalize a particular parameter space. However, it is not necessary to derive those labels from scratch a priori every single time. Instead, through incremental building on previous efforts, it is possible to improve on even coarse estimates. As the parameter space gets further explored, it may be possible to add other properties to the analysis (e.g., other abundance labels that may be more meaningful in analyzing the chemical content of star-forming regions, such as α/H), and extend the grid to other types of stars (e.g., with $T_{eff} > 8000$ K), to produce a fully unified spectral model. Furthermore, if different large spectroscopic surveys with different instruments (having different resolutions or different wavelength coverages) have observations of a few sources in common, it would be possible to cross-reference them relative to one another, allowing one to extract stellar parameters in a way that would have fewer systematic offsets.

Appendix A
DNN Background

DNNs are parameterized functions that map from an input domain to an output domain, the parameters of which are learned through a training process. While a DNN can be used for both regression and classification tasks, this work is principally concerned with its utility for performing regressions, as stellar parameters are continuous variables.

We are utilizing a particular architecture of DNN known as a CNN. The CNN is particularly well-suited for extracting local patterns in spatial data. It has been widely used to achieve state-of-the-art results in various image recognition tasks (Krizhevsky et al. 2012; Simonyan & Zisserman 2014; Szegedy et al. 2015; He et al. 2016). A CNN typically consists of a sequence of convolutional and pooling layers, followed by a set of fully connected layers. A convolutional layer transforms a subsection of the input, a pooling layer downsamples the input, and a fully connected layer is a classic DNN or multilayer perceptron. Broadly speaking, the convolutional layers can be interpreted as feature extractors, and the fully connected layers as the predictors. However, this is just an approximation, as the model prediction error is passed back across all its parameters through backpropagation.

A.1. Training via Backpropagation

To train a DNN model through backpropagation, it is necessary to use an optimizer and evaluation criteria. In our case, we will be using classic stochastic gradient descent for our optimizer. For our evaluation criterion, we will be using MSE loss.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{A}(Y_{ij} - \hat{Y}_{ij})^2 \qquad (5)$$

in which N is the number of training data points, A is the number of stellar parameters that are considered, Y is a target label, and $\hat{Y}$ is the corresponding model prediction. This loss is then propagated back through the model to produce gradients with respect to each model weight, which are used to update the weights to slightly reduce loss; after enough steps in the direction of the negative gradient, a local minimum is reached.

While the MSE loss sums over all training data points, it is inefficient, and our case infeasible, to load all of the training data into memory at once. Instead, data is usually divided into minibatches, and at each iteration the MSE loss is approximated by summing only over the data points in the minibatch. Using gradient descent with these approximate losses is known as stochastic minibatch gradient descent. A full pass through the training set is called an epoch. Typically, the training set is shuffled after each epoch and re-partitioned into minibatches. In addition to the memory consideration, smaller minibatch sizes and shuffling the training set also have the added benefit of helping to improve model generalization; the randomness in the gradient helps to avoid settling into shallow local minima. However, the smaller the minibatch size is, the more often the parameter weights need to be updated, increasing computational expense.

A.2. Hyperparameters

Machine-learning models typically have design parameters that cannot be learned by the network and must be chosen outside of the training process. These "hyperparameters" are typically adjusted by evaluating the model's performance against the held out development data. The number of data points in a minibatch is one such hyperparameter.

Another hyperparameter is the learning rate (also known as step size), the coefficient scaling the gradient used in gradient descent. If the learning rate is too large, the model may fail to converge to a good local minimum. If it is too small, training can be prohibitively slow.

We also treat the dropout rate as a hyperparameter. Dropout is a technique that helps with model generalization by using a mask to zero out a random subset of hidden units during each gradient calculation. This forces the model to distribute the learned representation of the input across all of its hidden units, yielding a more robust internal representation. A dropout rate of 0.1 implies that on any given training pass, 10% of the hidden units on the fully connected layer are masked out: they cannot contribute to the prediction, and gradients will not propagate through them.

Appendix B
APOGEE Net Model Code

class APOGEE_Net(nn.Module):

```
def _init_(self, num_layers, num_targets, drop_p = 0.0):
        super(Net, self)._init_()
        # 3 input channels, 6 output channels, convolution
        # kernel
        self.conv1 = nn.Conv1d(num_layers, 8, 3, padding = 1)
        self.conv2 = nn.Conv1d(8, 8, 3, padding = 1)
        self.conv3 = nn.Conv1d(8, 16, 3, padding = 1)
        self.conv4 = nn.Conv1d(16, 16, 3, padding = 1)
        self.conv5 = nn.Conv1d(16, 16, 3, padding = 1)
        self.conv6 = nn.Conv1d(16, 16, 3, padding = 1)
        self.conv7 = nn.Conv1d(16, 32, 3, padding = 1)
        self.conv8 = nn.Conv1d(32, 32, 3, padding = 1)
        self.conv9 = nn.Conv1d(32, 32, 3, padding = 1)
        self.conv10 = nn.Conv1d(32, 32, 3, padding = 1)
        self.conv11 = nn.Conv1d(32, 64, 3, padding = 1)
        self.conv12 = nn.Conv1d(64, 64, 3, padding = 1)

        #an affine operation: y = Wx + b
        self.fc1 = nn.Linear(64*133*1, 512)
        self.fc1_dropout = nn.Dropout(p = drop_p)
        self.fc2 = nn.Linear(512, 512)
        self.fc3 = nn.Linear(512, num_targets)

def forward(self, x):
        # Max pooling over a (2) window

        x = F.max_pool1d(F.relu(self.conv2(F.relu(self.conv1(x)))), 2)
        x = F.max_pool1d(F.relu(self.conv4(F.relu(self.conv3(x)))), 2)
        x = F.max_pool1d(F.relu(self.conv6(F.relu(self.conv5(x)))), 2)
        x = F.max_pool1d(F.relu(self.conv8(F.relu(self.conv7(x)))), 2)
        x = F.max_pool1d(F.relu(self.conv10(F.relu(self.conv9(x)))), 2)
        x = F.max_pool1d(F.relu(self.conv12(F.relu(self.conv11(x)))), 2)
        x = x.view(-1, self.num_flat_features(x))
        x = F.relu(self.fc1_dropout(self.fc1(x)))
        x = F.relu(self.fc1_dropout(self.fc2(x)))
        x = self.fc3(x)
        return x
```

(Continued)

```
def num_flat_features(self, x):
    size = x.size () [1:] # all dimensions except the batch dimension
    num_features = 1
    for s in size:
    num_features *= s
    return num_features
```

ORCID iDs

Marina Kounkel ⬤ https://orcid.org/0000-0002-5365-1267
Matthew T. Scoggins ⬤ https://orcid.org/0000-0002-0748-9115
K. R. Covey ⬤ https://orcid.org/0000-0001-6914-7797
Keivan G. Stassun ⬤ https://orcid.org/0000-0002-3481-9052

References

Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42
Anders, F., Chiappini, C., Minchev, I., et al. 2017, A&A, 600, A70
Bailer-Jones, C. A. L. 2000, A&A, 357, 197
Bailer-Jones, C. A. L., Irwin, M., Gilmore, G., & von Hippel, T. 1997, MNRAS, 292, 157
Baraffe, I., Homeier, D., Allard, F., & Chabrier, G. 2015, A&A, 577, A42
Bazarghan, M., & Gupta, R. 2008, Ap&SS, 315, 201
Beccari, G., Petr-Gotzens, M. G., Boffin, H. M. J., et al. 2017, A&A, 604, A22
Bell, C. P. M. 2012, PhD thesis, Univ. Exeter
Benedict, G. F., Henry, T. J., Franz, O. G., et al. 2016, AJ, 152, 141
Birky, J., Hogg, D. W., Mann, A. W., & Burgasser, A. 2020, ApJ, 892, 31
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Bovy, J., Rix, H.-W., Schlafly, E. F., et al. 2016, ApJ, 823, 30
Buder, S., Asplund, M., Duong, L., et al. 2018, MNRAS, 478, 4513
Cargile, P. A., Conroy, C., Johnson, B. D., et al. 2019, arXiv:1907.07690
Choi, J., Dotter, A., Conroy, C., et al. 2016, ApJ, 823, 102
Cottaar, M., Covey, K. R., Foster, J. B., et al. 2015, ApJ, 807, 27
Cottaar, M., Covey, K. R., Meyer, M. R., et al. 2014, ApJ, 794, 125
Covey, K. R., Agüeros, M. A., Law, N. M., et al. 2016, ApJ, 822, 81
Da Rio, N., Tan, J. C., Covey, K. R., et al. 2016, ApJ, 818, 59
Da Rio, N., Tan, J. C., Covey, K. R., et al. 2017, ApJ, 845, 105
D'Orazi, V., Biazzo, K., & Randich, S. 2011, A&A, 526, A103
D'Orazi, V., Randich, S., Flaccomio, E., et al. 2009, A&A, 501, 973
Doppmann, G. W., Greene, T. P., Covey, K. R., & Lada, C. J. 2005, AJ, 130, 1145
Fabbro, S., Venn, K. A., O'Briain, T., et al. 2018, MNRAS, 475, 2978
Foster, J. B., Cottaar, M., Covey, K. R., et al. 2015, ApJ, 799, 136
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
García Pérez, A. E., Allende Prieto, C., Holtzman, J. A., et al. 2016, AJ, 151, 144
Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332
Hacar, A., Alves, J., Forbrich, J., et al. 2016, A&A, 589, A80
Hacar, A., Tafalla, M., Forbrich, J., et al. 2018, A&A, 610, A77
Hayden, M. R., Bovy, J., Holtzman, J. A., et al. 2015, ApJ, 808, 132
He, K., Zhang, X., Ren, S., & Sun, J. 2016, arXiv:1512.03385
Hejazi, N., De Robertis, M. M., & Dawson, P. C. 2015, AJ, 149, 140
Holtzman, J. A., Shetrone, M., Johnson, J. A., et al. 2015, AJ, 150, 148
Kounkel, M., Covey, K., Moe, M., et al. 2019, AJ, 157, 196
Kounkel, M., Covey, K., Suárez, G., et al. 2018, AJ, 156, 84
Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in Advances in Neural Information Processing Systems 25 (NIPS 2012), ed. F. Pereira, C. J. C. Burges, & L. Bottou (Red Hook, NY: Curran Associates, Inc.), 1097
Kunder, A., Kordopatis, G., Steinmetz, M., et al. 2017, AJ, 153, 75
Leung, H. W., & Bovy, J. 2019, MNRAS, 483, 3255
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Mann, A. W., Feiden, G. A., Gaidos, E., Boyajian, T., & von Braun, K. 2015, ApJ, 804, 64
Marigo, P., Girardi, L., Bressan, A., et al. 2017, ApJ, 835, 77
Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16
Nidever, D. L., Holtzman, J. A., Allende Prieto, C., et al. 2015, AJ, 150, 173
Paszke, A., Gross, S., Chintala, S., et al. 2017, Automatic differentiation in PyTorch, https://openreview.net/pdf?id=BJJsrmfCZ
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Sharma, K., Kembhavi, A., Kembhavi, A., et al. 2020, MNRAS, 491, 2280
Simonyan, K., & Zisserman, A. 2014, arXiv:1409.1556
Stassun, K. G., Collins, K. A., & Gaudi, B. S. 2017, AJ, 153, 136
Stauffer, J. R., Schultz, G., & Kirkpatrick, J. D. 1998, ApJL, 499, L199
Szegedy, C., Liu, W., Jia, Y., et al. 2015, arXiv:1409.4842
Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2019, ApJ, 879, 69
Veyette, M. J., Muirhead, P. S., Mann, A. W., et al. 2017, ApJ, 851, 26
Wilson, J. C., Hearty, F., Skrutskie, M. F., et al. 2010, Proc. SPIE, 7735, 77351C
Yao, Y., Meyer, M. R., Covey, K. R., Tan, J. C., & Da Rio, N. 2018, ApJ, 869, 72
Zasowski, G., Cohen, R. E., Chojnowski, S. D., et al. 2017, AJ, 154, 198