Draft version July 19, 2024
Typeset using LaTeX **twocolumn** style in AASTeX62

Transfer Learning for Improved Stellar Parameter Estimation

Anya Phillips,[1, 2] Cecilia Garraffo,[1] Joshua Wing,[1] and Phillip Cargile[1]

[1] *Center for Astrophysics | Harvard & Smithsonian*
60 Garden St.
Cambridge, MA 02138, USA
[2] *Department of Astronomy, The Ohio State University*
140 West 18th Avenue
Columbus, OH 43210, USA

ABSTRACT

Accurately determining stellar ages and masses is crucial to studying many astrophysical phenomena. We often estimate these parameters by fitting observations to theoretical stellar evolutionary tracks. This can be computationally expensive and unreliable. StelNet is a Hierarchical Bayesian model of Deep Neural Networks that determines stellar mass and age given effective temperature and luminosity. StelNet has been trained on synthetic stellar evolution models and it performs best within that domain. Stellar evolution is a predictive theory but it is not perfect. Each implementation has its own systematic errors. As such, we do not expect a model of StelNet trained on synthetic data to perform well on real observations. In this work, we implement transfer learning with several catalogues of stars with reliable characterizations to adapt StelNet for improved performance on real observations. We show that transfer learning improves StelNet's performance on both the catalogues from which we drew training data as well as two catalogues outside of the training data. The resulting model is robust against systematic errors and allows us to quickly, automatically, and accurately characterize stars from large data sets. This will prove timely for the next generation of observatories and provide a new data set of stellar age and mass estimates for future studies.

Keywords: Stellar properties — Stellar ages — Stellar masses — Computational methods

1. INTRODUCTION

Accurate stellar age and mass estimates offer a window into a wide range of astrophysical problems. This includes the evolution of the stars themselves, the planetary systems that surround them, and the galaxy as a whole. For example, "ground truth" age and mass estimates for stars can help constrain models of main sequence spin down (e.g., McQuillan et al. 2014). Additionally, the age of a star can be used to determine irradiation and space weather conditions around it, permitting an assessment of the stability of the atmospheres of orbiting planets (e.g., Garraffo et al. 2022). Further still, analysis of large samples of stellar ages also allow us to study Galactic evolution over cosmic time (e.g., Bovy et al. 2019).

Given initial mass and metallicity, stellar evolution models predict the radius, luminosity, and effective temperature of a star as a function of its age (e.g., Choi et al. 2016; Bressan et al. 2012). If the distance to a star is known, effective temperature and luminosity can be derived from standard photometry, after which mass and age can be estimated by fitting observed temperatures and luminosities to the predictions of a theoretical evolutionary track or isochrone (e.g., Jørgensen & Lindegren 2005; Jurić et al. 2008; Breddels et al. 2010; Burnett & Binney 2010; Binney et al. 2014; Angus et al. 2019).

However, this method of age and mass determination can be computationally expensive and at times unreliable. In some regions of the H-R diagram (i.e., temperature-luminosity space), different stellar evolutionary tracks cross each other, meaning that a single instance of temperature and luminosity may correspond to multiple ages or masses (Sahlholdt et al. 2019). This results in variable precision of model-based age and mass predictions depending on location within the H-R Di-

Corresponding author: Anya Phillips
anya.phillips@cfa.harvard.edu

agram. Moreover, due to differences in evolutionary timescales for stars of different masses and the stellar initial mass function, certain combinations of temperature and luminosity are more commonly observed than others. Age estimates based on a best-fitting model do not account for this, introducing a bias toward older derived ages (Pont & Eyer 2004). On top of the difficulties with model accuracy and precision, interpolation within model stellar evolutionary tracks or isochrones is computationally expensive and impractical for large data sets. This is especially true when we implement probabilistic methods to estimate the model's uncertainty.

As an alternative to traditional model fitting methods, Garraffo et al. (2021) introduced StelNet, a Hierarchical Bayesian model of Deep Neural Networks to estimate the masses and ages of solar-metallicity stars given their luminosities and effective temperatures. Degeneracies in stellar evolutionary tracks arise primarily between times before and after reaching zero-age main sequence (ZAMS). StelNet combats this by hierarchically combining estimates from both pre- and post-ZAMS models, assigning a probability to each possible output based on the initial mass function and the time a star of the predicted mass spends before and after ZAMS (see Section 3.2 of Garraffo et al. 2021 for further detail). StelNet also quantifies the uncertainty in its outputs. By passing inputs through multiple pre- and post-ZAMS models, each trained on bootstraps of the original training data set, it can output the posterior probability distribution of age and mass estimates from each of the bootstrapped models. Once trained, StelNet is computationally efficient in performing inferences for large data sets.

StelNet is trained on solar metallicity evolutionary tracks from the Modules for Experiments in Stellar Astrophysics (MESA) Isochrones and Stellar Tracks (MIST; Choi et al. 2016; Dotter 2016). While it performs in this regime, StelNet is not expected to predict age and mass based on real photometry as accurately as it does for MIST data. First, stellar evolution models are imperfect, and each implementation, including MIST, has its own systematic errors. In addition, real observations of stars include observational uncertainties, which synthetic training data do not possess. This means that the fitted stellar parameters for an observed star are expected to be offset from those predicted by MIST-based models.

To mitigate this issue, Garraffo et al. (2021) suggest implementing transfer learning with well-characterized stars to calibrate StelNet for optimal performance on real data. Transfer learning involves re-training an already-trained model on a smaller data set and for a smaller number of iterations compared to the original training process. Doing so takes advantage of the knowledge gained from training with the original data set and adjusts the model for performance on a similar but distinct data set by fine-tuning its weights and biases. Garraffo et al. (2021) applied transfer learning to StelNet, re-training it with data from David & Hillenbrand (2015) (hereafter D&H), whose catalogue contains measurements of the physical parameters $T_{\rm eff}$, mass, and age for early-type (BAF) stars (all within 0.5 dex of solar metallicity). The re-trained model performed more accurately than the baseline on stars in the D&H catalogue. Nonetheless, while D&H determine ages and masses using a reliable probabilistic model fitting approach, they use a different set of isochrones from MIST, introducing a systematic that will weaken this model's performance on other datasets.

In this work, we further calibrate StelNet using transfer learning with the following data sets of both pre- and post-ZAMS stars in addition to the sample from D&H:

- A subsample of near-solar metallicity post-ZAMS stars in the *Gaia* FGK benchmark sample from the European Southern Observatory (ESO; Blanco-Cuaresma et al. 2014). These stars have temperature and luminosity determined through largely model-independent means, as well as readily available age and mass estimates.

- A sample of subgiants from Godoy-Rivera et al. (2021). The authors of this work demonstrate that where a precise luminosity can be determined, subgiants are ideal targets for precise age and mass estimates from isochrone fitting due to their rapid evolution at almost constant luminosity.

- The catalogue of pre-ZAMS Orion Nebula Cluster (ONC) stars from Hillenbrand (1997), which includes an individual age estimate for each star, rather than a single cluster age.

- The catalogue of pre-ZAMS stars in the NGC 6530 cluster from Henderson & Stassun (2012). Like the ONC catalogue, this one also includes an individual age estimate for each star.

Our approach avoids overfitting to a particular systematic of any one stellar evolution model by including catalogues with ages and masses determined by fitting to multiple sets of isochrones (see Section 2 for further detail).

We provide additional details of the data used for transfer learning in Section 2 and comment on transfer learning procedures in Section 3. We evaluate the resulting models in Section 4 and discuss conclusions and future directions in Section 5.

2. DATA

In order to train StelNet with non-synthetic data, we require a sample of stars with well-determined effective temperature ($T_{\rm eff}$), luminosity, age, and mass. Because the baseline StelNet models are trained on solar metallicity MIST evolutionary tracks, we constrain any additional data to be within 0.5 dex of solar metallicity, consistent with previous training using the D&H catalogue in Garraffo et al. (2021).

The locations of each new training set on the H-R Diagram are shown in Figure 1. Regions populated with training data are where we expect to improve StelNet's performance through transfer learning. The goal is to adjust performance in these regions while leaving the model's performance in other regions unchanged.

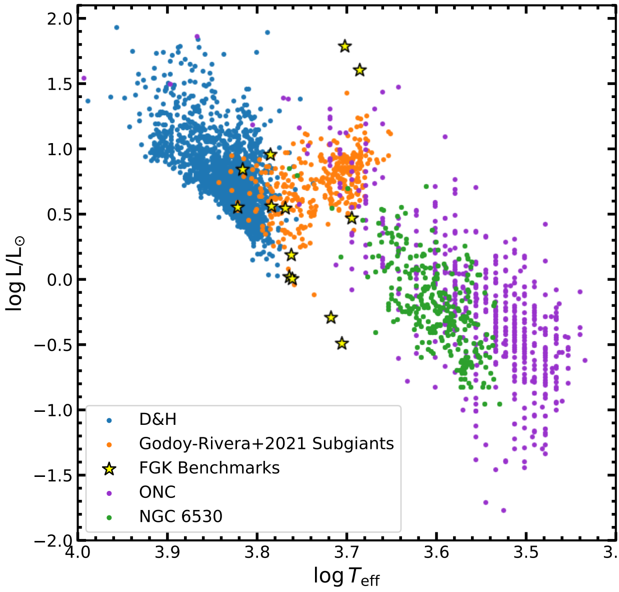

Figure 1. Distributions in $T_{\rm eff}$ and luminosity of the three post-ZAMS (including the D&H catalogue) and the two pre-ZAMS data sets to be used for transfer learning in the present work. Populated regions are where we expect to make the biggest improvements to StelNet's performance on actual data, while leaving performance in other regions as they are.

The ages and masses of our training samples are determined through fitting a range of different stellar evolutionary models (see Sections 2.1 and 2.2 for further details). We expect systematic biases to be introduced due to the difference in isochrone choice for each sample, but our goal in this work is to create a model that is robust against systematic errors and biases. Ultimately, we expect to compromise precision with the goal of a more general model that avoids overfitting to any particular systematic.

2.1. *Post-ZAMS training sets*

First, we use the assembled catalogue from Garraffo et al. (2021) of D&H BAF type stars, which include the age, mass, and $T_{\rm eff}$ estimates from D&H, and luminosity from the TESS input catalogue (Stassun et al. 2019). The authors performed Bayesian analysis with solar-metallicity PARSEC isochrones (Bressan et al. 2012) to determine ages and masses for this sample, and all stars in it have $|[{\rm Fe/H}]| < 0.5$. The full catalogue has 1869 stars with estimates for all of our required parameters.

Second, we use the *Gaia* FGK benchmark stars from the ESO (Blanco-Cuaresma et al. 2014). Out of the 34 stars in this catalogue, 20 have $|[{\rm Fe/H}]| < 0.5$ based on metallicity estimates from Jofré et al. (2014). We use $T_{\rm eff}$, luminosity, and mass estimates from Heiter et al. (2015). These come from averaging estimates from *Padova* (Bertelli et al. 2008, 2009) and *Yonsai-Yale* (Yi et al. 2003; Demarque et al. 2004) isochrones. For age estimates, we employ the results of Sahlholdt et al. (2019), who determine upper and lower limits (all of which have ranges within 3 Gyr) for 16 of the ESO benchmark stars, 12 of which have $|[{\rm Fe/H}]| < 0.5$. In accordance with the authors' recommendation, we adopt the median of the reported age range as a single age value for each of these stars, with the range given as the uncertainty. We include the Sun in this sample, resulting in a catalogue of 13 stars with estimates for all of our required parameters.

Third, we use the catalogue of subgiants from Godoy-Rivera et al. (2021). Though this catalogue does not include [Fe/H], we expect the mean of their distribution in [Fe/H] to be between -0.5 and 0.5, given that their distances are all $<$ 1 kpc. Godoy-Rivera et al. (2021) obtained parameters for these stars through SED fitting and MIST-based modeling software. While this sample occupies a limited regime of the post-ZAMS H-R diagram, its estimated ages are determined with unique precision due to the rapid evolution of subgiants at near-constant luminosity. In total this catalogue has 340 stars with estimates for all of our required parameters.

2.2. *Pre-ZAMS training sets*

Our data for training pre-ZAMS models consists fully of open clusters within 0.5 dex of solar metallicity. We use the catalogue of Orion Nebula Cluster (ONC) stars from Hillenbrand (1997), who determine ages and masses by fitting to isochrones from D'Antona & Mazzitelli (1994) and Swenson et al. (1994). This catalogue has 551 stars with estimates of all our required parameters. In addition, we use the catalogue of stars belonging to NGC 6530 from Henderson & Stassun (2012), who list masses and ages determined using isochrones from Siess et al. (2000). Henderson & Stassun (2012) list all of our required parameters for 580 stars.

3. TRANSFER LEARNING

The baseline StelNet model consists of 20 "pre-ZAMS" and 20 "post-ZAMS" neural networks, where each of the neural networks are trained on a different bootstrap (sample with replacement) of the MIST training data. The pre- and post-ZAMS bootstraps are selected from pre- and post-ZAMS portions of the MIST evolutionary tracks, respectively. Passing inputs through multiple neural networks trained on bootstrapped data results in a Bayesian model, where the combined output of the networks is a posterior probability rather than a single estimate, quantifying the uncertainty of the model (Gal & Ghahramani 2016).

In traditional machine learning approaches, it is assumed that future data to be fed to a model belong to the same data set as the training data. However, this assumption does not hold for baseline StelNet models, which are trained on synthetic evolutionary tracks but are meant to be used on actual observations. Transfer learning leverages knowledge acquired from a pre-existing machine learning model to enhance performance on a new data set (see Pan & Yang 2010 for a more complete review). Beginning the training process from scratch with observations of well-characterized stars would require enormous amounts of new training data. Instead, transfer learning involves re-training the baseline MIST models of StelNet with comparatively little new data, effectively calibrating the baseline models for better performance on actual observations by fine-tuning the original model's weights and biases.

In the original training process Garraffo et al. (2021) began with a neural network with randomized weights and biases. With transfer learning, we begin training by loading a MIST-trained StelNet bootstrap network and train it for relatively few iterations on the new observational data.

We implement transfer learning separately for StelNet's pre- and post-ZAMS models, so baseline pre-ZAMS models are retrained only on pre-ZAMS data and baseline post-ZAMS models are retrained only on post-ZAMS data. We begin with a single network from the baseline models and retrain it with 20 different bootstraps of the new training data, resulting in 20 neural networks each for pre- and post-ZAMS.

For the post-ZAMS training, we include data from the FGK benchmarks, the D&H sample, and the Godoy-Rivera et al. (2021) subgiants. Because the FGK benchmark stars have the most reliable parameter estimates, we give them more weight in the training process by making five copies of each star in the sample (yielding 65 stars), and randomly selecting roughly 50 for training. While it is not common practice to use such a high fraction of the available data for training, our aim is to prioritize the model's performance on the FGK benchmarks, testing the resulting models on other observations. We also include roughly 30 randomly selected stars each from D&H and from Godoy-Rivera et al. (2021).

For the pre-ZAMS training, we include roughly 30 randomly selected stars each from the Hillenbrand (1997) catalogue of ONC stars and the Henderson & Stassun (2012) catalogue of stars in the NGC 6530 cluster.

For both the pre- and post-ZAMS models, we also include additional MIST data points from the pre- and post-ZAMS portions of the evolutionary tracks, respectively. The observations selected for transfer learning do not uniformly populate the space of $T_{\rm eff}$, luminosity, age, and mass, so by including MIST data points, we avoid overfitting to observations and anchor the model to the predictions of stellar evolution where we do not have new training data. We randomly select 30 points from the entire MIST data set to include in training and then select additional MIST data outside the mass range of the transfer learning data. We add 60 additional MIST points (30 with masses lower than the lowest-mass new data point and 30 with masses higher than the highest-mass new data point). We train each pre- and post-ZAMS bootstrap on the new training data set for 100 gradient descent steps.

4. MODEL EVALUATION

In this section, we compare the performance of the baseline StelNet models and our pre- and post-ZAMS models after transfer learning with the specifications detailed in the previous section. In addition to visual comparison of the predicted and true parameters of each data set (see Figures 2 and 3), we can quantitatively validate our models through two methods.

First, we combine all of the data from which each training sample was selected and compute the Pearson correlation coefficient between the StelNet predicted parameters and their true values. The Pearson coefficient offers a metric of how well two parameters fit some linear relationship. While it does not test for whether the two parameters follow the one-to-one correspondence we strive for in this work, we use it in tandem with visual inspection to get a sense of the effects of transfer learning (see the discussion of Figures 2 and 3).

Additionally, we follow the prescription in Verde et al. (2013), who employ the tension parameter, τ, to evaluate the consistency between two distributions in a Bayesian approach. We can calculate the tension be-

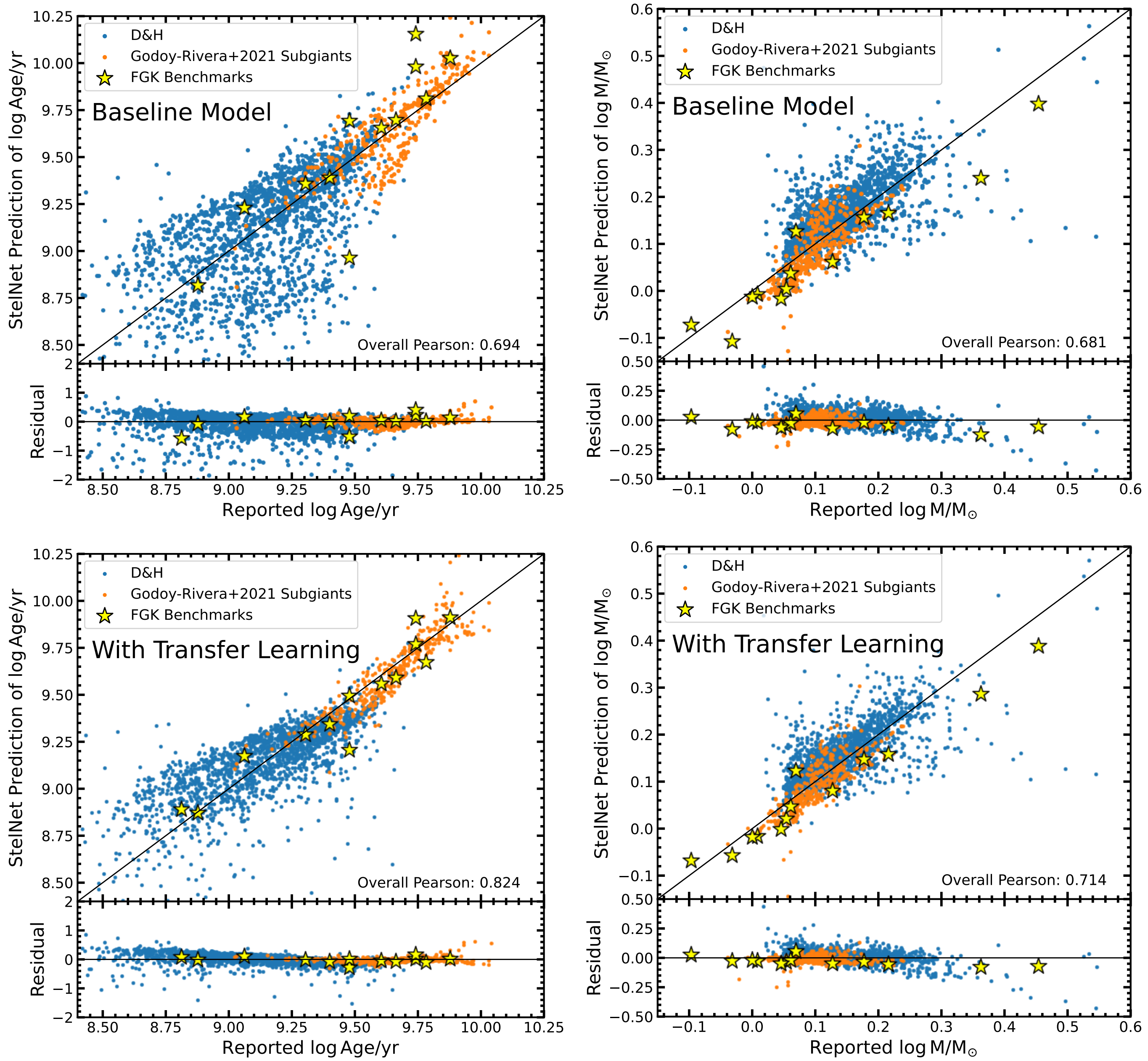

Figure 2. StelNet predictions for age (left panels) and mass (right panels) versus the reported value for the post-ZAMS training datasets, which include the D&H sample, the Godoy-Rivera et al. (2021) subgiants, and the FGK benchmark stars. The predictions in the top panels were made with the baseline StelNet models and the predictions in the bottom panels were made with StelNet after transfer learning with the post-ZAMS training sets. The Pearson coefficient for the predictions and true values of the combined data sets is displayed in the lower right of each panel. The bottom panels show the residuals.

tween two distributions, P_{pred}, the posterior predicted by StelNet with mean μ_p and standard deviation σ_p, and P_{true}, a Gaussian whose mean μ_t is the observed value of a stellar parameter and whose standard deviation σ_t is it's reported error. Given these two Gaussian distributions, we can calculate the Bayesian evidence, E, or the area of overlap between the two distributions, as described in Garraffo et al. (2021):

$$E = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_p^2 - \sigma_t^2}} \exp\left[-\frac{1}{2} \frac{(\mu_p - \mu_t)^2}{\sigma_p^2 - \sigma_t^2} \right]. \quad (1)$$

We then compute the normalized tension:

$$\tau = \frac{\bar{E}|_{\mathrm{maxA=maxB}}}{E}, \quad (2)$$

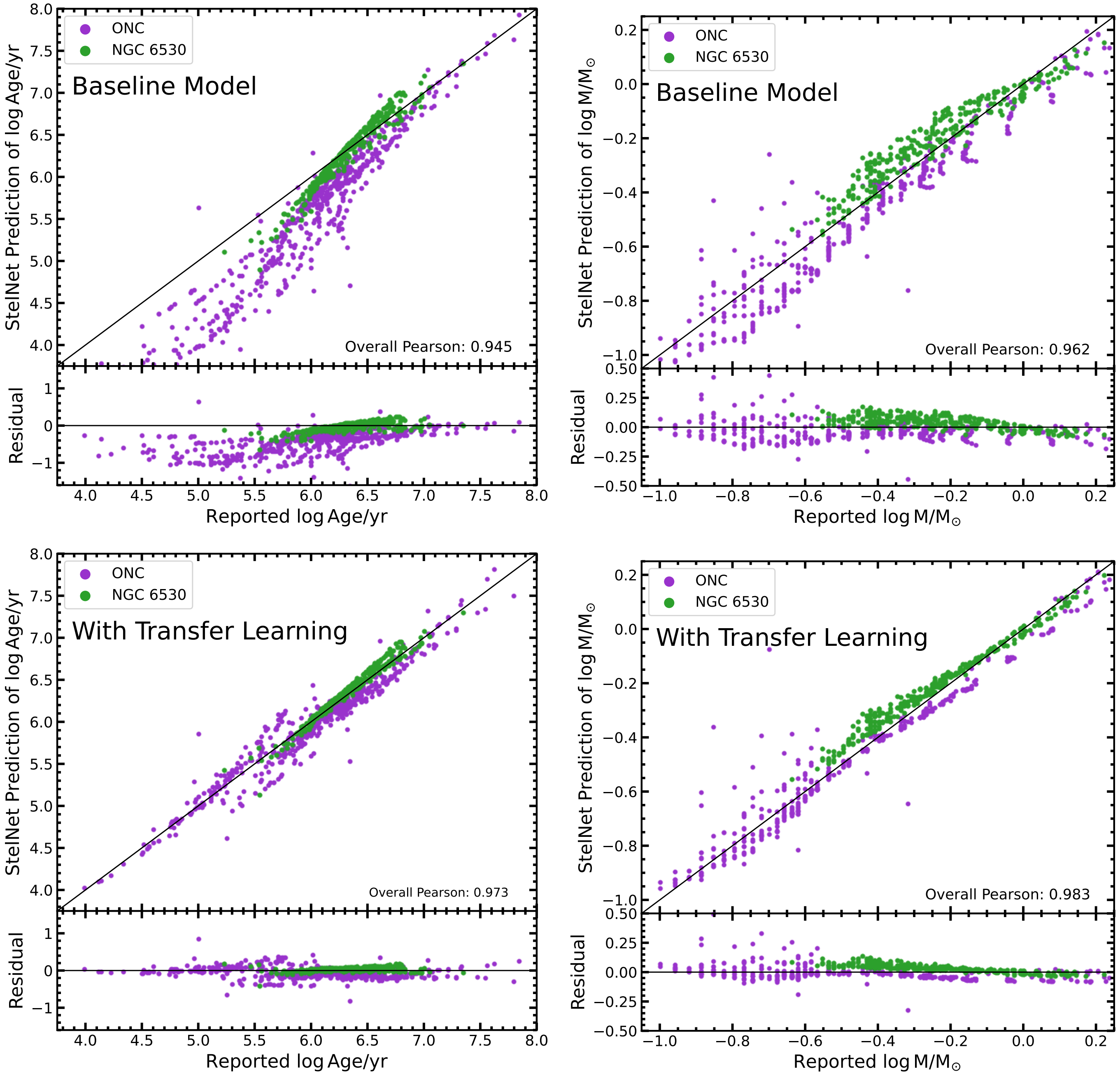

Figure 3. StelNet predictions for age (left panel) and mass (right panel) versus the reported value for the pre-ZAMS training sets, which include the ONC data from Hillenbrand (1997) and the the NGC 6530 data from Henderson & Stassun (2012). The Pearson coefficient for the predictions and true values of the combined data sets is displayed in the lower right of each panel. The predictions in the top panels were made with the baseline StelNet models and the predictions in the bottom panels were made with StelNet after transfer learning with the pre-ZAMS training sets. The bottom panels show the residuals.

where $\bar{E}|_{\mathrm{maxA=maxB}}$ is the area of overlap between the two distributions if the difference in their means is set to zero ($\mu_p = \mu_t$). Thus, as the two distributions become less similar, the tension τ between them increases. We interpret the tension using the Jeffreys scale (Jeffreys 1973), which asserts that $\log \tau < 1$ means that the difference between the two distributions is not significant, so we consider our predictions consistent with the ground truth only if $\log \tau < 1$.

As a minor caveat, while StelNet outputs a Gaussian mean and standard deviation to characterize its error as a model, the errors reported for stellar age and mass are generally asymmetric and non-Gaussian (in particular, this is the case for the data from D&H and Godoy-Rivera

Table 1. Fractions of the FGK benchmark sample, the D&H sample, and the Godoy-Rivera et al. (2021) (G-R) subgiant sample that have $\log\tau < 1$. The top entries use the tension τ between the reported distribution and the baseline StelNet estimate while the bottom entries use the tension between the reported distribution and the StelNet estimate after transfer learning. We include tensions in both age and mass distributions. We consider StelNet's predictions to be consistent with the ground truth when $\log\tau < 1$, so the fractions displayed in this table are the fractions of the sample with accurate StelNet characterizations.

Sample	Fraction with $\log\tau < 1$	
Baseline Model	Age	Mass
FGK Benchmarks	0.667	0.333
D&H	0.592	0.689
G-R Subgiants	0.856	1.000
After Transfer Learning	Age	Mass
FGK Benchmarks	0.917	0.583
D&H	0.793	0.750
G-R Subgiants	0.938	1.000

et al. 2021). In order to calculate a tension for such observed distributions, we conservatively use the narrower side of each error as σ_t, which minimizes the width of the assumed Gaussian. As a result, the tensions we compute can be considered upper limits, since we take the error that assumes the two distributions are as dissimilar as possible.

4.1. *Performance on Training Data Sets*

Figure 2 compares the StelNet prediction to the reported value for the data in each of the catalogues from which we drew post-ZAMS training data for both the baseline StelNet model and the model after transfer learning. It also includes the Pearson correlation coefficient between the predicted and true values of all of the data sets combined for mass and age in each model.

The transfer learning age estimates improve significantly with the Pearson coefficient increasing from 0.694 to 0.824. Notably, the model after transfer learning performs accurately on all of the FGK benchmark stars. The visual improvement between masses estimated with the baseline model and the model after transfer learning is less obvious, particularly for the D&H data, but the Pearson coefficient increases from 0.681 to 0.714.

We compute the tension of each prediction with the reported values in age and mass for the three catalogues used in post-ZAMS transfer learning. The fraction of each sample that have $\log\tau < 1$ are displayed in Table 1. We see that all of the fractions improve from the baseline to the transfer learning model.

Figure 3 compares the predictions from StelNet to the reported values for all of the data in both catalogues from which we drew pre-ZAMS training data for the baseline StelNet model and the model after transfer learning. Each panel includes the Pearson coefficients between ground truth and predicted values for both models. For the age estimates, we see that while the baseline model yields a relatively high Pearson coefficient of 0.945, it systematically underestimates the ages of younger stars. This effect is corrected after transfer learning, where the Pearson coefficient is raised to 0.973. For the masses, we decrease the residuals and increase the overall Pearson coefficient from 0.962 to 0.983 through transfer learning. Unfortunately, the Hillenbrand (1997) and Henderson & Stassun (2012) catalogues for the ONC and NGC 6530 do not report the uncertainty in their age and mass estimates, so we are unable to compute tensions between the true and predicted distributions as we did with the post-ZAMS data sets.

4.2. *Performance on Single-Age Open Clusters*

We showed in the previous section that transfer learning improves StelNet's performance for stars in the catalogues that include its new training data. However, our goal in transfer learning is to create a model that performs well on observations outside these catalogues as well. For additional test data we use the catalogue from Wright et al. (2011), which includes the temperatures, luminosities, and masses of stars belonging to several open clusters, including Praesepe and NGC 2547. We choose these two clusters for testing because all of the stars within each are in the same evolutionary stage. Figure 4 shows these clusters' ages versus the mass of each member. The figure also displays the ZAMS age as a function of mass. This boundary is given by setting the equivalent evolutionary point (EEP) parameter in the MIST tables to 202 (see Garraffo et al. 2021). We see that Praesepe consists only of post-ZAMS stars and NGC 2547 consists only of pre-ZAMS stars. Other clusters in the Wright et al. (2011) catalogue either have fewer stars or are of intermediate ages such that their lower-mass members are pre-ZAMS and their higher-mass members are post-ZAMS.

The authors determine stellar masses in this catalogue by fitting to isochrones from Siess et al. (2000). Praesepe and NGC 2547 also have cluster ages determined in Bossini et al. (2019) using Bayesian fitting methods to PARSEC isochrones, so while we do not have individual age estimates for each star, we can compare the age output of StelNet for a star in either of these clusters to the cluster age reported in this catalogue.

In the left panel of Figure 5, we show the distribution in predicted ages for Praesepe members from

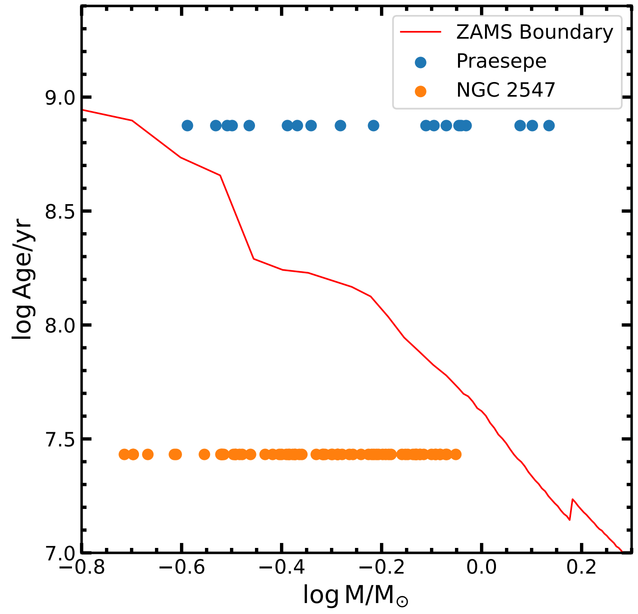

Figure 4. Distribution in log (age/yr) and log (M/M$_\odot$) for Praesepe (blue) and NGC 2547 (orange). A red line indicates the ZAMS age as a function of initial mass, showing that Praesepe consists fully of post-ZAMS stars and NGC 2547 consists fully of pre-ZAMS stars.

StelNet compared to the Praesepe age estimate from Bossini et al. (2019) (log (age/yr) = 8.875, with a lower limit of 8.871 and an upper limit of 8.877). The right panel of Figure 5 compares the StelNet predicted masses for Praesepe members to their true values reported in Wright et al. (2011). Both the baseline StelNet model and the model after post-ZAMS transfer learning yield overestimates of stellar age, though the model after transfer learning shifts slightly toward younger ages. Both models perform reasonably well for stellar mass estimates, though the model after transfer learning tends to slightly overestimate the masses of stars with true mass log (M/M$_\odot$) $\lesssim -0.3$. We note our post-ZAMS training sample contained very few real stars with M $\lesssim$ 1M$_\odot$. We aimed to anchor the model to MIST predictions at low masses by including extra low-mass MIST data points but MIST predictions are clearly biased toward higher ages as evidenced by the baseline model's performance in the left panel.

Figure 6 similarly compares StelNet's age and mass predictions for NGC 2547 to their reported values in Bossini et al. (2019) and Wright et al. (2011). In this pre-ZAMS case, transfer learning widens the distribution in age estimates but counteracts the systematic age overestimation seen in the predictions of the baseline models. Both the baseline model and the model after pre-ZAMS transfer learning perform reasonably well for stellar mass estimates, though the model after transfer learning seems to perform better than the baseline at low masses (log (M/M$_\odot$) $\lesssim -0.2$). Our pre-ZAMS training data tended to have lower masses (see the lower right panel of Figure 3), a bias introduced by the fact that lower mass stars spend more time on the pre-main sequence, so it is again unsurprising that our pre-ZAMS transfer learning improves StelNet's performance on low-mass stars.

Bossini et al. (2019) report an age for NGC 2547 of log (age/yr) = 7.432 with a lower limit of 7.414 and an upper limit of 7.449, and the distribution in predicted ages for the StelNet model after pre-ZAMS transfer learning is considerably wider than this range. However, in Figure 7 we show the distributions of individual literature ages for ONC and NGC 6530 that we used for transfer learning, both of which have considerably more spread than the range reported by Bossini et al. (2019) for NGC 2547. It is not unreasonable for StelNet to have estimated a wider range of ages than the margin of error in NGC 2547's age from Bossini et al. (2019).

5. CONCLUSIONS

We have separately implemented transfer learning to calibrate the pre- and post-ZAMS components of StelNet for improved performance on real data. We summarize our results below:

- The transfer learning models result in better correlations between predicted stellar parameters and their true values for the catalogues from which the training data were drawn.
- Fewer stars in the catalogues from which the post-ZAMS training data were selected have true distributions in high tension with their StelNet predictions in the transfer learning models than in the baseline StelNet models.
- We highlight that StelNet's performance on low-mass (< 1M$_\odot$) post-ZAMS stars still needs further improvement, as evidenced by its performance on members of the Praesepe cluster. Our post-ZAMS training sample consisted overwhelmingly of stars with M > 1M$_\odot$, so additional genuine low-mass training data (as opposed to simply augmenting the MIST data in thie regime) is required to resolve this issue.
- Transfer learning with ONC and NGC 6530 modestly improves StelNet's performance on low-mass pre-ZAMS stars. Our new model estimates a distribution of stellar ages in NGC 2547 more consistent with the literature age of the cluster than

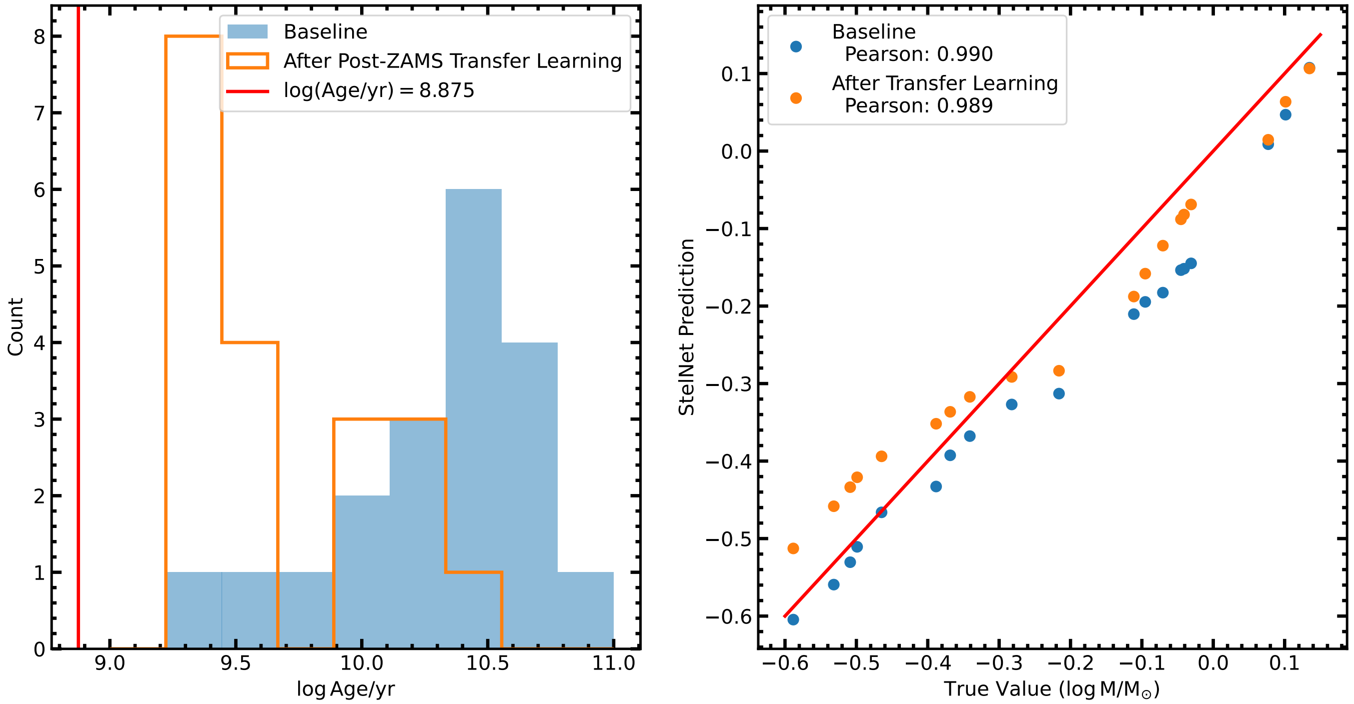

Figure 5. (Left panel) Distribution in StelNet age predictions for Praesepe members from the baseline model (blue) and the model after transfer learning (orange). A vertical red line indicates the age of the cluster as reported in Bossini et al. (2019). (Right panel) Comparison of StelNet predictions and true values of Praesepe members for the baseline model (blue) and the model after transfer learning (orange). Both models overestimate stellar age, though the model after transfer learning shifts to younger estimates. The model after transfer learning systematically overestimates the masses of stars with $\log{(\mathrm{M}/\mathrm{M}_\odot)} \lesssim -0.1$, likely due to the lack of higher-mass post-ZAMS training data.

the baseline StelNet model. While our pre-ZAMS training set did not consist of more massive stars ($\log{(\mathrm{M}/\mathrm{M}_\odot)} \gtrsim 0.2$), the amount of time a star spends on the pre-main sequence decreases with its initial mass, so the likelihood of observing a high-mass pre-ZAMS star is relatively low (a fact taken into account when StelNet's full hierarchical model is run).

There is a massive scope for expanding and continuing to improve StelNet going forward, both in our training sample, and in the architecture of the model itself.

In this work, we implemented training with age and mass estimates from many different stellar evolution models, all of which have their own systematics. Future transfer learning might employ homogeneous training sample, with ages and masses all estimated using the same model (ideally, MIST). This would yield a precise model up to a single systematic, rather than one that is robust against systematics at the cost of its precision.

Alternatively, StelNet could benefit from a training sample with age and mass estimates from different methods customized to the variety of the star in question. For example, asteroseismology is considered a highly reliable method of age determination for red giant stars but is not a feasible technique for main-sequence stars, whose oscillations are impossible to detect at standard observational cadences. On the other hand, gyrochronology models are precisely calibrated to main sequence stars (particularly in open clusters), but are not a viable age determinant in giant stars, which lack consistent age-rotation relationships. Lithium depletion has a well-understood relationship with stellar age but only for pre-ZAMS stars, before primordial lithium has been completely depleted. Lastly, abundances of radioactive isotopes provide accurate age estimates, but such measurements require extremely high-resolution spectroscopy. This data is available for very few stars. Ultimately, such a training sample may be inhomogeneous but could result in a StelNet model that is more accurate and avoids the systematic limitations of training with age and mass estimates from model fitting.

Apart from its training sample, there are multiple avenues of improvement to the architecture of StelNet itself. First, while StelNet provides an estimate of its epistemic uncertainty by bootstrapping multiple trained models (see Section 3.4 of Garraffo et al. 2021 for details), it does not currently take into account uncertainty in the initial inputs. Going forward, we hope to modify

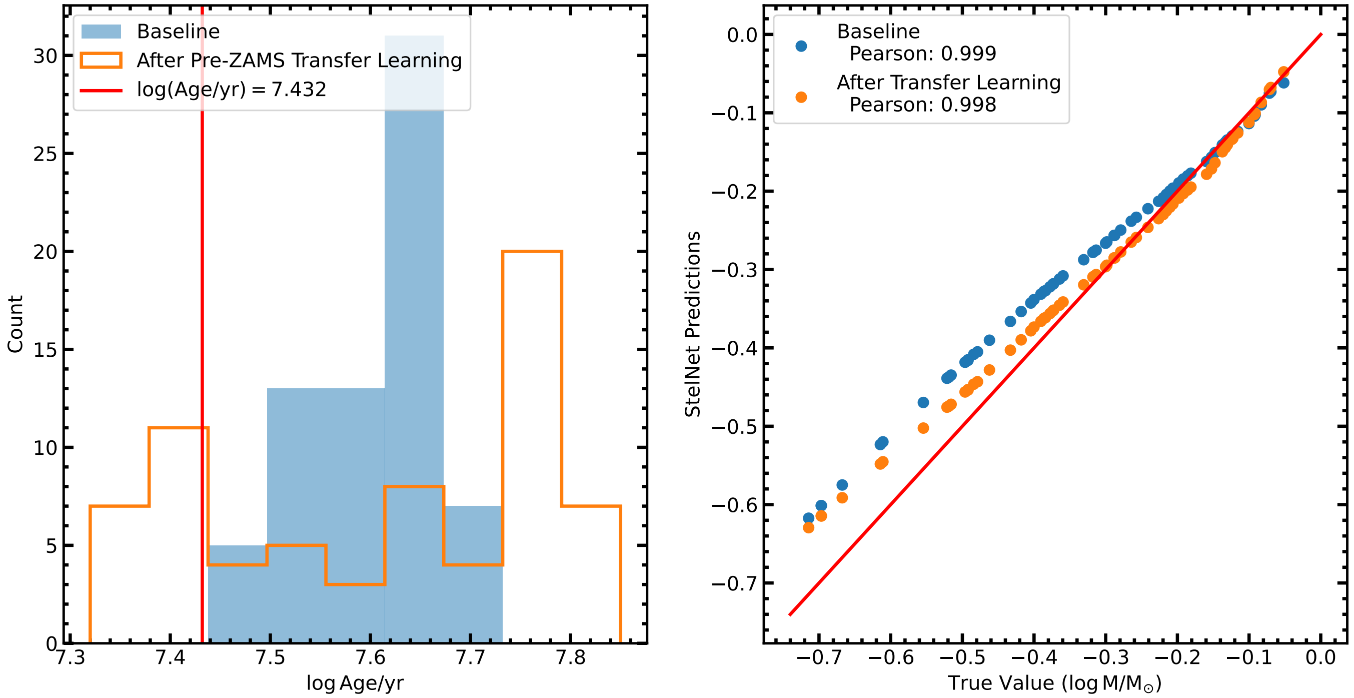

Figure 6. (Left panel) Distribution in StelNet age predictions for NGC 2547 members from the baseline model (blue) and the model after transfer learning (orange). A vertical red line indicates the age of the cluster as reported in Bossini et al. (2019). (Right panel) Comparison of StelNet predictions and true values of NGC 2547 members for the baseline model (blue) and the model after transfer learning (orange). Transfer learning counteracts the systematic overestimates of stellar age seen in the baseline. Both models yield reasonably good mass estimates, though the model after transfer learning seems to perform better than the baseline at low masses.

StelNet to sample posterior distributions in $T_{\rm eff}$ and luminosity, pass each sample through the existing model, and return the resulting distribution in mean StelNet estimates. This would allow us to propagate observational uncertainty through StelNet, yielding more realistic uncertainties in mass and age. Modifying StelNet to take distributions as inputs would allow us to use it in conjunction with models that output posteriors, such as ThePayne (Ting et al. 2019).

The ability to perform age and mass estimates directly from photometry (rather than from the derived $T_{\rm eff}$ and luminosity) would make StelNet a more useful tool to the astronomical community at large. MIST supplies synthetic evolutionary tracks in the *Gaia* photometric bands. We plan to use these to train new StelNet models before implementing transfer learning with stars where *Gaia* photometry and reliable age and mass estimates are available.

Another clear next step is to add metallicity as an output dimension to StelNet. Including isochrones of different metallicities in StelNet's training will introduce more degeneracies into the input space. We can circumvent this with the same hierarchical strategy StelNet already use between the pre- and post-ZAMS.

The SAO REU program is funded in part by the National Science Foundation REU and Department of Defense ASSURE programs under NSF Grant no. AST-2050813, and by the Smithsonian Institution.

REFERENCES

Angus, R., Morton, T., & Foreman-Mackey, D. 2019, The Journal of Open Source Software, 4, 1469, doi: 10.21105/joss.01469

Bertelli, G., Girardi, L., Marigo, P., & Nasi, E. 2008, A&A, 484, 815, doi: 10.1051/0004-6361:20079165

Bertelli, G., Nasi, E., Girardi, L., & Marigo, P. 2009, A&A, 508, 355, doi: 10.1051/0004-6361/200912093

Binney, J., Burnett, B., Kordopatis, G., et al. 2014, MNRAS, 437, 351, doi: 10.1093/mnras/stt1896

Blanco-Cuaresma, S., Soubiran, C., Jofré, P., & Heiter, U. 2014, A&A, 566, A98, doi: 10.1051/0004-6361/201323153

Bossini, D., Vallenari, A., Bragaglia, A., et al. 2019, A&A, 623, A108, doi: 10.1051/0004-6361/201834693

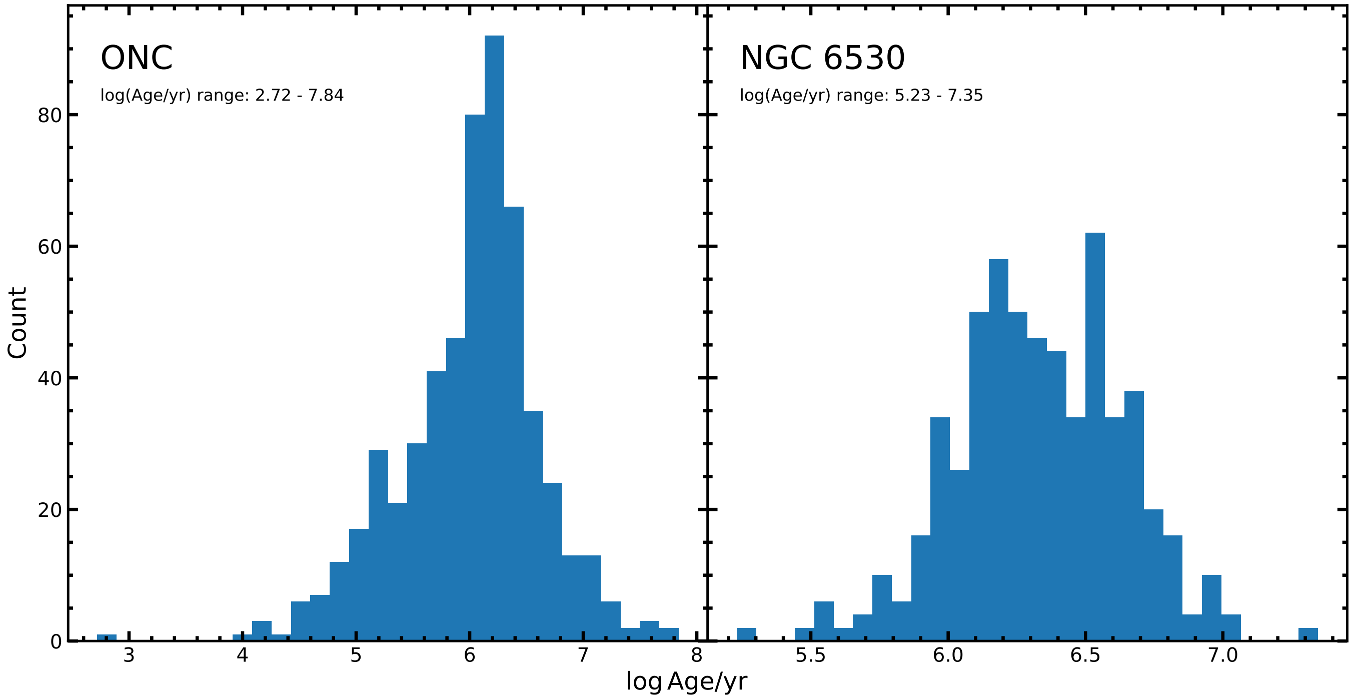

Figure 7. Distributions in individual log (age/yr) for the ONC (left panel) and NGC 6530 (right panel) from Hillenbrand (1997) and Henderson & Stassun (2012) respectively. We report the range the age estimates in each cluster in the top left corner of each panel. We note a wide range in ages for members of each single cluster, in contrast to the narrow error margin for the cluster ages of Praesepe and NGC 2547 reported in Bossini et al. (2019).

Bovy, J., Leung, H. W., Hunt, J. A. S., et al. 2019, MNRAS, 490, 4740, doi: 10.1093/mnras/stz2891

Breddels, M. A., Smith, M. C., Helmi, A., et al. 2010, A&A, 511, A90, doi: 10.1051/0004-6361/200912471

Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127, doi: 10.1111/j.1365-2966.2012.21948.x

Burnett, B., & Binney, J. 2010, MNRAS, 407, 339, doi: 10.1111/j.1365-2966.2010.16896.x

Choi, J., Dotter, A., Conroy, C., et al. 2016, ApJ, 823, 102, doi: 10.3847/0004-637X/823/2/102

D'Antona, F., & Mazzitelli, I. 1994, ApJS, 90, 467, doi: 10.1086/191867

David, T. J., & Hillenbrand, L. A. 2015, ApJ, 804, 146, doi: 10.1088/0004-637X/804/2/146

Demarque, P., Woo, J.-H., Kim, Y.-C., & Yi, S. K. 2004, ApJS, 155, 667, doi: 10.1086/424966

Dotter, A. 2016, ApJS, 222, 8, doi: 10.3847/0067-0049/222/1/8

Gal, Y., & Ghahramani, Z. 2016, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. https://arxiv.org/abs/1506.02142

Garraffo, C., Protopapas, P., Drake, J. J., Becker, I., & Cargile, P. 2021, AJ, 162, 157, doi: 10.3847/1538-3881/ac0ef0

Garraffo, C., Jeremy, D., Alvarado Gómez, J. D., et al. 2022, in 44th COSPAR Scientific Assembly. Held 16-24 July, Vol. 44, 593

Godoy-Rivera, D., Tayar, J., Pinsonneault, M. H., et al. 2021, ApJ, 915, 19, doi: 10.3847/1538-4357/abf8ba

Heiter, U., Jofré, P., Gustafsson, B., et al. 2015, A&A, 582, A49, doi: 10.1051/0004-6361/201526319

Henderson, C. B., & Stassun, K. G. 2012, ApJ, 747, 51, doi: 10.1088/0004-637X/747/1/51

Hillenbrand, L. A. 1997, AJ, 113, 1733, doi: 10.1086/118389

Jeffreys, H. 1973, Scientific inference (Cambridge University Press)

Jofré, P., Heiter, U., Soubiran, C., et al. 2014, A&A, 564, A133, doi: 10.1051/0004-6361/201322440

Jørgensen, B. R., & Lindegren, L. 2005, A&A, 436, 127, doi: 10.1051/0004-6361:20042185

Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864, doi: 10.1086/523619

McQuillan, A., Mazeh, T., & Aigrain, S. 2014, ApJS, 211, 24, doi: 10.1088/0067-0049/211/2/24

Pan, S. J., & Yang, Q. 2010, IEEE Transactions on Knowledge and Data Engineering, 22, 1345, doi: 10.1109/TKDE.2009.191

Pont, F., & Eyer, L. 2004, MNRAS, 351, 487, doi: 10.1111/j.1365-2966.2004.07780.x

Sahlholdt, C. L., Feltzing, S., Lindegren, L., & Church, R. P. 2019, MNRAS, 482, 895, doi: 10.1093/mnras/sty2732

Siess, L., Dufour, E., & Forestini, M. 2000, A&A, 358, 593, doi: 10.48550/arXiv.astro-ph/0003477

Stassun, K. G., Oelkers, R. J., Paegert, M., et al. 2019, AJ, 158, 138, doi: 10.3847/1538-3881/ab3467

Swenson, F. J., Faulkner, J., Rogers, F. J., & Iglesias, C. A. 1994, ApJ, 425, 286, doi: 10.1086/173985

Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2019, ApJ, 879, 69, doi: 10.3847/1538-4357/ab2331

Verde, L., Protopapas, P., & Jimenez, R. 2013, Physics of the Dark Universe, 2, 166, doi: 10.1016/j.dark.2013.09.002

Wright, N. J., Drake, J. J., Mamajek, E. E., & Henry, G. W. 2011, ApJ, 743, 48, doi: 10.1088/0004-637X/743/1/48

Yi, S. K., Kim, Y.-C., & Demarque, P. 2003, ApJS, 144, 259, doi: 10.1086/345101