# Chemical-abundance analysis of main-sequence stars in GALAH DR4 and the NASA exoplanet database

Author: Sara Drapkin Junyent, sdrapkju7@alumnes.ub.edu
*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.*

Advisors: Friedrich Anders, Chloé Padois; {fanders,chloe.padois}@fqa.ub.edu

**Abstract:** In this work we analyze the elemental abundances of a selected main-sequence stars data set from the GALAH DR4 survey by means of a dimensionality reduction technique (UMAP) and a clustering method (Agglomerative Clustering with Ward linkage). This will allow us, with some future work, to define robust and interpretable chemical-abundance groups that are more fine-grained than the usual thin vs. thick disk dichotomy. Using the subsample of known exoplanet host stars within the GALAH DR4 sample, we explore possible trends of planet occurrence rate with the abundance group, finding little statistical evidence.
**Keywords:** Exoplanets, main-sequence stars, elemental abundances, fundamental star parameters, data analysis, data normalization.
**SDGs:** Quality education.

## I. INTRODUCTION

The fundamental question of whether we are alone in the universe has been the catalyst for many scientific projects and research: finding life beyond our Solar System, or at least other planets, the so-called exoplanets. However, this search is not an easy task. Nowadays, exoplanets can be detected by many different approaches such as recognizing transits in front of stars, studying the gravitational microlensing effect or using astrometry to study the change of position and movement of stars [18].

Even though the first exoplanet orbiting a Sun-like star was not discovered until 1995 [2], in the last 30 years the number of detected exoplanets has increased and is currently at 5,811 [13], with up to 7,300 more candidates yet to be confirmed. But, among the ones already known to this date, only 70 are potentially habitable planets. This number is that small due to the characteristics of Earth-like exoplanets: quite close to their star, with a radius between 0.5 and 1.6 Earth radii, rocky composition and capable of having liquid water on its surface [19]. Detecting this type of exoplanets is tough, since they are not sufficiently massive or big to alter the movements and position of its star or to present a deep transit.

The initial aim of this project was to study the relationship between the detailed chemical composition of stars and their likelihood of hosting exoplanets, as an alternative method to be used when searching for exoplanets. For this reason, we selected a sample of main-sequence stars and used clustering methods to create groups of stars with similar chemical composition. Then, highlighting the stars known to be exoplanet host stars, we could check for patterns or correlations between the two ideas: chemical-abundance pattern of the stars and presence of exoplanets.

In Section 2 and 3 we present a description of the data used and its processing. Then, in Section 4 we discuss the obtained results regarding abundance dependencies, cluster characterization and exoplanet location in said clusters. Finally, Section 5 contains the conclusions drawn from this work.

## II. DATA

We used data from the fourth data release of the Galactic Archaeology with HERMES survey (GALAH DR4) [1][12]. This data release contains 1,085,520 spectra of 917,588 stars in the Milky Way, observed with the HERMES spectrograph at the Anglo-Australian Telescope between December 2013 and August 2023. The GALAH DR4 catalogue provides one-dimensional spectra measured with resolution $R \sim 28,000$ [1], stellar atmospheric parameters and up to 30 elemental abundances per star.

Since our ultimate goal is to find potentially habitable exoplanets, hence Sun-like host stars, we restricted the data corresponding to main-sequence stars. Therefore, we selected those stars with $\log(g) \geq 4$ and $5000\ \text{K} < T_{\text{eff}} < 6500\ \text{K}$.

In addition, following the recommended flag values, we only considered stars with parameters `flag_sp=0` and `sn_px_ccd3>30`, the major spectroscopic quality flag and the average signal-to-noise ratio per pixel of the third CCD sensor, respectively. Lastly, our aim was to work with those metallicities with parameters `flag_X_fe=0` for at least 90% of our sample.

We ended up with a sample of 190,509 main-sequence stars with good-quality abundances of 12 elements: Fe, O, Na, Mg, K, Ca, Sc, Cr, Mn, Ni, Cu, La.

### A. Cross-match with NASA Exoplanet Archive

The NASA Exoplanet Archive [13] is an astronomical exoplanet and stellar catalogue that gathers astronomical data and information on exoplanets and their host

stars, providing tools to work with it. Thanks to the fact that every star has a unique identifier given by Gaia [14], it was possible to cross-match the stars in our chosen sample with the Exoplanet Archive.

As a result, we came to a selection of 149 exoplanets orbiting 101 stars present in our sample from GALAH DR4. From there, the goal was to determine if there was some relationship between those stars and their elemental abundances.

## III. DATA PROCESSING

Before the use of most machine learning estimators, a preprocessing of the data set is commonly needed, as these algorithms are sensitive to the scale of the input variables. In our case, we standardized our sample data with *Robust Scaler* [15]. Usually, the standardization is done by removing the mean and scaling to unit variance, but then outliers can have a significant negative influence to the sample mean or deviation values.

However, algorithms such as *Robust Scaler* use the median and the interquartile range instead of the mean and standard deviation, which makes the algorithm less sensitive to the outliers, giving better results.

For these reasons, we used *Robust Scaler* to standardize our metallicity data, the 11 ratios [X/Fe], X being the different elements stated in the previous section.

### A. Dimensionality reduction

After some quality checks explicitly explained previously, we worked with metallicity data for 11 elements, given as ratios [X/Fe], a 11 dimensional data set. Therefore, due to the complexity of representing multidimensional spaces, we decided to reduce the dimensionality of our data by means of the Uniform Manifold Approximation and Projection (UMAP) [20].

When using UMAP, a projection of a multidimensional space is created, where the distances between points are correlated with the distances in the multidimensional space. The embedding is created by constructing a weighted graph in the high-dimensional space, to capture local relationships between points, and then adjusting a lower-dimensional graph so it closely matches these relationships by minimizing a cost function. Therefore, the global structure of the data set is preserved. In our case, we chose to create a 2 dimensional projection of our 11 dimensional abundance space, in order to visualize our data shown in Figure 1.

### B. Clustering Methods

Once we selected our sample data, the next step was to find groups of data with similar abundance characteristics. We approached this issue from two different angles: applying clustering methods directly to the UMAP projection and applying clustering methods to our 11-dimensional abundance space and then visualizing them in the UMAP projection.

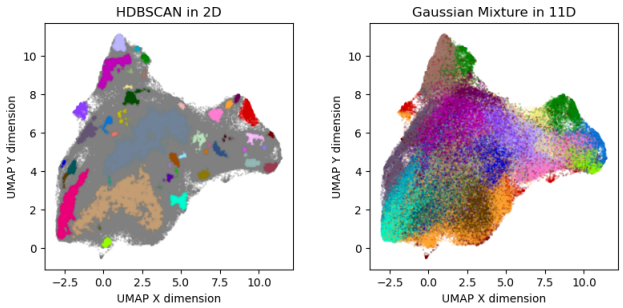


FIG. 1: Visualization of the clusters obtained with HDBSCAN applied directly to the 2 dimensional space created by UMAP and the clusters obtained with Gaussian Mixture applied to the 11 dimensional abundance space, respectively. Note the amount of points labeled as noise (in gray) in the first plot.

First of all, we applied a Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [16] method to the UMAP projection. HDBSCAN depends on various parameters, including the following three: the minimum number of samples in a group for that group to be considered a cluster (`min_cluster_size`), the maximum number of samples allowed in a cluster (`max_cluster_size`) and the parameter k used to calculate the distance between a point and its k-th nearest neighbor (`min_samples`). After trying different values for these parameters, we concluded the best result was obtained with `min_cluster_size=min_samples=100` and `max_cluster_size=50,000`, since not limiting the size of the clusters implied getting just 2 huge clusters. As a result, we obtained 47 clusters which contained less than half of our sample, the rest was labeled as noise. We conclude that this method is not satisfactory, as we expect the majority of the stars to be classified [10].

We applied HDBSCAN to the 11 dimensional space, trying out up to a hundred different values for the parameters stated above, but did not obtain better results than previously. Therefore, we tried other clustering methods. The Ordering Points To Identify the Clustering Structure (OPTICS) method [16] gave similar results to the ones obtained with HDBSCAN. We finally tried applying an Agglomerative Clustering method [16] and a Gaussian Mixture model [17] [11]. The first merges pairs of clusters of the sample data recursively, in our case using Ward's method as linkage distance, which minimizes the variance of the clusters being merged (from now on indicated as Ward). The second is a probabilistic model that assumes all our sample data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Both methods include

the parameter `n_components`, which we arbitrarily set at 20.

As can be seen in the right plot of Figure 1 and the center plot of Figure 2, the results obtained applying both methods are quite similar. We plotted the 2 dimensional abundances plots of each element, as the ones surrounding the UMAP projection in Figure 2. We observed that the clusters obtained with Ward followed the expected patterns more clearly than the Gaussian Mixture's clusters, and that their standard deviation from the mean value was smaller. For these reasons, we decided to continue working with the Ward clusters.

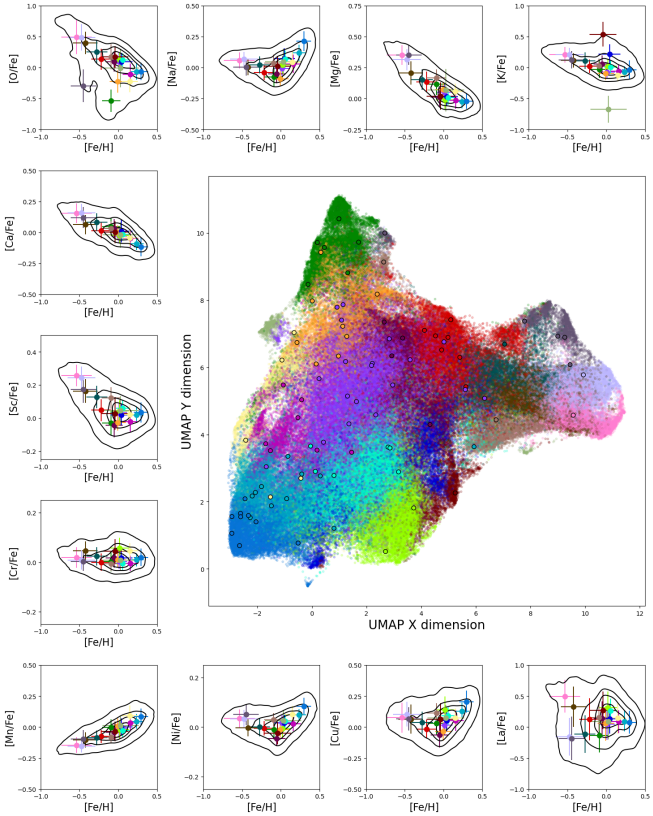


FIG. 2: UMAP projection plot showing the clusters obtained with Ward and highlighting the location of the exoplanet host stars. In addition, 2-D abundances plots of the elements used presented as the mean value of each cluster and its standard deviation, along with their kernel density estimate plot.

## IV. RESULTS

### A. Abundance subplots

In order to evaluate if our abundance data is correct, we need to understand the origin of each element we used and how its ratio [X/Fe] relates with the ratio [Fe/H].

#### 1. *α-elements: O, Mg, Ca*

These elements are produced during the nuclear fusion reactions by which helium is transformed into other heavier elements. This type of process usually occurs in high-mass stars and during supernovae.

Our results are consistent with the expected patterns [3]. However, the green and slate gray clusters deviate from the others for the oxygen, something we associate with problems in the abundance determination of this element, since these clusters do not present any other anomaly for the rest of the elements.

#### 2. *Iron-peak elements: Sc, Cr, Mn, Ni, Cu + Na*

These elements are the ones with atomic number near Fe. Nevertheless, they do not all behave the same way.

Sc is a Mg-like element, which means that its dependence with the ratio [Fe/H] is similar to the one of the $\alpha$-elements [6], as can be seen in Figure 2. This has been explained by suggesting that Sc is mostly made by high mass stars.

Cr, Ni, Cu, in addition of Na, are Ni-like elements, which behave alike for thick and thin disk stars [5], as can be seen in Figure 2. There are several studies related to the addition of Na in this group, and our results are consistent with some studies that suggest an increment of [Na/Fe] with [Fe/H] for stars of [Fe/H]$>0$ and an apparent slow rise in [Na/Fe] with decreasing [Fe/H] for stars with [Fe/H]$<0$ [5].

The trend of [Mn/Fe] appears to mirror that of the $\alpha$-elements, and some studies have shown that nucleosynthesis of Mn in massive stars with metallicity dependent output can explain this trend relatively well [6]. Therefore, our results are consistent with these studies.

#### 3. *Potassium K*

This element dependence with [Fe/H] is consistent with some studies showing that [K/Fe] gradually increases with a decrease of [Fe/H] over the range of $-0.8 \leq$ [Fe/H] $\leq 0.0$ [4]. However, as seen in Figure 2, even though our results are consistent with this, the maroon and pistachio clusters deviate from the others. Since these clusters don't present any other anomaly, we agreed it can be associated to a measurement problem.

#### 4. *Neutron-capture element: La*

La is an s-process element, that means its nucleosynthesis consists of slow neutron captures [7]. Some studies have shown [La/Fe] decreases for higher [Fe/H] [7], as can be seen in Figure 2, so our results are consistent. Nevertheless, for this element there is a lot of dispersion.

### B. Cluster Characterization

In light of the above, it is possible to characterize some of the clusters obtained with Ward.

Table I shows the hot pink, lilac, slate gray and dark brown clusters can be considered thick disk stars, since they have a high abundance ratio per low-metallicity stars for the O, Mg, Ca, Sc. In addition, notice these clusters correspond with the ones located in the emerging bubble on the right of the center plot in Figure 2. This means our characterization is consistent, being an illustration of the UMAP preservation of our data set global structure.

We considered the rest of the clusters to be part of the thin disk, although we would need to check more parameters to verify this statement. Also, it would be necessary to compare the spectra with the fits for each element, in order to corroborate the abundances are measured correctly. In addition, notice the columns $[X/Fe]_{max}$ and $[X/Fe]_{min}$ only contain the most significant deviations from the patterns. These alterations depend on each element covered range induced by the normalization, so it is possible there are more significant discrepancies we cannot see as a consequence of how the data is plotted.

It can be useful to work with violin plots, such as Figure 3, in order to visualize each cluster relative abundance tendency for each element used. Notice that in Figure 3 the top plot contains the thick disk clusters. Examining and comparing it with the other three, we can see that their relative abundance is higher than the general tendency for the $\alpha$-elements and Sc and lower for the Mn, as expected. Besides, their relative abundance is considerably similar to the other clusters for the Ni-like elements and Na, also as expected.

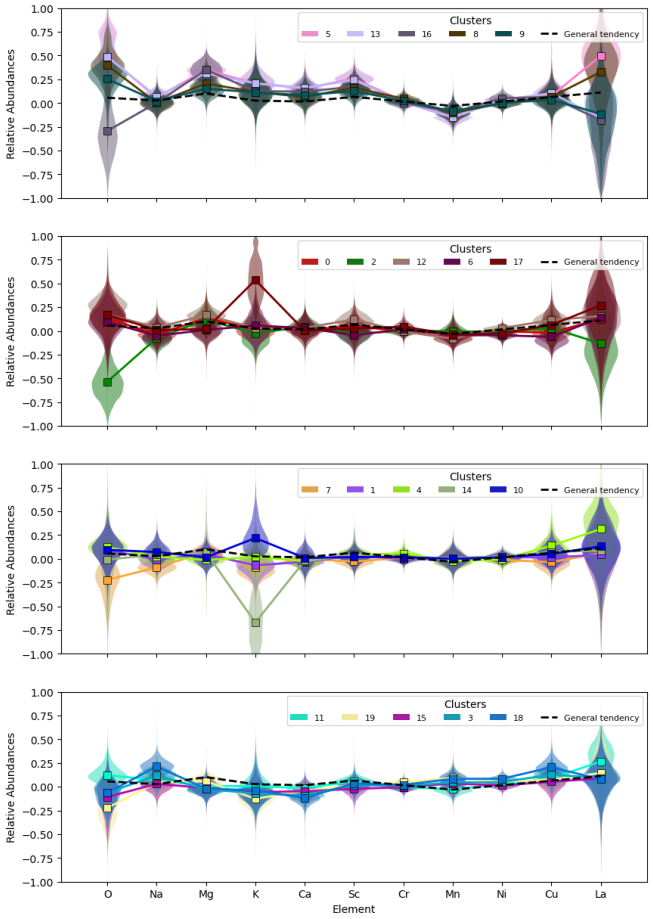


FIG. 3: Relative chemical abundances of the clusters defined by Ward, divided into bins of metallicity for a better visualization. The mean values of each cluster, its tendency and the general tendency of the whole used data set is also shown. Note the x axis represents the chemical elements.

### C. Exoplanets location in clusters

After highlighting the stars known to host exoplanets, we generated Table I in order to visualize more clearly our data.

The cluster with the highest number of exoplanets is also the one with the highest weight, which is coherent. Also notice the first four clusters, the sum of their weights being 37.6%, contain approximately half our host stars sample (48.5%). This is in agreement with the fact that thin disk stars are younger, slower and metal-richer, which is key to the formation of exoplanets [8].

The clusters we characterized as thick disk stars, which weigh 11.5%, contain around 7% of the host stars sample. This result is consistent with the fact that thick disk stars are older, faster and metal-poorer, which impacts negatively the formation of exoplanets around them [8].

Looking at the center plot of Figure 2 and Table I, it is clear there is no significant chemical-abundance characteristic for the cluster with the most host stars, other than belonging to the thin disk and following the expected abundance trends.

## V. CONCLUSIONS

Throughout this work we have seen that clustering methods such as HDBSCAN or OPTICS are inefficient to study large data sets in high-dimensional spaces such as an 11 dimensional abundance space with around 190,500 stars selected from GALAH DR4 survey.

We have also seen that Agglomerative Clustering with Ward linkage or Gaussian Mixture clustering models produce similar clusters for our type of data set. However, subtle differences, such as cluster visualization clarity in the UMAP projection and more definite tendencies for the ratios [X/Fe] as functions of [Fe/H], have tipped the scales to the use of the clusters obtained with Ward.

As discussed in the previous section, we have not found a clear relation between our data set clusters abundance ratios and the existence of host stars in them. In fact, there are some studies suggesting that exoplanets with radius inferior to four Earth radii form around host stars with a wide range of chemical abundances, which indi-

TABLE I: Interpretation of the 20 groups with their respective weights, number of host stars and their percentage respect to the overall number of stars in the group, metallicity and main maximum and minimum inconsistency with the expected patterns.

| # | Colour | Population | Weight | Host stars | % | [Fe/H] | [Mg/Fe] | $[X/Fe]_{max}$ | $[X/Fe]_{min}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Blue Violet | Thin Disk | 13.3% | 19 | 0.075% | $0.01 \pm 0.10$ | $0.05 \pm 0.06$ | – | – |
| 11 | Turquoise | Thin Disk | 9.0% | 6 | 0.035% | $0.06 \pm 0.09$ | $0.01 \pm 0.06$ | – | – |
| 0 | Red | Thin Disk | 8.0% | 13 | 0.085% | $-0.22 \pm 0.13$ | $0.13 \pm 0.08$ | – | – |
| 3 | Cyan | Thin Disk | 7.3% | 11 | 0.079% | $0.24 \pm 0.09$ | $-0.03 \pm 0.05$ | – | – |
| 6 | Dark Magenta | Thin Disk | 6.9% | 4 | 0.03% | $-0.05 \pm 0.11$ | $0.01 \pm 0.08$ | – | – |
| 4 | Lime | Thin Disk | 6.6% | 2 | 0.016% | $0.02 \pm 0.12$ | $-0.01 \pm 0.07$ | – | – |
| 15 | Magenta | Thin Disk | 5.6% | 7 | 0.066% | $0.16 \pm 0.10$ | $-0.01 \pm 0.07$ | – | – |
| 2 | Green | Thin Disk | 5.5% | 6 | 0.057% | $-0.09 \pm 0.12$ | $0.12 \pm 0.09$ | – | [O/Fe]=-0.53 |
| 18 | Dodger Blue | Thin Disk | 5.0% | 9 | 0.095% | $0.30 \pm 0.09$ | $-0.02 \pm 0.06$ | – | – |
| 10 | Blue | Thin Disk | 4.5% | 0 | 0.0% | $0.04 \pm 0.14$ | $0.02 \pm 0.07$ | [K/Fe]=0.22 | – |
| 7 | Orange | Thin Disk | 4.5% | 8 | 0.092% | $-0.01 \pm 0.11$ | $0.07 \pm 0.06$ | – | – |
| 9 | Teal | Thin Disk | 3.8% | 1 | 0.014% | $-0.28 \pm 0.14$ | $0.15 \pm 0.08$ | – | [La/Fe]=-0.12 |
| 12 | Rosy Brown | Thin Disk | 3.7% | 2 | 0.028% | $-0.09 \pm 0.12$ | $0.17 \pm 0.07$ | – | – |
| 13 | Lilac | Thick Disk | 3.5% | 1 | 0.015% | $-0.47 \pm 0.18$ | $0.31 \pm 0.09$ | – | [La/Fe]=-0.15 |
| 5 | Hot Pink | Thick Disk | 3.2% | 1 | 0.016% | $-0.54 \pm 0.2$ | $0.35 \pm 0.08$ | – | – |
| 8 | Dark Brown | Thick Disk | 3.1% | 0 | 0.0% | $-0.42 \pm 0.16$ | $0.21 \pm 0.09$ | – | – |
| 19 | Yellow | Thin Disk | 2.5% | 5 | 0.106% | $0.15 \pm 0.10$ | $0.06 \pm 0.06$ | – | – |
| 16 | Slate Gray | Thick Disk | 1.8% | 5 | 0.144% | $-0.45 \pm 0.16$ | $0.35 \pm 0.08$ | – | [O/Fe]=-0.29 |
| 17 | Maroon | Thin Disk | 1.5% | 1 | 0.035% | $-0.04 \pm 0.17$ | $0.02 \pm 0.11$ | [K/Fe]=0.54 | – |
| 14 | Pistachio | Thin Disk | 0.6% | 0 | 0.0% | $0.03 \pm 0.23$ | $0.07 \pm 0.17$ | – | [K/Fe]=-0.67 |

cates that Earth-like exoplanets may be widespread in the disk of the Milky Way [9].

In conclusion, there is still some work to be completed such as ensuring the abundances are correctly measured and performing a deeper study with stellar parameters (as age, mass or kinematics) in order to achieve a more accurate characterization of the abundance subpopulations. In order to confirm possible trends of the exoplanet occurrence rate with stellar abundances, a much larger spectroscopic sample of exoplanet host stars is required.

**Acknowledgments**

I would like to thank my advisors Friedrich Anders and Chloé Padois for their constant guidance, insights and for having faith in me from the start. I would also like to thank all my family and friends that have patiently supported me during the last few months.

[1] Buder, S. et al. "The GALAH Survey: Data Release 4", ApJ, subm. arXiv:2409.19858v1 (2024).
[2] Mayor, M. and Queloz, D. "A Jupiter-mass companion to a solar-type star", Nature **378**, 355–359 (1995)
[3] Koch, A. et al. "Detailed chemical abundance analysis of the thick disk star cluster Gaia 1", A&A **609**: A13 (2018).
[4] Zhang, H.W. et al. "Potassium abundances in nearby metal-poor stars", A&A **457**: 645–650 (2006).
[5] Reddy, B.E. et al. "Elemental Abundance Survey of The Galactic Thick Disk", MNRAS, **367**: 1329–1366 (2006).
[6] Nissen, P.E. et al. "Sc and Mn abundances in disk and metal-rich halo stars", A&A **353**: 722–728 (2000)
[7] Pompéia, L. et al. "Detailed analysis of nearby bulge-like dwarfs stars III. $\alpha$ and heavy-element abundances", AJ **592**: 1173–1185 (2003).
[8] Bashi, D. et al. "Exoplanets in the Galactic context: Planet occurrence rates in the thin disk, thick disk and stellar halo of Kepler stars", MNRAS, **510**: 3449–3459 (2021).
[9] Buchhave, L.A. et al. "An abundance of small exoplanets around stars with a wide range of metallicities", Nature **486**: 375–377 (2012).
[10] Dolcet. J, "Unsupervised machine learning techniques for chemical analysis in spectroscopic stellar surveys", TFG (2021), https://hdl.handle.net/2445/180269
[11] Cuevas, A. "Analysing the red-clump star population of the Milky Way with abundance-space extreme deconvolution", TFG (2022), https://hdl.handle.net/2445/189706
[12] https://www.galah-survey.org/dr4/the_catalogues/
[13] https://exoplanetarchive.ipac.caltech.edu/
[14] https://www.cosmos.esa.int/gaia
[15] https://scikit-learn.org/stable/api/sklearn.preprocessing
[16] https://scikit-learn.org/stable/api/sklearn.cluster
[17] https://scikit-learn.org/stable/api/sklearn.mixture
[18] https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet
[19] https://phl.upr.edu/hwc
[20] https://umap-learn.readthedocs.io/en/latest/

# Anàlisi de l'abundància química d'estrelles de la seqüència principal de GALAH DR4 i de la base de dades d'exoplanetes de la NASA

Author: Sara Drapkin Junyent, sdrapkju7@alumnes.ub.edu
*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.*

Advisors: Friedrich Anders, Chloé Padois; {fanders,chloe.padois}@fqa.ub.edu

**Resum:** En aquest treball analitzem les abundàncies químiques d'un conjunt de dades corresponents a estrelles de la seqüència principal, sel·leccionat del catàleg de GALAH DR4, mitjançant tècniques de reducció de dimensionalitat (UMAP) i mètodes d'agrupament (Agglomerative Clustering amb nexe tipus Ward). Això ens permetrà, amb l'ajut de més treball futur, definir grups clars i interpretables d'abundàncies químiques, amb més detall que l'usual dicotomia entre disc prim i gruixut. Utilitzant una submostra d'estrelles hostes d'exoplanetes ja conegudes dins la mostra de GALAH DR4, explorem les possible tendències en la taxa d'aparició de planetes en funció del grup d'abundància, trobant poca evidència estadística.
**Paraules clau:** Exoplanetes, estrelles en la seqüència principal, abundàncies elementals, paràmetres estelars fonamentals, anàlisi de dades, normalització de dades.
**ODSs:** Aquest TFG està relacionat amb l'Objectius de Desenvolupament Sostenible (SDGs) número 4, educació de qualitat.

**Objectius de Desenvolupament Sostenible (ODSs o SDGs)**

| | | | |
|---|---|---|---|
| 1. Fi de la pobresa | | 10. Reducció de les desigualtats | |
| 2. Fam zero | | 11. Ciutats i comunitats sostenibles | |
| 3. Salut i benestar | | 12. Consum i producció responsables | |
| 4. Educació de qualitat | X | 13. Acció climàtica | |
| 5. Igualtat de gènere | | 14. Vida submarina | |
| 6. Aigua neta i sanejament | | 15. Vida terrestre | |
| 7. Energia neta i sostenible | | 16. Pau, justícia i institucions sòlides | |
| 8. Treball digne i creixement econòmic | | 17. Aliança pels objectius | |
| 9. Indústria, innovació, infraestructures | | | |