Metallicity determination in galactic open star clusters by exploring multiwavelength surveys

Eduardo Machado Pereira[1,2], Simone Daflon[1], Vinicius Placco[2]

eduardopereira@on.br

[1]Observatório Nacional, [2]NSF NOIRLab

CONTEXT

Usually obtained by spectroscopy, effective temperature (Teff), surface gravity (log g), and metallicity ([Fe/H]) are among the main parameters of interest in the context of stellar and galactic astrophysics. Photometry, however, is considerably less expensive and can be exploited for selecting interesting objects for a variety of scientific purposes. We explore machine learning to build photometry-based models to estimate these parameters for members of open clusters (OCs) in the footprint of the Javalambre-Photometric Local Universe Survey (J-PLUS). By taking advantage of J-PLUS 12-filter system, and after a comprehensive feature engineering step, our models show competitive results for all parameters. Moreover, our main goal is to provide [Fe/H] for these clusters, particularly aiming at enabling subsequent cluster and galactic studies on e.g. membership analysis, multiple stellar populations, stream formation and member evaporation.

DATA

Table 1: Cleaning steps to build sample used in models development. Cross matches were made from top to bottom. Photometric data in this work is composed of J-PLUS DR3 (López-Sanjuan et al., 2024), Gaia DR3 (Gaia Collaboration, 2022) and CatWISE (Marocco et al., 2021). Spectroscopic data was taken from LAMOST (Cui et al., 2012; DR8). Final sample underwent through additional cuts for quality selection, resulting in **88 421 stars**, split into train (70%), validation (20%) and test (10%) subsamples.

Catalog	Sample size	Fraction	Notes
J-PLUS DR3	5 114 494	1.00000	-
CatWISE	5 027 068	0.98291	For $W1$ and $W2$
Gaia DR3	5 026 927	0.98288	For G, B_P, R_P
Gaia EDR3	5 004 860	0.97856	For distances
Gaia AP	5 004 738	0.97854	For extinctions
LAMOST DR8	529 565	0.10354	For spectroscopic targets

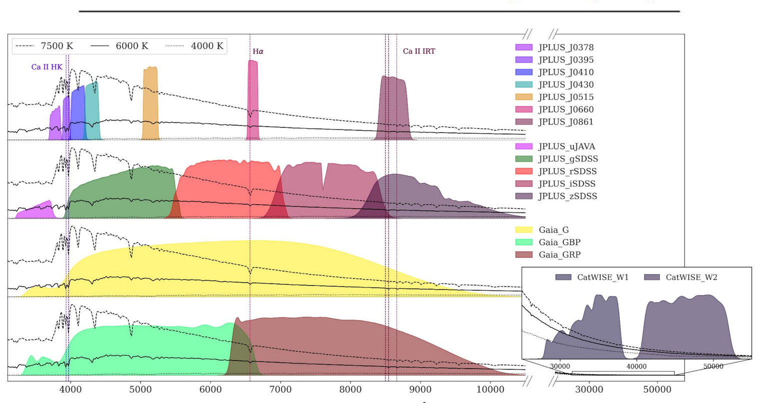


Fig. 1: Transmission curves of the filters of the three catalogs used in this work, plotted against three examples of spectral energy distributions of log g = 4.5, [Fe/H] = 0, [α/Fe] = 0. J-PLUS and Gaia cover the optical portion of the electromagnetic spectrum, while CatWISE gathers infrared information. Note that J-PLUS relies on seven narrow-/intermediate-band filters strategically located.

Fig. 2: Distributions of atmospheric parameters as taken from LAMOST for the final sample to be used in models development.

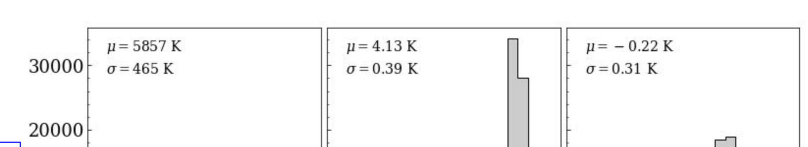
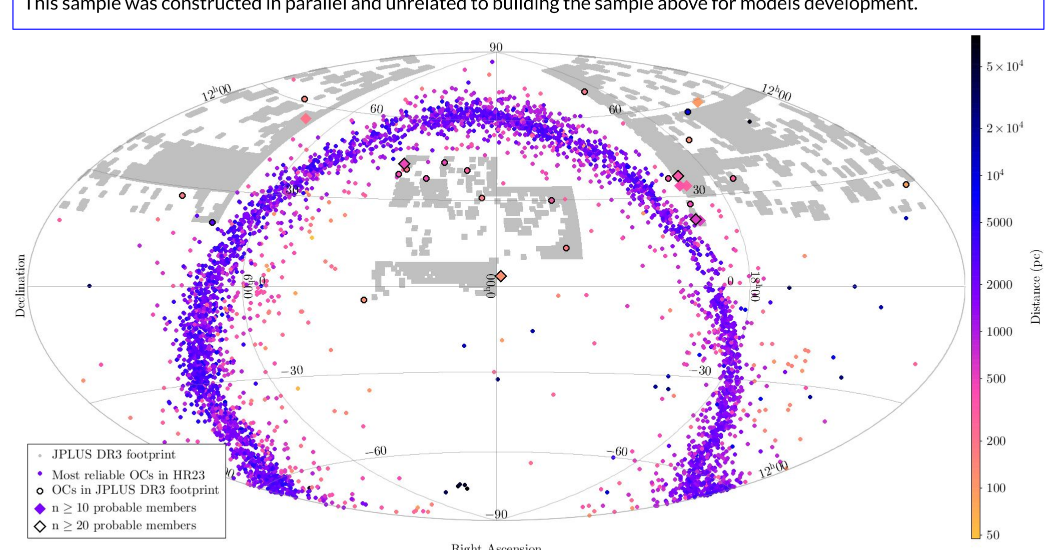


Fig. 3: Cross match for J-PLUS DR3 stars (gray regions) and the OCs sample of Hunt & Reffert (HR23, 2023; distance-colored points), along with the ones found in the J-PLUS DR3 footprint (distance-colored points with black outlines). Clusters with at least 10 probable members are shown as diamonds, while those with at least 20 members are diamonds with black outlines. The color bar indicates distance (in parsecs) of the OCs.
This sample was constructed in parallel and unrelated to building the sample above for models development.



METHOD

shap-hypetune
LightGBM

Machine learning:
map correlations in data to build models capable of predicting the targets, provided the features

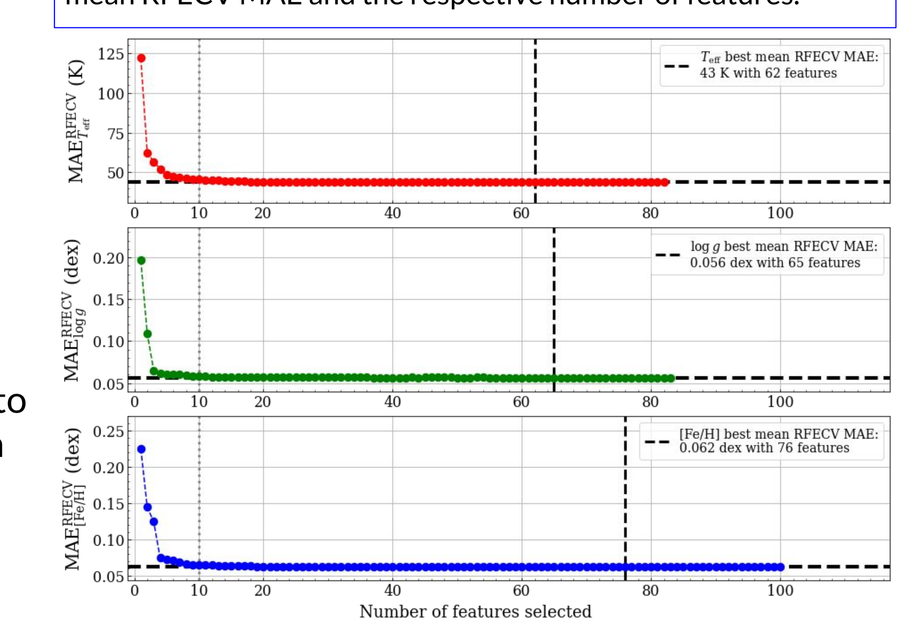
- Photometric data → 153 **features**
  - 136 colors (combination of 17 magnitudes)
  - 17 absolute magnitudes (12 J-PLUS + 2 CatWISE + 3 Gaia)
- Spectroscopic data → 3 **targets** (LAMOST)
  - Teff
  - log g
  - [Fe/H]

LightGBM + shap-hypetune (see top right QR codes): gradient boosting technique with optimization framework to select best features through recursive feature elimination (RFE) and best hyperparameters through bayesian search

Number of features selected from shap-hypetune RFE
Teff: 82 log g: 83 [Fe/H]: 100

After further RFE with cross validation (RFECV)
Teff: 62 log g: 65 [Fe/H]: 76

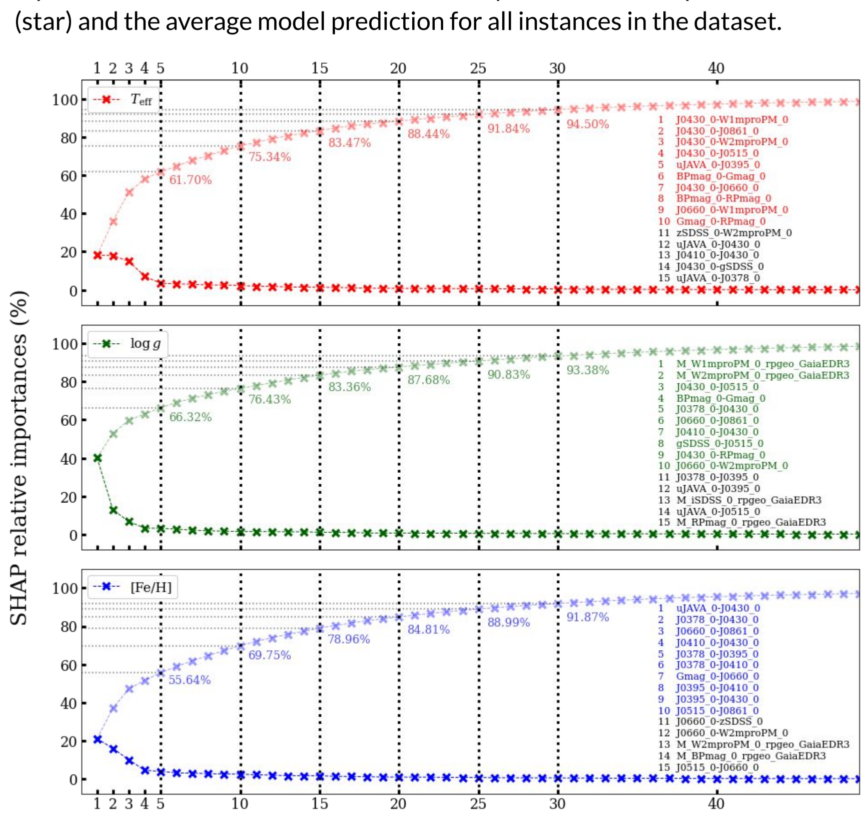
Fig. 4: RFECV mean absolute errors (MAE) as a function of the number of optimal features selected by shap-hypetune optimization. The black dashed lines in each panel mark the best mean RFECV MAE and the respective number of features.



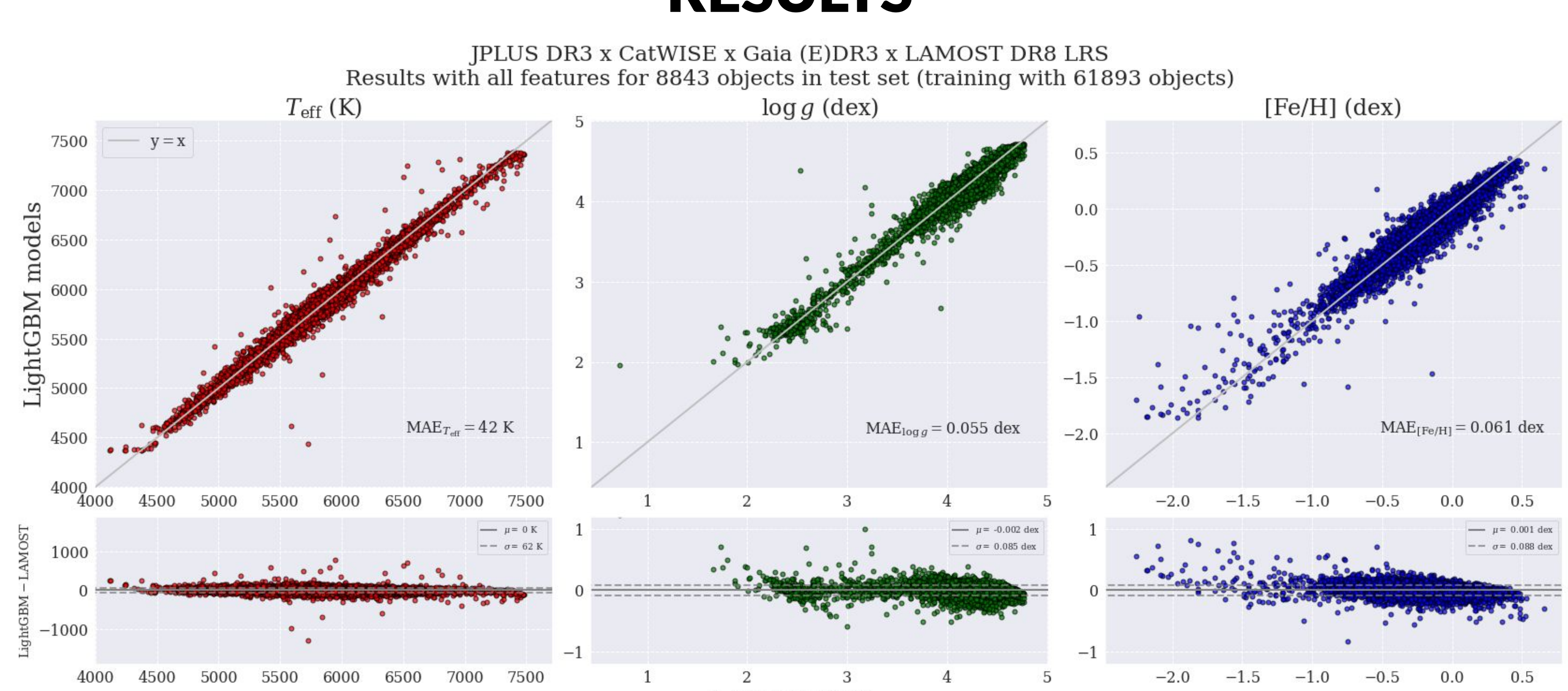
Starting with all 153 features, shap-hypetune first selected best hyperparameters along with an optimal subset of features, which was further refined by a cross validation step.

RESULTS

SHapley Additive exPlanations (SHAP) importances (Lundberg & Lee, 2017): the sum of the individual importances of each feature for a given model is equal to the difference between the model prediction for a specific instance (star) and the average model prediction for all instances in the dataset.



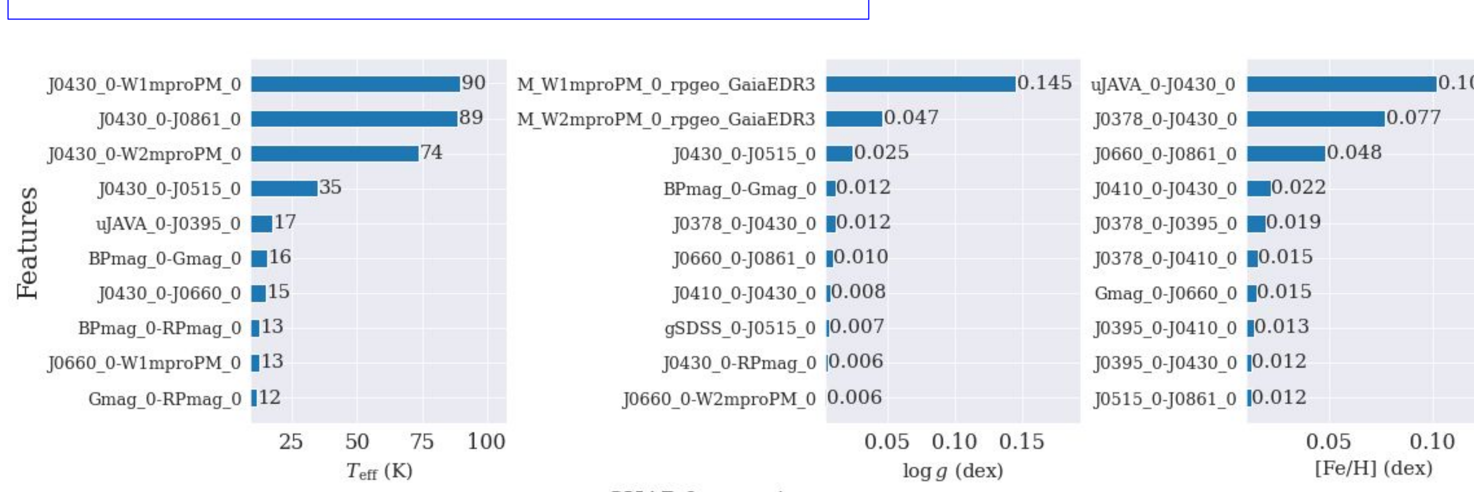
Relative importances as a funtcion of their respective rankings show that ~70% of the models' variabilities can be explained by using the top 10 most important features, as indicated above (Fig. 5) and quantitatively exhibited below (Fig. 6).



Although predictions show visible scatter, the [Fe/H] model is well behaved in the usual regime where OCs are found, around [Fe/H] ~ 0.

Fig. 7: Predictions for Teff, log g and [Fe/H] using the blind test set of our sample show MAEs of 42 K, 0.055 dex and 0.061 dex, respectively. Residuals attest essentially zero average difference in all cases and only weak trends for extreme values of log g and [Fe/H], where data is scarce.

Fig. 8 (below): MAEs for training models using selections of most important features only, instead of all selected from RFECV (see Fig. 4). MAEs rapidly approach near-optimal values before reaching optimal number of features.
In this work, all features were used (see Fig. 7) since all magnitudes appear in top 15 most important features (which explain ~80% of models' variabilities), such that no features could be dropped, and no benefit was noticed in terms of computing time by using less features.



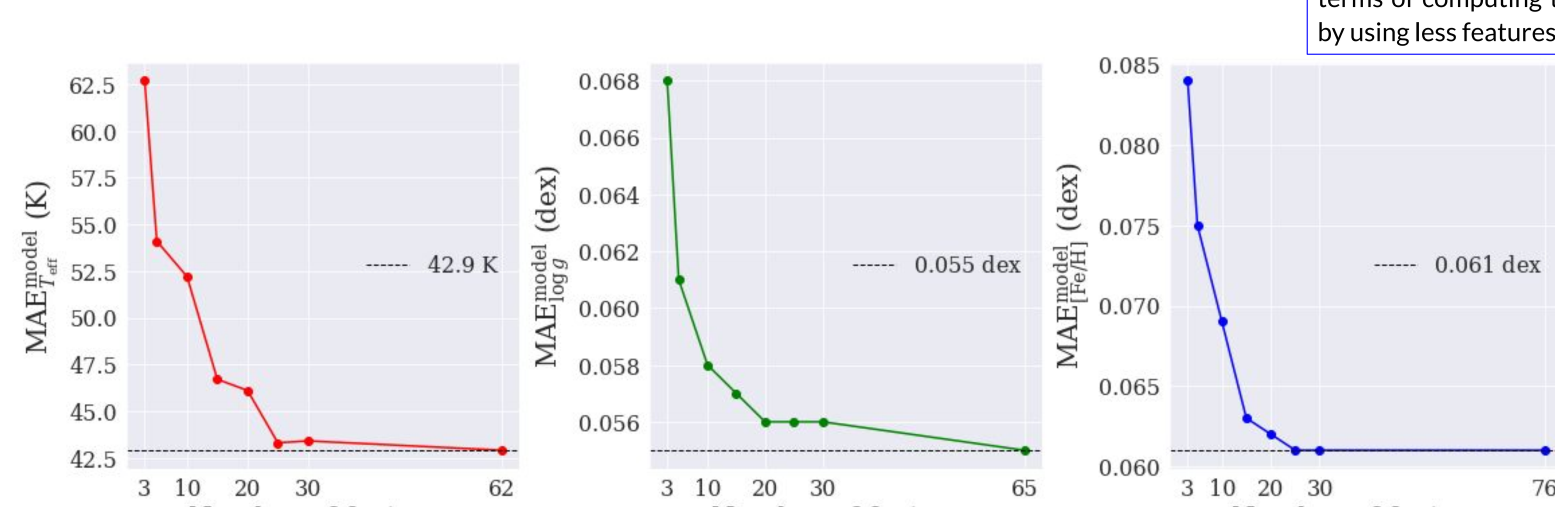
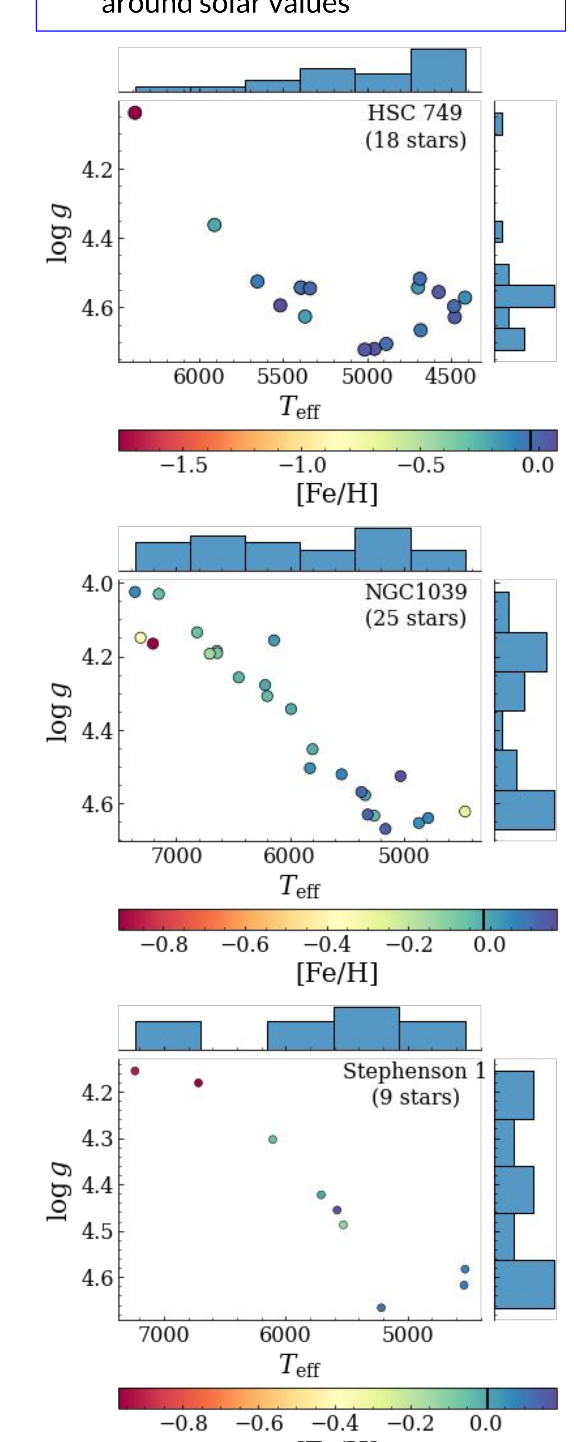


Fig. 9: **Preliminary results:**
- ➔ Main sequence reasonably recovered
- ➔ Scattered [Fe/H] with clear outliers
- ➔ Median [Fe/H] (marked in colorbar) around solar values



NEXT STEPS

- Validate models via predictions for other datasets (e.g. APOGEE, SEGUE, GALAH)
- Estimate parameters for members of 6 clusters with available photometry and at least 10 members
- Further analyze preliminary results
- MCMC for uncertainties
- Isochrone fitting for clusters
  - estimate parameters for clusters (e.g. distance, extinction, age, isochronal [Fe/H])
- Compare with J-PLUS only photometry
- Explore photometric cluster membership

References

Bayo et al. (2008). VOSA: virtual observatory SED analyzer. An application to the Collinder 69 open cluster. A&A, 492, 277
Bailer-Jones et al. (2021). Estimating Distances from Parallaxes. V. Geometric and Photogeometric Distances to 1.47 Billion Stars in Gaia Early Data Release 3. AJ, 161, 147
Coelho (2014). A new library of theoretical stellar spectra with scaled-solar and α-enhanced mixtures. MNRAS, 440, 1027
Cenarro et al. (2019). J-PLUS: The Javalambre Photometric Local Universe Survey. A&A 622, A176
Cui et al. (2012). The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). Res. Astron. Astrophys. 12, 1197.
Friedman (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29, 1189-1232.
Gaia Collaboration, Prusti, de Bruijne et al. (2016b). The Gaia mission. A&A 595, pp. A1.
Gaia Collaboration, Vallenari et al. (2022). Gaia DR3: data release content and main properties. A&A 649, A1 (2021).
Hunt & Reffert (2023). Improving the open cluster census II. An all-sky cluster catalogue with Gaia DR3. A&A 673, A114
López-Sanjuan et al. (2024). J-PLUS: Toward a homogeneous photometric calibration using Gaia BP/RP low-resolution spectra. A&A, 683, A29
Lundberg & Lee (2017). A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems, 30
Marocco et al. (2021). The CatWISE2020 Catalog. ApJS, 253, 8

Acknowledgements

E. M.-P. acknowledges support from CAPES (proc. 88887.605761/2021-00 and 88881.846754/2023-01). E. M.-P. acknowledges support from NSF NOIRLab with registration fee. The work of E. M.-P. and V.M.P. is supported by NSF NOIRLab, which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation. The work of E. M.-P. and S.D. is supported by Observatório Nacional / MCTI (Brazil).