Highlights of Spanish Astrophysics XI, Proceedings of the XV Scientific Meeting of the Spanish Astronomical Society held on September 4–9, 2022, in La Laguna, Spain. M. Manteiga, L. Bellot, P. Benavidez, A. de Lorenzo-Cáceres, M. A. Fuente, M. J. Martínez, M. Vázquez- Acosta, C. Dafonte (eds.), 2023

Parameters for >300 million *Gaia* stars: Bayesian inference vs. machine learning.

Anders, F.[1], Khalatyan, A.[2], Queiroz, A.[2,3], and Nepal, S.[2], and Chiappini, C.[2]

[1] Institut de Ciències del Cosmos (IEEC-UB), Dept. Física Quàntica i Astrofísica (FQA), Universitat de Barcelona, C Martí i Franqués, 1, 08028 Barcelona, Spain
[2] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
[3] Institut für Physik und Astronomie, Universität Potsdam, Haus 28 Karl-Liebknecht-Str. 24/25, D-14476 Golm, Germany

Abstract

The *Gaia* Data Release 3 (DR3), published in June 2022, delivers a diverse set of astrometric, photometric, and spectroscopic measurements for more than a billion stars. The wealth and complexity of the data makes traditional approaches for estimating stellar parameters for the full *Gaia* dataset almost prohibitive. We have explored different supervised learning methods for extracting basic stellar parameters as well as distances and line-of-sight extinctions, given spectro-photo-astrometric data (including also the new *Gaia* XP spectra). For training we use an enhanced high-quality dataset compiled from *Gaia* DR3 and ground-based spectroscopic survey data covering the whole sky and all Galactic components. We show that even with a simple neural-network architecture or tree-based algorithm (and in the absence of *Gaia* XP spectra), we succeed in predicting competitive results (compared to Bayesian isochrone fitting) down to faint magnitudes. We will present a new *Gaia* DR3 stellar-parameter catalogue obtained using the currently best-performing machine-learning algorithm for tabular data, `XGBoost`, in the near future.

1 Introduction

The *Gaia* mission [14] has triggered an enormous increase in astronomical data that is revolutionising not only stellar and Galactic science but also has implications for cosmology and fundamental physics. The latest *Gaia* data release, DR3 [17], for the first time contains also low-resolution spectra, taken with *Gaia*'s blue and red photometer (BP/RP), for 219 million sources. These so-called XP spectra [10] represent the biggest homogeneous spectroscopic dataset (albeit at very low resolution; $R \sim 25$) available to date. Efficient methods to extract information from this dataset are starting to appear (e.g. [28, 3]).

2 Bayesian inference of stellar parameters with `StarHorse`

In the past years, our group has been developing an efficient Bayesian isochrone-fitting code, `StarHorse` [27, 24, 25], to infer stellar parameters, distances, and extinctions, to be able to analyse the ever-growing stellar spectroscopic survey datasets.

In [1], we applied the code for the first time to the *Gaia* data without spectroscopic information. The first experiments proved to be so promising that a big-data analysis run (285 million *Gaia* DR2 stars with $G < 18$, cross-matched with the Pan-STARRS1, 2MASS, and WISE photometric catalogues) was carried out using the computing cluster of the Leibniz-Institut für Astrophysik Potsdam (AIP). Our results doubled the number of Gaia DR2 sources with astrophysical parameters, improved the accuracy of the geometric distances, and revealed the presence of the Galactic bar in the Gaia data in a direct and completely unexpected manner (see Fig. 1).

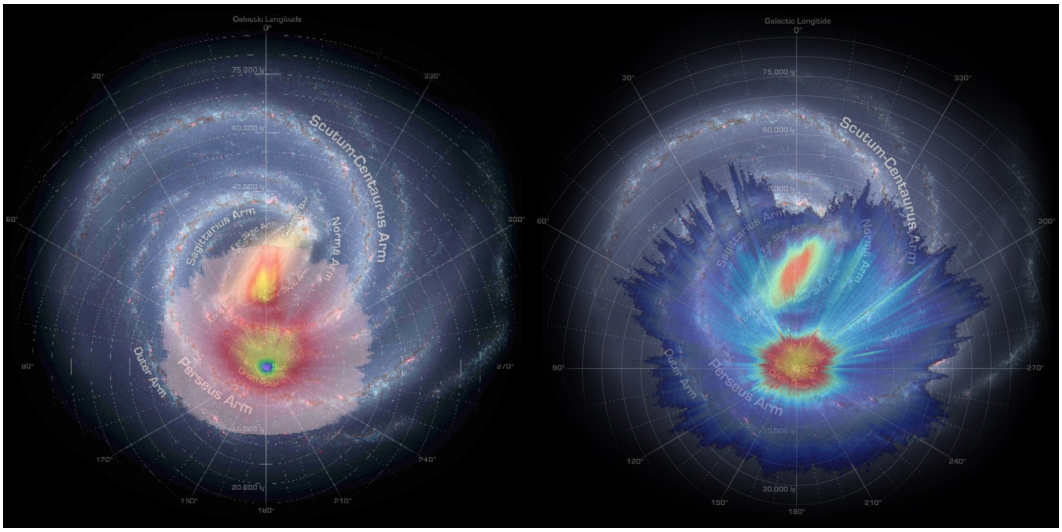

Figure 1: Comparison of the 5-year mission expectation for the Galactic coverage of the *Gaia* data before launch (left; [23]) with the results from `StarHorse` for *Gaia* DR2 ($G < 18$, 22 months of observation, right; [1]).

After the release of *Gaia*'s Early Data Release 3 in 2021 [16], and building on the success of the previous `StarHorse` catalogue, we ran our code on Gaia EDR3 data coupled with other large-area photometric surveys (now also including SkyMapper data). This resulted in an improved catalogue of 362 million stars, down to magnitude $G < 18.5$ (published in [2]). Thanks to the more precise EDR3 parallaxes and drastically reduced systematics [22], the Galactic density maps derived from it now probe a much greater volume than in [1], extending to regions beyond the Galactic bar and to Local Group galaxies, with a larger total number density.

The code was also sped up by using a less dense stellar model grid and a new computing cluster, which improved the CO_2 footprint of the project by factor ~ 6, but still consumed

around 1 month of computing time on a 1000-core cluster. At latest with the upcoming LSST surveys, this type of endeavour will therefore become unfeasible. In addition, `StarHorse` is unable to process spectroscopic measurements (such as the XP spectra) directly, which is an obvious drawback given that these data contain valuable information on the stellar parameters [13].

3 Predicting stellar parameters with a simple neural network

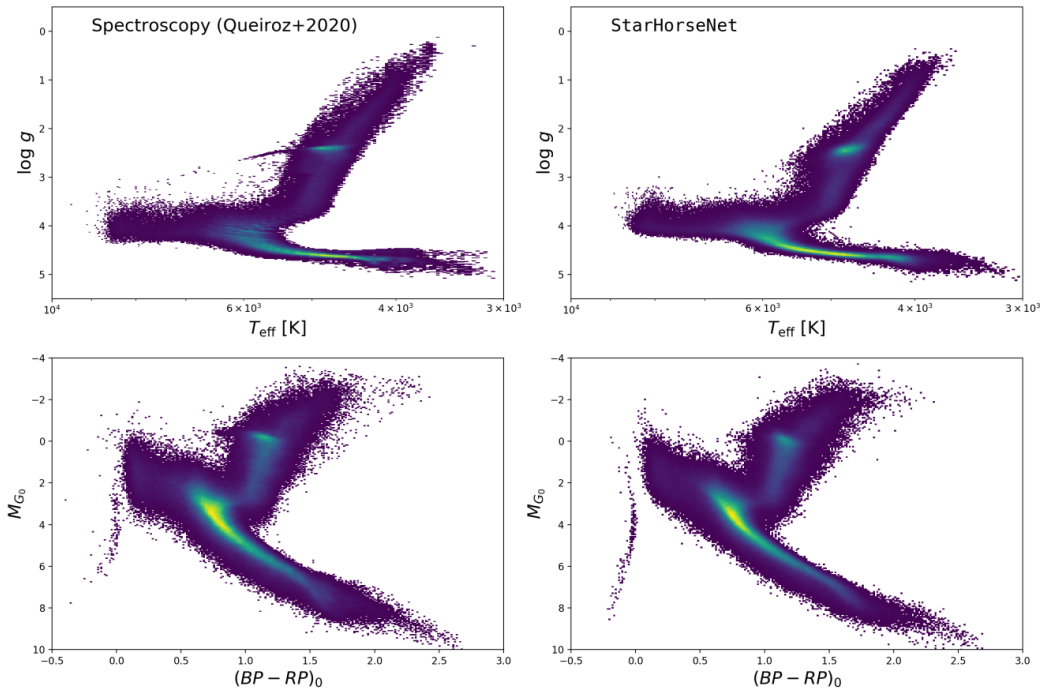

Figure 2: Results of the neural-network approach to stellar parameter estimation. Comparison between the colour-magnitude and Kiel Diagrams produced by the ANN model (right) with the spectroscopic test dataset [25]. Plot from the MSc thesis of Rany Assaad (University of Surrey, 2021).

In view of the computational drawbacks of traditional isochrone-based methods, in 2020 we started testing artificial neural networks (ANN) to predict stellar parameters for *Gaia* DR2 stars together with R. Assaad, an Erasmus+ MSc student visiting from the University of Surrey. As a training and test dataset, we used the `StarHorse` labels for various large-scale spectroscopic stellar surveys available at the time [25] (see Fig. 2).

This experiment was also surprisingly successful: with respect to [1] the ANN technique (even a very basic multi-layer perceptron architecture) doubled the number of stars (now more tha 300 million) considered to be of good quality. The full pipeline (including training

and prediction) could now be run on a 48-core machine within only 3 days, and it produced competitive posterior uncertainties (using a Monte-Carlo drop-out technique). The median uncertainties amounted to 16% in distance, 0.15 mag in V-band extinction A_V, 155 K in effective temperature $T_{\rm eff}$.

After cleaning the results of our Gaia DR2 run for poor input and output data, a sample of 373 million converged stars remained, for which we achieve a median uncertainty of 16% in distance, 0.15 mag in V-band extinction, and 155 K in effective temperature. Our experiment showed that even with a simple neural-network structure, one can succeed in predicting competitive results based on *Gaia* DR2 down to fainter magnitudes compared to classical isochrone or spectral-energy distribution fitting methods. Independent recent work [12] reached similar concluisons.

4 Improved results with tree-based algorithms

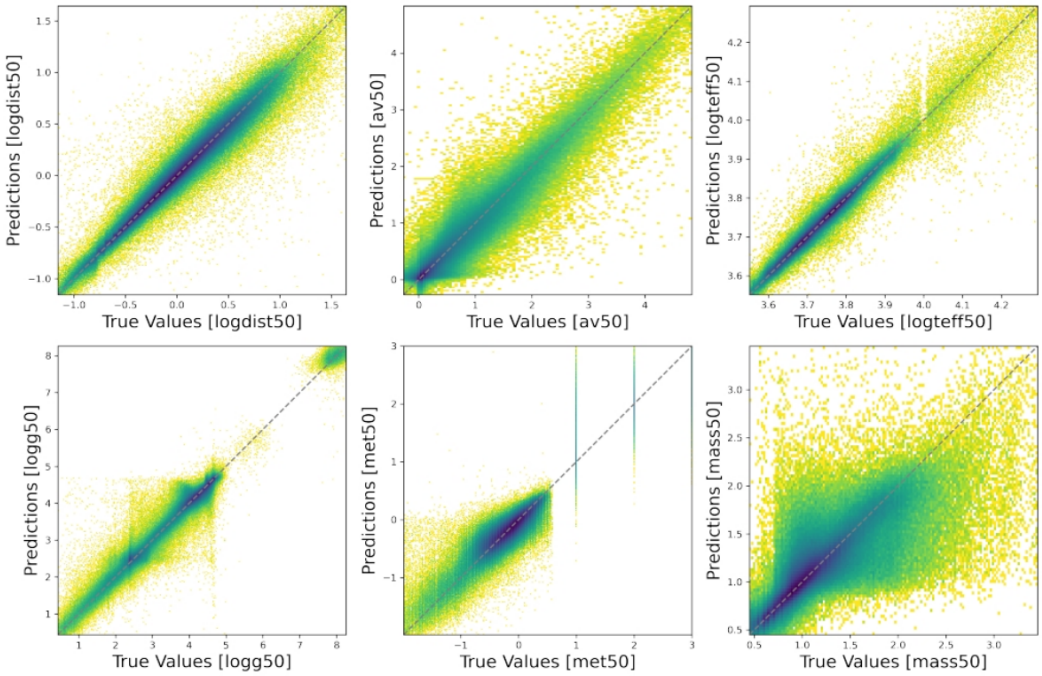

Figure 3: First results (June 2022) of the `XGBoost` approach to stellar parameter estimation for the case of stars with available *Gaia* DR3 XP spectra. One-to-one comparisons between the predicted stellar parameters and the spectroscopic test dataset [26] (labelled "true values").

The *Gaia* DR3 XP spectra are not delivered in the form of classical spectra, but come in the form of internally-calibrated spectral coefficients [6]. They can therefore be treated as tabular data and easily fed to a supervised learning technique. [5] benchmark-tested several

regression algorithms (including several neural network architectures) for tabular data and found that the most efficient and accurate technique was the tree-based algorithm Extreme Gradient-Boosted Trees (`XGBoost`, [7]). This technique is implemented in the `python` package `xgboost` that is widely used for classification tasks in astronomy (e.g. [4, 29, 20, 9]). Examples for the use of `XGBoost` for regression tasks in astronomy are sparser, but also start to appear, e.g. for photometric redshifts [8, 21], predicting the number of sunspots [11], or determining spectroscopic stellar ages [19].

For our first tests, we assembled a training set of 4 million *Gaia* DR3 stars with XP spectra that have also been observed by spectroscopic stellar surveys and have `StarHorse` stellar parameters [26]. As training columns we used the normalised XP coefficients, as well as *Gaia* astrometry and broad-band optical and infra-red photometry. The training set was complemented with spectroscopically observed white dwarfs from [18]. The predicted labels include distance, A_V, $T_{\rm eff}$, surface gravity $\log g$, metallicity [M/H], and stellar mass.

The first tests yielded the following precisions and accuracies for unseen test data: $\Delta \log d = -0.002 \pm 0.075, \Delta A_V = +0.001 \pm 0.145, \Delta \log T_{\rm eff} = 0.000 + -0.011$ (see Fig. 3). The computational cost is even lower than in the case of the simple neural network (Sect. 3), and robust to both the choice of hyperparameters and the scaling of the input data. For the first time, our group is also able to deliver acceptable stellar parameters for white dwarfs.

5 Conclusions

While traditional (and genuinely astrophysical) methods continue to be both useful and necessary to understand the new astronomical data, machine-learning techniques are needed to handle the sheer amounts of present and future data, and to lower the total CO_2 budget of astronomical research. Here we showed that both neural-network and tree-based algorithms, once sufficiently well trained, can be successfully employed to infer stellar parameters, distances, and extinctions in the absence of high-resolution spectroscopic data. In the near future, we will present a new `StarHorse`-like catalogue for *Gaia* DR3 stars based on `XGBoost` trained on high-resolution spectroscopic data.

Acknowledgments

This work was (partially) funded by the Spanish MICIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" by the "European Union" through grant RTI2018-095076-B-C21, and the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia 'María de Maeztu') through grant CEX2019-000918-M. FA acknowledges financial support from MICINN (Spain) through the Juan de la Cierva-Incorporcion programme under contract IJC2019-04862-I.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`http://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `http://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

References

[1] Anders, F., Khalatyan, A., Chiappini, C., et al. 2019, A&A, 628, A94

[2] Anders, F., Khalatyan, A., Queiroz, A. B. A., et al. 2022, A&A, 658, A91

[3] Andrae, R., Fouesneau, M., Sordo, R., et al. 2023, A&A, in press, arXiv:2206.06138

[4] Bethapudi, S. & Desai, S. 2018, Astronomy and Computing, 23, 15

[5] Borisov, V., Leemann, T., Seßler, K., et al. 2021, arXiv e-prints, arXiv:2110.01889

[6] Carrasco, J. M., Weiler, M., Jordi, C., et al. 2021, A&A, 652, A81

[7] Chen, T. & Guestrin, C. 2016, arXiv e-prints, arXiv:1603.02754

[8] Chong, K. & Yang, A. 2019, arXiv e-prints, arXiv:1901.07544

[9] Cunha, P. A. C. & Humphrey, A. 2022, A&A 666, A87

[10] De Angeli, F., Weiler, M., Montegriffo, P., et al. 2023, A&A, in press, arXiv:2206.06143

[11] Dang, Y., Chen, Z., Li, H., Shu, H. 2022, Applied Artificial Intelligence, 2022, 36, arXiv:2203.05757

[12] Fallows, C. P. & Sanders, J. S. 2022, MNRAS 516, 5521

[13] Fouesneau, M., Frémat, Y., Andrae, R., et al. 2023, A&A, in press, arXiv:2206.05992

[14] Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A 595, A1

[15] Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A 616, A1

[16] Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, A&A 649, A1

[17] Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2021, A&A, in press, arXiv:2208.00211

[18] Gentile Fusillo, N. P., Tremblay, P. E., Cukanovaite, E., et al. 2021, MNRAS, 508, 3877

[19] Hayden, M. R., Sharma, S., Bland-Hawthorn, J., et al. 2022, MNRAS 517, 5325

[20] Li, C., Zhang, Y., Cui, C., et al. 2021, MNRAS, 506, 1651

[21] Li, C., Zhang, Y., Cui, C., et al. 2022, MNRAS, 509, 2289

[22] Lindegren, L., Bastian, U., Biermann, M., et al. 2021, A&A 649, A4

[23] Luri, X., Palmer, M., Arenou, F., et al. 2014, A&A, 566, A119

[24] Queiroz, A. B. A., Anders, F., A., Santiago, B. X., et al. 2018, MNRAS 476, 2556

[25] Queiroz, A. B. A., Anders, F., A., Chiappini, C., et al. 2020, A&A, 638, A76

[26] Queiroz, A. B. A., Anders, F., A., Chiappini, C., et al. 2023, A&A, subm.

[27] Santiago, B. X.; Brauer, D. E., Anders, F., et al. 2016, A&A 585, A42

[28] Weiler, M., Carrasco, J. M., Fabricius, C., Jordi, C. 2022, A&A, accepted, arXiv:2211.06946

[29] Yi, Z., Chen, Z., Pan, J., et al. 2019, ApJ, 887, 241