AI for dating stars: a benchmarking study for gyrochronology

Andrés Moya
Universidad Politécnica de Madrid
a.moya@upm.es

Jarmi Recio-Martínez
University of Alcalá
jarmi.recio@edu.uah.es

Roberto J. López-Sastre
University of Alcalá
robertoj.lopez@uah.es

Abstract

In astronomy, age is one of the most difficult stellar properties to measure, and gyrochronology is one of the most promising techniques for the task. It consists in dating stars using their rotational period and empirical linear relations with other observed stellar properties, such as stellar effective temperature, parallax, and/or photometric colors in different passbands, for instance. However, these approaches do not allow to reproduce all the observed data, resulting in potential significant deviations in age estimation. In this context, we propose to explore the stellar dating problem using gyrochronology from the AI perspective. Technically, we replace other linear combinations and traditional techniques with a machine learning regression approach. For doing so, we introduce a thorough benchmarking study of state-of-the-art AI regression models trained and tested for stellar dating using gyrochronology. Our experiments reveal promising results, where some models report a mean average error < 0.5 Gyr, which can be considered as an outstanding breakthrough in the field. We also release a dataset and propose a set of simple assessment protocols to aid research on AI for dating stars as part of this study. Code and data to reproduce all our results are available at https://github.com/gramuah/ai4datingstars.

1. Introduction

The accurate estimation of stellar ages is critical to study many different astrophysical problems. For example, in searching for life outside our solar system, we know that the impact of life in the atmosphere of the hosting exoplanet evolves with time, making it potentially detectable through biomarkers. Therefore, dating detected exoplanets is critical, and it can only be done via dating its hosting star. In Galactic archaeology, or how our galaxy evolves with time, dating stars is the only way for mapping this evolution.

Among all the stellar characteristics, age is the most difficult variable to be obtained, since it is the only one that cannot be *directly* observed in any case. It must be inferred using diverse methods. Soderblom presented two excellent surveys describing most of these techniques [54, 55].

Within the group of approaches for stellar dating, one of the most promising statistical techniques is gyrochronology or dating stars using their rotational period. Stars are born with a range of rotational velocities for different physical reasons, see [6]. Then, the process called magnetic braking reduces the rotational velocity of the star with time [49]. Since late 60's [10, 26, 37, 38, 59], we know that rotational braking, or braking law, is a function of the rotational period. That is, the larger the rotation, the larger the braking. Therefore, with time, all the stars tend to converge to a similar rotation velocity for a given age and mass. This effect was empirically shown for the first time by Skumanich [53]. Gyrochronology is the stellar dating technique that exploits this fact [9, 54]. Traditionally, gyrochronology has been applied by fitting a linear expression containing the stellar age, its rotational period and some proxies of its mass, to observations. In recent works, Angus *et al.* [2, 4] have proposed novel classic linear relations for gyrochronology combined with other dating techniques, such as isochrones fitting or stellar kinematics, in a Bayesian framework for a more accurate age estimation.

However, possibly because of the difficulty of collecting data samples that are sufficient and accurate, no attempt has been made to approach the problem of age estimation of stars as a regression exercise using artificial intelligence techniques. In this work, we explore this novel perspective, and our main contributions are as follows.

First, we have gathered a dataset for gyrochronology. In general, our knowledge about gyrochronology rests mainly on the shoulders of stellar clusters, the only systems with a large number of stars with an accurate age determination. For this work we have constructed a comprehensive and accurate sampling with stars coming from clusters but also from asteroseismology [1, 39, 56]. The appearance of the asteroseismology as an efficient tool for dating stars has allowed us to add a large number of field stars to the sample. All of the stars have precise effective temperature (T_{eff}), metallicity ([Fe/H]), logarithm of the surface gravity (log g), luminosity (L), mass (M), radius (R), age, and rotational period (P_{rot}). Section 3 details how the data sample has been constructed, analyzing its strengths and weaknesses. We are facing a major challenge when we want to estimate stellar ages using gyrochronology, mainly because we do not fully understand the physics and the physical dependencies of the rotational braking phenomenon yet. However, we believe that this is an interesting problem for machine learning techniques to offer their best.

Second, we present, for the first time, a benchmarking study of state-of-the-art AI regression models using our dataset (see Section 4). We include in the analysis an heterogeneous and complete set of approaches, covering: classical kNN and linear or bayesian regression; ensemble models (Random Forest, Stacking); decision trees; Support Vector Regression machine; Gaussian process; and Neural Networks. We propose up to three different benchmarks, where all these models are compared, to understand their precision, robustness and generalization capabilities.

Finally, in Section 5 our experiments reveal promising results, where some models achieve a mean average error < 0.5 Gyr dating stars, which can be considered a significant advancement in the field. Moreover, so as to encourage further research into this problem, we publicly release the data sample, the codes to reproduce our experiments, and all the evaluation protocols designed.

2. Related work

Gyrochronology The main benefit of gyrochronology is that the required observational inputs are quite easy to obtain: photometric or spectroscopic proxies to the stellar mass and the rotational period. On the other hand, it has a number of uncertainties that makes it a not very accurate dating technique, and it has a relatively narrow application range in terms of stellar mass. In [6, 7], the authors published, for the first time, an empirical equation for the estimation of the stellar age as a function of its rotational period and its color index. This can be regarded as the origin of gyrochronology as a practical technique. This empirical relation has been revised, as in [31], [8], and [1], the latest using an asteroseismic data sample.

But real life is usually not that simple. Rotational braking is theoretically linked to the existence of a bipolar stellar magnetic field. The best candidate for the dynamo generating this field is a well developed outer convective zone. Therefore, only stars with masses below the so-called Kraft limit [28], around 1.3 $M_{\odot}$, can reduce its rotational velocity with time following known breaking laws. But [61] showed that A stars also have rotational velocities evolving with time. This can be thanks to the fact that strong magnetic fields, or at least their signatures, can also be found in these stars [5] and, therefore, above the Kraft limit, the rotational braking may still exist. In this mass range, above 1.3 $M_{\odot}$, the exact braking mechanism is almost unknown.

For young stars, gyrochronology can only be applied to stellar clusters or moving groups [18]. For individual stars, the uncertainties make it only possible a general statement like "this star is/isn't young". On the other hand, [56] found, comparing a set of 21 stars dated using asteroseismology with rotational braking models [57], that from a certain value of the Rossby number, the rotational braking seems to stop and the rotational velocity remains in a steady-state, making traditional gyrochronology almost useless. In any case, this change in the rotational braking regime is still under debate, with works following the line of its confirmation [23, 27, 39, 40, 58], and some others concluding that it is not observed, especially in the case of solar-twins in terms of mass and metallicity [9, 30]. In any case, the existence of an age range where the convergence is not reached yet, and the existence of this change in the braking law, makes classical linear relations for gyrochronology a simplistic approximation to a complex and heterogeneous problem. It is precisely in this context where machine learning can help.

AI for dating stars models We can find few works that address the problem of age estimation using artificial intelligence [3, 4, 19, 48]. Sanders and Das introduce in [48] a catalogue with distances, masses, and ages for ~ 3 million stars from the second Gaia data release. Ages are estimated following a Bayesian fitting to stellar structure and evolution models, to characterize their probability density functions. Angus *et al.* [3, 4] propose a combination of linear relations for gyrochronology with isochrones fitting under a Bayesian estimation framework. Their approach uses a Markov Chain Monte Carlo process to infer the ages. It is in [48] where we find for the first time a Bayesian Neural Network applied to the problem of age estimation. Technically, Sanders and Das [48] train a multi-layer perceptron, with a single hidden layer (using 10 neurons), to generate predictive posterior distributions for the mass, age, and distance of the stars. This way, their model is able to replace the reliance of isochrones technique. The differences of all these works with our own are the following. All previous models address the problem from a Bayesian perspective, focusing on the predictions of posterior distributions of the ages. Instead, we propose here a pure regression problem, where models are faced with the estimation of a particular value for the age, and they are evaluated accordingly. Moreover, for this reason, a direct comparison with these previous works is not an easy task. Finally, with our study we offer the community clear benchmarking scenarios together with a data sample, so that others can compare with our approaches.

3. The Data

We have gathered a sample of 1464 stars with accurate ages coming from asteroseismology or cluster belonging. The asteroseismic sample consists of 312 entries for which accurate fundamental stellar observables (effective temperature $T_{\rm eff}$, logarithm of the surface gravity log g, mass M, radius R, and stellar Iron to Hydrogen content [Fe/H]) have been inferred from a combination of photometric and spectroscopic observations and asteroseismology. The most significant contribution comes from [50], with 224 entries. Others were obtained from [15, 17, 51, 52].

In terms of rotational periods ($P_{\rm rot}$), of the 312 entries obtained from asteroseismology, 293 periods were taken from [20]. The remaining periods were taken from [15, 32, 33, 34, 42, 46]. All of them are Kepler's targets. The autocorrelation function (ACF) method is used in [32, 33, 34] for obtaining the rotational period. Garcia *et al.* [20] analyzed the surface rotation rate in the subset of Kepler solar-like stars studied in [17]. The same analysis methods implemented in [20] is adopted in [15]. In [42], on the other hand, Nielsen *et al.* computed a Lomb-Scargle (LS) periodogram. They chose as $P_{\rm rot}$ the median value of all the recorded peaks of maximum power measured over several quarters of Kepler's data (quarters 2 through 9, corresponding to two years of observations in total). In [46] also the LS periodogram is used, but restricting the analysis exclusively to Kepler's quarter 3 long cadence data.

We complemented the sample with studies from clusters performed by the Kepler/K2 mission. We collected a total of 1152 entries taken from [9, 22, 35, 36, 45]. In [9, 22] the authors studied the 4.2 ± 0.7 Gyr old cluster M67, analyzing data of Kepler/K2 Campaign 5 light curves. M67 is an interesting target for gyrochronology, given that it is about the same age, and shares a similar chemical composition, as the Sun. M67 is also the oldest cluster in our sample. Barnes *et al.* [9] derived surface rotation periods using a combination of four methods: phase dispersion minimization, minimum string length, the Bayesian period signal detection method, and the autocorrelation function. In [22] the Lomb-Scargle periodogram analysis method was used. The age of the cluster was settled by [9] and it agrees with the chromospheric [21] and isochrone [11] derived ages. Meibom *et al.* in [35] studied NGC 6811 (1.0 ± 0.2 Gyr, see [24]) and in [36], they reported periods for stars in NGC 6819 (2.5 ± 0.3 Gyr, see [25]). These two clusters bridge the gap in age between Praesepe and M67. These authors used the LS periodogram method for obtaining the rotational periods. In addition, for all reported $P_{\rm rot}$, they visually examined the periodogram and light curves, and they also checked the periods independently using the CLEAN algorithm. Praesepe was observed during Kepler/K2 Campaign 5. In [45] the authors identified the surface rotation periods applying the Lomb-Scargle periodogram. They took the period corresponding to the strongest peak in the periodogram as the rotation period (with some exceptions). The study produced periods for over 80% of all Praesepe light curves. The age estimation of this cluster has been a subject of some debate, with the most recent value set at 790 ± 60 Myr by [12].

Feature	Unit	Description
id	-	Star ID
$T_{\tt eff}$	K	Stellar effective temperature
logg	dex	Logarithm of the surface gravity
M	Solar masses	Stellar mass
R	Solar radii	Stellar radius
[Fe/H]	dex	Stellar Iron over Hydrogen content
$P_{\rm rot}$	days	Stellar rotational period
Age	Gyr	Stellar age

Table 1: Features provided in our sample for every star.

In order to work with the highest accuracy and precision, we have completed the clusters' sample with masses and radii derived with a machine-learning Random Forest model of the `empiricalRelationsMR` R-package [41].

Our sample is, therefore, a mix of four clusters with ages 0.79, 1, 2.5, and 4.2 Gyrs gathering a total of 1152 stars, with masses and radii estimated using machine learning and empirical data, on the one hand. And 312 stars with masses, radii, and ages determined using asteroseismology, and precise rotational periods, on the other. Finally, we have selected those FGK and Main Sequence stars with rotational periods below 50 days. The reason of this filtering is that the physics behind gyrochronology occurs mainly in these stellar types, and therefore is limited to a mass range of $[0.7, 2]$ $M_\odot$. In addition, rotational periods larger than 50 days can be hardly explained by current stellar structure and evolution models. Table 1 shows all the features that we release for every star.

We are going to set aside 32 non-clustered stars (including the Sun) dated using asteroseismology for testing purposes (See Benchmark C in Section 4.2 below). Therefore, we work with a final sample consisting of 397 stars in clusters plus 240 stars studied using asteroseismology, that is, a total of 637 stars.

Unfortunately, this sample exhibits two important biases: 1) the asteroseismic sample is biased to massive and old stars; and 2) the cluster-based sample is age-quantified and biased towards young ages. We show in our experiments that, despite these biases, machine learning techniques are able to extract reliable information from the dataset for estimating stellar ages. With time, specially with current and future space missions, these biases will be progressively mitigated, with the consequent improvement on estimations.

3.1. Understanding the sample

In Fig. 1 we show the position in the HR diagram of all the selected MS stars. We also show which of them are members of a cluster and which of them have been characterized using asteroseismology. Here we can see that asteroseismic stars cover the more massive and/or evolved zone

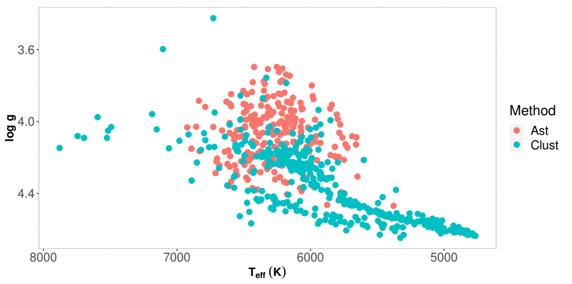


Figure 1: HR diagram showing the MS sample.

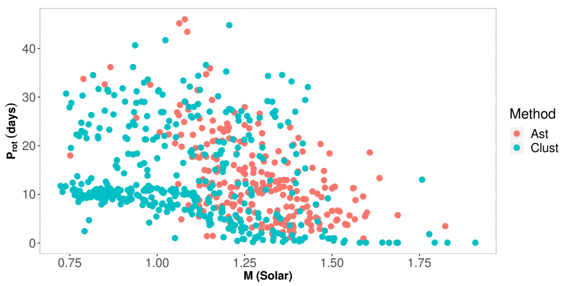


Figure 2: M - $P_{\rm rot}$ for the selected stars.

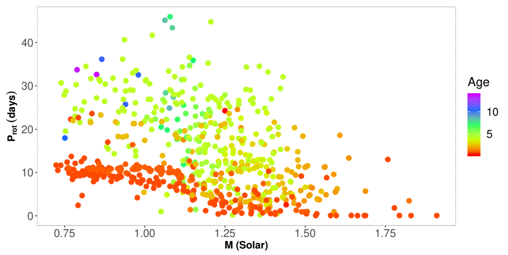


Figure 3: M - $P_{\rm rot}$ for the selected stars. The age is shown in color scale.

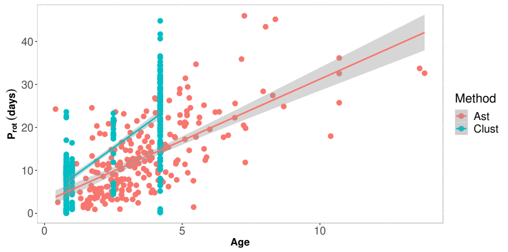


Figure 4: Age - $P_{\rm rot}$ for the selected stars. The asteroseismic and cluster stars are shown in different colours. We also present the linear regression obtained using these two groups.

of the HR diagram, and cluster stars are younger and also cover the low mass region.

The classic idea of gyrochronology is to estimate stellar ages using any proxy of the stellar mass, and the stellar rotational period. Since we have a mass estimation for all our training sample, we can directly represent M vs. $P_{\rm rot}$ (Fig. 2) avoiding using that proxy. In this plot, we also differentiate between asteroseismic and cluster stars. There are no clear differences between them except the already mentioned bias of the asteroseismic sample towards more massive stars. The Kraft limit is clear around 1.2M$_\odot$.

If we add the information of the stellar age to the M vs. $P_{\rm rot}$ plot we obtain Fig. 3. Here we can see that, with a large dispersion, the larger the $P_{\rm rot}$, the older the age. It is remarkable that it is also true for massive stars, above the Kraft limit. That is, even in the absence of a developed outer convective envelope, the reduction of the rotational velocity with time also occurs.

If we represent the stellar age vs. its rotational period, we obtain Fig. 4. In general, we confirm that the largest the age, the largest the rotational period, with a large dispersion, mainly for stars in clusters due to the mass dependence of the rotation braking. We have also fit a linear regression to this relation, differentiating between asteroseismic and cluster stars to guide the eye. Here we can see one of the biases of the sample. These linear relations have the same tendency but they are barely different for each subgroup. In any case, the dispersion is really large and we cannot ensure that these regressions can be used for age estimation purposes. This is the reason why we move to more sophisticated AI based analysis methods for generating models suitable for stellar age estimations.

4. Benchmarking AI regression models for dating stars

Previous works in dating stars from gyrochronology data and machine learning models suffer from non-standard testing and training paradigms, making direct comparisons of certain algorithms difficult (*e.g.* [3, 4, 19, 48]).

In this section we detail all the regression models trained to predict the age of the stars (see Section 4.1). Later, in Section 4.2 we introduce the designed benchmarking procedure using the data described in Section 3.

4.1. Models

Linear regressor We first train a star dating model assuming that the age is a linear combination of the features provided in the data sample. Let $\hat{a}$ be the predicted age value, then

$$\hat{a}(\mathbf{w},\mathbf{x}) = w_0 + w_1 x_1 + ... + w_D x_D \ , \qquad (1)$$

where $\mathbf{x}$ and $\mathbf{w}$ encode the input feature vector and the coefficient of the linear regressor, respectively. We follow the least squares approach for learning the model, solving a problem of the form: $\min_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \mathbf{a}||_2^2$. $\mathbf{X}$ is a matrix containing all the training vectors, and $\mathbf{a}$ is ages vector.

Bayesian Regression We implement here a Bayesian Ridge regression to estimate a probabilistic model of the stellar dating estimation problem with the form:

$$p(a|\mathbf{X}, \mathbf{w}, \alpha) = \mathcal{N}(a|\mathbf{X}\mathbf{w}, \alpha) \ , \qquad (2)$$

where the output age a is assumed to be a Gaussian distribution over $\mathbf{X}\mathbf{w}$. α is estimated directly from data being treated as a random variable. The regularization parameter is tuned to the data available, introducing over the hyper parameters of the model, *i.e.* $\mathbf{w}$, the following spherical Gaussian $p(\mathbf{w}|\lambda) = \mathcal{N}(\mathbf{w}|0, \lambda^{-1}\mathbf{I}_D)$.

Decision Tree Regressor We use a decision tree model for regression [14]. During learning, we let our model to optimize its internal hyperparameters via a grid search process with cross validation. Specifically, we adjust: the strategy to choose the split at each tree node (best or random); the minimum number of samples required to be at a leaf node (5, 10, 50, 100); and the function to measure the split quality (mean squared error, mean squared error with Friedman's improvement score for potential splits, mean absolute error or reduction in Poisson deviance to find splits).

Random Forest Regressor This regression model is implemented following the original work of Breiman [13], where the decision trees are built from data samples drawn with replacement from the training set. We again use a grid search plus cross validation process to tune the following hyperparameters: number of trees (5, 10, 50, 100); minimum number of samples required to be at a leaf node (5, 10, 50, 100); and the function for measuring the quality of the split (mean squared error or mean absolute error).

Support Vector Regression We include in our study this regression model, based on the LibSVM implementation for regression [16], following the ϵ-SVR approach. We employ the RBF kernel, and perform a grid search with cross validation to adjust parameter C (1, 10, 100, 1000).

Gaussian Process We train a Gaussian Process regressor [44], where the prior mean is assumed to be constant and zero. For the prior's covariance we use a kernel that is the sum of the following three kernels. RBF, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2l^2}\right)$, where $d(\cdot, \cdot)$ is the Euclidean distance, and l is the length-scale parameter. Dot product, $k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_0^2 + \mathbf{x}_i \cdot \mathbf{x}_j$. It is a non-stationary kernel that is obtained from linear regression by putting priors on the coefficients of $x_i (i = 1, \ldots, N)$ and a prior of $N(0, \sigma_0^2)$ on the bias. And finally, White kernel, $k(x_i, x_j) = noise\_level$ if $x_i == x_j$ else 0. The White kernel's key application is as part of a sum-kernel, where it describes the input's noise component. Tuning its parameter corresponds to estimating the noise-level then.

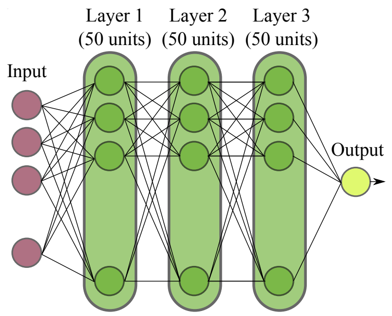


Figure 5: Architecture of the neural network implemented. 3 hidden layers have been used, with 50 units each of them followed by a ReLU. The output layer has no activation function, to directly perform the star date estimation.

kNN We include also in the benchmarking a kNN regressor. We basically tune k parameter (1, 5, 10, 15, 20, 50) and the type of weight function used to scale the contribution of each neighbor. Two types of weight functions are explored: a) uniform, where all points in each neighborhood are weighted equally; and b) distance, which weights points by the inverse of their distance to a query point.

Neural Network The architecture implemented is depicted in Figure 5. It consists in a multi layer perceptron. Our model includes an input layer, a set of 3 hidden fully-connected layers with 50 units each, and the output layer in charge of the regression of the ages of the stars. Every hidden unit is followed by a ReLU activation function. We use as loss function the square error, $Loss(\hat{a}, a, W) = \frac{1}{2}||\hat{a} - a||_2^2 + \frac{\alpha}{2}||W||_2^2$, where W encodes the weights of the network, and α is the regularization parameter. Backpropagation [29] is used for learning the model with SGD [47] optimizer. During learning we fix $\alpha = 0.01$, and use an adaptive learning rate policy, starting with a value of 0.1.

Stacking Finally, we use a machine learning ensemble method known as stacked generalization (or stacking) [60]. In a regression problem like ours, the predictions of each individual regressor (at level 0) are stacked and fed into a final estimator (at level 1), which calculates the age prediction. For this work, in level 0, we integrate: the neural network, and the Gaussian process. For the last layer (level 1), we incorporate a linear regressor. During training, and to generalize and avoid over-fitting, the final estimator of level 1 is trained on out-samples (taken from the training set), following a cross-validation methodology.

4.2. Benchmarks

Our data sample, detailed in Section 3, offers more than 600 stars, with precise ages, where we propose to construct the following experimental evaluation scenarios.

Stellar dating regression problem (Benchmark A) In this setup, we propose to evaluate the different regression models following a classical training/test data splitting scheme. From the data sample distribution, we release a training set and a testing set, where 80 % and 20 % of the stars have been randomly included, respectively.

Generalization capability (Benchmark B) We first provide a generalization study for the models, where we train the approaches on *young* stars, and evaluate their performance on *old* stars. This scenario, named Benchmark B1, is interesting in order to evaluate the ability of artificial intelligence models to work with stars whose age range falls outside that of the training set. The training and testing age ranges are $[0.4, 4.2]$ and $[4.2, 13.8]$ Gyr, respectively.

We propose a second scenario to evaluate the generalization capability of the models when they are trained only with stars belonging to clusters. In practice, gyrochronology is used to date individual stars that can have any possible age. Stars in clusters are usually dated thanks to its belonging to the cluster. In this Benchmark B2, we propose to train our models only using stars in clusters. The rest of the stars are included in the test set. In total, this scenario leaves us with 397 and 240 samples for training and testing, respectively.

Stellar dating regression in a control data sample (Benchmark C) We propose here to examine the age estimation performance of all the models on a control data sample composed only of stars not belonging to any cluster, and with a more realistic age distribution. Specifically, we evaluate in this Benchmark C how the models trained using all our data sample perform over a set of novel non-clustered 32 stars, including the Sun. This independent control set has been obtained from the same asteroseismology based sources used to gather the information for the sample described in Section 3. We have set aside 32 stars from these sources, before building the sample. The age range in this set spans from 1.2 to 10.1 Gyr. This range overlaps with the age range used in the training sample ($[0.4, 13.8]$ Gyr), and includes some novel and unobserved ages. We will focus our attention on the precision of the models for the estimation of the age of the Sun (4.6 Gyr) in the experiments.

5. Experiments

We present the experimental setup and report results of benchmarking all the stellar age prediction models detailed in Section 4.1. This benchmarking is carried out through the 3 described scenarios: the effect of the different models (Benchmark A); the generalization behavior (Benchmark B); and the performance in a control data sample (Benchmark C);

5.1. Experimental setup

Implementation In order to better perform apples-to-apples comparisons, we have built all the regression models in Python, using the excellent scikit-learn library [43]. We publicly release all the codes (from training to testing) to reproduce the detailed experiments[1]. We use the following acronyms to identify the different models implemented: Neural Network (nnet), Linear regressor (lr), Decision Tree (dtr), Random Forest (rf), Support Vector Regressor (svm), Bayesian Regression (bayes), kNN (knn), Gaussian Process (gp) and Stacking (stacking).

Evaluation metrics The main evaluation metric used is the Mean Absolute Error, $MAE = \frac{\sum_{i=1}^{N} |a_i - \hat{a}_i|}{N}$, where a_i and $\hat{a}_i$ are the age provided in our dataset, and the age estimated by a regression model, respectively. Since our dataset provides information on the precision of the age of each star, in the form of error bounds, we also propose to use as a precision evaluation metric the percentage of star age predictions that fall within the confidence interval provided by our dataset.

5.2. Benchmark A: Stellar dating problem

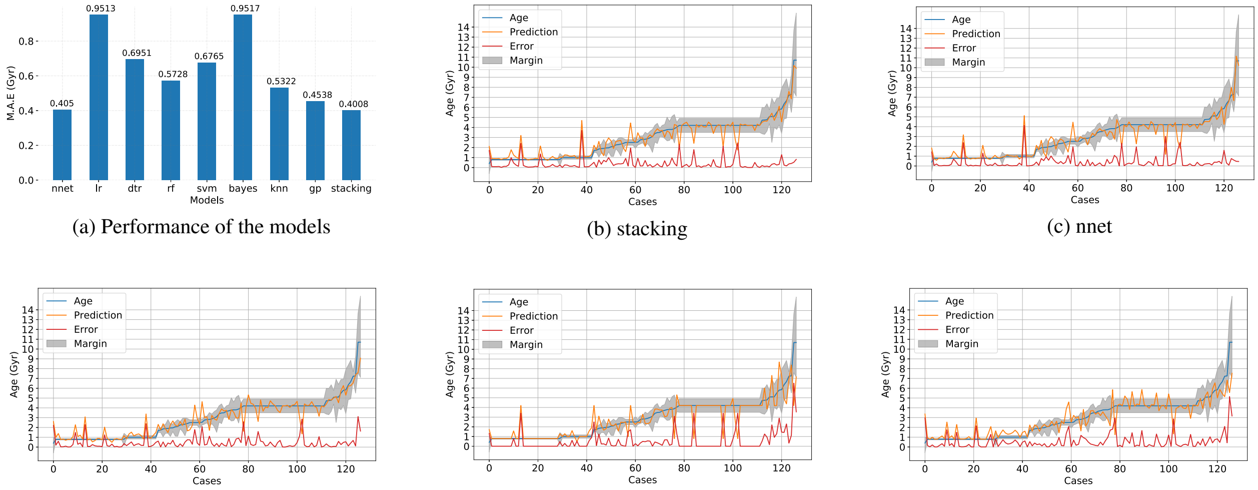
Figure 6a shows the performance of all the models, comparing their corresponding MAE. In this scenario it is interesting to note that we have 2 models, plus their stacking, that present the best results, establishing a margin with respect to all the others. They are the Neural Network and the Gaussian Process. Their stacking slightly reduces the best MAE of the Neural Network from 0.405 to 0.400. Figures 6b–6f offer a detailed analysis of the predictions for the top-5 methods for all the test samples[2]. Interestingly, most models present difficulties in age estimation for older stars. Also relevant in this benchmark is the result provided by a model as simple as a kNN, with a MAE of just 0.53 Gyr. A possible explanation is that the database has a lot of elements concentrated in 2 very specific age vales, which allows models such as kNN to make very accurate estimates on testing stars with these ages. This fact is corroborated in Table 2, where it is observed that the best precision is achieved by the kNN model, followed by the neural network. Overall, except for lr and bayes, all models offer a MAE below 0.86 Gyr, which is a breakthrough in the field of stellar dating.

5.3. Benchmark B: Generalization behaviour

We here analyze the generalization capability of all the models for the problem of stellar dating. Figures 7a and 7b

[1] https://github.com/gramuah/ai4datingstars

[2] In the supplementary material we provide these graphs for all the models



(a) Performance of the models

(b) stacking

(c) nnet

(d) gp

(e) knn

(f) rf

Figure 6: Benchmark A. (a) Performance of the models as a function of the MAE. In this setting, the two best approaches are the Neural Network (nnet) and the Gaussian Process (gp), and therefore their stacking. (b)-(f) Detailed prediction for the top-5 models. We show the ground truth age (in blue), the prediction of the models (in orange), and the corresponding error (in red).

Benchmark	nnet	lr	dtr	rf	svm	bayes	knn	gp	stacking
A	73.23	35.43	63.78	60.63	46.46	32.28	**77.95**	62.20	62.99
B1	**59.84**	29.92	51.18	48.03	37.79	28.35	55.91	51.97	n/a
B2	28.45	30.96	21.34	25.94	**33.05**	29.71	25.52	30.96	n/a
C	34.37	21.87	9.37	21.87	12.50	21.87	15.62	**40.62**	34.37

Table 2: Precision of all the methods in the different benchmarks. Precision is measured as the percentage of age estimations that fall within confidence margin offered in the dataset for each star.

show the performance in the benchmarking scenarios B1 and B2, respectively.

Recall that benchmark B1 focuses on analyzing how accurate the models are when they must estimate the ages of stars older than the ones they have seen during training. Obviously, the average error reported by all methods has increased. Best performance is reported here by our neural network, with a MAE of 1.47 Gyr. Its detailed estimations can be seen in Figure 7d, where it is worth noting how the neural network is able to generalize up to 6 Gyr, casting stellar age estimations that mostly fall within the confidence margin. Table 2 confirms this fact, where the Neural Network also exhibits the highest precision for B1 benchmark.

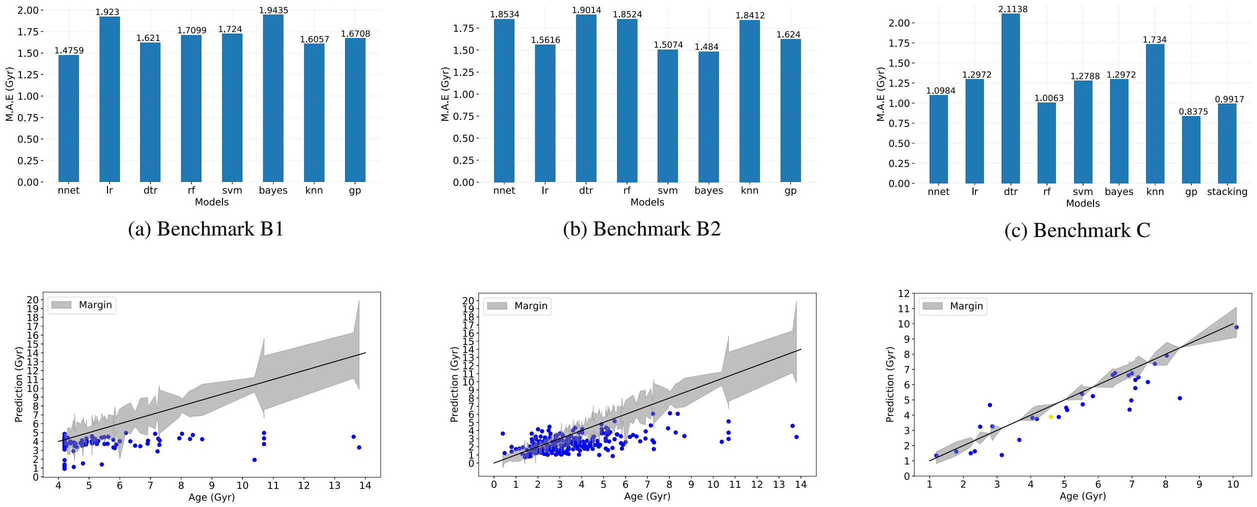
The story is completely different in Benchmark B2. The neural network is not able to generalize the best. Its error rises up to 1.85 Gyr, being surpassed by lr, rf, svm, knn, gp and bayes. The latter turns out to be the best model now, with a MAE of 1.48 Gyr. From Figure 7e we conclude that a bayesian regressor is able to provide accurate ages in the range from 1 Gyr to 5 Gyr, that is, the range covered by the training sample in this case. In terms of precision, Table 2 reveals that the svm for regression is the winner, closely followed by both gp and lr. The bayesian model, although has the lowest MAE, reports a precision of 29.71 %.

Finally, in both scenarios we have observed that all models tend to underestimate the ages, although that effect is more pronounced in the second benchmark. This is not a surprise since in both cases ages above 4.2 Gyrs are not covered by the training samples. We encourage the reader to consult the supplementary material, where more details for all the models have been included.

5.4. Benchmark C: Performance in a control sample

We show the MAE and the precision for every method for this control benchmark in Figure 7c and Table 2, respectively. Who is the winner here? The three best methods are gp, nnet and their stacking. Among these three, the one that excels is the Gaussian process. In Figure 7f we can inspect our stellar age predictions using this method vs reference ages for the stars in this set. gp is the method with the highest precision: 40.62% of the predictions fall within the confidence margin of the data set. For the Sun, gp predicts an age of 3.88 versus an accepted age of 4.6 Gyr, underestimating it. Actually, the model tends to slightly underestimate most of the ages. We have observed in the experiments that this is the common behaviour for all the models, except for kNN and dtr. In the supplementary material we extend the results by providing more details for all models.

With respect to the accuracy of our estimations for the age of our Sun, 4.6 Gyr, we provide in Table 3 the specific



(a) Benchmark B1

(b) Benchmark B2

(c) Benchmark C

(d) Best model in Benchmark B1 (nnet)

(e) Best model in Benchmark B2 (bayes)

(f) Best model in Benchmark C (gp)

Figure 7: Results for Benchmarks B1, B2 and C. We report the MAE for all the models in figures (a), (b) and (c). Figures (d), (e) and (f) show the detailed performance for the best performing method for every benchmark. Note in figure (f) the yellow dot where we depict the age estimation for the Sun.

Methods	nnet	lr	dtr	rf	svm	bayes	knn	gp	stacking
Sun (4.6 Gyr)	4.02	3.92	4.49	4.76	3.81	3.91	4.49	3.88	4.02

Table 3: Sun age estimates in Gyr. The most precise method is rf.

age reported by all the models. All methods do underestimate, like we just pointed above, and our Random Forest obtains the closest prediction, followed by dtr and kNN.

Comparing results of Benchmarks A (Figure 6a) and C (Figure 7c) one can conclude that: a) winning models are common (nnet, gp and stacking); b) kNN and dtr suffer a considerable degradation of their performance, as they are highly dependent on the data distribution used during their training; and c) bayes and lr go hand in hand, showing a similar increase in their corresponding MAEs, but at the same time offering some stability between scenarios.

6. Conclusion

We have presented a thorough benchmarking study of state-of-the-art AI regression models trained and tested for stellar dating using gyrochronology. Based on this study, we report the following findings on the performance and generalization capability of the stellar age estimation regression models analyzed.

First, is there a winning model? Our study demonstrates that a model for stellar dating based on a neural network such as ours, provides results that are remarkable in terms of trade-off between generalization and accuracy. The experiments reveal that the performance of the nnet model is good in Benchmarks A and C, but also in B1, where it has to generalize to unseen ages. It is a simple model, an MLP with 3 hidden layers, which could be extended (*e.g.* using 1D CNN layers) possibly offering better results. We leave this door open for future research. If practitioners look for the solution with the lowest error, then our take-home message is: use a stacking model combining Gaussian processes and neural networks. Finally, we recommend a Bayesian Ridge regression if the available training data belong mostly to stellar clusters. Second, *all* models tend to underestimate in general, and not only in Benchmark B1, where it would be natural. Therefore, we think that an interesting line of future work is to reduce this bias. And third, the study reveals promising results for this stellar dating problem. An error < 0.5 Gyr can be considered as an outstanding breakthrough in the field.

For this study, we have released a dataset and proposed a set of evaluation protocols to assist the research on AI for dating stars. We hope that our study may help to improve on the state of the art for robust star dating AI based solutions.

Acknowledgements This project has been funded by: i) European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 749962 (project THOT); ii) Spain's Ministry of Science and Innovation under projects PID2019-107061GB-C65 and PID2019-104323RB-C31 (AIRPLANE). Authors also want to acknowledge Federico Zuccarino (RIP) for his help building the dataset.

References

[1] Ruth Angus, Suzanne Aigrain, Daniel Foreman-Mackey, and Amy McQuillan. Calibrating gyrochronology using Kepler asteroseismic targets. *MNRAS*, 450(2):1787–1798, June 2015.

[2] Ruth Angus, Angus Beane, Adrian M. Price-Whelan, Elisabeth Newton, Jason L. Curtis, Travis Berger, Jennifer van Saders, Rocio Kiman, Daniel Foreman-Mackey, Yuxi Lucy Lu, Lauren Anderson, and Jacqueline K. Faherty. Exploring the Evolution of Stellar Rotation Using Galactic Kinematics. *AJ*, 160(2):90, Aug. 2020.

[3] Ruth Angus, Timothy D. Morton, and Daniel. Foreman-Mackey. stardate: Combining dating methods for better stellar ages. *Journal of Open Source Software*, 41(4):1469, 2019.

[4] Ruth Angus, Timothy D. Morton, Daniel Foreman-Mackey, Jennifer van Saders, Jason Curtis, Stephen R. Kane, Megan Bedell, Rocio Kiman, David W. Hogg, and John Brewer. Toward Precise Stellar Ages: Combining Isochrone Fitting with Empirical Gyrochronology. *AJ*, 158(5):173, Nov. 2019.

[5] L. A. Balona. Starspots on A stars. *MNRAS*, 467(2):1830–1837, May 2017.

[6] Sydney A. Barnes. On the Rotational Evolution of Solar- and Late-Type Stars, Its Magnetic Origins, and the Possibility of Stellar Gyrochronology. *ApJ*, 586(1):464–479, Mar. 2003.

[7] Sydney A. Barnes. Ages for Illustrative Field Stars Using Gyrochronology: Viability, Limitations, and Errors. *ApJ*, 669(2):1167–1189, Nov. 2007.

[8] Sydney A. Barnes. A Simple Nonlinear Model for the Rotation of Main-sequence Cool Stars. I. Introduction, Implications for Gyrochronology, and Color-Period Diagrams. *ApJ*, 722(1):222–234, Oct. 2010.

[9] Sydney A. Barnes, Joerg Weingrill, Dario Fritzewski, Klaus G. Strassmeier, and Imants Platais. Rotation Periods for Cool Stars in the 4 Gyr old Open Cluster M67, The Solar-Stellar Connection, and the Applicability of Gyrochronology to at least Solar Age. *ApJ*, 823(1):16, May 2016.

[10] J. W. Belcher and K. B. MacGregor. Magnetic acceleration of winds from solar-type stars. *ApJ*, 210:498–507, Dec. 1976.

[11] A. Bellini, L. R. Bedin, B. Pichardo, E. Moreno, C. Allen, G. Piotto, and J. Anderson. Absolute proper motion of the Galactic open cluster M 67. *A&A*, 513:A51, Apr. 2010.

[12] Timothy D. Brandt and Chelsea X. Huang. The Age and Age Spread of the Praesepe and Hyades Clusters: a Consistent, ~800 Myr Picture from Rotating Stellar Models. *ApJ*, 807(1):24, July 2015.

[13] L. Breiman. Random forests. *Machine Learning*, 5:5–32, 2001.

[14] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[15] T. Ceillier, J. van Saders, R. A. García, T. S. Metcalfe, O. Creevey, S. Mathis, S. Mathur, M. H. Pinsonneault, D. Salabert, and J. Tayar. Rotation periods and seismic ages of KOIs - comparison with stars without detected planets from Kepler observations. *MNRAS*, 456(1):119–125, Feb. 2016.

[16] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[17] W. J. Chaplin, S. Basu, D. Huber, A. Serenelli, L. Casagrande, V. Silva Aguirre, W. H. Ball, O. L. Creevey, L. Gizon, R. Handberg, C. Karoff, R. Lutz, J. P. Marques, A. Miglio, D. Stello, M. D. Suran, D. Pricopi, T. S. Metcalfe, M. J. P. F. G. Monteiro, J. Molenda-Żakowicz, T. Appourchaux, J. Christensen-Dalsgaard, Y. Elsworth, R. A. García, G. Houdek, H. Kjeldsen, A. Bonanno, T. L. Campante, E. Corsaro, P. Gaulme, S. Hekker, S. Mathur, B. Mosser, C. Régulo, and D. Salabert. Asteroseismic Fundamental Properties of Solar-type Stars Observed by the NASA Kepler Mission. *ApJS*, 210(1):1, Jan. 2014.

[18] Jason L. Curtis, Marcel A. Agüeros, Eric E. Mamajek, Jason T. Wright, and Jeffrey D. Cummings. TESS Reveals that the Nearby Pisces-Eridanus Stellar Stream is only 120 Myr Old. *AJ*, 158(2):77, Aug. 2019.

[19] Payel Das and Jason L Sanders. MADE: a spectroscopic mass, age, and distance estimator for red giant stars with Bayesian machine learning. *Monthly Notices of the Royal Astronomical Society*, 484(1):294–304, 10 2018.

[20] R. A. García, T. Ceillier, D. Salabert, S. Mathur, J. L. van Saders, M. Pinsonneault, J. Ballot, P. G. Beck, S. Bloemen, T. L. Campante, G. R. Davies, Jr. do Nascimento, J. D., S. Mathis, T. S. Metcalfe, M. B. Nielsen, J. C. Suárez, W. J. Chaplin, A. Jiménez, and C. Karoff. Rotation and magnetism of Kepler pulsating solar-like stars. Towards asteroseismically calibrated age-rotation relations. *A&A*, 572:A34, Dec. 2014.

[21] Mark S. Giampapa, Jeffrey C. Hall, Richard R. Radick, and Sallie L. Baliunas. A Survey of Chromospheric Activity in the Solar-Type Stars in the Open Cluster M67. *ApJ*, 651(1):444–461, Nov. 2006.

[22] Guillermo Gonzalez. Variability among stars in the M 67 field from Kepler/K2-Campaign-5 light curves. *MNRAS*, 459(1):1060–1068, June 2016.

[23] Tyler A. Gordon, James R. A. Davenport, Ruth Angus, Daniel Foreman-Mackey, Eric Agol, Kevin R. Covey, Marcel Agüeros, and David Kipping. Stellar Rotation in the K2 Sample: Evidence for Broken Spindown. *arXiv e-prints*, page arXiv:2101.07886, Jan. 2021.

[24] Kenneth Janes, Sydney A. Barnes, Søren Meibom, and Sadia Hoq. NGC 6811: An Intermediate-age Cluster in the Kepler Field. *AJ*, 145(1):7, Jan. 2013.

[25] R. D. Jeffries, Tim Naylor, N. J. Mayne, Cameron P. M. Bell, and S. P. Littlefair. A lithium depletion boundary age of 22 Myr for NGC 1960. *MNRAS*, 434(3):2438–2450, Sept. 2013.

[26] Steven D. Kawaler. Angular Momentum Loss in Low-Mass Stars. *ApJ*, 333:236, Oct. 1988.

[27] Leonid Kitchatinov and Alexander Nepomnyashchikh. How supercritical are stellar dynamos, or why do old main-sequence dwarfs not obey gyrochronology? *MNRAS*, 470(3):3124–3130, Sept. 2017.

[28] Robert P. Kraft. Studies of Stellar Rotation. V. The Dependence of Rotation on Age among Solar-Type Stars. *ApJ*, 150:551, Nov. 1967.

[29] Y. LeCun, L. Bottou, G. Orr, and K.-R. Müller. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, 2012.

[30] Diego Lorenzo-Oliveira, Jorge Meléndez, Jhon Yana Galarza, Geisa Ponte, Leonardo A. dos Santos, Lorenzo Spina, Megan Bedell, Iván Ramírez, Jacob L. Bean, and Martin Asplund. Constraining the evolution of stellar rotation using solar twins. *MNRAS*, 485(1):L68–L72, May 2019.

[31] Eric E. Mamajek and Lynne A. Hillenbrand. Improved Age Estimation for Solar-Type Dwarfs Using Activity-Rotation Diagnostics. *ApJ*, 687(2):1264–1293, Nov. 2008.

[32] Tsevi Mazeh, Hagai B. Perets, Amy McQuillan, and Eyal S. Goldstein. Photometric Amplitude Distribution of Stellar Rotation of KOIs—Indication for Spin-Orbit Alignment of Cool Stars and High Obliquity for Hot Stars. *ApJ*, 801(1):3, Mar. 2015.

[33] A. McQuillan, T. Mazeh, and S. Aigrain. Stellar Rotation Periods of the Kepler Objects of Interest: A Dearth of Close-in Planets around Fast Rotators. *ApJ*, 775(1):L11, Sept. 2013.

[34] A. McQuillan, T. Mazeh, and S. Aigrain. Rotation Periods of 34,030 Kepler Main-sequence Stars: The Full Autocorrelation Sample. *ApJS*, 211(2):24, Apr. 2014.

[35] Søren Meibom, Sydney A. Barnes, David W. Latham, Natalie Batalha, William J. Borucki, David G. Koch, Gibor Basri, Lucianne M. Walkowicz, Kenneth A. Janes, Jon Jenkins, Jeffrey Van Cleve, Michael R. Haas, Stephen T. Bryson, Andrea K. Dupree, Gabor Furesz, Andrew H. Szentgyorgyi, Lars A. Buchhave, Bruce D. Clarke, Joseph D. Twicken, and Elisa V. Quintana. The Kepler Cluster Study: Stellar Rotation in NGC 6811. *ApJ*, 733(1):L9, May 2011.

[36] Søren Meibom, Sydney A. Barnes, Imants Platais, Ronald L. Gilliland, David W. Latham, and Robert D. Mathieu. A spin-down clock for cool stars from observations of a 2.5-billion-year-old cluster. *Nature*, 517(7536):589–591, Jan. 2015.

[37] L. Mestel. Magnetic braking by a stellar wind-I. *MNRAS*, 138:359, Jan. 1968.

[38] L. Mestel and H. C. Spruit. On magnetic braking of late-type stars. *MNRAS*, 226:57–66, May 1987.

[39] Travis S. Metcalfe and Ricky Egeland. Understanding the Limitations of Gyrochronology for Old Field Stars. *ApJ*, 871(1):39, Jan. 2019.

[40] Travis S. Metcalfe and Jennifer van Saders. Magnetic Evolution and the Disappearance of Sun-Like Activity Cycles. *Sol. Phys.*, 292(9):126, Sept. 2017.

[41] Andy Moya, Federico Zuccarino, William J. Chaplin, and Guy R. Davies. Empirical Relations for the Accurate Estimation of Stellar Masses and Radii. *ApJS*, 237(2):21, Aug. 2018.

[42] M. B. Nielsen, L. Gizon, H. Schunker, and C. Karoff. Rotation periods of 12 000 main-sequence Kepler stars: Dependence on stellar spectral type and comparison with v sin i observations. *A&A*, 557:L10, Sept. 2013.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[44] Carl Eduard Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[45] L. M. Rebull, J. R. Stauffer, L. A. Hillenbrand, A. M. Cody, J. Bouvier, D. R. Soderblom, M. Pinsonneault, and L. Hebb. Rotation of Late-type Stars in Praesepe with K2. *ApJ*, 839(2):92, Apr. 2017.

[46] Timo Reinhold, Ansgar Reiners, and Gibor Basri. Rotation and differential rotation of active Kepler stars. *A&A*, 560:A4, Dec. 2013.

[47] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

[48] Jason L Sanders and Payel Das. Isochrone ages for ~ 3 million stars with the second Gaia data release. *Monthly Notices of the Royal Astronomical Society*, 481(3):4093–4110, 2018.

[49] E. Schatzman. A theory of the role of magnetic activity during star formation. *Annales d'Astrophysique*, 25:18, Feb. 1962.

[50] Aldo Serenelli, Jennifer Johnson, Daniel Huber, Marc Pinsonneault, Warrick H. Ball, Jamie Tayar, Victor Silva Aguirre, Sarbani Basu, Nicholas Troup, Saskia Hekker, Thomas Kallinger, Dennis Stello, Guy R. Davies, Mikkel N. Lund, Savita Mathur, Benoit Mosser, Keivan G. Stassun, William J. Chaplin, Yvonne Elsworth, Rafael A. García, Rasmus Handberg, Jon Holtzman, Fred Hearty, D. A. García-Hernández, Patrick Gaulme, and Olga Zamora. The First APOKASC Catalog of Kepler Dwarf and Subgiant Stars. *ApJS*, 233(2):23, Dec. 2017.

[51] V. Silva Aguirre, G. R. Davies, S. Basu, J. Christensen-Dalsgaard, O. Creevey, T. S. Metcalfe, T. R. Bedding, L. Casagrande, R. Handberg, M. N. Lund, P. E. Nissen, W. J. Chaplin, D. Huber, A. M. Serenelli, D. Stello, V. Van Eylen, T. L. Campante, Y. Elsworth, R. L. Gilliland, S. Hekker, C. Karoff, S. D. Kawaler, H. Kjeldsen, and M. S. Lundkvist. Ages and fundamental properties of Kepler exoplanet host stars from asteroseismology. *MNRAS*, 452(2):2127–2148, Sept. 2015.

[52] Víctor Silva Aguirre, Mikkel N. Lund, H. M. Antia, Warrick H. Ball, Sarbani Basu, Jørgen Christensen-Dalsgaard, Yveline Lebreton, Daniel R. Reese, Kuldeep Verma, Luca Casagrande, Anders B. Justesen, Jakob R. Mosumgaard, William J. Chaplin, Timothy R. Bedding, Guy R. Davies, Rasmus Handberg, Günter Houdek, Daniel Huber, Hans Kjeldsen, David W. Latham, Timothy R. White, Hugo R. Coelho, Andrea Miglio, and Ben Rendle. Standing on the Shoulders of Dwarfs: the Kepler Asteroseismic LEGACY Sample. II.Radii, Masses, and Ages. *ApJ*, 835(2):173, Feb. 2017.

[53] A. Skumanich. Time Scales for Ca II Emission Decay, Rotational Braking, and Lithium Depletion. *ApJ*, 171:565, Feb. 1972.

[54] David Soderblom. Astrophysics: Stellar clocks. *Nature*, 517(7536):557–558, Jan. 2015.

[55] David R. Soderblom. The Ages of Stars. *ARA&A*, 48:581–629, Sept. 2010.

[56] Jennifer L. van Saders, Tugdual Ceillier, Travis S. Metcalfe, Victor Silva Aguirre, Marc H. Pinsonneault, Rafael A. García, Savita Mathur, and Guy R. Davies. Weakened magnetic braking as the origin of anomalously rapid rotation in old field stars. *Nature*, 529(7585):181–184, Jan. 2016.

[57] Jennifer L. van Saders and Marc H. Pinsonneault. Fast Star, Slow Star; Old Star, Young Star: Subgiant Rotation as a Population and Stellar Physics Diagnostic. *ApJ*, 776(2):67, Oct. 2013.

[58] Jennifer L. van Saders, Marc H. Pinsonneault, and Mauro Barbieri. Forward Modeling of the Kepler Stellar Rotation Period Distribution: Interpreting Periods from Mixed and Biased Stellar Populations. *ApJ*, 872(2):128, Feb. 2019.

[59] Edmund J. Weber and Jr. Davis, Leverett. The Angular Momentum of the Solar Wind. *ApJ*, 148:217–227, Apr. 1967.

[60] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[61] J. Zorec and F. Royer. Rotational velocities of A-type stars. IV. Evolution of rotational velocities. *A&A*, 537:A120, Jan. 2012.