

MAISTEP: A new grid-based machine learning tool for inferring stellar parameters

I. Ages of giant planet host stars

J. Kamulali^{1,2,3,*}, B. Nsamba^{2,1,3}, V. Adibekyan^{3,4}, A. Weiss², T. L. Campante^{3,4}, and N. C. Santos^{3,4}

¹ Department of Physics, Faculty of Science, Kyambogo University, P.O. Box 1, Kyambogo, Kampala, Uganda

² Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, D-85748 Garching, Germany

³ Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, Rua das Estrelas, 4150-762 Porto, Portugal

⁴ Departamento de Física e Astronomia, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

Received 2 December 2024 / Accepted 30 January 2025

ABSTRACT

Context. Our understanding of exoplanet demographics partly depends on their corresponding host star parameters. With the majority of exoplanet-host stars having only atmospheric constraints available, robust inference of their parameters (including ages) is susceptible to the approach used.

Aims. The goal of this work is to develop a grid-based machine learning tool capable of determining the stellar radius, mass, and age using only atmospheric constraints. We also aim to analyse the age distribution of stars hosting giant planets.

Methods. Our machine learning approach involves combining four tree-based machine learning algorithms (random forest, extra trees, extreme gradient boosting, and CatBoost) trained on a grid of stellar models to infer the stellar radius, mass, and age using effective temperatures, metallicities, and *Gaia*-based luminosities. We performed a detailed statistical analysis to compare the inferences of our tool with those based on seismic data from the APOKASC (with global oscillation parameters) and LEGACY (with individual oscillation frequencies) samples. Finally, we applied our tool to determine the ages of stars hosting giant planets.

Results. Comparing the stellar parameter inferences from our machine learning tool with those from the APOKASC and LEGACY, we find a bias (and a scatter) of -0.5% (5%) and -0.2% (2%) in radius, 6% (5%) and -2% (3%) in mass, and -9% (16%) and 7% (23%) in age, respectively. Therefore, our machine learning predictions are commensurate with seismic inferences. When applying our model to a sample of stars hosting Jupiter-mass planets, we find the average age estimates for the hosts of hot Jupiters, warm Jupiters, and cold Jupiters to be 1.98 Gyr, 2.98 Gyr, and 3.51 Gyr, respectively.

Conclusions. Our machine learning tool is robust and efficient in estimating the stellar radius, mass, and age when only atmospheric constraints are available. Furthermore, the inferred age distributions of giant planet host stars confirm previous predictions – based on stellar model ages for a relatively small number of hosts, as well as on the average age-velocity dispersion relation – that stars hosting hot Jupiters are statistically younger than those hosting warm and cold Jupiters.

Key words. astronomical databases: miscellaneous – planet-star interactions – stars: fundamental parameters – stars: solar-type – stars: statistics

1. Introduction

The proper characterisation of exoplanet-host stars plays a vital role in our understanding of exoplanetary systems. This is based on the synergy between stellar and planetary science. For instance, the planet's radius and mass, determined through transits and radial velocity analysis, rely on the known radius and mass of the host star, respectively (e.g. Seager & Mallen-Ornelas 2003; Torres et al. 2008; Mortier et al. 2013; Perryman 2018; Hara & Ford 2023). Stellar ages have also been used in investigating changes in the exoplanet demographics over time (e.g. Berger et al. 2020; Chen et al. 2023). They are also expected to be relevant in distinguishing between the different theories regarding the formation and evolution of hot Jupiters (Dawson & Johnson 2018).

To determine robust stellar radius, mass, and age, various methods have been developed. Torres et al. (2010) devised poly-

nomial functions based on a sample of non-interacting eclipsing binary systems with model-independent radii and masses available, and with measured parameters such as effective temperature, T_{eff} , surface gravity, $\log g$, and metallicity, $[\text{Fe}/\text{H}]$. These relations have been adopted to uniformly estimate similar parameters of planet host stars (e.g. Santos et al. 2013; Hamer & Schlaufman 2019). Furthermore, Moya et al. (2018) used an extended sample of stellar targets with radius and mass estimates obtained through the analysis of detached eclipsing binary star systems, interferometric measurements, and seismic inferences. They reported 38 empirical relations for estimating mass and radius, which were obtained from linear combinations of stellar density, ρ , luminosity, L , $[\text{Fe}/\text{H}]$, T_{eff} , and $\log g$. However, the relations by both Torres et al. (2010) and Moya et al. (2018) cannot provide any information about the stellar interior structure and ages.

Stellar ages are typically inferred through empirical approaches, such as gyro-chronology (e.g. Angus et al. 2015;

* Corresponding author; kamulali@mpa-garching.mpg.de

Meibom et al. 2015; Barnes et al. 2016; Claytor et al. 2020), chemical clocks (e.g. Da Silva et al. 2012; Nissen 2015, 2016; Feltzing et al. 2016; Titarenko et al. 2019; Mena et al. 2019; Espinoza-Rojas et al. 2021; Moya et al. 2022), and stellar activity (e.g. Noyes et al. 1984; Mamajek & Hillenbrand 2008). Model-dependent techniques, such as isochrone fitting (e.g. Liu & Chaboyer 2000; Sandage et al. 2003; VandenBerg et al. 2013; Berger et al. 2020) and asteroseismology are also used (e.g. Silva Aguirre et al. 2011, 2013, 2017; Campante et al. 2015, 2017; Nsamba et al. 2018, 2019; Moedas et al. 2020; Deal et al. 2021; Moedas et al. 2024). Each of these methods works well in particular regions of the Hertzsprung–Russell (HR) diagram. For a comprehensive summary of these techniques, we refer to the review by Soderblom (2010, 2015).

Grid-based methods involve finding the best match of stellar observables to model observables through processes, such as iterative optimisation (Gai et al. 2011), Markov chain Monte Carlo (MCMC; Bazot et al. 2012) sampling, and genetic optimisation algorithms (Metcalfe et al. 2009). In iterative optimisation, stellar models are specially constructed for the region of parameter space where the solution is believed to lie. A subsequent optimisation procedure is then carried out to find the best-fitting model based on the available observable parameters of the star. The MCMC algorithm is an augmentation of iterative technique that allows for an extensive probabilistic exploration of the parameter space and yields accurate estimates of stellar parameters with credible confidence intervals (e.g. Bazot et al. 2012; Gruberbauer et al. 2012; Lund & Reese 2017; Claytor et al. 2020). In a genetic optimisation algorithm, stellar tracks have to be run each time a new target is to be characterised. Additionally, the algorithm must be initialised from different starting points to ensure convergence to a global minimum. Thus, to optimise the efficacy of these methods, we must run dense model grids.

With large datasets, machine learning (ML) techniques have been efficiently utilised in the characterisation of stars. The application of ML algorithms for regression tasks in astrophysics dates back to the earlier works of Pulone & Scaramella (1997), in which an artificial neural network was built to determine stellar ages based on the position of the star on the HR diagram. Using a combination of classical and asteroseismic observations, Verma et al. (2016) and Bellinger et al. (2016) characterised sun-like stars through the application of neural networks and a random forest, respectively. Despite the tremendous progress of ML codes in utilising both seismic and classical constraints to estimate stellar radius, mass, and age, majority of stars do not have seismic data. This opens an opportunity to explore the effectiveness of ML codes in the absence of seismic data.

Moya & López-Sastre (2022) used atmospheric constraints to determine the capabilities of various machine learning algorithms in inferring stellar mass and radius. They demonstrated that combining (stacking) the high-performing algorithms produces a more robust stellar model than any of the individual algorithms. In addition, they trained their ML algorithms on data from eclipsing binary stars, interferometry, and asteroseismology. However, only inferences of the stellar mass and radius were conducted in their work.

In this study, we employed a stacking approach, similar to Moya & López-Sastre (2022), but combine (stack) tree-based ensemble algorithms (RF, XT, CatBoost, and XGBoost). Tree-based algorithms are computationally efficient (i.e. require relatively low computational resources) compared to other algorithms, such as neural networks (Hastie et al. 2009). In addition, they are effective in developing regression models that cap-

ture complex non-linear relationships (Hastie et al. 2009; Baron 2019), particularly in linking observed stellar parameters to their fundamental properties (such as mass, radius, age, etc.). To achieve this, we train our algorithms directly on grids of stellar models to determine not only the radius and the mass, but also the stellar age using atmospheric constraints. Furthermore, this approach makes it possible to explore the impact of stellar model physics on the predictions made. The rest of the paper is structured as follows. In Sect. 2 we present our ML tool and its underlying mechanisms. This is followed by an evaluation of the tool in Sect. 3. We then apply the technique to characterise planet host stars in Sect. 4. Finally, we provide the summary and conclusions of our results in Sect. 5.

2. MAISTEP code

Machine learning Algorithm for Inferring STEllar Parameters (MAISTEP) is a tool which combines Random Forest (RF; Breiman 2001), eXtremely randomised Trees/eXtra Trees (XT; Wehenkel et al. 2006), eXtreme Gradient Boosting (XGBoost; Chen & Guestrin 2016), and Categorical Boosting (CatBoost; Prokhorenkova et al. 2018) algorithms, trained on a grid of stellar models to predict the radius, mass, and age of stars using atmospheric constraints. The predictions from the four algorithms are combined using a weighted sum to produce an optimal prediction. Combining algorithms with diverse architectures increases computational demand and adds complexity, but it contributes to building a generalised highly accurate predictor. This approach known as stacking generalisation (Wolpert 1992; Breiman 1996) allows different algorithms to capture different patterns in the data, resulting in improved performance on unseen data compared to using a single algorithm. Figure 1 highlights the vital sections of MAISTEP and their details are summarised in Sects. 2.1–2.4.

2.1. Training and test data

The data for training and testing the algorithms in MAISTEP were generated using the open-source one-dimensional (1D) stellar evolution code, Modules for Experiments in Stellar Astrophysics (MESA; Paxton et al. 2018, 2019), version r12778. The parameter space of the stellar grid spans a range in initial mass, $[0.7\text{--}1.6] M_{\odot}$ in steps of $0.05 M_{\odot}$, initial metallicity, $[\text{Fe}/\text{H}]$, $[-0.5\text{--}0.5]$ dex in steps of 0.1, and enrichment ratios, $\Delta Y/\Delta Z$, $[0.4\text{--}2.4]$ in steps of 0.4. The initial helium mass fraction, Y_i was calculated following

$$Y_i = \frac{\Delta Y}{\Delta Z} Z_i + Y_o, \quad (1)$$

where, Z_i is the initial metal mass fraction and the offset, $Y_o = 0.2484$, is the primordial helium mass fraction (Cyburt et al. 2003). The grid consists of tracks evolved from the zero-age main sequence (ZAMS) to end of the main sequence (defined to be a point when the central hydrogen mass fraction drops below 10^{-3}). Microscopic diffusion was included considering only the gravitational settling component for models that do not show an unrealistic over-depletion of surface elements; thus, it was applied only to models with a ZAMS mass below $\lesssim 1.2 M_{\odot}$. In addition, for models with convective cores, an exponential overshoot scheme was implemented with a fixed efficiency value of 0.01. We assume that our target stars are slow rotators and, thus, the effects of rotation were neglected. Our grid employs a single solar-calibrated value of $\alpha_{\text{MLT}} = 1.71$. The impact of using a

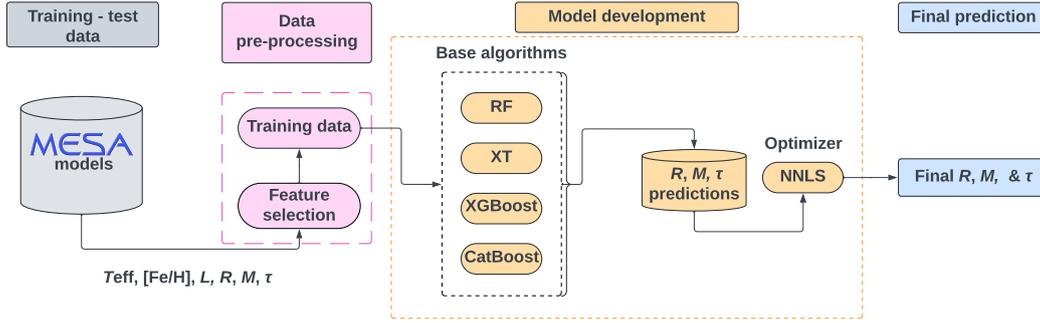


Fig. 1. Schematic representation of MAISTEP. The training data undergoes pre-processing before being passed to the different base algorithms under the model development phase, ultimately predicting the radius, mass, and age. See text for details.

Table 1. Main components of our grid of stellar models.

Parameter	Range
Initial mass in M_{\odot}	[0.7–1.6], $\Delta M = 0.05$
Initial metallicity in dex	[−0.5–0.5], $\Delta[\text{Fe}/\text{H}] = 0.1$
Enrichment ratio, $\Delta Y/\Delta Z$	[0.4–2.4], $\Delta(\Delta Y/\Delta Z) = 0.4$
Input physics	Prescription
Nuclear reaction rates	1, 2
Solar chemical mixtures	3
OPAL equation of state	4
OPAL opacities	5, 6
Atmosphere model	7
Atomic diffusion	8
Core overshooting, $f_{\text{ov}} = 0.01$	9
$\alpha_{\text{MLT}} = 1.71$	10

Notes. (1) Angulo et al. (1999); (2) Imbriani et al. (2005); (3) Asplund et al. (2009); (4) Rogers & Nayfonov (2002); (5) Iglesias & Rogers (1996); (6) Ferguson et al. (2005); (7) Krishna Swamy (1966); (8) Thoul et al. (1994); (9) Herwig (2000); (10) Cox & Giuli (1968).

single mixing length parameter in stellar grids on inferred stellar properties has been explored in Silva Aguirre et al. (2015). Table 1 summarises the main aspects of our grid of stellar models. For a detailed description of the global physical inputs adopted, we refer to Nsamba et al. (2025).

2.2. Data pre-processing

We selected 50 models that are nearly evenly spaced in the central hydrogen mass fraction along each evolutionary track. We followed the same technique as in Bellinger et al. (2016) to set up and solve an optimisation problem using `cvxpy` library (Boyd & Vandenberghe 2004) in Python. This involves finding a subset of models along each track that match a desired set of evenly spaced values in the core hydrogen mass fraction. This approach ensures a well-balanced and representative sample of stellar models during the main sequence phase.

The dataset is randomly split into 80% for training and 20% for testing our algorithms. While the training set is passed to each algorithm, the testing set is held back until the final model evaluation phase. We note that our training features only took into account T_{eff} , $[\text{Fe}/\text{H}]$, and L .

2.3. Model development

The training set was further split by employing the k -fold cross-validation (Hastie et al. 2009) approach. Each algorithm (i.e. RF, XT, CatBoost, and XGBoost) trains on $k - 1$ folds (subsets) and then makes predictions for the validation set. The process is repeated k times with each fold (subset) being used as a validation set once. In our case, we consider $k = 10$.

We employed Optuna (Akiba et al. 2019) to auto-tune the hyperparameters (i.e. parameters that control the structure and learning process of a model) of each base algorithm. Optuna incorporates cross-validation and explores various hyperparameter configurations to identify the ones that deliver the best model performance, as measured by a defined metric. The library includes an in-built class to sample the user-defined parameter space, along with a pruning algorithm that monitors the intermediate result of each trial and terminates the unpromising ones, thereby accelerating the exploration process.

Given the large number of hyperparameters associated with our base algorithms, we initially used Optuna to determine which hyperparameters most significantly influence each algorithm’s ability to predict stellar radius, mass, and age. We then focused on tuning these key hyperparameters to streamline the hyperparameter search while maintaining each algorithm’s predictive performance.

The mean squared error metric is employed in Optuna to evaluate the performance of a model and guide the optimisation process in selecting the optimal set of hyperparameters. A maximum of 50 trials are specified to balance computational efficiency while allowing for a thorough exploration of the hyperparameter space. The selected optimal hyperparameters (see Table 2) are then used to generate cross-validation predictions which served as training input (meta-features) for the non-negativity constrained least squares (nnls) optimiser. For a detailed description of each base algorithm, we refer to Breiman (2001), Wehenkel et al. (2006), Chen & Guestrin (2016), Breiman (2017), and Prokhorenkova et al. (2018).

2.4. Stacking approach

The predictions from the base algorithms, generated during cross-validation, were assigned non-negative weights, with the most accurate predictions typically receiving higher weights. The weights were determined by the least squares optimisation method (Lawson & Hanson 1995), which adjusts them to minimise the overall prediction error. Essentially, given the base algorithms labelled, $j = 1, 2, 3, 4$, each making predictions p_{ij} , for the i^{th} data point in a sample of a size, n , we sought to find

Table 2. Optimal hyperparameters (hp) returned by Optuna for the radius, mass, and age from each base model.

hp	RF			XT			XGBoost			CatBoost		
	Radius	Mass	Age	Radius	Mass	Age	Radius	Mass	Age	Radius	Mass	Age
learning_rate	–	–	–	–	–	–	0.29	0.3	0.3	0.64	0.97	0.95
depth / max_depth	20	20	20	20	20	20	6	6	6	5	5	5
n_estimators / iterations	300	224	75				1000	1000	1000	1438	1172	1495

Notes. “def” signifies that the default value was considered. Default values were also used for any hp that are excluded here.

the weights, β_j , as:

$$\min \sum_{i=1}^n (\hat{y}_i - y_i)^2, \quad (2)$$

with

$$\hat{y}_i = \sum_{j=1}^k \beta_j \cdot p_{ij}, \quad (3)$$

subject to $\sum \beta_j = 1$, $\beta_j \geq 0$. Here, y_i is the actual value of the i^{th} data point and \hat{y}_i is the associated weighted prediction defined by Eq. (3). This approach allows for a clear assessment of how much each base algorithm contributes to the final stacked model’s prediction, with the weights directly reflecting each algorithm’s relative importance. Since the weights are constrained to be non-negative and sum to 1, the interpretation is straightforward: higher weights indicate a greater contribution to the final prediction. The optimal weights generated are adopted for producing the final predictions during the test and the application phase, via the stacking approach. We note that we imported the nnls optimiser from the `scipy.optimize` library into our Python code.

3. Evaluation of MAISTEP

We conducted a double assessment of our code using both artificial data from stellar models and real observed stars. We utilised the root mean squared error (RMSE) to evaluate and compare the performance of the base algorithms and our weight-based model on test data. We chose RMSE because it retains the same unit as the target variable, making it easy to interpret the bias.

3.1. Performance on artificial data

We recall that our data is divided into 80% training and 20% test sets. We further partitioned the training set into subsets ranging from 10% to 100% in increments of 10%. For each subset, optimal weights used to stack the predictions generated through cross-validation (see Sect. 2.3) by the base algorithms were determined using nnls, as described in Sect. 2.4. These weights are shown in Fig. 2. Overall, we observe that higher weights are consistently assigned to XT predictions in estimating the radius, mass, and age, across all training sizes. This suggests that XT’s predictions are generally more accurate compared to those from other algorithms, at least within the scope of our tuned hyperparameters. Furthermore, while variations in weights are observed at different training sizes, they tend to stabilise as training data approach full size, except in the case of mass.

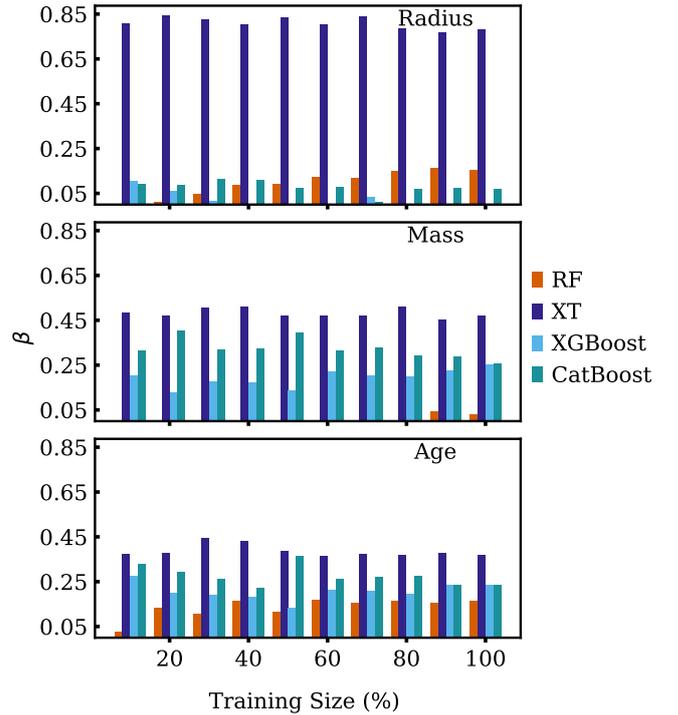


Fig. 2. Distribution of the weights (β) derived using nnls optimiser, which are employed to combine predictions of stellar radius (top panel), mass (middle panel), and age (bottom panel) from the base algorithms. See text for details.

To evaluate the generalisation capability of our base algorithms, we trained each algorithm on each subset. We then made predictions for the 20% test set. This evaluation assumes that the input features (i.e. T_{eff} , $[\text{Fe}/\text{H}]$, and L) used to predict the radius, mass, and age of the test set are free from errors and bias. Finally, we compared the performance of each base model on the test set to that of the stacking model.

Figure 3 shows that as more training examples become available, each algorithm learns and improves its ability to predict the radius, mass, and age, resulting in a reduced bias on the test set. For radius (top panel of Fig. 3), the learning continues until a stable regime (saturation point) is reached with about 50% of the training samples; beyond this, no additional improvement occurs within the framework of our hyperparameter tuning. The early attainment of the saturation point could be due to two of the training features (i.e. T_{eff} and L) directly constraining the radius through the Stefan-Boltzmann relation (Planck 1900). In addition, we note that the stellar model luminosity, L , is derived using the same relation with the T_{eff} and R as inputs, in stellar

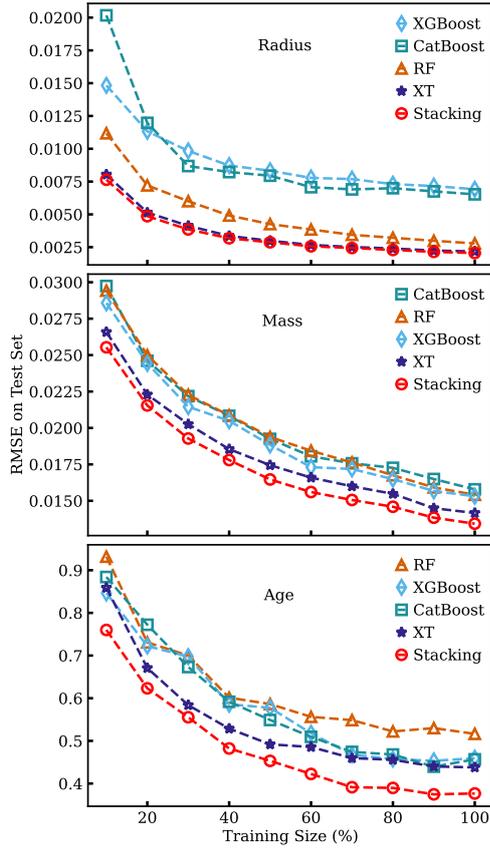


Fig. 3. Evaluation metric on the test sample as a function of the training size. The legend is organised in the order of performance, with the worst performers at the top and the best performers at the bottom.

evolutionary codes. Therefore, these findings may be consistent due to the non-independence of L , yet they still demonstrate that an accurate L coupled with T_{eff} , can effectively recover the stellar radius using any of our ML algorithms.

In contrast, the RMSE curves for the mass (middle panel of Fig. 3) show that learning continues up to 100% training size, while for the age (bottom panel of Fig. 3), they flatten out at $\sim 90\%$. In particular, for the mass, this could suggest that the algorithms required more samples (information) to explore in order to achieve adequate training. To test this, we trained our algorithms on the entire dataset and made predictions for some observed stars through stacking. We find that increasing the training sample had almost no influence on the resulting stellar masses.

Overall, within the scope of our hyperparameter tuning, XT stands out as the best performing base algorithm for our regression tasks in predicting stellar radius, mass, and age. It is also weighted the most in our stacking approach across all training sizes as seen in Fig. 2. XT’s good performance likely stems from its highly randomised approach to feature selection and choosing split points, which allows for an effective balance between the bias and variance while maintaining a competitive accuracy. Therefore, we recommend it for use when computational resources are limited (we refer the reader to Bellinger et al. 2016, for a full exploration of XT). However, we observe enhanced performance when XT is combined with the other three algorithms, particularly in reducing the bias in age predictions, resulting into an increase in accuracy of $\sim 7\%$ (compared to XT) at full training size (see Fig. 3). Therefore, our stacking approach

demonstrates a distinct advantage in performance, as it effectively combines the strength of the base algorithms, making it the preferred method of choice for inferring the radius, mass, and age of observed stars.

3.2. Performance on real stars

We applied MAISTEP in its stacking approach setting, along with the associated weights to characterise real stars in the APOKASC (Serenelli et al. 2017) and LEGACY (Silva Aguirre et al. 2017; Lund et al. 2017) samples. We select targets that fall within the parameter space of the stellar models employed in the training process. We used the observations of T_{eff} , $[\text{Fe}/\text{H}]$, and L , to predict radius, mass, and age. To check for consistency, we also compared the predicted stellar properties to those obtained in the two data sources (APOKASC and LEGACY sample), which employ different methods and incorporate both seismic and spectroscopic constraints in their optimisation processes.

The APOKASC catalogue consists of 415 (main sequence and sub-giant) stars whose properties were determined utilising global asteroseismic observables (i.e. large frequency separation, $\Delta\nu$ and frequencies at maximum power, ν_{max}) complemented with effective temperatures, T_{eff} and metallicities, $[\text{Fe}/\text{H}]$. Of the targets, 167 have stellar properties within the parameter space of our training grid. This excludes stars that overlap with the LEGACY sample, since they are accounted for in the LEGACY comparisons. In addition, we also exclude stars identified as binaries according to SIMBAD and/or with astrometric solution above the favoured renormalised unit weight error (RUWE) threshold of ≤ 1.4 in *Gaia* DR3¹ (Lindgren 2018). Serenelli et al. (2017) present two sets of results based on different temperature scales. We compare our findings to those obtained using the photometric T_{eff} scale from the SDSS *griz* bands, as recommended by the authors.

The LEGACY sample is composed of 66 main-sequence *Kepler* targets (Silva Aguirre et al. 2017; Lund et al. 2017). The stellar radii, masses, and ages were determined using either individual mode frequencies or frequency ratios together with spectroscopic effective temperature and metallicity compiled from different literature sources (i.e. Ramírez et al. 2009; Pinsonneault et al. 2012; Huber et al. 2013; Pinsonneault et al. 2014; Chaplin et al. 2013; Casagrande et al. 2014; Buchhave & Latham 2015). Silva Aguirre et al. (2017) demonstrates a disagreement between parallaxes obtained from asteroseismic distances (which they computed from a combination of infrared flux measurements of angular diameter and asteroseismic radius) and *Gaia* parallaxes, with discrepancies exceeding 3σ for some of the stellar targets. They emphasised that the flux of those stars could be contaminated by their companions, and since that directly impacts the luminosity determinations, we excluded them. These include: KIC 9025370, KIC 7510397, KIC 8379927, KIC 10454113, KIC 9139151, KIC 9965715, KIC 7940546, KIC 4914923, and KIC 12317678. Furthermore, we excluded KIC 10068307 and the spectroscopic binary component KIC 6933899. Hence, a total of 55 stars from the LEGACY sample are considered in our analysis.

3.2.1. Luminosity determination

Both the APOKASC and LEGACY data sources provide measurements of metallicities and effective temperatures. We took

¹ <https://gea.esac.esa.int/archive/>

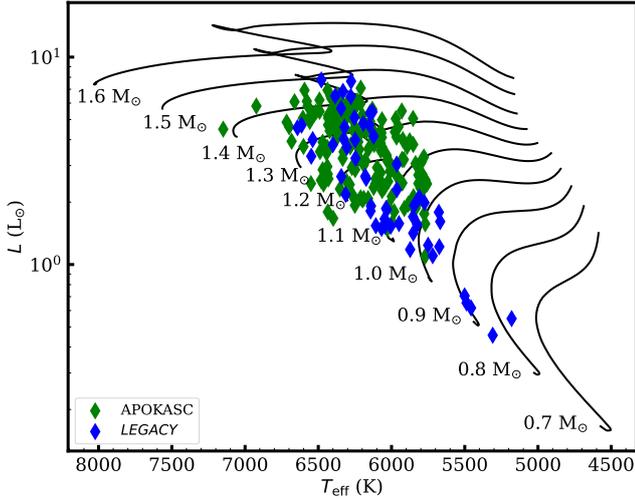


Fig. 4. HR diagram showing our selected sample, which include 167 stars from APOKASC catalogue plotted in green and 55 stars from the LEGACY sample shown in blue diamonds. In addition, we over-plot a few tracks at solar metallicity, $[Fe/H] = 0$, extending up to the bottom of the red-giant branch.

advantage of the precise parallaxes from *Gaia* data in DR3 and followed Pijpers (2003) to determine the corresponding luminosities using the expression

$$L = 10^{4.0 + 0.4M_{bol,\odot} - 2\log\pi(\text{mas}) - 0.4(m_g - A_g + BC_g)}, \quad (4)$$

where π and m_g represent the *Gaia* parallax and apparent g -band magnitudes (Gaia Collaboration 2023), respectively. The bolometric corrections BC_g for the sample were computed using coefficients from Andrae et al. (2018) and we adopted $M_{bol,\odot} = 4.74$, as recommended by Mamajek et al. (2015). We correct for the effects of interstellar absorption by computing the extinction, defined by

$$A_g = R_g \times E(B - V), \quad (5)$$

where $E(B - V)$ is the color excess which was queried using 3D dustmaps² python package (Green 2018). We adopt an extinction coefficient, $R_g = 3.303$, according to (Schlafly & Finkbeiner 2011). The dust maps do not provide uncertainties on the reddening. Therefore, to compute the lower and upper uncertainty levels on luminosity, we employed standard error propagation considering only the uncertainty in parallax and apparent magnitude. In Fig. 4, we display the approximate positions of our selected targets on the HR diagram with evolutionary tracks constructed at solar metallicity running from ZAMS to the bottom of the red giant branch.

To propagate the uncertainties in T_{eff} , $[Fe/H]$, and L to our inferred properties, we randomly generate 10 000 realisations based on the standard deviations of these observables. These realisations are then input into the trained algorithms, resulting in 10 000 predictions for the radius, mass, and age of each star. From the distributions, we compute the median and quartiles.

3.2.2. Comparison with APOKASC

In Fig. 5, we compare the estimates of the stellar radii, masses, and ages obtained using MAISTEP to those in APOKASC. We

² <https://dustmaps.readthedocs.io/en/latest/>

find a good agreement between the MAISTEP radius values, R_{ML} , with those from APOKASC, R_{SEIS} (see top panel of Fig. 5). We observe a scatter of 5% and an offset/bias of -0.5% . The average relative uncertainty in radius from our tool is 3%, which is higher than the 1% uncertainty in the APOKASC radius estimations.

The middle panel of Fig. 5 shows that the masses inferred using MAISTEP are systematically higher by 6% and we obtain a scatter of 5%. The average statistical relative uncertainty in mass from MAISTEP is 4%, compared to 3% from APOKASC. Furthermore, this translates into MAISTEP yielding systematically lower ages with a mean offset of -9% and a scatter of 16% (see the bottom panel of Fig. 5). The average relative uncertainty in age is 23% and higher than the 17% based on estimations in the APOKASC. The observed offsets could stem from variations in the model physics and the sets of constraints employed.

3.2.3. Comparison with LEGACY

Silva Aguirre et al. (2017) provides seven sets of results generated by different optimisation codes adopting grids with different stellar model physics. Here, we compare our results with those inferred using the Asteroseismic Inference on a Massive Scale (AIMS: Rendle et al. 2019) pipeline. AIMS infers stellar radii, masses, or ages by combining the likelihood function with prior distributions on those properties, followed by an exploration of parameter space using MCMC to obtain the resulting distributions. The grid of models adopted in AIMS by Silva Aguirre et al. (2017) was constructed by Coelho et al. (2015) using the MESA stellar evolutionary code, without taking into account atomic diffusion. In the optimisation process, individual oscillation mode frequencies were complemented with spectroscopic effective temperatures and metallicities.

The top panel of Fig. 6 presents a comparison between the radius estimates obtained using MAISTEP and AIMS. We find a bias of -0.2% and a scatter of 2%, demonstrating an excellent agreement in radius determinations by the two approaches. The statistical average relative uncertainties in radius from MAISTEP and AIMS are consistent, namely, 2% and 1%, respectively. In the middle panel of Fig. 6, we observe a bias of -2% and a scatter of 3% in mass with MAISTEP estimations being systematically lower than those from AIMS. The average relative uncertainty in mass for the MAISTEP and AIMS estimations is 4% and 3%, respectively. For the age, estimates obtained using MAISTEP are systematically higher by 7% and a scatter of 23% (see bottom panel of Fig. 6). ML estimates also have an average relative age uncertainty of 28%, compared to 9% from AIMS. The systematically higher ages inferred by MAISTEP contrast with the expected age estimates from models with and without atomic diffusion. The observed discrepancy may be explained by the systematically lower masses inferred by MAISTEP, which could arise from the differences in the observational constraints used in MAISTEP, compared to the LEGACY analyses. Overall, MAISTEP predicts stellar radii, masses, and ages that are commensurate with seismic inferences.

4. Ages of giant-planet hosts

Over the past two decades, the discovery of exoplanets has consistently defied our expectations set by the Solar System, uncovering a much greater diversity of planetary systems than ever thought. One of the most remarkable discoveries is the existence of close-in Jupiter-mass planets (the hot Jupiters), with orbital periods of just a few days (Mayor & Queloz 1995; Rasio & Ford

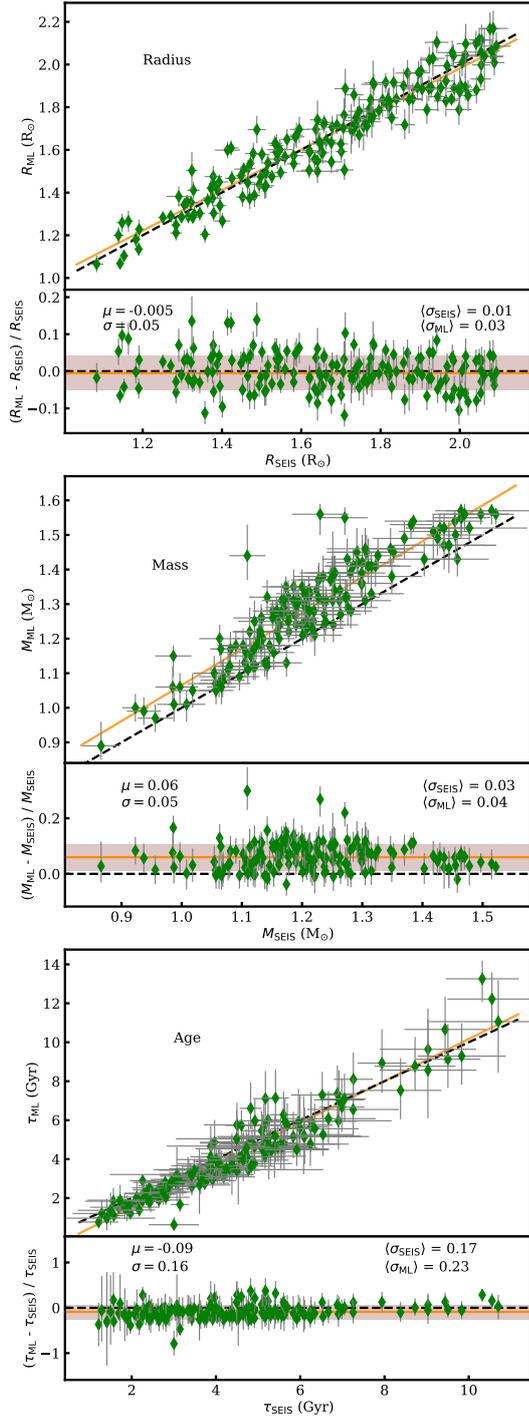


Fig. 5. One-to-one relation and fractional differences from MAISTEP with respect to APOKASC as the reference values, for radius (top), mass (middle), and age (bottom). The black dashed lines in all panels represent the unity relation. The orange line in the one-to-one plots indicates the best-fit line while in the fractional difference plots, it represents the mean offset/bias, μ . The brown shaded region highlights the associated scatter, $\pm\sigma$. Additionally, we give the average statistical relative uncertainties in angle brackets.

1996), which has continually raised questions about their formation and longevity. More recently, with the substantial amount of exoplanet detections and our growing understanding of the star-planet connections, some statistical studies have focused on analysing the ages of the host stars to gain better insights into the

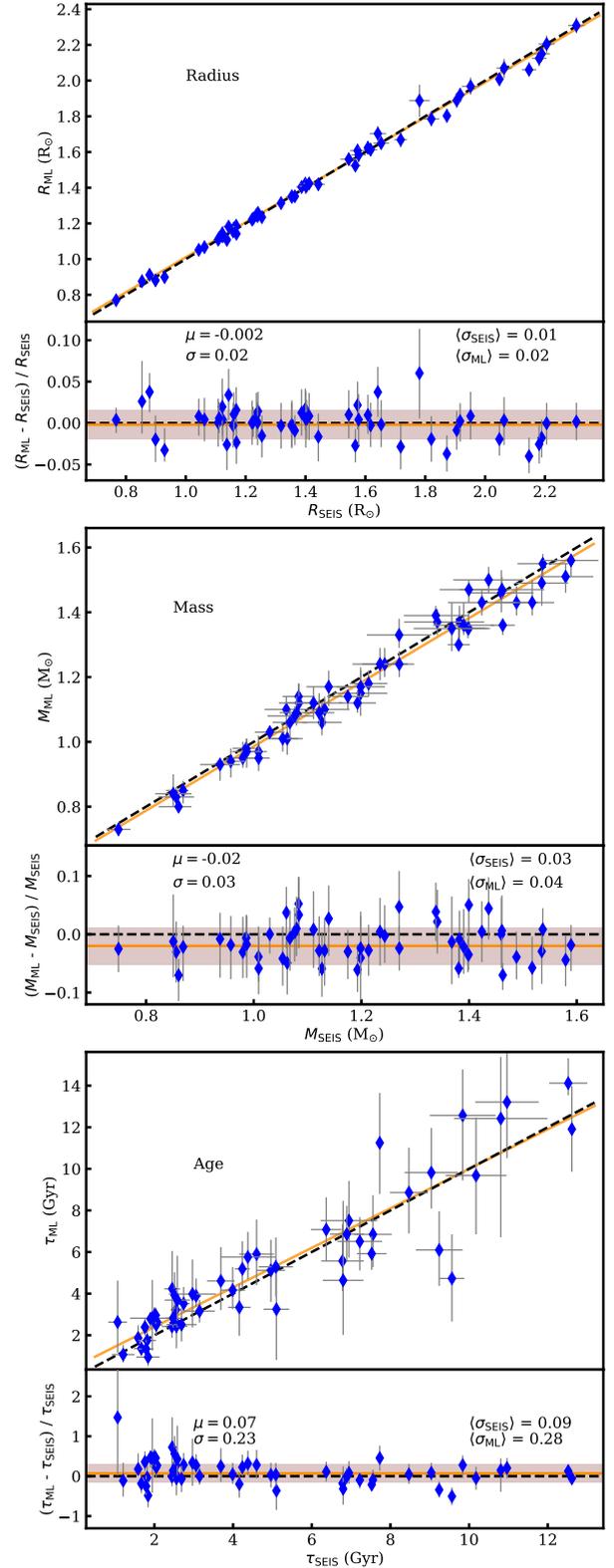


Fig. 6. Same as Fig. 5, but comparing estimates from MAISTEP and AIMS for the stars in the LEGACY sample.

evolution of the hot Jupiters. For instance, Miyazaki & Masuda (2023) used ages inferred from stellar models (isochrones) for a sample of 382 Sun-like stars, which includes 124 giant planets, from the California Legacy Survey catalogue to demonstrate that hot Jupiters are preferentially hosted by relatively

younger stars and that their number decreases with stellar age. Hamer & Schlaufman (2019) and Chen et al. (2023) illustrate their findings using age proxies, namely, the Galactic space velocities and the average age-velocity dispersion relation, respectively, and reach similar conclusions. The explanation provided is that tidal interactions between hot Jupiters and their host stars during the main sequence phase lead to orbital decay over time, causing these planets to spiral inward and ultimately be engulfed. As a result, the number of hot Jupiters decreases at the later stages of the main-sequence, suggesting that they are typically found around young stellar hosts. We aim to apply MAISTEP to infer the ages of individual stars from a relatively large sample consisting solely of planet hosts, to test the above claims.

We collect planet data from the NASA Exoplanet Archive³ and the Extrasolar Planets Encyclopaedia⁴. We selected stars orbited by giant planets with the best available mass estimate in the following order of preference: M_p , $M_p \times \sin i / \sin i$, or $M_p \times \sin i$, depending on data availability. The selected planets span a range in Jupiter mass of 0.3 to 13 ($13M_J$ is the lower limiting mass of brown dwarfs; Des Etangs & Lissauer 2022). We identified hot Jupiters as giant planets orbiting very close to their host stars with periods of ≤ 10 days. Our control sample consists of stars hosting the warm and cold Jupiters, with the same mass range as the hot Jupiters. The warm Jupiters have orbital periods of $10 < P(\text{days}) \leq 365$, while the cold Jupiters have orbital periods of $1 < P(\text{years}) \leq 10$ (Miyazaki & Masuda 2023). We then cross-matched the hosts that have confirmed Jupiter-mass planets with SWEET-Cat⁵, selecting only stars with homogeneously derived spectroscopic T_{eff} and $[\text{Fe}/\text{H}]$ (Santos et al. 2013).

The stellar luminosities were derived using Eq. (4). We select main sequence stars with observable properties within the parameter space of the ML training grid. Our sample consists of 427 stars hosting 470 giant planets, including 208 hot Jupiters, 96 warm Jupiters, and 152 cold Jupiters. Of these stars, 387 host single planets, 37 host double planets, and 3 host three planets. We clarify that, unless stated otherwise, for stars hosting multiple planets, if the planets share a similar range in orbital periods, the star is counted only once. However, if the planets have orbital periods that place them in different categories, the star is included in each of those categories.

Following the same procedures as discussed in Sect. 3.2, we calculate the ages of the giant-planet hosts. In the top panel of Fig. 7, we show the resulting age distributions of stellar populations hosting hot Jupiters, warm Jupiters, and cold Jupiters, which are constructed solely from the median ages. The data show distinct peaks at ages of 1.98, 2.98, and 3.51 Gyr, with corresponding mean ages of 3.52, 4.41, and 5.13 Gyr, respectively.

To quantify the age differences further, we conducted statistical tests on the distributions between the categories. Specifically, we tested the null hypothesis that the distributions are identical and that the observed mean age differences are not significant. We used Anderson-Darling (Anderson & Darling 1954; Scholz & Stephens 1987) and Kolmogorov-Smirnov (Kolmogorov 1933; Smirnov 1939) for k samples to assess the distributions, along with the t-test (Student 1908) to compare the mean values. These functions were sourced from `scipy.stat` in Python 3.10.12. We summarise the results in Table. 3, for three categories: A – hot Jupiter hosts versus warm Jupiter hosts; B – hot Jupiter hosts versus cold Jupiter hosts; and C – warm Jupiter hosts versus cold Jupiter hosts.

³ <https://exoplanetarchive.ipac.caltech.edu/>

⁴ <http://www.exoplanet.eu/>

⁵ <https://sweetcat.iastro.psu/>

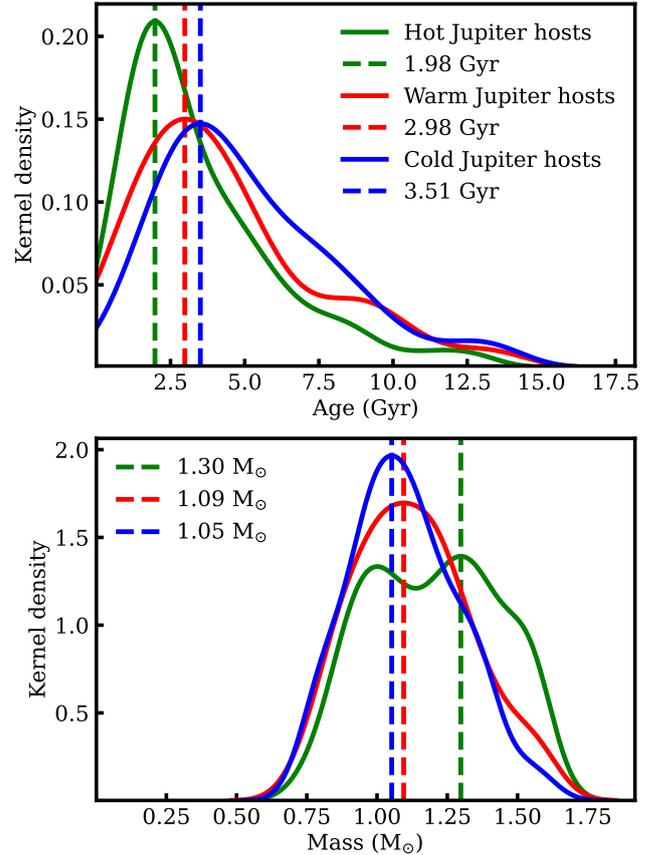


Fig. 7. Kernel density illustrating the age (top panel) and mass (lower panel) distributions of planet hosting stars.

We found low p -values at the 5% significance level for categories A and B, indicating that the distributions differ and the mean age differences are statistically significant. For category C, both the t and the KS tests fail to reject the null hypothesis, but the AD test rejects with a $p = 0.025$, suggesting that there may be a noticeable difference between the two distributions that the other two tests did not detect. This could be because of the AD’s sensitivity to the tails of the distributions. In addition, we conducted a test that excluded stars hosting planets in different categories. The results showed that the warm and cold Jupiter hosts are identical, with p -values of 0.05, 0.22, and 0.09. Generally, the statistical tests confirm that hot Jupiters are hosted by stars that are relatively younger than those hosting the warm and cold Jupiters. This is consistent with previous findings, which examined a relatively small sample of host stars (Miyazaki & Masuda 2023), as well as with kinematic results (e.g. Hamer & Schlaufman 2019; Chen et al. 2023).

In the bottom panel of Fig. 7, our sample shows that the hot-Jupiter hosts are statistically more massive than stars hosting the warm and cold Jupiters, which may help explain the observed age differences. The dip in the mass distribution of hot Jupiters could be a selection effect, as our study considers only hosts with homogeneously derived spectroscopic properties in the SWEET-cat catalogue. The p -values from the AD test, the KS and the t-test, for categories A and B, are all below the critical value of 0.05 (see Table 3) suggesting that the mass distributions are not identical. For category C, all tests do not reject the null hypothesis, with p -values greater than 0.05. Therefore, the warm and cold Jupiter mass distributions and associated mean values are

Table 3. Statistical test results from the age and mass distributions.

Category	AD test		KS test		t-test	
	AD	<i>p</i> -value	KS	<i>p</i> -value	t	<i>p</i> -value
Age						
A	7.81	$<10^{-3}$	0.23	8×10^{-4}	3.39	8×10^{-4}
B	30.5	$<10^{-3}$	0.36	3.07×10^{-11}	6.28	1.17×10^{-9}
C	2.74	0.025	0.16	0.07	1.57	0.12
Mass						
A	4.4	0.006	0.19	0.01	2.9	0.004
B	11.6	$<10^{-3}$	0.36	1.0×10^{-4}	4.6	5.8×10^{-6}
C	-0.53	0.25	0.08	0.81	0.98	0.33

Notes. The Anderson-Darling (AD) test, Kolmogorov-Smirnov (KS) test, and t-test for three categories: A – hot Jupiter hosts versus warm Jupiter hosts; B – hot Jupiter hosts vs cold Jupiter hosts; and C – warm Jupiter hosts vs cold Jupiter hosts.

similar, but they are statistically different when compared with the hot Jupiter host masses. Further, to assess the sensitivity of the statistics to the orbital period threshold set for the hot Jupiter hosts, we varied the period over a range of 5 to 15 days in steps of 1 day. The results show no significant departure from the conclusions drawn based on a period threshold of $P \leq 10$ days.

5. Summary and conclusions

We developed a grid-based machine learning tool (MAISTEP) for estimating robust stellar radius, mass, and age using atmospheric constraints. The code relies on transfer learning in which various machine algorithms (i.e. RF, XT, CatBoost, XGBoost) pre-trained on data from stellar models are used to make predictions for real stars. Most importantly, MAISTEP, has enhanced generalisation capabilities producing highly accurate stellar radius, mass, and age, in contrast to the typical use of single algorithms.

We explored the robustness of MAISTEP by conducting a detailed comparison with inferences from APOKASC and LEGACY samples. In terms of radius, we observe a bias of -0.5% with a scatter of 5% when compared to the APOKASC, and a bias of -0.2% with a scatter of 2% when compared to the LEGACY sample. A bias of 6% and -2% , with scatter values of 5% and 3% , is obtained in mass when compared to the APOKASC and LEGACY sample results, respectively. For age, the bias is -9% and 7% , with scatter values of 12% and 23% when compared to the APOKASC and LEGACY samples, respectively. Accordingly, our results demonstrate that we can infer stellar masses and radii to a precision relevant for exoplanetary studies, in cases where seismic information is not available. Our findings also emphasise the need for seismic information in order to yield robust ages. In addition, in cases where it's necessary to explore the internal structures of stars, asteroseismology offers a solution.

Lastly, we used MAISTEP to estimate ages of planet-hosting stars that have homogeneously derived spectroscopic effective temperatures and metallicities, which we complemented with *Gaia*-based luminosities. We obtain consistent results with previous findings determining that hot Jupiter planets are preferentially hosted by relatively a younger and massive stellar population compared to the warm and cold Jupiter hosts.

Our study primarily focused on the stars in the main sequence; however, we plan to extend it to the sub-giant and red-giant branches, where there is a growing number of detections of high-mass planet-hosting stars, supported by their slow surface

rotations and cooler atmospheres – both of which are advantageous for radial velocity measurements. In addition, we plan to use our tool to investigate the impact of stellar model physics on an ensemble of planet-host stars.

Acknowledgements. J.K. acknowledges funding through the Max-Planck Partnership group – SEISMIC Max-Planck-Institut für Astrophysik (MPA) – Germany and Kyambogo University (KyU) – Uganda. B.N. acknowledges post-doctoral funding from the “Branco Weiss fellowship – Society in Science” through the SEISMIC stellar interior physics group. T.L.C. is supported by Fundação para a Ciência e a Tecnologia (FCT) in the form of a work contract (CEECIND/00476/2018). V.A. and N.C.S. acknowledge support from Fundação para a Ciência e a Tecnologia through national funds and by FEDER through COMPETE2020 – Programa Operacional Competitividade e Internacionalização by these grants: UIDB/04434/2020; UIDP/04434/2020; 2022.06962.PTDC. Funded/Co-funded by the European Union (ERC, FIERCE, 101052347). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. J.K. and B.N. also acknowledge funding from the UNESCO-TWAS programme, “Seed Grant for African Principal Investigators” financed by the German Ministry of Education and Research (BMBF). J.K. also thanks Andreas W.N., Pedro A.C.C., Thibault B. at IA-CAUP, and Earl P.B. at Yale University, for the useful discussions regarding machine learning. We thank the referee for the helpful and constructive remarks. Software: Data analysis in this manuscript was carried out using the Python 3.10.12 libraries; scikit-learn 1.5.1 (Pedregosa et al. 2011), pandas 2.2.2 (McKinney et al. 2010), NumPy 1.26.4 (Van Der Walt et al. 2011), and SciPy 1.14.1 (Virtanen et al. 2020).

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*, 2623
- Anderson, T. W., & Darling, D. A. 1954, *J. Am. Stat. Assoc.*, **49**, 765
- Andrae, R., Fouesneau, M., Creevey, O., et al. 2018, *A&A*, **616**, A8
- Angulo, C., Arnould, M., Rayet, M., et al. 1999, *Nucl. Phys. A*, **656**, 3
- Angus, R., Aigrain, S., Foreman-Mackey, D., & McQuillan, A. 2015, *MNRAS*, **450**, 1787
- Appourchoux, T., Antia, H., Ball, W., et al. 2015, *A&A*, **582**, A25
- Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, *ARA&A*, **47**, 481
- Barnes, S. A., Weingrill, J., Fritzewski, D., Strassmeier, K. G., & Platais, I. 2016, *ApJ*, **823**, 16
- Baron, D. 2019, ArXiv e-print [arXiv:1904.07248]
- Bazot, M., Bourguignon, S., & Christensen-Dalsgaard, J. 2012, *MNRAS*, **427**, 1847
- Bellinger, E. P., Angelou, G. C., Hekker, S., et al. 2016, *ApJ*, **830**, 31
- Berger, T. A., Huber, D., Gaidos, E., van Saders, J. L., & Weiss, L. M. 2020, *AJ*, **160**, 108
- Boyd, S. P., & Vandenberghe, L. 2004, *Convex optimization* (Cambridge University Press)
- Breiman, L. 1996, *Mach. Learn.*, **24**, 49
- Breiman, L. 2001, *Mach. Learn.*, **45**, 5
- Breiman, L. 2017, *Classification and regression trees* (Routledge)

- Buchhave, L. A., & Latham, D. W. 2015, *ApJ*, 808, 187
- Campante, T. L., Barclay, T., Swift, J. J., et al. 2015, *ApJ*, 799, 170
- Campante, T. L., Veras, D., North, T. S., et al. 2017, *MNRAS*, 469, 1360
- Casagrande, L., Portinari, L., Glass, I., et al. 2014, *MNRAS*, 439, 2060
- Chaplin, W., Basu, S., Huber, D., et al. 2013, *ApJS*, 210, 1
- Chen, T., & Guestrin, C. 2016, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785
- Chen, D.-C., Xie, J.-W., Zhou, J.-L., et al. 2023, *Proc. Nat. Acad. Sci.*, 120, e2304179120
- Claytor, Z. R., van Saders, J. L., Santos, A. R., et al. 2020, *ApJ*, 888, 43
- Coelho, H., Chaplin, W., Basu, S., et al. 2015, *MNRAS*, 451, 3011
- Cox, J. P., & Giuli, R. T. 1968, *Principles of Stellar Structure: Physical Principles* (Gordon and Breach), 1
- Cyburt, R. H., Fields, B. D., & Olive, K. A. 2003, *Phys. Lett. B*, 567, 227
- Da Silva, R., de Mello, G. P., Milone, A., et al. 2012, *A&A*, 542, A84
- Dawson, R. I., & Johnson, J. A. 2018, *ARA&A*, 56, 175
- Deal, M., Cunha, M., Keszthelyi, Z., Perraut, K., & Holdsworth, D. L. 2021, *A&A*, 650, A125
- Des Étangs, A. L., & Lissauer, J. J. 2022, *New Astron. Rev.*, 94, 101641
- Espinoza-Rojas, F., Chanamé, J., Jofré, P., & Casamiquela, L. 2021, *ApJ*, 920, 94
- Feltzing, S., Howes, L. M., McMillan, P. J., & Stokutė, E. 2016, *MNRAS*, 465, L109
- Ferguson, J. W., Alexander, D. R., Allard, F., et al. 2005, *ApJ*, 623, 585
- Gai, N., Basu, S., Chaplin, W. J., & Elsworth, Y. 2011, *ApJ*, 730, 63
- Gaia Collaboration (Bailer-Jones, C., et al.) 2023, *A&A*, 674, A41
- Green, G. 2018, *J. Open Source Soft.*, 3, 695
- Gruberbauer, M., Guenther, D. B., & Kallinger, T. 2012, *ApJ*, 749, 109
- Hamer, J. H., & Schlaufman, K. C. 2019, *AJ*, 158, 190
- Hara, N. C., & Ford, E. B. 2023, *Annual Review of Statistics and Its Application*, 10, 623
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. 2009, *The elements of statistical learning: data mining, inference, and prediction* (Springer), 2
- Herwig, F. 2000, *A&A*, 360, 952
- Huber, D., Chaplin, W. J., Christensen-Dalsgaard, J., et al. 2013, *ApJ*, 767, 127
- Iglesias, C. A., & Rogers, F. J. 1996, *ApJ*, 464, 943
- Imbriani, G., Costantini, H., Formicola, A., et al. 2005, *Eur. Phys. J. A-Hadrons Nucl.*, 25, 455
- Kolmogorov, A. 1933, *Giorn. Ist. Ital. Attuari*, 4, 1
- Krishna Swamy, K. 1966, *ApJ*, 145, 174
- Lawson, C. L., & Hanson, R. 1995, *Solving least squares problems* (SIAM)
- Lindgren, L. 2018, *Gaia Data Processing and Analysis Consortium, Tech. Rep., GAIA-C3-TN-LU-LL-124*
- Liu, W. M., & Chaboyer, B. 2000, *ApJ*, 544, 818
- Lund, M. N., & Reese, D. R. 2017, ArXiv e-print [arXiv:1711.01896]
- Lund, M. N., Aguirre, V. S., Davies, G. R., et al. 2017, *ApJ*, 835, 172
- Mamajek, E. E., & Hillenbrand, L. A. 2008, *ApJ*, 687, 1264
- Mamajek, E., Torres, G., Prsa, A., et al. 2015, ArXiv e-print [arXiv:1510.06262]
- Mayor, M., & Queloz, D. 1995, *Nature*, 378, 355
- McKinney, W., van der Walt, S., & Millman, J. 2010, *Proceedings of the 9th Python in Science Conference*
- Meibom, S., Barnes, S. A., Platias, I., et al. 2015, *Nature*, 517, 589
- Mena, E. D., Moya, A., Adibekyan, V., et al. 2019, *A&A*, 624, A78
- Metcalfe, T. S., Creevey, O. L., & Christensen-Dalsgaard, J. 2009, *ApJ*, 699, 373
- Miyazaki, S., & Masuda, K. 2023, *AJ*, 166, 209
- Moedas, N., Nsamba, B., & Clara, M. T. 2020, *Dynamics of the Sun and Stars: Honoring the Life and Work of Michael J. Thompson* (Thompson: Springer), 259
- Moedas, N., Bossini, D., Deal, M., & Cunha, M. S. 2024, *A&A*, 684, A113
- Mortier, A., Santos, N., Sousa, S., et al. 2013, *A&A*, 558, A106
- Moya, A., & López-Sastre, R. J. 2022, *A&A*, 663, A112
- Moya, A., Zuccarino, F., Chaplin, W. J., & Davies, G. R. 2018, *ApJS*, 237, 21
- Moya, A., Sarro, L. M., Delgado-Mena, E., et al. 2022, *A&A*, 660, A15
- Nissen, P. E. 2015, *A&A*, 579, A52
- Nissen, P. 2016, *A&A*, 593, A65
- Noyes, R., Hartmann, L., Baliunas, S., Duncan, D., & Vaughan, A. 1984, *ApJ*, 279, 763
- Nsamba, B., Campante, T., Monteiro, M., et al. 2018, *MNRAS*, 477, 5052
- Nsamba, B., Campante, T. L., Monteiro, M. J., Cunha, M. S., & Sousa, S. G. 2019, *Front. Astron. Space Sci.*, 6, 25
- Nsamba, B., Weiss, A., & Kamulali, J. 2025, *MNRAS*, 536, 2558
- Paxton, B., Schwab, J., Bauer, E. B., et al. 2018, *ApJS*, 234, 34
- Paxton, B., Smolec, R., Schwab, J., et al. 2019, *ApJS*, 243, 10
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *J. Mach. Learn. Res.*, 12, 2825
- Perryman, M. 2018, *The exoplanet handbook* (Cambridge University Press)
- Pijpers, F. P. 2003, *A&A*, 400, 241
- Pinsonneault, M. H., An, D., Molenda-Žakowicz, J., et al. 2012, *ApJS*, 199, 30
- Pinsonneault, M. H., Elsworth, Y., Epstein, C., et al. 2014, *ApJS*, 215, 19
- Planck, M. 1900, *Entropie*, 144, 164
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulín, A. 2018, *Advances in neural information processing systems*, 31
- Pulone, L., & Scaramella, R. 1997, *Neural Nets WIRN VIETRI-96: Proceedings of the 8th Italian Workshop on Neural Nets, Vietri sul Mare, Salerno, Italy, 23–25 May 1996* (Springer), 231
- Ramírez, I., Melendez, J., & Asplund, M. 2009, *A&A*, 508, L17
- Rasio, F. A., & Ford, E. B. 1996, *Science*, 274, 954
- Rendle, B. M., Buldgen, G., Miglio, A., et al. 2019, *MNRAS*, 484, 771
- Rogers, F., & Nayfonov, A. 2002, *ApJ*, 576, 1064
- Sandage, A., Lubin, L. M., & Vandenberg, D. A. 2003, *PASP*, 115, 1187
- Santos, N., Sousa, S., Mortier, A., et al. 2013, *A&A*, 556, A150
- Schlafly, E. F., & Finkbeiner, D. P. 2011, *ApJ*, 737, 103
- Scholz, F. W., & Stephens, M. A. 1987, *J. Am. Stat. Assoc.*, 82, 918
- Seager, S., & Mallen-Ornelas, G. 2003, *ApJ*, 585, 1038
- Serenelli, A., Johnson, J., Huber, D., et al. 2017, *ApJS*, 233, 23
- Silva Aguirre, V., Ballot, J., Serenelli, A., & Weiss, A. 2011, *A&A*, 529, A63
- Silva Aguirre, V., Christensen-Dalsgaard, J., Chaplin, W., et al. 2013, *ApJ*, 769, 2
- Silva Aguirre, V., Davies, G., Basu, S., et al. 2015, *MNRAS*, 452, 2127
- Silva Aguirre, V., Lund, M. N., Antia, H., et al. 2017, *ApJ*, 835, 173
- Smirnov, N. 1939, *NS*, 6, 3
- Soderblom, D. R. 2010, *ARA&A*, 48, 581
- Soderblom, D. 2015, *Nature*, 517, 557
- Student, 1908, *Biometrika*, 6, 1
- Thoul, A. A., Bahcall, J. N., & Loeb, A. 1994, *ApJ*, 421, 828
- Titarenko, A., Recio-Blanco, A., de Laverny, P., Hayden, M., & Guiglion, G. 2019, *A&A*, 622, A59
- Torres, G., Winn, J. N., & Holman, M. J. 2008, *ApJ*, 677, 1324
- Torres, G., Andersen, J., & Giménez, A. 2010, *A&ARv*, 18, 67
- Van Der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, *Comput. Sci. Eng.*, 13, 22
- Vandenberg, D. A., Brogaard, K., Leaman, R., & Casagrande, L. 2013, *ApJ*, 775, 134
- Verma, K., Hanasoge, S., Bhattacharya, J., Antia, H., & Krishnamurthi, G. 2016, *MNRAS*, 461, 4206
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nat. Methods*, 17, 261
- Wehenkel, L., Ernst, D., & Geurts, P. 2006, *Robust Methods for Power System State Estimation and Load Forecasting*
- Wolpert, D. H. 1992, *Neural Net.*, 5, 241

Appendix A: Additional tests

A.1. Comparison with other LEGACY results

We conducted additional comparisons between the estimates from MAISTEP and those derived from other pipelines in LEGACY analyses. Here, we report only the results from the two pipelines that show the best and worst agreement in age compared to MAISTEP. The Bayesian STellar Algorithm (BASTA; [Silva Aguirre et al. 2015](#)) and GOEttingen (GOE; [Appourchaux et al. 2015](#)) pipelines yield consistent estimates of radius and mass in comparison to MAISTEP, with bias (and scatter) values of 0.7% (2%) and 0.6% (3%), 0.4% (4%), and -1% (6%), as illustrated in the top and middle panels of Fig. A.1 and Fig. A.2, respectively. In age, we find a slight disagreement with bias (and scatter) values of 6% (23%) and 20% (32%) as shown in the bottom panels of Fig. A.1 and Fig. A.2, respectively. We note that the primary difference in the results obtained using BASTA and GOE stem from the sets of constraints employed during optimisation: frequency ratios in BASTA and individual oscillation frequencies for GOE.

A.2. Comparison of results from two sets of constraints

We performed an additional test by training our algorithms with T_{eff} , $[\text{Fe}/\text{H}]$ and $\log g$, instead of T_{eff} , $[\text{Fe}/\text{H}]$, and L . Applying the model to the LEGACY sample, we obtain consistent results as shown in Fig. A.3. The resulting bias (and scatter) is -1% (2%) in radius, -1% (2%) in mass, and 1% (5%) in age.

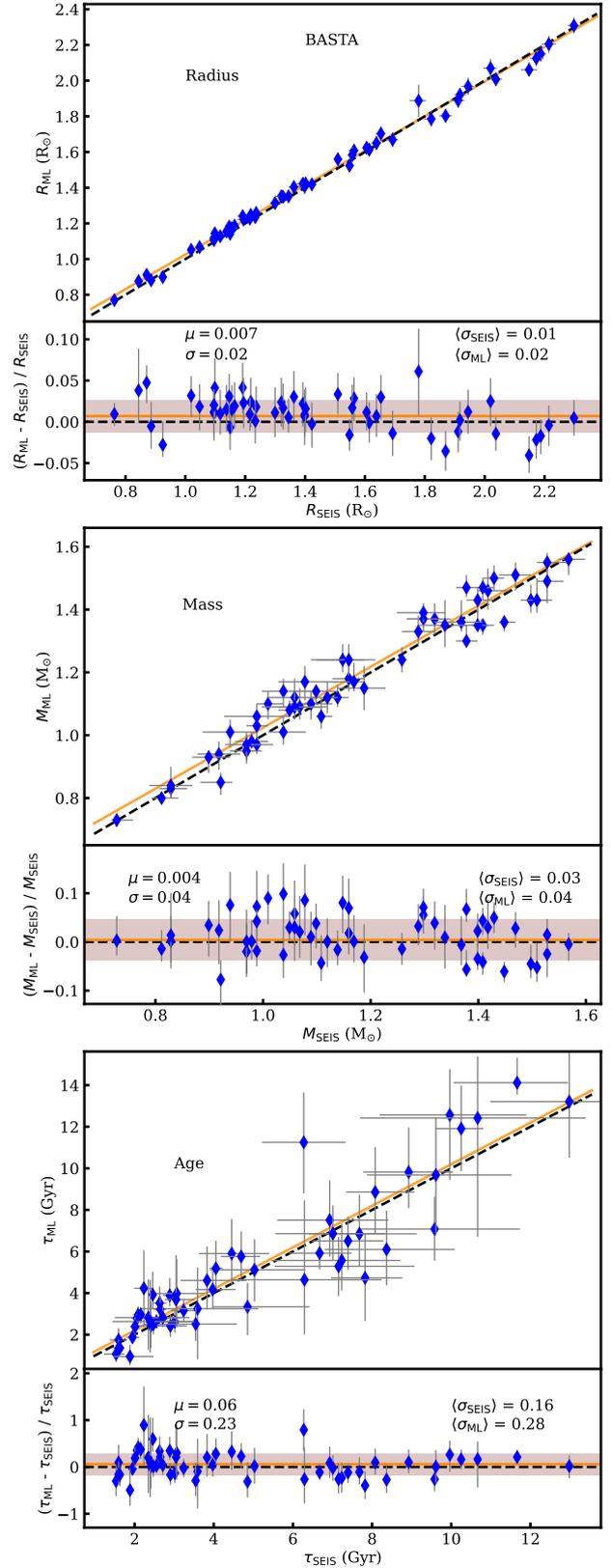


Fig. A.1. Same as Fig. 6, but for results from the BASTA pipeline.

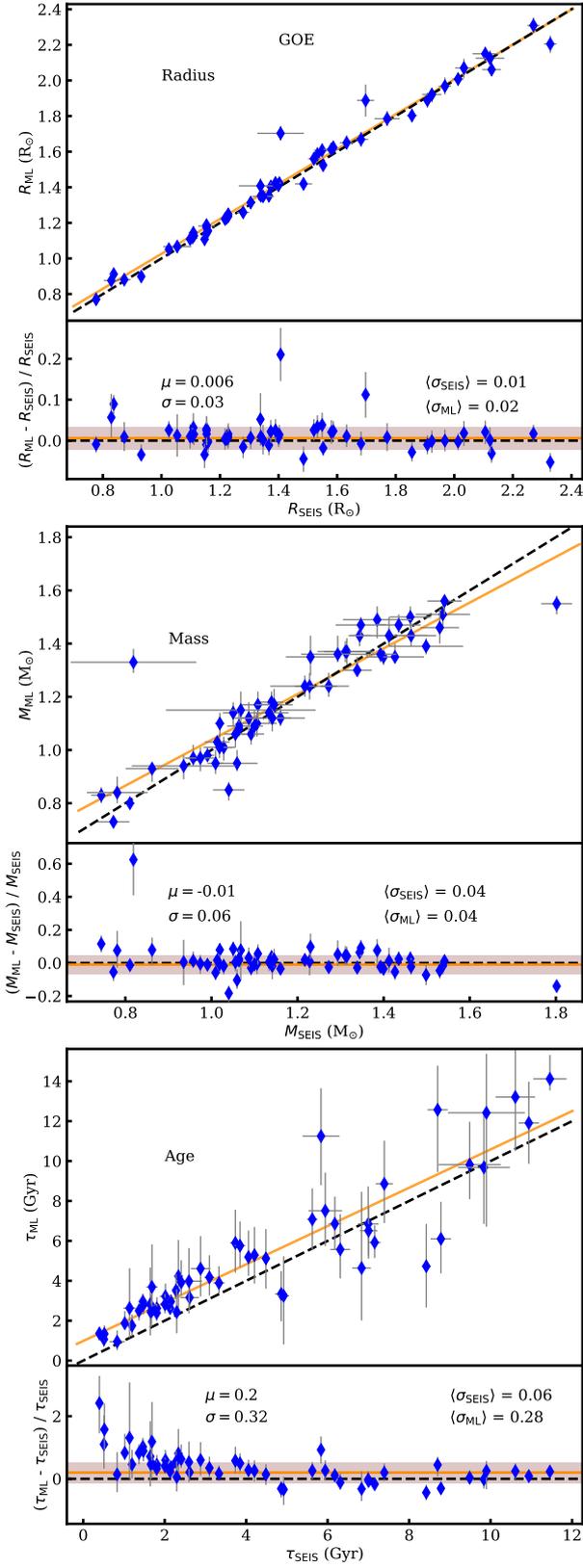


Fig. A.2. Same as Fig. 6, but for results from the GOE pipeline.

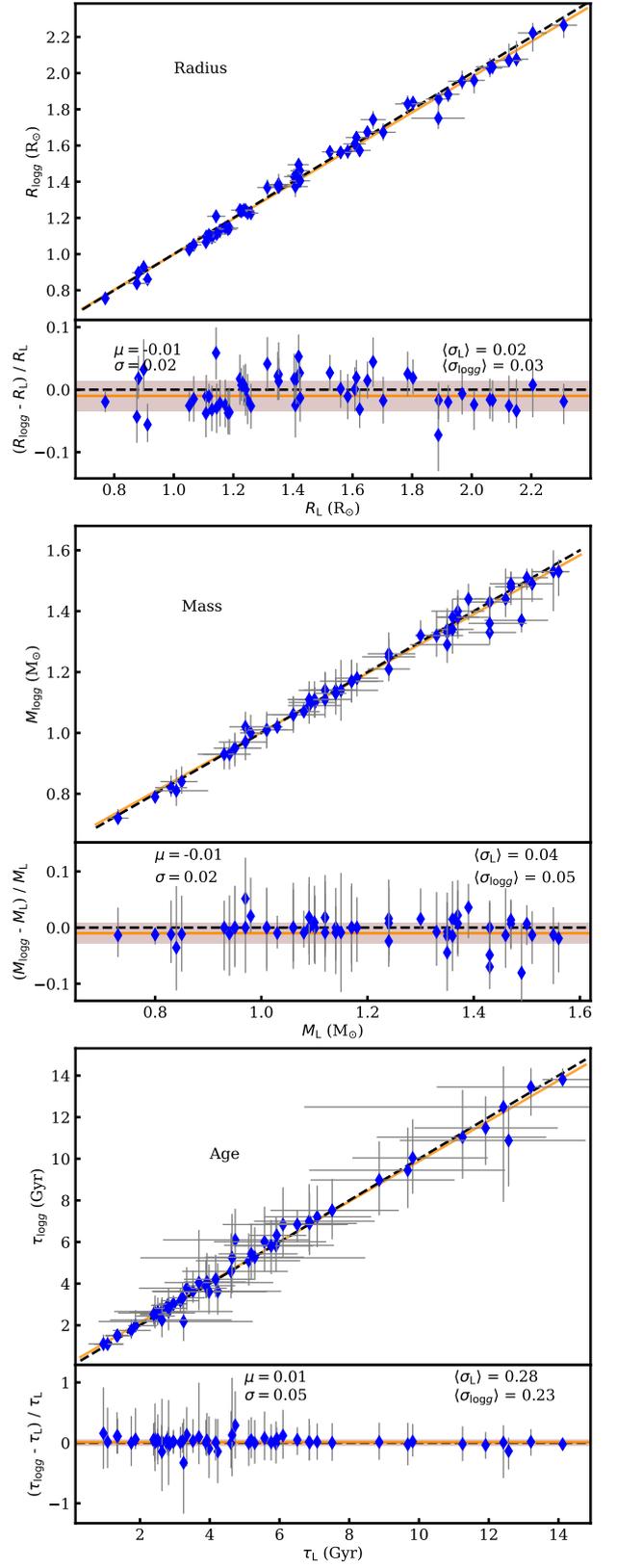


Fig. A.3. Same as Fig. 6, but for the combinations of T_{eff} , $[\text{Fe}/\text{H}]$, and $\log g$ versus T_{eff} , $[\text{Fe}/\text{H}]$, and L as constraints.