JOURNAL OF
EMERGING INVESTIGATORS

Article

Differentiating characteristics in exoplanet host stars

Daniel Adibi[1], Sarah Kane[2], Bhuvnesh Jain[2]

[1] The Episcopal Academy, Newtown Square, Pennsylvania

[2] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania

SUMMARY

With technological advancements now allowing for precise measurements of stars, surveys are discovering thousands of exoplanets—planets outside of our solar system. We now have data on not just the kinematics and stellar chemistry (a star's chemical makeup and evolutionary stage) of host stars (the stars around which exoplanets orbit) but also on exoplanets' positions, sizes, and chemical compositions. However, while past studies have explored specific host star trends, no study has comprehensively analyzed how the stellar properties measured across these surveys differ in host stars. This understanding is important for exoplanet studies, as it can help astronomers understand the conditions favorable for exoplanet development and the exoplanets themselves better. In this study, we hypothesized that stellar chemistry, classification, and kinematics would differ significantly between exoplanet host stars and the galactic stellar population, as well as between host star subpopulations based on the type of planet hosted. To test this hypothesis, we analyzed data from recent surveys, including Gaia, the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, the Transiting Exoplanet Survey Satellite (TESS), and the NASA Exoplanet Archive. While we found that stellar chemistry was a significant differentiator between the exoplanet host star and the general population, we could not draw conclusive results about stellar classification or kinematics due to significant bias in TESS's selection. However, when comparing exoplanet host star subpopulations, we found that both stellar chemistry and kinematics yielded significant differences. These findings can be used to further test planetary formation models and indicate which stars could be more likely to host exoplanets for future exoplanet surveys.

INTRODUCTION

To date, astronomers have discovered just over 5,500 exoplanets—planets that orbit stars other than the Sun (1). As with our own diverse solar system, there are many types of exoplanets distinguished by their sizes and orbital periods (how long the planet takes to make one complete revolution around its host star) (2). Astronomers have five methods to detect these planets; of the five, transit photometry and Doppler measurements (or radial velocity) are the most utilized (3). Most known exoplanets were discovered using transit photometry, the process of scanning potential host stars for transits—periodic dips in brightness that occur when an exoplanet passes through our line of sight to its host star (4). NASA's Kepler spacecraft used transit photometry to survey hundreds of thousands of stars in a small region of the sky, and NASA's Transiting Exoplanet Survey Satellite (TESS) mission is performing the same procedure across a wider area (5). Doppler measurements rely on the principle that both an exoplanet and its host star orbit around their respective center of mass, causing the star to "wobble"—to orbit a point somewhere between its center and its planet (3). This movement causes periodic color changes in the star due to Doppler redshift and blueshift, which can be recorded to detect an exoplanet (3).

Factors such as temperature, mass, and behavior of stars can affect their likelihood to host exoplanets and the types of exoplanets they host (6). In particular, we used metallicity ([Fe/H], the iron-to-hydrogen ratio, indicating how metal-rich a star is; unitless), effective surface temperature (T_{eff}, the surface temperature of a star; in Kelvin), absolute magnitude (how bright a star is from 10 parsecs away; unitless), and galactocentric velocities (providing an indication of where a star lies in the galaxy; in km/s) to examine the properties of exoplanet host stars. A greater metallicity means a star is more metal-rich, and a greater absolute magnitude means a star is less luminous. Surveys such as the European Space Agency's Gaia and NAOC's Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) provide kinematic and chemical characteristics for large numbers of stars (7, 8). Understanding how host star conditions affect exoplanet development is an important step in forming planetary formation models and discovering more properties about exoplanets themselves, as measuring exoplanet properties is physically reliant on understanding their host stars using current measurement techniques (6, 9).

While current research into host stars' effects on planetary development is plentiful, to our knowledge, no study has yet viewed all known host stars comprehensively and statistically examined which parameters differ significantly from the general stellar population, which is what we intended to do. Instead, most studies have focused on using simulated or observed data to examine a specific factor or correlation between aspects of planetary development. Perhaps the most studied trend has been which aspects of host stars are conducive to gas giant formation with both simulated and observed data (10-13). These studies have demonstrated a positive correlation between metallicity and gas giant formation probability and suggested that gas giants are more likely to form around smaller stars (10-13).

https://doi.org/10.59720/24-018

Some studies have examined the host star population as a whole. For example, Tang et al. examined the properties of host star distributions from observed data (14). They found that stellar metallicity affected gas giant formation rate for stars with smaller surface gravity, metallicity had no effect on the formation of other types of planets, and that the surface gravity of the star had a positive correlation with planet formation for a given orbital period (14). However, unlike our study, Tang et al. examined these properties in isolation rather than comparing them to distributions of general stellar populations to understand defining differences (14). We drew from methods used by previous studies to study large stellar populations, most significantly from Carrillo et al., which aimed to characterize TESS targets to achieve this goal of comparing general and host star populations to understand significant differentiators between the two (15).

In our study, we tested how the various stellar properties mentioned differ between exoplanet host star populations using data from Gaia DR3, LAMOST, and the NASA Exoplanet Archive. We divided our experiment into two main parts: comparing exoplanet host stars to the general galactic stellar population and comparing different subpopulations of host stars based on the type of planet they host. We originally hypothesized that metallicity, kinematics, and stellar classification would all significantly differ between host stars and the general population, as well as between host stars with different types of planets. However, while we found that metallicity and kinematic distributions yielded significant differences between host star subpopulations based on the type of planet, only metallicity revealed significant differences between host stars and the general stellar population. Overall, this study highlighted that the metallicity of stars should be a central component of planet formation models. It also demonstrated that kinematics and stellar classification can reveal important trends in host stars, but we need less biased observational data from exoplanet surveys in order to perform this analysis.

RESULTS

To test our hypothesis that we would observe significant differences between host star populations in stellar chemistry, metallicity, and kinematics, we analyzed data from Gaia, TESS, and the NASA Exoplanet Archive. For each stellar characteristic, we examined trends and statistical features for a random sample of 1,000,000 stars in Gaia DR3, TESS-Gaia crossmatches, host stars, and host star subpopulations based on the type of exoplanet hosted. The TESS-Gaia crossmatches served as our control group; comparing them with the host star distribution revealed whether we could conclude that the observed trends are intrinsic to host stars. In addition, comparing TESS-Gaia data to the random sample revealed whether TESS's target selection was biased for a particular stellar property.

Comparing exoplanet host stars to the general stellar population

We first used H-R diagrams (graphs that provide a way to classify stars' sizes and evolutionary stages) to compare exoplanet host stars to the general stellar population. We performed this analysis by juxtaposing an H-R diagram for a random sample of 1,000,000 stars with a diagram for known exoplanet host stars and TESS-Gaia crossmatches (**Figure 1**). On an H-R diagram, main sequence stars—stars fusing hydrogen into helium in their cores—appear in a diagonal from the top left to bottom right. Giant stars—larger stars towards the end of their evolutionary life cycle—appear in a branch off the main sequence towards the top right (**Appendix**). In our H-R diagram, almost all exoplanet host stars were main sequence stars, but the random sample contained a much larger proportion of giant stars (**Figure 1A, 1B**). In addition, Gaia host stars had a larger percentage of K and M type dwarf stars (the reddest types of main sequence stars) than the random sample of Gaia stars did (**Figure 1A, 1B**). While there was a clear difference between the random Gaia stars and host stars, we also found that there was significant observational bias, i.e., potential bias in the selection sample of surveys. TESS intentionally selects smaller main sequence stars because detecting exoplanets around them is easier, which creates a nonrandom sample of stars and a biased control group (5). When comparing the H-R diagram of exoplanet host stars with that of TESS-Gaia crossmatches, we found the two distributions to be very similar (**Figure 1B, 1C**). Both contained almost completely main sequence stars, implying that TESS stars are more likely to be main sequence than Gaia stars and that the exoplanet trend could result from biased TESS data.

To quantify the trends we observed in the H-R diagrams, we compared the three distributions using absolute magnitude. The average absolute magnitude was 3.06 for the random Gaia sample, 4.94 for the exoplanet host star sample, and 4.22 for TESS-Gaia crossmatches. Thus, exoplanet host stars were dimmer on average and thus less likely to be giant stars or massive main sequence stars, which agreed with what we observed in the H-R diagrams. The difference in average absolute magnitude between exoplanet host stars and Gaia stars was significant (*t*-test: *t*-statistic=21.03, *p*-value<0.001), but the difference between host stars and TESS-Gaia crossmatches, while evident, was not significant enough to rule out observational bias (*t*-test: *t*-statistic=1.892, *p*-value=0.059).

The next astrophysical property we examined was metallicity. For each stellar population's metallicity distribution (MDFs), we graphed the distribution, determined its basic statistical information, and ran Kolmogorov-Smirnov (KS) and Anderson-Darling tests to determine if the distributions were significantly different. We used LAMOST metallicity data to compare LAMOST crossmatches from our random sample with host stars in LAMOST. While the random sample of LAMOST crossmatches appeared to follow a normal distribution of metallicities, exoplanet host star metallicities followed a negatively skewed distribution and were much greater on average (**Figure 2**, **Table 1**). The statistical significance of this difference was confirmed when we ran the statistical tests on 500 stars from each distribution (KS test: *p*-value<0.001, statistic=0.4, Anderson-Darling test: *p*-value<0.001, statistic=124.51).

However, we observed similar observational bias with respect to metallicity in the TESS dataset. The Gaia-LAMOST (i.e., random Gaia sample crossmatched with LAMOST) and TESS-Gaia-LAMOST crossmatch distributions were significantly different (KS test: *p*-value<0.001, statistic=0.346, Anderson-Darling test: *p*-value<0.001, statistic=84.92; **Figure 2, Table 1**). On average, TESS-Gaia-LAMOST crossmatches had higher metallicities than the Gaia-LAMOST crossmatches,

https://doi.org/10.59720/24-018

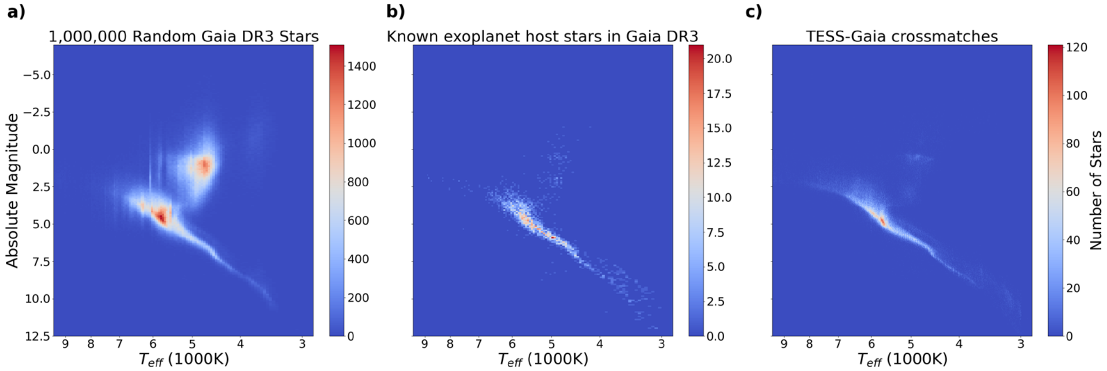

Figure 1. Comparative Hertzsprung-Russell (H-R) diagrams of three stellar samples. Distributions of a) 1,000,000 random Gaia DR3 stars, **b)** Gaia DR3 host stars (Gaia DR3 crossmatched with NASA Exoplanet Archive stars), and **c)** TESS-Gaia crossmatches from our random Gaia selection. T_{eff} (effective surface temperature) refers to the temperature of a star at its surface and gives an indication of the star's color (a hotter star is blue, and a cooler star is red). Absolute magnitude refers to how bright a star is from a set distance away (10 parsecs) (see **Appendix** for a full explanation of an H-R diagram).

indicating that TESS favored higher-metallicity stars in its target dataset (**Figure 2**, **Table 1**). However, the dissimilarity between TESS-Gaia-LAMOST crossmatches and the host star distribution implied that, while TESS exhibited a metallicity-related observational bias, it was likely not solely responsible for the observed differences in metallicity distributions (KS test: *p*-value=0.0059, statistic=0.115, Anderson-Darling test: *p*-value=0.0035, statistic=5.001; **Figure 2**).

Regarding host star kinematics, we examined Toomre diagrams using Gaia data on stars' positions, parallaxes, and velocities (**Figure 3, Appendix**). Graphically, we saw that host stars had a much tighter distribution than the random sample of Gaia stars, but the TESS-Gaia crossmatches also shared a similar distribution (**Figure 3**). To quantify these trends, we classified stars as either residing in the galactic disk or halo (the two main populated regions of the Milky Way; **Appendix**). By splitting the dataset, we achieved a numerical understanding of the percentage of stars in each distribution lying in each division of the Milky Way. We found 7.15% of the stars in our random sample to reside within the galactic halo, 0.082% of exoplanet host stars (2 stars), and 0.27% of TESS-Gaia crossmatches.

To further quantify how stellar kinematics differ in these stellar populations, we examined the distributions of v_{tot} (**Appendix**) across the star samples. We found that the Gaia-TESS crossmatch distribution was very similar to that of the Gaia host stars, leading to observational bias in the data (KS test of 500 stars from each distribution: *p*-value=0.935, statistic=0.034, Anderson-Darling test: *p*-value=0.250, statistic=-0.752). Considering that the distributions of TESS targets and host stars were almost identical, we could not conclude that any difference between host stars and the general stellar population was not due to bias. However, we still observed a significant difference between host stars and the general population with respect to the percentage of stars in the halo, so more research with a less biased dataset would be needed to fully confirm or reject kinematics as a defining trait of host stars.

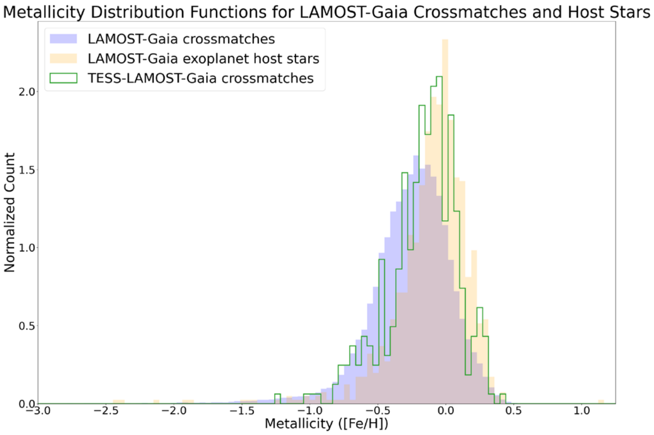

Figure 2. Metallicity distribution [Fe/H] functions. Metallicity ([Fe/H]) gives an indication of how metal-rich a star is, with higher values corresponding to a higher metal abundance. N ≈ 74,000 for Gaia-LAMOST crossmatches, N ≈ 1,000 for Gaia-LAMOST host stars, and N ≈ 400 for TESS-Gaia-LAMOST crossmatches.

Comparing host star subpopulations

Using the exoplanet classifications outlined in Zhu and Dong, we assigned each exoplanet from the NASA Exoplanet Archive a category and used that to classify stars based on the type of exoplanet they host (**Table 2**) (2). As with comparing host stars to our random sample, the first approach we took in comparing the host star subpopulations was by graphing H-R diagrams of each group of stars (**Figure 4**). From the diagram, we saw noticeable differences between the various graphs, but we could not confirm whether these differences were intrinsic to the properties of host stars or to the limitations of exoplanet detection techniques.

Regarding metallicities, we examined the exoplanet

Population	Sample Size	Mean	Standard Deviation	Min	Lower Quartile	Median	Upper Quartile	Max
Random sample	73,385	-0.267	0.353	-5.708	-0.415	-0.229	-0.063	1.272
Exoplanet host stars	963	-0.123	0.398	-4.775	-0.220	-0.063	0.062	1.164
TESS crossmatches	383	-0.174	0.309	-3.519	-0.297	-0.132	0.003	0.440

Table 1: Distribution features of metallicity distribution functions based on stellar population. Since all measurements were taken with LAMOST data, all populations were crossmatched with LAMOST; the measurements that pertain to all stars in each sample listed are also in LAMOST.

https://doi.org/10.59720/24-018

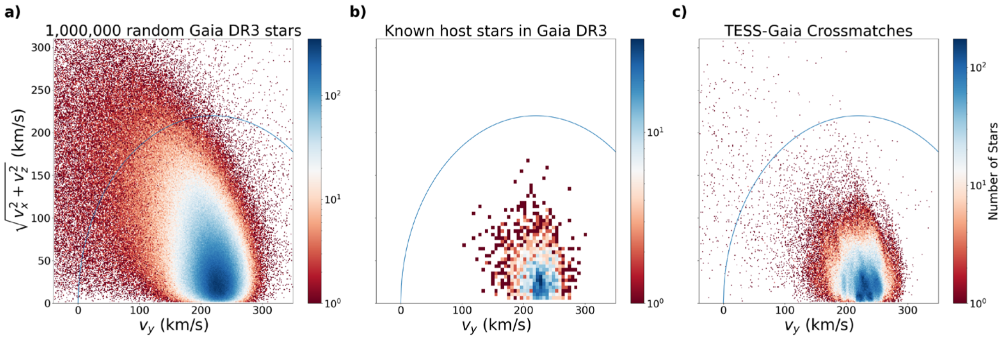

Figure 3. Comparative Toomre diagrams of three stellar samples. Distributions of a) 1,000,000 random Gaia DR3 stars, **b)** Gaia DR3 host stars (Gaia DR3 crossmatched with NASA Exoplanet Archive stars), and **c)** TESS-Gaia crossmatches from our random Gaia selection. On each graph, the blue line represents the thick disk/halo division (a rough divide between stars in the galactic disk and halo) that was proposed by Bonaca et al. (26). The horizontal axis of the diagram plots the tangential velocity of a star around the Milky Way (v_y), and the vertical axis plots the velocity of a star away from the center of the Milky Way (see Appendix for full explanation).

Classifications		
Subpopulation	**Planetary Radius ($R_\oplus$)**	**Orbital Period (days)**
Hot Jupiters	$8 \leq R_p < 20$	P < 10
Cold Jupiters	$8 \leq R_p < 20$	P ≥ 10
Hot Neptunes	$2 < R_p < 8$	P ≤ 4
Cold Neptunes	$2 < R_p < 8$	P > 4
Ultra-short-period planets (USPs)	$0.5 \leq R_p \leq 2$	P ≤ 1
Super Earths	$1 < R_p \leq 2$	P > 1
Small terrestrials	$R_p \leq 1$	P > 1

v_{tot} Distributions (km/s)							
Subpopulation	**Mean**	**Standard Deviation**	**Min**	**Lower Quartile**	**Median**	**Upper Quartile**	**Max**
Hot Jupiters	38.041	19.906	4.146	23.349	36.351	48.725	113.279
Cold Jupiters	38.873	19.652	7.383	22.264	35.493	52.236	90.562
Hot Neptunes	40.818	25.741	2.929	21.631	32.373	51.499	146.862
Cold Neptunes	45.470	24.784	3.501	26.872	41.191	61.472	156.170
USPs	50.618	28.421	7.938	30.969	44.564	61.779	144.912
Super Earths	46.001	24.249	5.983	27.633	41.050	60.544	167.974
Small terrestrials	45.771	26.719	9.112	25.873	42.178	56.844	146.429

Metallicity Distributions ([Fe/H])							
Subpopulation	**Mean**	**Standard Deviation**	**Min**	**Lower Quartile**	**Median**	**Upper Quartile**	**Max**
Hot Jupiters	-0.031	0.387	-2.141	-0.100	0.051	0.185	0.364
Cold Jupiters	-0.066	0.226	-0.517	-0.193	-0.047	0.055	0.286
Hot Neptunes	0.052	0.274	-0.944	-0.014	0.043	0.237	0.323
Cold Neptunes	-0.132	0.381	-4.775	-0.210	-0.074	0.027	0.409
USPs	-0.124	0.551	-2.375	-0.175	-0.028	0.125	0.301
Super Earths	-0.159	0.403	-3.860	-0.257	-0.094	0.042	0.357
Small terrestrials	-0.129	0.251	-1.051	-0.244	-0.107	0.037	0.235

Table 2: Table of host star subpopulation classifications as well as distribution features for metallicity ([Fe/H]) and v_{tot} (km/s) distributions. The classifications are the same as those used by Zhu and Dong, but we added a "small terrestrials" category to account for smaller terrestrials with longer orbital periods (such as Mercury and Venus in our own solar system) (2). $R_\oplus$ is the radius of the Earth. We used LAMOST data on metallicities and Gaia kinematic data to gather the data for the distributions.

subpopulations' metallicity distribution functions and ran KS tests to quantitatively compare them (**Figure 5**). In addition, we also determined basic statistical information for all the distributions (**Table 2**). We saw that stars with large, hot planets had significantly higher average metallicities and generally very different metallicity distributions compared to stars with terrestrial planets. When comparing those types of distributions, we found multiple significant differences (**Figure 5**). For example, when comparing hot Jupiters with small terrestrials, cold Jupiters, and cold Neptunes (opposites in size or distance from the host star), the hot Jupiter host stars had significantly higher metallicities in all three comparisons. By contrast, colder gas giant host stars shared very similar metallicities to terrestrial planet host stars (**Table 2**). Regarding our original hypothesis, we found that metallicity was a significant differentiator between host star subpopulations.

We then graphed the distributions of stellar kinematics for host star subpopulations through Toomre Diagrams (**Figure 6**). We saw that the subpopulations mostly followed very similar distributions, but we did notice that in a few cases, as with metallicity, distributions were significantly different. When we ran KS tests, all differences were nonsignificant

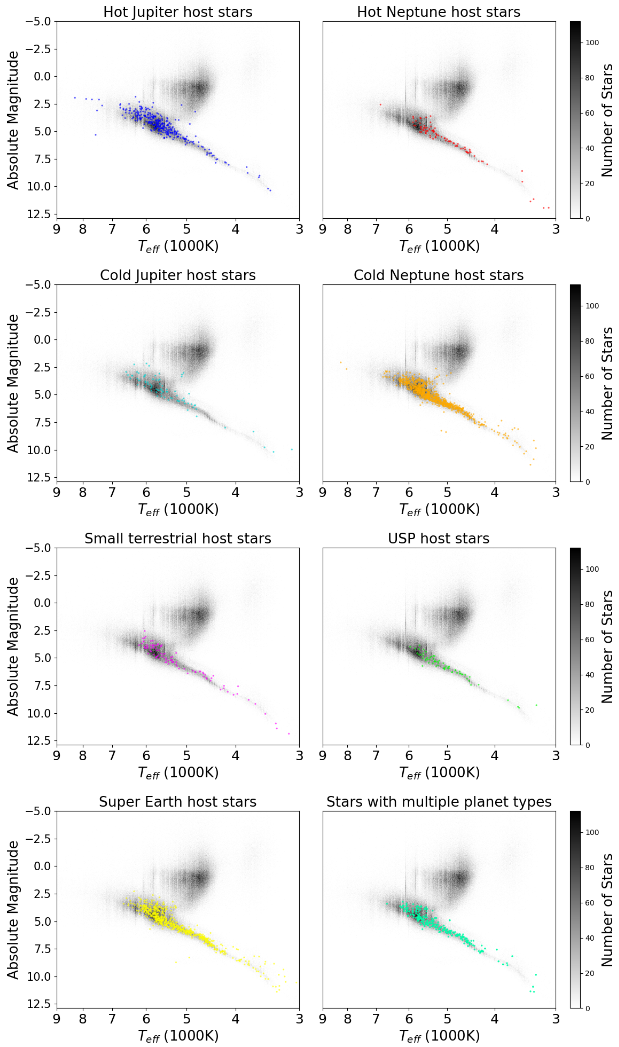

Figure 4: H-R diagrams of exoplanet host star subpopulations scattered on top of the H-R diagram for the random sample of Gaia stars. Effective surface temperature (T_{eff}) gives an indication of the star's color by calculating the temperature on the surface of the star. A hotter star is bluer, and a cooler star is redder. Absolute magnitude is a measure of how bright a star is from a fixed distance away (10 parsecs) (see **Appendix** for a full explanation of an H-R diagram).

except for when hot Jupiter host stars were compared to USP (p-value=0.008), super Earth (p-value<0.001), small terrestrial (p-value=0.037), and cold Neptune (p-value<0.001) host stars and when hot Neptune host stars were compared to USP (p-value=0.021), super Earth (p-value=0.015), and cold Neptune (p-value=0.028) host stars. In addition, the subpopulations were divided into two groups based on the standard deviations of their distributions—hot and cold Jupiters had standard deviations between 19 and 20 km/s, while the rest of the subpopulations had standard deviations between 24 and 29 km/s (**Table 2**). As with metallicity, we saw that the means and standard deviations of the distributions had some trends with respect to planet radius and distance from the host star. Stars with large-radius planets generally had smaller means and standard deviations than stars with smaller planets (e.g., comparing hot Jupiters with USPs), and stars with more distant planets had higher means than stars with closer planets (e.g., comparing hot Neptunes with cold Neptunes) (**Table 2**). We concluded that stellar kinematics was able to significantly differentiate host star subpopulations, supporting our original hypothesis.

DISCUSSION

We analyzed data from Gaia, LAMOST, and the NASA Exoplanet Archive to compare both host stars with the general stellar population as well as different host star subpopulations. We hypothesized that we would observe significant differences in stellar chemistry, classification, and kinematics. Regarding stellar classification, we found that currently discovered exoplanets primarily orbit dim main sequence stars, whereas the general stellar population contained a much larger proportion of giant stars. However, there was significant observational bias associated with this result. We saw that TESS surveyed very few giant stars, which could have been one of the reasons we did not observe many exoplanets around them. In addition, we noted that planets are harder to observe when orbiting larger stars using transit photometry or radial velocity, which could also lead to the observed lack of giant host stars (3).

We also found that H-R diagrams differentiated host star subpopulations through certain trends, such as how stars tended to be dimmer when the orbital period and planet radius decreased. However, as with comparing host stars to the general stellar population, some trends could also be a result of bias from detection technique limitations. For example, we observed that smaller planets tended to orbit dimmer and redder stars (e.g., USP, super Earth, and small terrestrial vs. hot Jupiter, cold Jupiter, and cold Neptune host star H-R diagrams). While smaller dwarf stars may be more suitable for smaller terrestrial planets, exoplanets around these stars are much more sensitive to measurement techniques such as transit photometry and radial velocity, which could result in a biased selection (3). This difference in sensitivity arises because exoplanet transits cause proportionately greater dips in light relative to a star's luminosity for smaller stars. Similarly, an orbiting planet would cause a smaller star to "wobble" more, causing a higher doppler shift for radial velocity methods (3). As a result, a small planet around a large main sequence star might simply not be detectable, perhaps explaining why we could see this trend. Overall, we could not draw conclusive results as to the differences between the populations because of the potential bias in the data arising from detection limitations.

Stellar kinematics also differentiated exoplanet host stars from the general stellar population. While the random sample of Gaia stars had ≈7% of stars in the galactic halo, there were only two exoplanet host stars (0.082% of the host star population). One explanation for this difference is the Milky Way's chemical composition, as the thin disk has significantly more metal-rich stars than the thick disk or halo. Therefore, much of what we see kinematically could be the result of underlying chemical properties in stars based on location (16). However, as with the H-R diagrams, we had to account for bias from TESS, as it surveyed very few halo stars (0.27%). Within exoplanet host star subpopulations, stellar kinematics also highlighted differences between distributions; the v_{tot} distributions divided the host star subpopulations into two main groups.

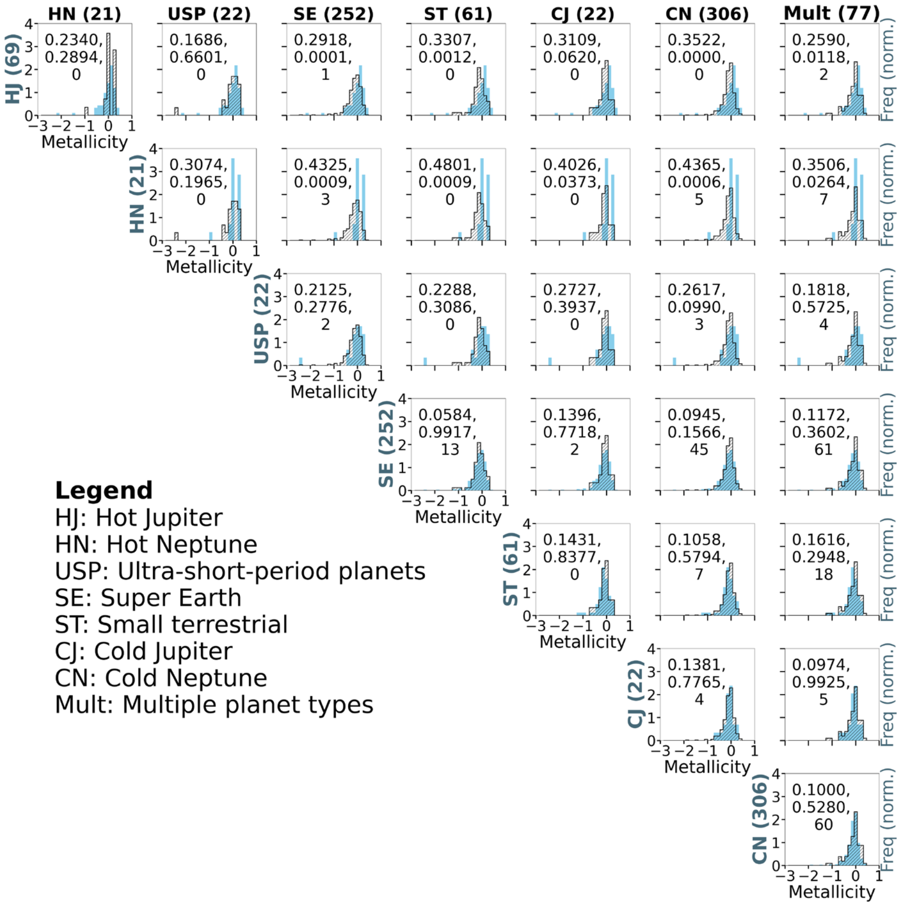

Figure 5: Metallicity distribution functions of exoplanet host star subpopulations. Plots are labeled by the type of planet stars have. The corner plot pairs every combination of two subpopulations together and graphs the metallicity distribution functions on top of each other, as well as providing KS test results and sample sizes. Metallicity ([Fe/H]) provides an indication of how metal-rich a star is, with a larger value reflecting a higher metal abundance. For the text inside each subplot, the first and second numbers refer to the statistic and *p*-value given by a KS test run on the two populations' metallicity distributions, respectively. The third number refers to the number of stars in our sample that host both types of planets.

Finally, metallicity was a significant differentiator between the different populations of stars in our study. Exoplanet host stars were significantly more metal-rich than the general stellar population. While we saw bias from TESS, the KS and Anderson-Darling tests suggested that TESS's metallicity distribution was still significantly different from that of host stars, indicating that the trend is intrinsic to host stars. This observation agrees with previous research regarding stellar metallicity to planet frequency correlations: planet frequency increases with host star metallicity since more metallic stars have more solid materials available for planet development (17).

Examining the similarity of metallicities of exoplanet host star subpopulations with KS tests revealed relationships between different types of host stars. We found that metallicities were similar for similar types of planets (e.g., hot Jupiters and hot Neptunes) and planets that were likely to be companion planets or formed around similar stars. An example of the latter case is outer gas giants (cold Neptunes or cold Jupiters) and small terrestrials. Our collection of data from the NASA Exoplanet Archive showed that 22.5% of all known cold Neptune planets orbit a star with an inner terrestrial companion (USP, super Earth, or small terrestrial), indicating that planetary systems with an outer giant and inner terrestrial are common, which could explain the similar metallicity distribution between the two types of planets' host stars. We found that a larger planet radius and shorter orbital period correlated with greater metallicities, which is in accordance with the core-accretion model of giant planet formation (18). In the core-accretion model, gas giant planets initially develop their inner core from the gradual coalescence of small solid particles orbiting a star, and this core begins to "accrete" gas once it has become sufficiently large (18). The higher abundance of metal in stars would lead to more solid particles that could develop into gas giants, and a higher concentration of particles would allow the gas giant cores to grow larger, allowing them to accrete more gas and have a larger planet radius.

Overall, examining H-R diagrams and stellar kinematics highlighted certain trends in exoplanet host stars but contained bias from TESS's target selection. By contrast, we discovered that metallicity was a very significant differentiator between exoplanet host stars and the general stellar population, as well as different host star subpopulations. This study can serve as a reference to future researchers focusing on the properties of exoplanet host star populations and can help test planet formation models. To further examine the significance of our data, we also recommend analyzing Kepler observational bias instead of just TESS bias. Since Kepler primarily observed just

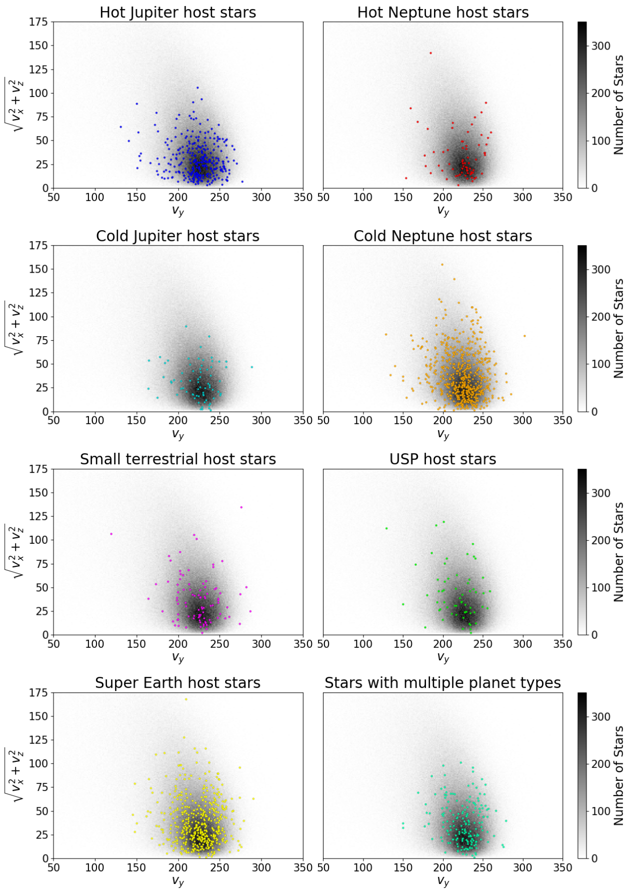

Figure 6: Toomre diagrams of exoplanet host star subpopulations scattered on top of Toomre diagram for random sample of Gaia stars. The horizontal axis of the diagram plots the tangential velocity of a star around the Milky Way (v_y), and the vertical axis plots the velocity of a star away from the center of the Milky Way (see Appendix for full explanation).

one portion of the sky, using its data could lead to less bias in stellar chemistry as it may not have been as selective of the evolutionary stage and metallicity of stars it observed (19). As more and more data become available, we recommend examining stellar kinematics more with complete thin disk, thick disk, halo, and bulge divisions (instead of just the disk/halo division used in this study), as well as improving our data on stellar chemistry with larger sample sizes from future data and even more confirmed exoplanets.

METHODS

Data and methodology

To obtain a random sample of stars as the general stellar population, we selected 1,000,000 random Milky Way stars from Gaia DR3 with Renormalised Unit Weight Error (RUWE, an indication of how much error is present overall for a star's measurements) < 1.2 and data on celestial position, parallax, proper motion, radial velocity, bp-rp (how blue or red a star is), and effective surface temperature. In addition, we used the NASA Exoplanet Archive (as of July 18, 2023) to obtain the Gaia source ids for exoplanet host stars. We also used the archive to classify the exoplanet host stars by the types of exoplanets they hosted, using categories outlined in Zhu and Dong (**Table 1**) (2). After classifying the exoplanets, we organized the Gaia host stars into 7 subpopulations based on these classifications, which had some overlap due to some stars having more than one type of exoplanet. To account for the overlapping stars, we added an 8th subpopulation of host stars, which consisted of stars with more than one type of exoplanet (e.g., a super Earth and a cold Neptune). In addition, we removed binary host stars from our dataset, as we wanted to focus on only single-star systems.

To make the comparisons for our study, we used six main datasets: a random sample of Gaia stars, TESS-Gaia crossmatches, known exoplanet host stars, Gaia-LAMOST crossmatches, TESS-Gaia-LAMOST crossmatches, and exoplanet host stars in LAMOST. When comparing the general population to exoplanet host stars, we compared the random sample dataset (general population) with the known host star dataset. Then, to verify that our findings were not due to observational bias—selection bias in surveys that can cause a nonrandom control group—we compared these two datasets with TESS-Gaia crossmatches. Since TESS is NASA's current flagship exoplanet detection mission, bias in the host stars it surveys could have an effect on the perceived trends of exoplanet host stars, making the additional comparison necessary. The bias affecting our results arises from two main reasons: detection technique limitations and intentionally biased target choices. The two main exoplanet detection techniques—transit photometry and Doppler measurements—are much more sensitive when monitoring less luminous and less massive stars (3). Thus, known host stars may be more likely to be small main sequence stars simply because of the limitations of our detection techniques. As a result of these limitations, TESS also intentionally surveys smaller stars, which further contributes to the bias (5). We crossmatched TESS targets with Gaia targets so that we could use the same measurements (Gaia measurements) for both datasets, ensuring that there is not a discrepancy in the measurements due to instrument differences. When examining metallicity trends, we followed the same principle of having a random sample, host star, and TESS target group, but instead of using Gaia metallicities, we used LAMOST metallicities because they are derived from higher resolution equipment (20-22). As such, we then crossmatched each dataset (random sample, host star, and TESS-target) with LAMOST in order to use LAMOST metallicities to compare the three. Our crossmatching was either performed with Gaia's built-in ADQL server or Topcat (23-25).

In the first part of our study (comparing host stars to the general stellar population), we used these datasets to compare our random sample of 1,000,000 stars in Gaia DR3 (Data Release 3) with all known exoplanet host stars using graphs and statistical tests. We then compared both datasets to TESS-Gaia crossmatches to determine if there was significant observational bias from TESS. In the second part of the study, we divided exoplanet host stars into overlapping populations based on the type of planet they hosted and compared those subpopulations. In both cases, we aimed to test whether the stellar properties differed between populations enough to be considered significant. The three main diagrams we used to differentiate the outlined populations of stars are Hertzsprung–Russell (H-R) diagrams, Toomre diagrams, and Metallicity Distribution Functions (MDFs) (**Appendix**).

When examining stellar kinematics, one method we used to obtain a quantitative measurement of kinematic distributions was to classify each star as either residing in the galactic disk or halo (the two main populated regions of the Milky Way, **Appendix**). We did this by dividing the diagram using a disk/halo division, a method to discern between disk and halo stars, proposed by Bonaca et al. (26). We identified halo stars as having a total galactocentric velocity v_{tot} with $|v_{tot}| > 220$km/s (the blue line in **Figure 3**, **Appendix**) (26).

Statistical analyses

Throughout this study, we used basic statistical features (e.g. mean, median, standard deviation, quartiles) to compare two distributions of data and advanced statistical tests (KS tests, Anderson-Darling tests, and occasionally *t*-tests) to determine if distributions were significantly different. The KS test measures the maximum distance between the empirical cumulative distribution functions of two data samples, which is then converted to a *p*-value, which gives the probability that the two samples are from the same distribution (30, 31). The Anderson-Darling test operates in the same way but places more emphasis on the tails of the distributions (32). On the other hand, *t*-tests use the standard deviation and sample sizes of two distributions to determine their means are significantly different (33). Throughout the entire study, we used a *p*-value cutoff of 0.05 to determine significance. We also reduced all sample sizes to 500 samples to both standardize our sample size and avoid the tests being overly sensitive (**Appendix**).

Programming script

Our code for our analysis (as a Jupyter notebook) and the survey data used for the study are included in the following link: github.com/daniel20082061/exoplanet_research_project.git.

Received: January 7, 2024
Accepted: May 28, 2024
Published: October 8, 2025

REFERENCES

1. "NASA Exoplanet Archive." *California Institute of Technology*. https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS. Accessed 6 Jan. 2024.
2. Zhu, W. and S. Dong. "Exoplanet Statistics and Theoretical Implications." *Annual Review of Astronomy and Astrophysics*, vol.59, no.1, Sep. 2021, pp. 291-336, https://doi.org/10.1146/annurev-astro-112420-020055.
3. Fischer, D. A., et al. "Exoplanet Detection Techniques." *Protostars and Planets VI*, Jan. 2014, https://doi.org/10.48550/arXiv.1505.06869.
4. "5 Ways to Find a Planet." *NASA*. exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/?intent=021. Accessed 7 Jan. 2024.
5. Richer, G. R., et al. "Transiting Exoplanet Survey Satellite." *Journal of Astronomical Telescopes, Instruments, and Systems*, vol.1, Jan. 2015, p. 014003, https://doi.org/10.1117/1.JATIS.1.1.014003.
6. Kubyshkina, D. and A. Vidotto. "How does the mass and activity history of the host star affect the population of low-mass planets?" *Monthly Notices of the Royal Astronomical Society*, vol.504, no.2, Mar. 2021, pp. 2034-2050, https://doi.org/10.1093/mnras/stab897.
7. Hourihane, A., et al. "The Gaia-ESO Survey: Homogenisation of stellar parameters and elemental abundances." *Astronomy&Astrophysics*, vol.676, 2023, p. A129, https://doi.org/10.1051/0004-6361/202345910.
8. Yan, H., et al. "Overview of the LAMOST survey in the first decade." *The Innovation*, vol.3, no.2, Mar. 2022, p. 100224, https://doi.org/10.1016/j.xinn.2022.100224.
9. Adibekyan, V., et al. "Characterization of Exoplanet-Host Stars." *Asteroseismology and Exoplanets: Listening to the Stars and Searching for New Worlds*, vol.49, Jul. 2017, pp. 225-238, https://doi.org/10.1007/978-3-319-59315-9_12.
10. Ida, S. and D. Lin. "The formation and retention of gas giant planets around stars with a range of metallicities." *The Astrophysical Journal*, vol.616, no.1, Nov. 2004, pp. 567-572, https://doi.org/10.1086/424830.
11. Kornet, K., et al. "Formation of giant planets around stars with various masses." *Astronomy&Astrophysics*, vol.458, no.2, Nov.2006, pp. 661-668, https://doi.org/10.1051/0004-6361:20053689.
12. Osborn, A. and D. Bayliss. "Investigating the planet–metallicity correlation for hot Jupiters." *Monthly Notices of the Royal Astronomical Society*, vol.491, no.3, Jan. 2020, pp. 4481–4487, https://doi.org/10.1093/mnras/stz3207.
13. Debra, A. and V. Valenti. "The Planet-Metallicity Correlation." *The Astrophysical Journal*, vol.622, Apr. 1, 2005, pp.1102-1117, https://doi.org/ 10.1086/428383.
14. Tang, Y., et al. "The Statistical Analysis of Exoplanet and Host Stars Based on Multi-Satellite Data Observations." *Universe*, vol.10, no.4, Apr. 2024, p. 182, https://doi.org/10.3390/universe10040182.
15. Carrillo, A., et al. "Know thy star, know thy planet: chemo-kinematically characterizing TESS targets." *Monthly Notices of the Royal Astronomical Society*, vol.491, no.3, Jan. 2020, pp.4365-4381, https://doi.org/10.1093/mnras/stz3255.
16. Bensby, T., et al. "The nature of the metal-rich thick disk." *From Stars to Galaxies: Building the Pieces to Build up the Universe ASP Conference Series*, vol.374, no.2, 2007, pp.181-186, https://doi.org/10.48550/arXiv.astro-ph/0612459.
17. Mortier, A., et al. "On the functional form of the metallicity-giant planet correlation." *Astronomy&Astrophysics*, vol.551, Mar. 2013, p. A112, https://doi.org/10.1051/0004-6361/201220707.
18. Wang, J. and D. Fischer. "Revealing a Universal Planet-Metallicity Correlation for Planets of Different Sizes around Solar-type Stars." *The Astronomical Journal*, vol.149, no.1, Jan. 2015, p.14, https://doi.org/10.1088/0004-6256/149/1/14.
19. "Kepler Field of View." *NASA*. https://science.nasa.gov/resource/kepler-field-of-view/. Accessed 7 Jan 2024.
20. Niu, Z., et al. "Internal Calibration of LAMOST and Gaia DR3 GSP-Spec Stellar Abundances." *The Astrophysical Journal*, vol.950, no.2, Jun. 2023, p.104, https://doi.org/10.3847/1538-4357/accf8b.
21. Gaia Collaboration, et al. "Gaia Data Release 3." *Astronomy&Astrophysics*, vol.674, p.A1, Jun. 2023, https://doi.org/10.1051/004-6361/202243940.
22. Andrae, R. "11.3.3 General Stellar Parametrizer from Photometry (GSP-Phot)." *European Space Agency*. https://gea.esac.esa.int/archive/documentation/GDR3/

Data_analysis/chap_cu8par/sec_cu8par_apsis/ssec_cu8par_apsis_gspphot.html. Accessed 7 Jan 2024.

23. Gaia Collaboration et al. “The Gaia mission.” *Astronomy&Astrophysics*, vol.595, Nov. 2016, p. A1, https://doi.org/10.1051/0004-6361/201629272.
24. Babusiaux, C., et al. “Gaia Data Release 3: Catalogue Validation.” *Astronomy&Astrophysics*, vol.674, Jun. 2023, p. A32, https://doi.org/10.1051/0004-6361/202243790.
25. Taylor, M. “TOPCAT & STIL: Starlink Table/VOTable Processing Software.” *Astronomical Data Analysis Software and Systems XIV*, vol.347, Dec. 2005, pp.29-33.
26. Bonaca, A., et al. “Gaia reveals a metal-rich in-situ component of the local stellar halo.” *The Astrophysical Journal*, vol.845, no.2, Aug. 2017, p.101, https://doi.org/10.48550/arXiv.1704.05463.
27. HOW BIG, WARM, OLD, ... ARE THE STARS? GAIA’S STELLAR PARAMETERS.” *European Space Agency.* https://www.cosmos.esa.int/web/gaia/dr3-how-big-or-warm-or-old-are-the-stars. Accessed 6 Jan. 2024.
28. Moore, R. “Metallicity of stars.” *Institute for Computational Cosmology.* https://icc.dur.ac.uk/~tt/Lectures/Galaxies/TeX/lec/node27.html. Accessed 11 Aug. 2024.
29. “Definitions of Magnitudes and Surface Brightness.” *Australia Telescope National Facility*. https://www.mso.anu.edu.au/~geoff/HEA/Magnitudes. Accessed 11 Aug. 2024.
30. “Kolmogorov-Smirnov Goodness-of-Fit Test.” *National Institute of Standards and Technology.* https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm. Accessed 10 Mar. 2024.
31. Lanzante, J. R. “Testing for differences between two distributions in the presence of serial correlation using the Kolmogorov–Smirnov and Kuiper’s tests.” *International Journal of Climatology*, vol.41, no.14, 2021, pp. 6314-6323, https://doi.org/10.1002/joc.7196.
32. “Anderson-Darling Test.” *National Institute of Standards and Technology*. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm. Accessed 10 Mar. 2024.
33. “Two-Sample t-Test for Equal Means.” *National Institute of Standards and Technology*. https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm. Accessed 10 May. 2024.
34. Mohd Razali, N. and Y. Wah. “Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests.” *Journal of Statistical Modeling & Analytics*., vol.2, pp.21-33, Jan. 2011.

Appendix A: Explanation of Scientific Measurements, Diagrams, Statistical Test Preparation, and Smearing Correction

The total galactocentric velocity v$_{tot}$ is defined as $v_{tot} = \sqrt{v_x^2 + \left(v_y - v_{y_{\mathrm{LSR}}}\right)^2 + v_z^2}$ (**Appendix**) and gives an object's velocity relative to the Local Standard of Rest, or the average movement of stars traveling around the Milky Way (15). $v_{y_{\mathrm{LSR}}}$ is equal to about 220 km/s and describes the average tangential velocities of stars around the Milky Way (15, 26).

A star's metallicity [Fe/H] is defined as $[\mathrm{Fe/H}] = log_{10}\left(\frac{N_{\mathrm{Fe}}}{N_{\mathrm{H}}}\right)_{\mathrm{star}} - log_{10}\left(\frac{N_{\mathrm{Fe}}}{N_{\mathrm{H}}}\right)_{\mathrm{sun}}$, where N_{Fe} is the number density of iron and N_{H} is the number density of hydrogen. The result is that [Fe/H] gives a logarithmic measurement of how much greater the iron-to-hydrogen ratio $\frac{N_{\mathrm{Fe}}}{N_{\mathrm{H}}}$ is compared to that of the sun. An [Fe/H] value of 1 means that a star has ten times more iron compared to its hydrogen than the sun does (28).

The magnitude scale provides a logarithmic measurement of a stars' brightness. A magnitude of 0 is a set brightness, equivalent to approximately how bright the star Vega appears from Earth. An increase in magnitude corresponds to a decrease in the measured brightness. Absolute magnitude, while still a measure of brightness, gives an indication of a stars' luminosity (J/s, the amount of electromagnetic energy radiated by a star per unit time) by fixing the observer distance to 10 parsecs away from the star. Its value can be derived using a stars' brightness from Earth along with the distance to the star (29).

H-R Diagram

An H-R diagram graphs either the color or surface temperature of the star on the horizontal axis and the absolute magnitude on the vertical axis. It allows astronomers to classify stars based on trends in the diagram. We can divide stars into spectral classes based on the surface temperature or color of the star (e.g. K and M, the reddest spectral classes) as well as group stars into "sequences" (trends on the H-R diagram that indicate the brightness and evolutionary stage of a star). The largest sequence, or main sequence, contains stars that are fusing hydrogen to helium inside their cores and are in their main stage of life. When a star approaches the end of its lifetime, it fuses heavier elements and swells, becoming brighter and redder in the process. As a result, it appears on a separate branch on an H-R diagram — the giant branch — which breaks of the main sequence and contains brighter and redder stars.

Toomre Diagram

A Toomre diagram is one of the most popular ways to view kinematic velocities of stars. For every Milky Way star, we can assign three Galactocentric velocities that describe its motion relative to the center of the Milky Way: v_x, the velocity away from the galactic center parallel to the galactic plane; v_y, the tangential velocity, or the linear velocity of a star as it is revolving around the Milky Way center; and v_z, the velocity away from the galactic center perpendicular to the galactic plane. In a Toomre diagram, we graph v_y on the horizontal axis and $\sqrt{v_x^2 + v_z^2}$ (the magnitude of the sum of v_x, and v_z) on the vertical axis.

Visualizations of galactic velocities such as the Toomre diagram allow us to understand where stars lie inside our Milky Way galaxy. The Milky Way is divided into four main regions: the galactic bulge, the thin disk, the thick disk, and the halo. The galactic bulge refers to stars close to the center of the Milky Way that usually have a very tight distribution of velocities. Thin disk stars are very close to the galactic plane, have a tight kinematic distribution, and are very metal-rich compared to thick disk stars. The galactic halo spans around and above the galaxy and contains stars in the galaxy's outer reaches; these stars have very different velocities relative to each other.

Metallicity Distribution Functions (MDFs)

In addition, Metallicity Distribution Functions can be used to demonstrate the distribution of stellar metallicities in a stellar population. The metallicity distribution contains values of metallicities for a certain dataset of stars, and the graphical element of MDFs is a histogram of these distributions.

Statistical Test Preparation

When comparing two 1D distributions with Kolmogorov-Smirnov and Anderson-Darling tests, we faced another limitation due to large sample sizes. Both tests tend to become extremely sensitive with larger sample sizes, as even a small difference in two distributions can cause a significant effect (34). This could negatively affect our results because even a minuscule difference between distributions could cause the test to reject the null hypothesis. Additionally, not having a standard sample size means that the tests could be more sensitive to certain distribution comparisons over others. While we wanted to reduce the sample size, we also wanted to avoid too small a sample size, as then we would risk the selected data points not being an accurate representation of the entire dataset. Choosing the exact right sample size is impossible, as there is no empirical way to determine the best sample size since results can differ based on the dataset being used. As a result, we used a sample size of 500 since it was the smallest data size we were confident would not risk misrepresentation. At a sample size of around 500, simulated power measurement (the probability of rejecting the null hypothesis when it is false) for both KS and Anderson-Darling exceeds 0.9 (for Laplace distributions), which is much higher than the conventional standard of 0.8 for research, allowing us to conclude it is most likely sufficiently high to avoid a high risk of misrepresentation (34).

Smearing Correction

One significant limitation of the data provided in the surveys that must be corrected for analysis is the "smearing" Milky Way stars can experience from gas clouds in their line of sight; stars' color (bp-rp) can be affected by these gas clouds, giving an inaccurate measurement. To bypass this problem, we used effective surface temperature (T_{eff}) in our H-R diagrams instead of bp-rp, as Gaia T_{eff} measurements calculate a star's intrinsic color before deriving T_{eff}, providing a more accurate H-R diagram (27).