Draft version June 10, 2025
Typeset using LaTeX default style in AASTeX631

Finding the Planetary Fertile Crescent

Oliver Altindag and Ryan Purviance

ABSTRACT

Understanding the chemical compositions of stellar hosts is critical to determining the conditions under which planets form around M dwarfs. We investigate whether stars with different elemental abundances demonstrate distinct planet-hosting capabilities. We use two statistically distinct chemical composition clusters to develop a population-wide analysis of compositional trends via Bayesian inference conducted with a Markov Chain Monte Carlo (MCMC) simulation. We then incorporate the rate of confirmed exoplanetary hosts within each cluster to provide insight into the role of stellar chemistry in planetary formation. We found that metal-rich M dwarfs result in a more favorable environment for planet formation. Our work contributes to the broader understanding of stellar composition's impact on the likelihood of planet formation across M dwarf systems.

1. BACKGROUND

Planetary systems originate within protoplanetary disks—large rotating clouds of gas and dust encircling young stars (Williams and Cieza 2011). Within these disks, particles collide and gradually stick together, forming larger and larger bodies: like a Russian nesting doll, layers building upon layers, eventually creating full-fledged planets (NASA, 2025). The chemical makeup of a protoplanetary disk is intimately linked to the chemical abundance of its host star, as the disk is leftover material from the star's formation. In particular, M dwarfs are the smallest and coolest stars in the universe; they are also the most common, comprising roughly 70% of all stars in the Milky Way (Dieterich 2020). Their long lifespans and stable luminosities make them prime candidates in the search for exoplanets.

In our project, we investigate how variations in the stellar chemical abundances–the relative proportions of chemical elements in a star's composition–of M dwarfs may influence planet formation. Specifically, we create two statistically distinct clusters, Cluster 0 containing metal-poor M dwarfs, and Cluster 1 containing metal-rich, and compare the frequency of exoplanet hosts between them. As the hunt for exoplanets accelerates, gaining insight into likely environments of planet formation is vital for advancing impactful exploration.

2. DATA

We obtained our data from the M dwarf Catalog, a compiled database of chemical properties for roughly 17,000 M dwarf stars observed in the Gaia DR3 mission. This catalog was assembled by a team of astronomers aiming to build a data-driven model of chemical abundances in M dwarfs (Behmard et al. 2025). The catalog includes a wide array of elemental abundance measurements, but for our analysis, we focused on five key variables: [Fe/H], [Mg/H], [Si/H], [C/H], and [O/H]. These elements were each selected for their known relevance in different planet-hosting environments, noted in earlier studies on chemical abundances impact on planet formation (Adibekyan, V. Zh. et al. 2012)(Fischer, and Valenti 2005)(Thiabaud, A. et al. 2015).

To identify which M dwarfs are confirmed planet hosts, we used data from the NASA Exoplanet Archive (Akeson et al. 2013), a well-established database that compiles information on confirmed exoplanetary systems. For our purposes, we extracted the Gaia DR2 source IDs of known stellar hosts to be used as a counter for our stellar host rate calculation.

Using Astroquery and the Gaia DR2 Neighborhood table, we matched the M dwarf source IDs from the Gaia DR3 mission to their DR2 counterparts. This cross-referencing process enabled the creation of a refined dataset containing only M dwarfs with confirmed exoplanetary companions. However, this subset proved small—only 69 M dwarfs matched across all data sources. The limited size of this confirmed stellar host sample necessitated the use of bootstrapping techniques in our statistical analysis to draw meaningful conclusions from such a small population.

Several potential sources of bias and error could influence our findings. Most notably, the small number of confirmed M dwarf stellar hosts limits the statistical robustness of our results and makes the analysis particularly sensitive to outliers. Furthermore, the matching process between Gaia DR2 and DR3 source IDs is not always available, with

38 unmatched M dwarfs in our dataset from an original 16590. This could potentially lead to systematic biases in our stellar host rates by excluding more viable hosts from one cluster. However, it is important to note that even with a large number of stellar hosts missing, we do not expect significant bias favoring one cluster in the Gaia neighborhood table. The validity of our research is therefore not compromised; we are internally comparing host rates between clusters and not searching for a numerically accurate, but rather a representative comparative stellar host rate between the two. Finally, uncertainties in the measurement of chemical abundances used in the M dwarf catalog can further complicate the accuracy and completeness of our analysis; however, this was factored into our MCMC modeling by using the measured uncertainty, reducing but not eliminating its systematic impact.

3. PROCEDURE

We began by normalizing the M dwarf stellar abundance data from the M dwarf catalog using the sklearn machine learning library. Specifically, we applied standard scaling so that each feature had a mean of zero and a standard deviation of one. This preprocessing step ensured that no single feature would dominate the analysis due to scale differences when using K-Means.

Once filtered, we applied the K-Means clustering algorithm to the normalized data to segment the stars into two statistically distinct groups. This method determines these clusters by performing statistical modeling with multiple parameters to find the minimal distance from each data point to the centroid values of each cluster. This machine learning step was central to our analysis, providing a reproducible and data-driven way to define clusters based on multivariate abundance patterns. It established the structural framework necessary for all subsequent comparisons, achieving a silhouette score–a metric used to evaluate the quality of clustering with -1 corresponding to poor separation and 1 corresponding to complete separation–of 0.51, signifying good separation across all abundances.

With the clusters identified, we applied Bayesian inference to model each cluster's abundance distribution. Specifically, we assumed the data followed a multivariate Gaussian framework, where each cluster is characterized by a mean vector, representing the average elemental abundances and a covariance matrix containing the dependencies between elements. The likelihood function derived from this model shows how well the mean vector and covariance matrix explain the observed data; we develop the model with the Mahalonobis distance and a Chokesky Factorization of the covariance matrix to achieve these results. Additionally, we incorporate measurement uncertainty propagation into the likelihood to increase robustness. Our priors were uniform, ensuring our results were data-driven. We used the emcee package to sample from the posterior distribution over these parameters, with our framework allowing us to estimate the mean abundance vector and the covariance matrix for each cluster improving upon sheer frequentist methods. From the converged walkers, shown in the Appendix, we removed 5000 burn-in steps. We then extracted the posterior samples to estimate the mean abundance vector and correlations between abundances.

Next, we filtered a new dataset from the NASA Exoplanet Archive dataset, containing only stellar hosts whose source IDs matched ones in our M dwarf subset.

Finally, we analyzed the dataset of known exoplanetary hosts with the M dwarf catalog, determining the number of known stellar hosts in each cluster. Since the number of known hosts was relatively small, we implemented a bootstrapping procedure to evaluate statistical significance. Using 10,000 simulations, randomly selecting 10,000 M dwarfs from our dataset with replacement and finding the number of selected stellar hosts, we constructed a distribution of stellar host frequencies for each cluster. We then applied a K-S test to assess whether the observed difference in distribution was significant for the frequency of stellar hosts between clusters.

4. ANALYSIS AND DISCUSSION

In Fig. 1, we explored the relationship between stellar chemical abundances by comparing every combination of normalized chemical abundances across our two clusters, represented by different colors. Overlaid on these plots are all stellar hosts, shown as red points. The first key insight from these visualizations was the confirmation that the stellar hosts are randomly distributed across the entire chemical abundance spectrum. This random distribution ensured that we could proceed with the population-wide analysis without any undue bias introduced by concentrated stellar hosts at the center of each abundance space. Additionally, it confirmed that the clustering had worked properly and that the two clusters barely overlapped in every abundance comparison.

Following this, we calculated each cluster's observed stellar host frequency rate. For Cluster 0, the rate was 0.003372, while for Cluster 1, the rate was 0.004095. We then performed a bootstrapped K-S test on the stellar host rate, comparing whether the simulated distribution of the stellar host frequency rate of the two samples is drawn from the

same continuous distribution. The result was a K-S statistic of 0.542 and a p-value of p ¡ 0.0001, which provided strong evidence against the null hypothesis that both populations come from the same distribution. We adopted a significance level of $\alpha = 0.05$, which is standard in astronomical research (Rice et al. 2024), but due to the very low p-value, we could also comfortably apply stricter significance thresholds if needed. The rejection of the null hypothesis indicates that Cluster 1 has a significantly higher rate of stellar hosts compared to Cluster 0.

Next, we analyzed the best-fit normalized mean abundance vectors for all stars in each cluster derived through MCMC. The results are shown as follows:

Ratio	**Cluster 0**	**Cluster 1**
[Fe/H]	-0.8102	0.7615
[Mg/H]	-0.7869	0.7393
[Si/H]	-0.8175	0.7682
[C/H]	-0.7888	0.7520
[O/H]	-0.8166	0.7689

Table 1: Normalized elemental abundance ratios for Cluster 0 and Cluster 1.

These reflect the overall trends within each cluster: Cluster 0 is metal-poor, whereas Cluster 1 is metal-rich. Since the abundances are normalized, these results cannot be directly compared to specific chemical abundance values; however, the values indicate the number of standard deviations away from the mean, which as demonstrated, is consistently different between the two clusters.

This difference suggests that metal-rich M dwarfs may provide a more favorable environment for planet formation. This is corroborated by the statistical finding in our K-S test, which suggested that the stellar host rates between clusters were significantly different.

We made a corner plot of the abundances, to demonstrate their relationships as 2-D kernel density plots, in Fig. 2,3. The plot revealed that the abundances remained well correlated to each other like in the unsampled data, further demonstrating a dependent relationship. Additionally, the individual abundance distributions were normally distributed around their calculated mean values from the MCMC.

In Fig. 4, we calculated the observed correlations seen in Fig. 2,3 between each pair of normalized chemical abundances within the clusters in a heat map. The results showed that while the two clusters had very different chemical abundances, the correlations between elements were generally similar. However, a notable exception emerged in the correlation between [Fe/H] and [Mg/H], where Cluster 1 exhibited a significantly higher correlation (0.17 greater) compared to Cluster 0. Given the higher stellar host frequency in Cluster 1, this could suggest a potential link between the ratio of magnesium to iron abundance [Mg/Fe] and the likelihood of planet formation: in particular, it could indicate that the [Mg/Fe] ratio plays a more important role in the planet formation process in metal-rich stars compared to metal-poor stars.

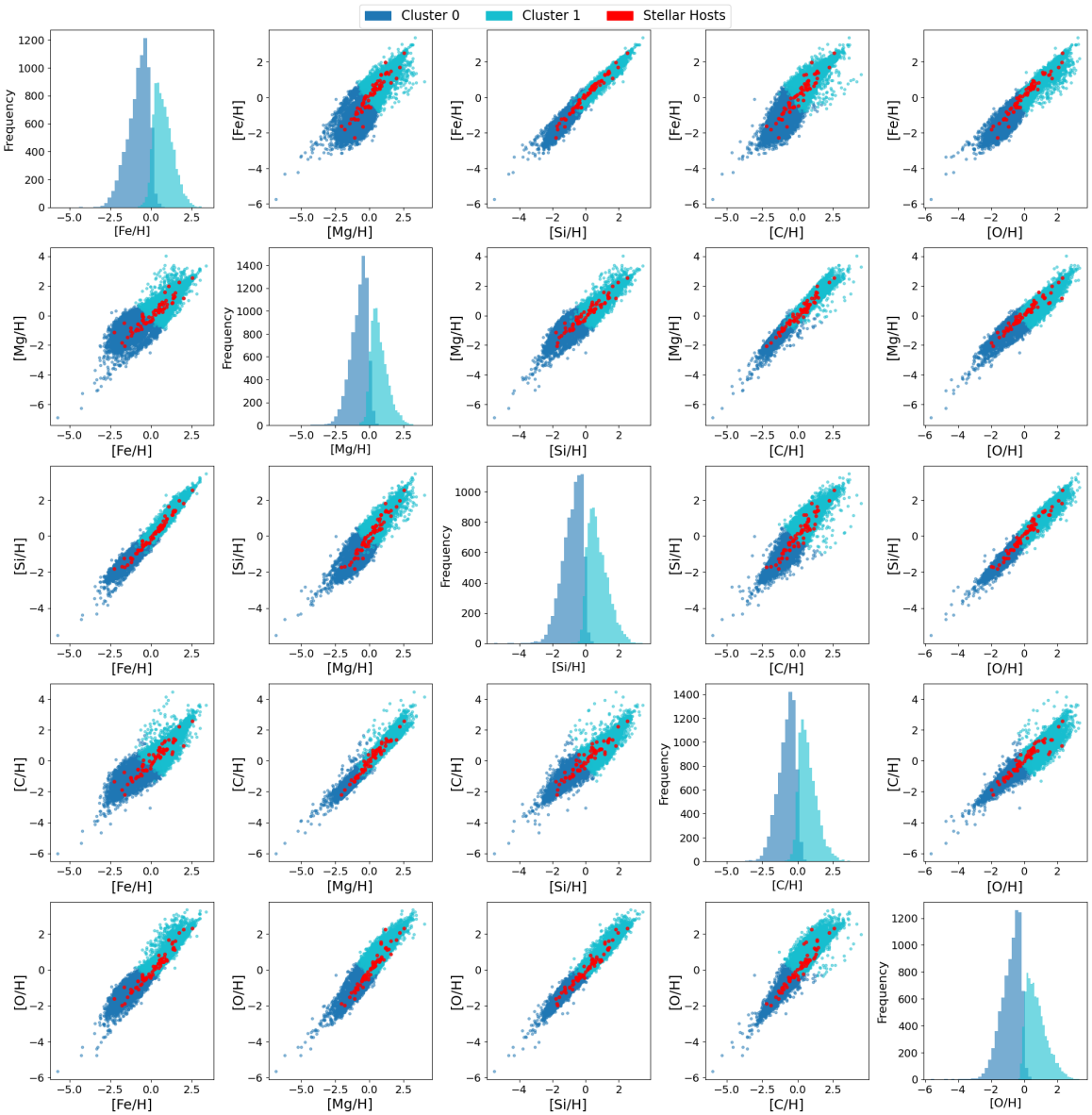

Figure 1: Matrix of scatter plots, demonstrating the relationship between each pair of normalized abundances. We can see they are all positively correlated, with stellar hosts randomly distributed across both clusters.

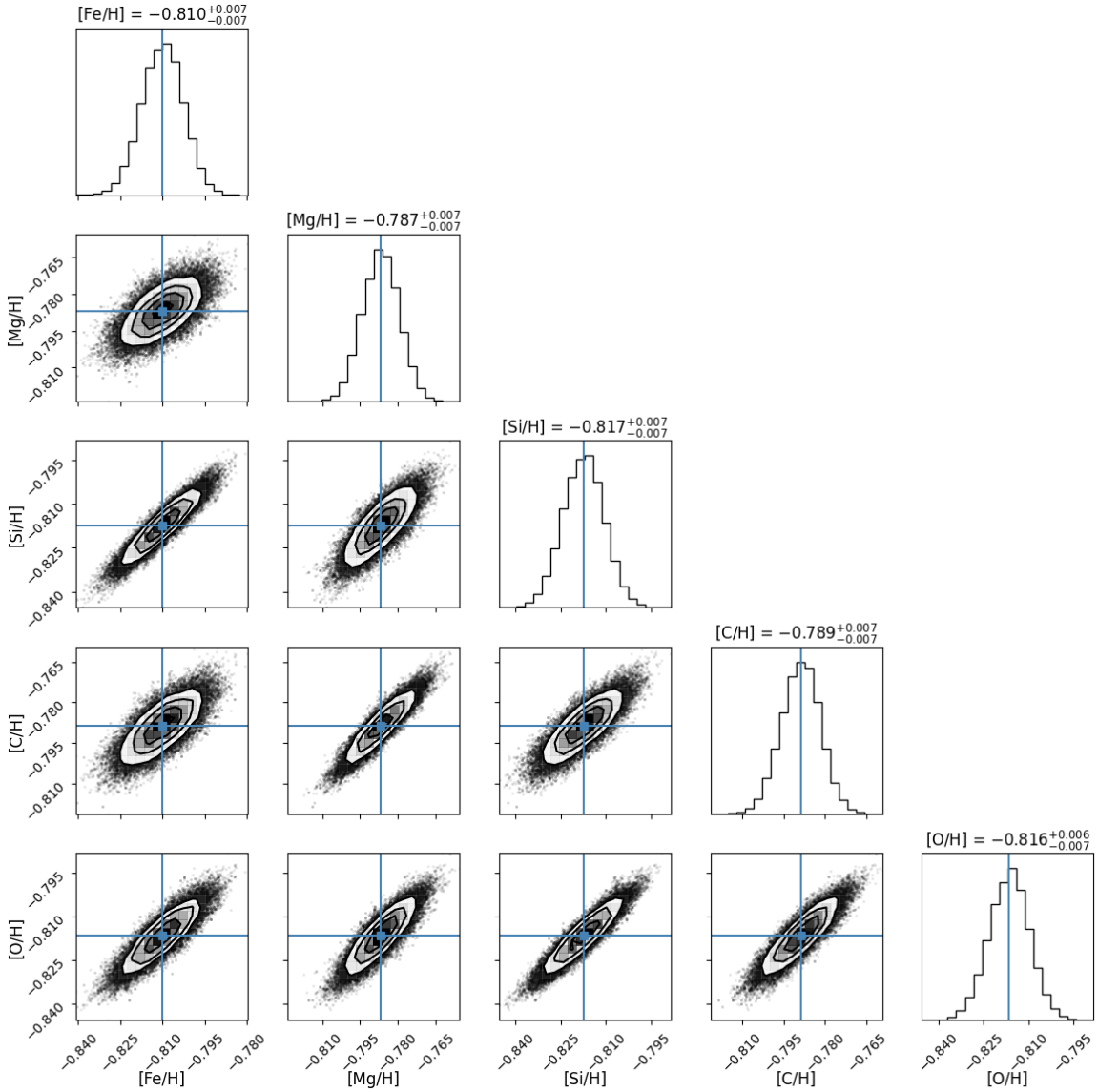

Figure 2: Corner plot of MCMC results for the posterior distributions of our chemical abundances, represented as a 2-D kernel density plot (KDE), between each abundance for Cluster 0. The oblong shape of each of the KDEs further demonstrates the correlations seen in Fig. 1 between the normalized abundances; the blue lines indicate the mean abundance values. Additionally, the histogram of each chemical abundance is normally distributed as dictated by the central limit theorem.

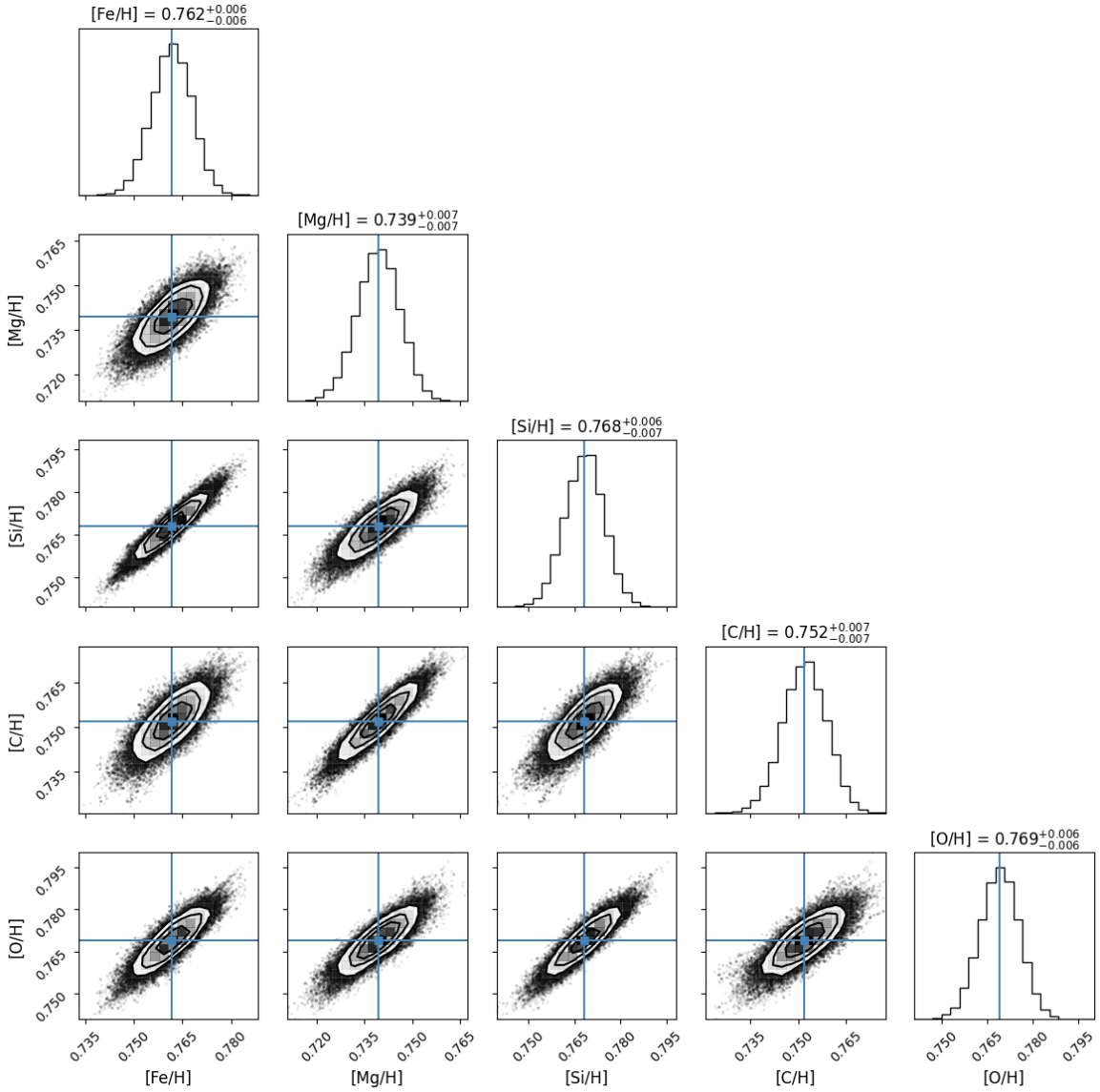

Figure 3: Similar to Fig. 2, this is a corner plot of MCMC results for the posterior distributions of our chemical abundances, represented as a 2-D KDE, between each abundance for Cluster 1. The oval shape of each of the KDEs further displays the correlations seen in Fig. 1 between the normalized M dwarf abundance data; the blue lines show the mean abundance values. The histogram of each chemical abundance is normally distributed, demonstrating the effects of the central limit theorem.

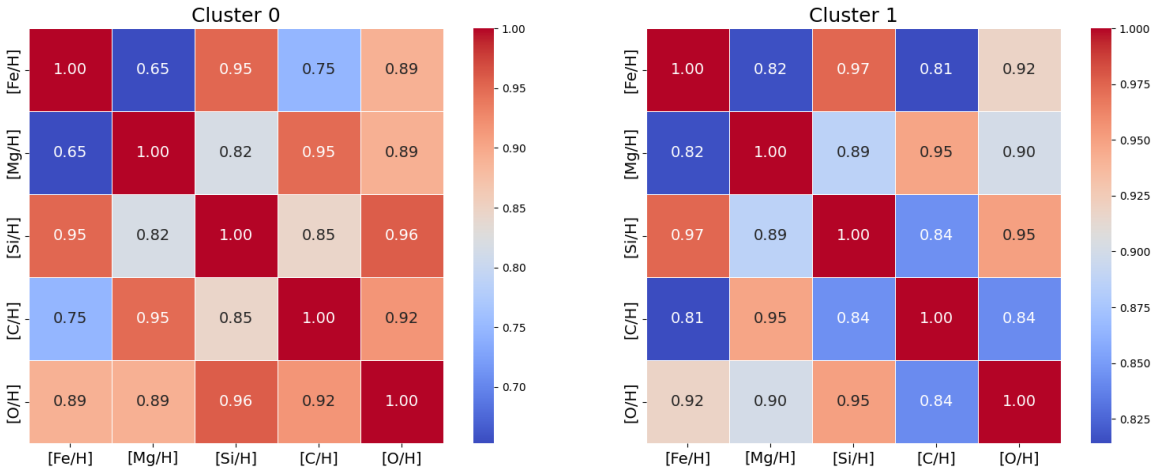

Figure 4: Heat map of the correlations between each combination of abundances, calculated from the covariance matrix from our MCMC; notice that the scaling of the color bar is different for the two clusters, showing an overall difference in the variability of their correlation.

5. CONCLUSIONS

In this study, we explored the relationship between stellar chemical abundances and the likelihood of planet formation around M dwarfs. By analyzing two distinct clusters—metal-poor and metal-rich M dwarfs—we identified significant differences in their stellar host rates. These differences suggest a higher likelihood of planet formation around metal-rich stars. This conclusion is supported by our K-S test and Bayesian inference, both of which revealed that metal-rich M dwarfs are significantly more likely to host planets compared to their metal-poor counterparts.

Furthermore, the notable increase in the correlation between [Fe/H] and [Mg/H] in the metal-rich cluster suggests that the ratio of magnesium to iron may be a critical factor in determining a star's ability to host planets, particularly in metal-rich M dwarfs.

Our work reinforces the idea that stellar composition plays a key role in the planetary formation process and provides valuable insights into how specific elemental abundances, like the [Mg/Fe] ratio, could influence the formation of exoplanets. This contributes to the growing body of knowledge about the conditions that make certain M dwarf systems more conducive to planet formation and expands our understanding of the role stellar composition plays in the formation of planetary systems.

6. APPENDIX

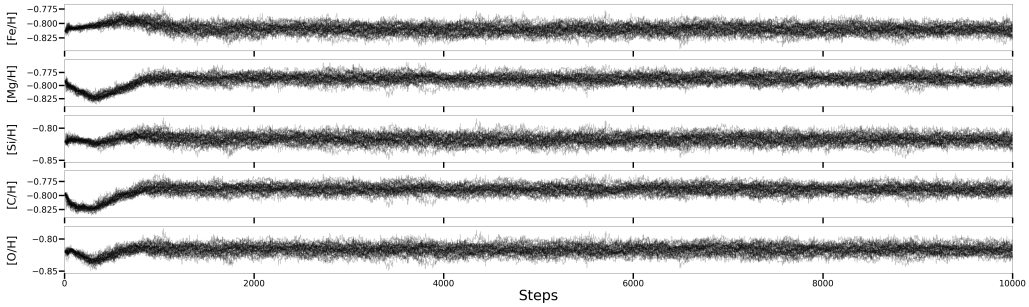

Figure A.1: Raw MCMC chains of the mean abundance vectors for Cluster 0, showing convergence.

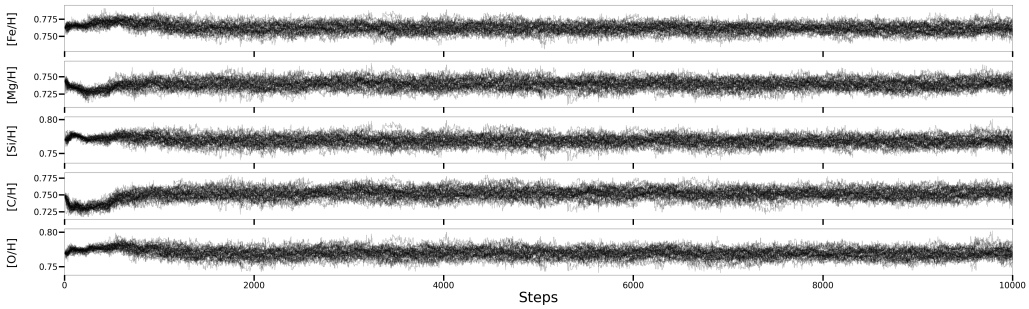

Figure A.2: Raw MCMC chains of the mean abundance vectors for Cluster 1, showing convergence.

REFERENCES

[1] Adibekyan, V. Zh. et al. "Chemical abundances of 1111 FGK stars from the HARPS GTO planet search program. Galactic stellar populations and planets." *Astronomy & Astrophysics*, vol. 545, id. A32. Sept, 2012. https://ui.adsabs.harvard.edu/abs/2012A%26A...545A..32A/abstract

[2] Akeson, R. L. et al. "The NASA Exoplanet Archive: Data and Tools for Exoplanet Research." *Publications of the Astronomical Society of the Pacific*, vol. 125, no. 930, p. 19. Aug, 2013. https://ui.adsabs.harvard.edu/abs/2013PASP..125..989A/abstract

[3] Baudin, Michaël and Régis Lebrun. "Linear algebra of linear and nonlinear Bayesian calibration". *Eccomas Proceedia*, pp. 339-353. June, 2021. https://files.eccomasproceedia.org/papers/uncecomp-2021/UC21_19074.pdf?mtime=20210921125442

[4] Behmard, Aida et al. "A Data-driven M Dwarf Model and Detailed Abundances for ∼17,000 M Dwarfs in SDSS-V." *The Astrophysical Journal*, vol. 982, no. 1, pp. 13-19. Mar, 2025. https://ui.adsabs.harvard.edu/abs/2025ApJ...982...13B/abstract

[5] Dieterich, S. "How Well Do We Understand M Dwarfs?" *STScI Newsletter*, vol. 37, no. 1, 2020. https://www.stsci.edu/contents/newsletters/2020-volume-37-issue-01/how-well-do-we-understand-m-dwarfs

[6] Fischer, Debra A, and Jeff Valenti. "The Planet-Metallicity Correlation." *The Astrophysical Journal*, vol. 622, no. 2, pp. 1102-1117. Apr, 2005. https://ui.adsabs.harvard.edu/abs/2005ApJ...622.1102F/abstract

[7] Foreman-Mackey, Daniel, et al. "emcee: The MCMC Hammer." *Publications of the Astronomical Society of the Pacific*, vol. 125, no. 925, p. 306. Mar, 2013. https://ui.adsabs.harvard.edu/abs/2013PASP..125..306F/abstract

[8] Ginsburg, Adam, et al. "astroquery: An Astronomical Web-querying Package in Python." *The Astronomical Journal*, vol. 157, no. 3, id. 98. Mar, 2019. https://ui.adsabs.harvard.edu/abs/2019AJ....157...98G/abstract

[9] Karthickaravindan, "K Means Clustering Project," last modified 2018. Accessed 1 May, 2025. https://www.kaggle.com/code/karthickaravindan/k-means-clustering-project/notebook

[10] NASA, "How Do Planets Form?" last modified 29 Oct, 2024. Accessed 20 Apr, 2025. https://science.nasa.gov/exoplanets/how-do-planets-form/

[11] Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825-2830. 2011. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[12] Rice, David R. et al. "The Distribution of Planet Radius in Kepler Multiplanet Systems Depends on Gap Complexity," *The Astrophysical Journal Letters*, vol. 973, no. 1. Sept. 10, 2024. https://iopscience.iop.org/article/10.3847/2041-8213/ad73db

[13] Thiabaud, A. et al. "Gas composition of the main volatile elements in protoplanetary discs and its implication for planet formation." *Astronomy & Astrophysics*, vol. 574, id. A138. Feb, 2015. https://ui.adsabs.harvard.edu/abs/2015A%26A...574A.138T/abstract

[14] Williams, Jonathan P, and Lucas A. Cieza. "Protoplanetary Disks and Their Evolution." *Annual Review of Astronomy and Astrophysics*, vol. 49, no. 3, pp. 67-117. 2 Mar, 2011. https://doi.org/10.1146/annurev-astro-081710-102548