The Astronomical Journal, 161:231 (12pp), 2021 May https://doi.org/10.3847/1538-3881/abee1d

The Stars Kepler Missed: Investigating the Kepler Target Selection Function Using Gaia DR2

Linnea M. Wolniewicz[1], Travis A. Berger[2], and Daniel Huber[2]
[1] Department of Astrophysical and Planetary Sciences, University of Colorado, 2000 Colorado Avenue, Boulder, CO 80305, USA
[2] Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Drive, Honolulu, HI 96822, USA
Received 2021 January 8; revised 2021 February 20; accepted 2021 March 10; published 2021 April 23

Abstract

The Kepler Mission revolutionized exoplanet science and stellar astrophysics by obtaining highly precise photometry of over 200,000 stars over 4 yr. A critical piece of information to exploit Kepler data is its selection function, since all targets had to be selected from a sample of half a million stars on the Kepler CCDs using limited information. Here we use Gaia DR2 to reconstruct the Kepler selection function and explore possible biases with respect to evolutionary state, stellar multiplicity, and kinematics. We find that the Kepler target selection is nearly complete for stars brighter than $Kp < 14$ mag and was effective at selecting main-sequence stars, with the fraction of observed stars decreasing from 95% to 60% between $14 < Kp < 16$ mag. We find that the observed fraction for subgiant stars is only 10% lower, confirming that a significant number of subgiants selected for observation were believed to be main-sequence stars. Conversely we find a strong selection bias against low-luminosity red giant stars ($R \approx 3$–$5R_{\odot}$, $T_{\rm eff} \approx 5500$ K), dropping from 90% at $Kp = 14$ mag to below 30% at $Kp = 16$ mag, confirming that the target selection was efficient at distinguishing dwarfs from giants. We compare the Gaia Re-normalized Unit Weight Error (RUWE) values of the observed and nonobserved main-sequence stars and find a difference in elevated (>1.2) RUWE values at $\sim\sigma$ significance, suggesting that the Kepler target selection shows some bias against either close or wide binaries. We furthermore use the Gaia proper motions to show that the Kepler selection function was unbiased with respect to kinematics.

Unified Astronomy Thesaurus concepts: Astronomy databases (83); Astronomy data analysis (1858); Multiple stars (1081); Exoplanet astronomy (486)

Supporting material: machine-readable table

1. Introduction

The Kepler mission (Borucki et al. 2010), officially retired in 2018, has left behind a legacy data set for stellar astrophysics and exoplanet science. One of the biggest breakthroughs enabled by Kepler was our understanding of exoplanet occurrence rates as a function of planet size, orbital period, and stellar type. For example, many planets observed around Kepler host stars have been found to have sizes between Earth and Neptune (Howard et al. 2012), a population that is absent in our own solar system. Dressing & Charbonneau (2013) found that for the M dwarf stars in the Kepler sample, the Earth-sized (0.5–1.4 $R_{\oplus}$) planetary occurrence rate is 0.51 planets per star for orbital periods less than 50 days, significantly higher than the 0.26 planetary occurrence rate found by Petigura et al. (2013) for Earth-sized planets around solar-type stars with orbital periods between 5 and 100 days. A large number of studies have since explored planet occurrence as a function of orbital period, planet size, and stellar spectral type using the Kepler sample (Youdin 2011; Dong & Zhu 2013; Foreman-Mackey et al. 2014; Burke et al. 2015; Garrett et al. 2018; Kopparapu et al. 2018; Mulders et al. 2018; Hsu et al. 2019; Pascucci et al. 2019; Zink & Hansen 2019; Bryson et al. 2020; Kunimoto & Matthews 2020). A complicating factor for many of these studies is the presence of stellar companions to Kepler targets (Adams et al. 2012; Lillo-Box et al. 2012; Dressing et al. 2014; Law et al. 2014; Baranec et al. 2016; Furlan et al. 2017; Ziegler et al. 2018), which can have significant effects on exoplanet demographics both by biasing planet radii (Teske et al. 2018) and through astrophysical effects such as the suppression of planet formation (Kraus et al. 2016). In addition to exoplanet demographics, a number of studies have used asteroseismology of red giants to explore stellar populations in the Kepler field (Miglio et al. 2013; Pinsonneault et al. 2014; Sharma et al. 2016).

A critical piece of information for Kepler exoplanet and stellar population studies is the process through which targets were selected. For example, most planet occurrence rate studies have so far assumed that the Kepler target selection function is unbiased with respect to stellar multiplicity (Murphy et al. 2018). However, Kepler was forced to select targets, as only 200,000 stars could be observed over the course of the mission. Previous attempts to recreate the Kepler target selection method found that binary stars were selected at similar rates as single-star systems and that the MS dwarf population was underestimated (Farmer et al. 2013). However, Farmer et al. (2013) reconstructed the Kepler selection function by creating a synthetic Kepler Input Catalog (defined below) using population synthesis tools and following the steps detailed in Batalha et al. (2010) to select stars for observation. The limitations of this study are that their conclusions were formed upon a synthetic population that may not accurately represent the stellar population in the Kepler field of view. As such, many of the underlying assumptions and biases of the selection function remain unexplored.

The basis for the Kepler target selection was the Kepler Input Catalog (KIC), which contains physical properties and photometric data for sources in the Kepler field of view (Brown et al. 2011). The primary goal of the KIC was to distinguish cool dwarf stars from red giants, with an expected reliability rate of 98% (Brown et al. 2011). The KIC used broadband

photometry to infer stellar parameters for all of their stars; however, the log(g) values were imprecise as they were only constrained by photometry using the D51 filter. Kepler selected the optimal targets for observation using the KIC, with a goal of selecting solar type stars that could host terrestrial-sized planets (Batalha et al. 2010). The highest priority targets were solar-type stars where it would be possible to detect an Earth-sized planet in the habitable zone (HZ). The next criterion of the selection process was to include stars that were brighter than fourteenth magnitude in the Kepler passband (Kp). The next targets were the brightest stars where it would be possible to detect an Earth-sized planet in the HZ, even if they were fainter than $Kp = 14$ mag. Finally, the criterion for detectable planets in the HZ was relaxed, allowing stars that would benefit from additional transit data to be observed. This created a list of 261,363 stars brighter than sixteenth magnitude in the Kepler passband, which was reduced to less than 200,000 stars due to mission constraints (Batalha et al. 2010).

The recent Gaia second data release (DR2) now provides a unique opportunity to look back at the Kepler target sample and better understand its selection function (Gaia Collaboration et al. 2018). In particular, Gaia DR2 has provided high-precision parallaxes for a total of 1,692,919,135 sources (Lindegren et al. 2018) and includes nearly all the stars in the Kepler field of view, including those not observed by Kepler. The Gaia DR2 parallaxes can be used to vastly improve the properties of stars in the Kepler field (Berger et al. 2018).

With Gaia DR2 we can now conduct a detailed investigation of the Kepler target selection function. In particular, we aim to (1) determine the degree to which Kepler's target selection matches the mission's priorities and (2) whether the selection of targets was biased with respect to stellar multiplicity. Understanding any potential biases in the selection function has important implications for exoplanet science and any future determinations of planetary occurrence rates using the Kepler target sample.

2. Methodology

2.1. Catalog Cross-matching

We started with a subset of 2.4 million targets within the KIC that are located in the Kepler field of view, which we downloaded from the Mikulski Archive for Space Telescopes (MAST).[3] We focus on Kepler stars with $Kp < 16$ mag, as $\approx 97\%$ of the Kepler targets were below this threshold (Batalha et al. 2010).

As a first step, we cross-matched the KIC with Gaia DR2 to obtain Gaia information for each star in the KIC. To do this, we used the Centre de Données astronomiques de Strasbourg (CDS) cross-match.[4] This service is provided by the Université de Strasbourg and joins any VizieR data, in this case Gaia DR2, with a private data table based on the R.A. and decl. of the stars.

We conducted a positional match with a matching radius of 5 arcseconds. We chose 5″ because the astrometric offsets between the KIC and Gaia have not been well characterized. Frequently multiple Gaia stars were matched to a single Kepler ID, as the stars were located at similar R.A. and decl. We removed duplicates by only selecting the Kepler and Gaia ID associated with the most similar magnitudes in the Kepler passband Kp and Gaia passband (G).

We then extracted Gaia Re-normalized unit weight error (RUWE) values for all sources. The unit weight error (UWE) values are a representation of the normalized chi-squared values resulting from the fitting of Gaia DR2 sources to single-star point-spread functions (PSFs). The RUWE value corresponds to a PSF fitting corrected for color-dependent biases. RUWE values center around 1.0, but can be large if the fit is not good or there is more noise than expected. A large RUWE value, such as 1.2 or higher, indicates a multistellar system where the presence of stellar companions increases the noise (Belokurov et al. 2020). RUWE values above 1.2 have been shown to be indicators of binaries that are closer than the typical $\sim 1''$ resolution limits of Gaia (Evans 2018).

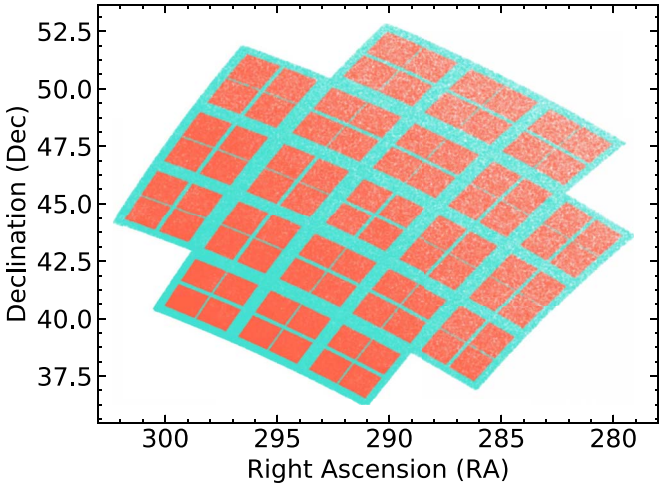

Figure 1. Spatial distribution of stars in the Kepler field. Turquoise points are all stars in the KIC near the Kepler CCDs, while red points are stars whose light did fall on the CCDs for all four seasons and are 8″ away from the CCD edges.

2.2. Downselection of Targets on the Kepler CCDs

The Kepler telescope required a 90° roll about its optical axis every three months—the length of a Kepler quarter—to keep its solar array pointed at the Sun. Because these rolls produced pointing discrepancies, we define observable stars as those located more than 8 arcseconds from the edges of the CCDs (two Kepler pixels) for all four Kepler quarters (Borucki et al. 2010). Figure 1 shows the spatial extent of the Kepler field for stars brighter than $Kp = 16$ mag, consisting of 580,000 stars. Of the 580,000 stars in the Kepler field of view, only 379,000 stars actually fell upon the Kepler charge-coupled devices (red points). The remaining 214,000 stars' fell in the cracks of the CCDs and were physically unobservable (blue points). Finally, we cross-matched the data set of observable stars with the Kepler target list (Batalha et al. 2010), a table possessing the Kepler ID's of the 208,712 stars observed by Kepler. We flagged all stars without matches as nonobserved stars.

Figure 2 shows the distributions of stars observed as a function of Kp mag. The features in the histogram reflect the Kepler target selection function (Batalha et al. 2010): at $Kp = 14$ mag the observed count drops, and likewise beyond $Kp = 16$ mag very few stars are observed. Stars fainter than $Kp = 16$ were selected based on more complex criteria and are not indicative of the general selection function and its biases. Therefore, we focus predominantly on stars brighter than

[3] https://archive.stsci.edu/missions/kepler/catalogs/
[4] http://cdsxmatch.u-strasbg.fr

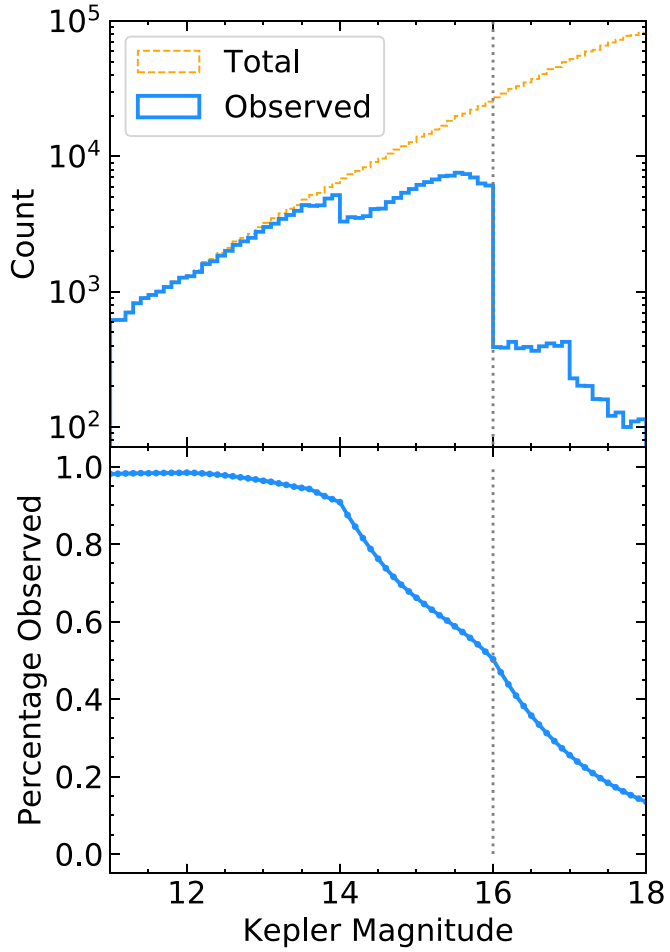

Figure 2. (a) Kepler magnitude for observed and nonobserved samples. Kepler observed 194,859 stars out of the 1,559,884 stars brighter than $Kp = 18$ mag on the CCDs. (b) Percentage observed vs. Kepler magnitude. The dotted gray line marks $Kp = 16$ mag.

$Kp < 16$ mag in Sections 3 and 4, and those with $Kp > 16$ mag in Section 5.

We calculated revised stellar parameters for all stars brighter than $Kp = 16$ mag following the method of Berger et al. (2020). We made this decision upon discovering that over 60,000 of our 379,000 stars lacked radii and luminosity values in the Gaia archive. In addition, Gaia DR2 effective temperatures do not account for interstellar extinction, which can affect our data strongly as many stars in the Kepler field of view are located near the galactic plane. We used isoclassify (Huber 2017) based on the grid model from Berger et al. (2020) with Gaia parallaxes modified by the zero-point offset of Lindegren (2018), Gaia G, B_p, and R_p photometry with uncertainties to derive revised stellar parameters for all 379,000 stars. This data set gave us access to uniform temperatures, radii, and luminosities that account for interstellar extinction for all stars. We subsequently reduced our data set to 327,849 stars by removing stars with parallax uncertainties larger than 20%. All properties are shown in Table 1.

2.3. Validation of Stellar Parameters

For the 172,019 stars in common between our data set and that of Berger et al. (2020), we compared stellar effective temperatures and radii. We found no systematic offset and an ~2% scatter in our effective temperatures, and an ~1% systematic offset and an ~4% scatter in stellar radii. This scatter roughly matches the median catalog uncertainties determined in Berger et al. (2020). As a function of both effective temperature and stellar radius, no strong trend exists in the differences between the two catalogs. Therefore, we are confident in the accuracy and precision of our derived stellar parameters.

3. Full Kepler Sample

3.1. HR Diagram

Figure 3 displays our derived radii and effective temperatures for subsets of the Kepler data with increasing upper limits of Kepler magnitude. The color corresponds to the percentage of stars that were observed for each effective temperature and radius bin. We observe that for Kepler magnitudes brighter than 14, nearly all stars were observed. This matches the selection function detailed in Batalha et al. (2010), as Kepler had the capacity to observe all stars brighter than $Kp < 14$ mag.

At fainter magnitudes, the HR diagrams show parameter-dependent patterns. We observe a strong selection bias against cool, low-luminosity red giants with $Kp > 14$ mag. We suspect that this is due to the fact that these stars could be more efficiently distinguished from cool dwarfs at the same temperature. Dwarfs are far more likely to host planets than red giant stars, and so the giants in this region were dropped from the target list. The observed fraction on the red giant branch is highest for the most luminous giants. This is most likely because these large giants have long pulsation periods that required the full 4 yr of Kepler data to resolve (Bányai et al. 2013; Stello et al. 2014; Yu et al. 2020) and because a significant number of cool giants were misclassified as dwarfs (Mann et al. 2012).

The main sequence is well observed for bright Kepler magnitudes, but decreases substantially at $Kp = 16$ mag. We see little difference in the observed fraction between the solar-type stars and their subgiant neighbors. This is likely because the KIC's broadband photometry was insensitive to the slight difference in $\log(g)$ values between subgiant and solar-type main-sequence stars (Verner et al. 2011; Everett et al. 2013; Gaidos & Mann 2013). As a consequence, many subgiant stars were observed because they were thought to be solar-type main-sequence stars, and many solar-type stars were not observed because their evolutionary state was unknown. The three bottom panels of Figure 3 support this claim, as we can see Kepler's broad selection of all the stars of a given temperature, whether they be subgiant or solar-type stars.

3.2. Evolutionary States

To quantify the percentage of observed stars as a function of evolutionary state we use `evolstate`.[5] These classifications assume solar-metallicity isochrones, which on average are adequate for the Kepler field (Dong et al. 2014). `evolstate` places each star, according to its effective temperature and radius, into one of three evolutionary states: main sequence, subgiant, and giant. We additionally define solar analogs as stars with $T_{\rm eff} = 5700$–5900 K and $R_\odot = 0.9$–1.1. Figure 4(a) shows an H-R diagram of our sample with the delineation of these evolutionary states marked with solid lines.

[5] http://github.com/danxhuber/evolstate

Table 1
Stellar Properties of All Stars That Fall on the Kepler CCDs

KIC ID	Gaia DR2 ID	obsFlag	hostFlag	Kp	$T_{\rm eff}$ K	Radius $R_\odot$	Distance Pc	RUWE	Velocity km s^{-1}	evolState
757076	2050233807328471424	1	0	11.7	5135	4.08	652	0.947	44.98	2
757099	2050233601176543104	1	0	13.2	5448	0.98	368	2.173	6.85	0
891968	2050246795316089088	0	0	14.7	5632	0.99	816	1.006	38.16	0
892010	2050234975566082176	1	0	11.7	4572	15.15	1826	1.014	97.71	2
892107	2050234696381511808	1	0	12.4	4904	4.52	937	0.940	65.46	2
892119	2050235113005074304	0	0	15.2	4830	6.30	4555	1.007	116.34	2
892195	2050234735047928320	1	0	13.8	5371	0.97	479	1.122	18.95	0
892202	2050236521754351360	0	0	15.7	5997	1.16	1723	1.014	13.01	0
892203	2050236521754360832	1	0	13.6	5690	1.06	554	1.245	13.56	0
892212	2050233876054461056	0	0	14.4	5405	0.88	1281		45.48	0

Note. The first 10 rows of the data set used for our analysis. The full table, in machine-readable format, can be found online. obsFlag: 1 is observed, 0 is not. hostFlag: 1 is a host star, 0 is not. evolState: 0 is main sequence, 1 is subgiant, 2 is red giant.

(This table is available in its entirety in machine-readable form.)

The bottom panels of Figure 4 show the percentage observed as a function of Kepler magnitude, colored by evolutionary state. The dotted lines of these panels are stars observed for more than eight quarters, or one half, of the Kepler mission. Similarly, the solid lines are stars observed at any point in the mission. Figure 4(b) confirms the conclusion of Figure 3 that Kepler observed nearly all stars brighter than $Kp = 14$ mag. The main sequence is the most observed evolutionary state of Figure 4(b), dropping to 60% observed at $Kp = 16$ mag. The subgiant stars closely resemble the main-sequence stars, although the cumulative observed percentage drops to a lower 50% at $Kp = 16$ mag. This is most likely due to the Kepler selection function's inability to distinguish subgiant stars from solar-type main-sequence stars, and it has been shown that the Kepler mission preferentially selected subgiant stars for observation (Huber et al. 2014).

From $Kp = 14$–15 mag, the fraction of observed red giants drops steeply from ∼80% to ∼50%, with only ∼40% of red giants at $Kp = 15$ mag being observed for more than eight quarters. The large separation of the red dotted line from the red solid line in Figure 4(b) is most likely because the goal of the mission was to observe solar-type stars, and as a result many red giants were dropped from the target list after being observed for one quarter.

Figure 4(c) breaks the main sequence into three subsections: solar analogs, MS stars cooler than the Sun, and MS stars hotter than the Sun. At $Kp = 16$ mag, the fraction of observed stars cooler than the Sun is ∼65%, and the fraction of observed solar analog stars drops steeply to ∼55%. We suspect this ∼10% difference in observation percentages is because the small, cool dwarf stars could more easily be distinguished from giants. In addition, Dressing & Charbonneau (2013) showed that M dwarfs host a lot of small planets, which in turn led to many M dwarfs being added to the target list during later stages of the Kepler mission. In summary, our analysis shows that Kepler successfully targeted 90% of all solar analogs brighter than $Kp < 14$ mag, decreasing to ∼80% at $Kp < 15$ mag, and ∼55% at $Kp < 16$ mag.

3.3. Effective Temperatures

Figure 5 shows histograms of the observed and nonobserved samples separated by evolutionary state. As anticipated, in Figure 5(a) the observed sample peaks around solar temperature, confirming that Kepler prioritized the observation of solar-type stars. The second peak corresponds to red giants, which were mainly observed to perform asteroseismology.

In Figures 5(b) and (c), we see that for both the main sequence and subgiants the curves peak around 5800 K. This is because, as discussed previously, it was hard for Kepler to distinguish between main sequence and subgiant stars around the solar temperature. As such, all the stars with temperatures close to $T_\odot$ were observed without knowledge of their evolutionary states. For main-sequence stars with temperatures below 5800 K, the percentage of stars observed is larger than it is around the solar temperature. For subgiant stars the curves are similar for all $T_{\rm eff}$.

Finally, in Figure 5(d) the red giant nonobserved curve is significantly larger than the observed curve. This suggests that the red giant stars were not as well observed as stars of other evolutionary states, a conclusion supported by Figures 3 and 4(b).

3.4. Kinematics

McTier & Kipping (2019) investigated galactocentric velocities of Kepler host stars and found the host stars to be moving significantly slower than the rest of the Kepler targets. Further analysis showed that this difference was due to a selection bias in the Kepler host sample, leading to the conclusion that Kepler planet occurrence is independent of galactocentric velocity. Here, we investigate whether this conclusion also holds for Kepler targets with respect to the background population.

To do this we used our derived distances to convert Gaia DR2 proper motions into R.A. and decl. space motions in units of km s^{-1}, and then added these space motions in quadrature to determine stellar sky-plane velocities. Radial velocities were

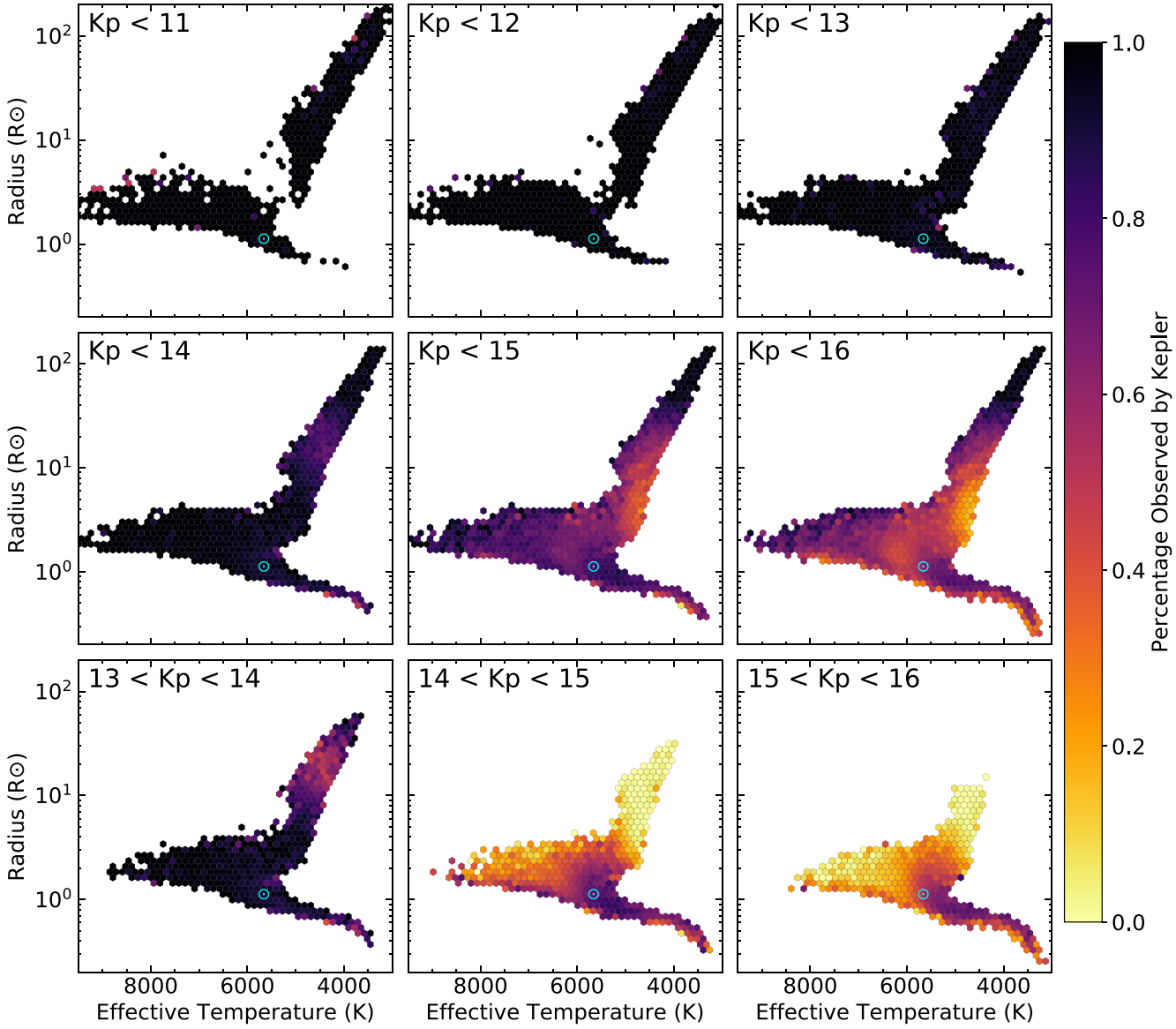

Figure 3. Stellar radius vs. effective temperature for stars on the Kepler CCDs. The first six panels are for cumulative Kepler magnitudes and the last three panels are for binned Kepler magnitudes. The color of each bin corresponds to the percentage of stars observed by Kepler in that bin. The Sun is shown as the teal circled dot in each panel.

only available for a small subset of stars and so we do not include them in our stellar sky-plane velocities. We reduced our investigation of kinematics to Kepler magnitudes between 14 and 16, as this is where the Kepler team was forced to make selection decisions.

The different distance distributions of the observed and nonobserved samples have a strong effect on the proper motions of the samples, and therefore their sky-plane velocities. To account for this we match the nonobserved sample distance distribution to the observed sample individually for each evolutionary state to the best of our ability. The distributions of sky-plane velocity for each evolutionary state is shown in Figure 6.

The solar analog and main-sequence stars are the only panels of Figure 6 showing noticeable differences between the observed and nonobserved samples. The nonobserved samples possess more stars with large sky-plane velocities. We attribute these differences to imperfect matches of the distance distributions of both samples, with the nonobserved stars systematically extending to larger distances and thus larger velocities.

The subgiant and red-giant distributions extend to higher velocities, as expected for stars at larger distances and similar proper motions, and from differences in kinematics for stars in different galactic populations such as the thick disk (Fuhrmann 1998). In summary, we conclude that the Kepler target selection function appears to be unbiased with respect to kinematics, supporting the conclusions by McTier & Kipping (2019) that Kepler planet occurrence is unbiased with respect to galactocentric velocities.

3.5. Stellar Multiplicity

Understanding biases in the target selection is important for studies investigating the effects of stellar multiplicity on planet formation using Kepler. For example, Kraus et al. (2016) used

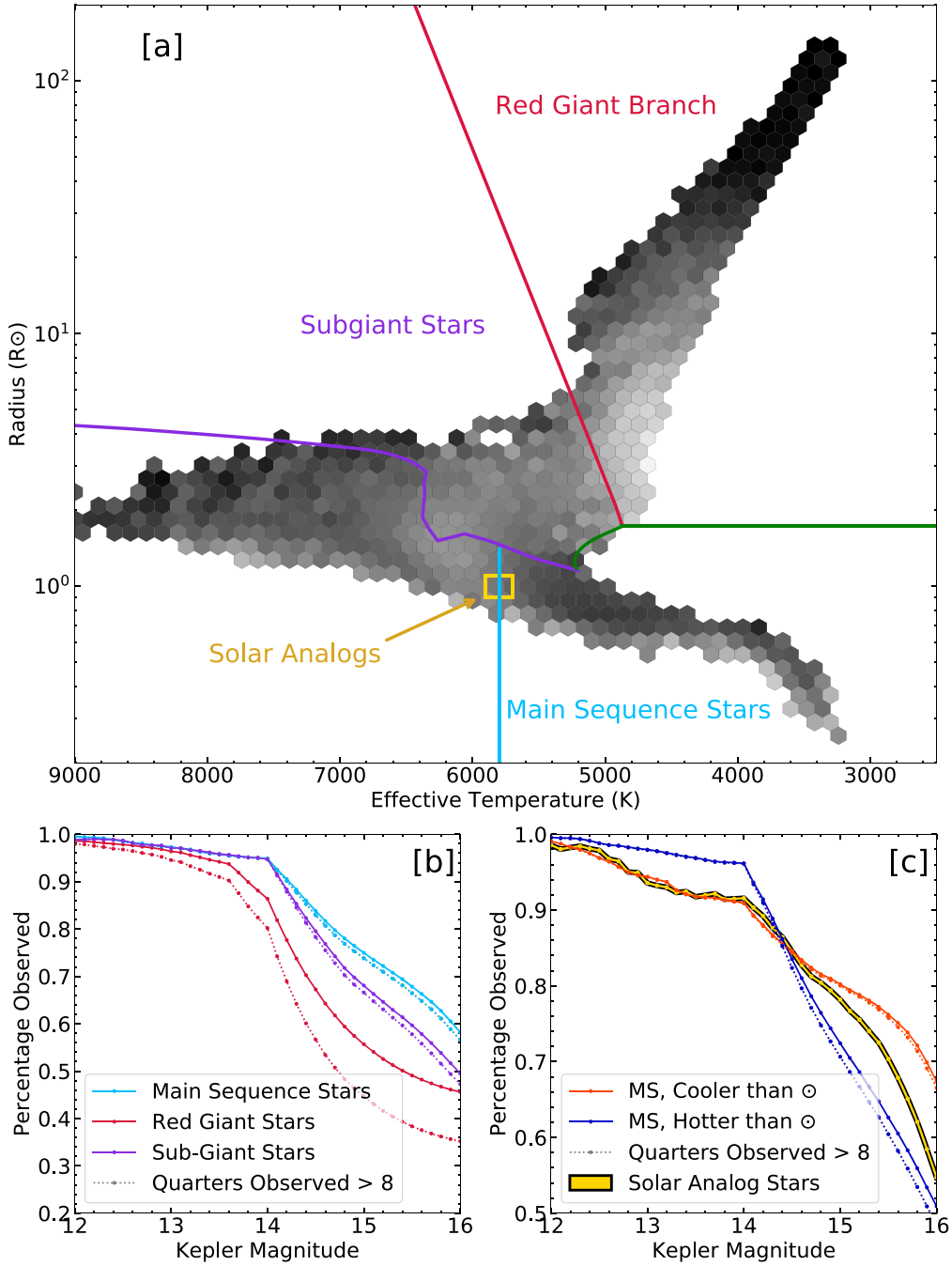

Figure 4. (a): Radius vs. effective temperature for stars with $Kp < 16$ mag. The shading of each bin corresponds to the percentage of stars observed by Kepler in that bin, with darker shadings corresponding to higher percentages observed. The lines separate and define evolutionary state/stellar property ranges. Red separates the giant and subgiants, magenta the subgiant and main-sequence (MS) branch, blue the MS stars cooler and hotter than the Sun, and yellow the solar analogs (5700 K $< T_{\rm eff} <$ 5900 K, $0.9 < R_{\odot} < 1.1$). (b) and (c): Percentage observed at each Kp magnitude for different evolutionary states, as defined in (a). Dotted lines represent stars observed for more than eight quarters of the Kepler mission.

AO imaging of the Kepler host star sample to conclude that the planet occurrence rate in close binary systems ($\lesssim$0.″1, $\lesssim$50 au) is ∼70% lower than that of wider (∼0.″1–1″, ∼ 50–500 au) binaries, and thus a fifth of all solar-type stars in the Milky Way are disallowed from hosting planetary systems due to the influence of a binary companion. While the differential suppression factor derived by Kraus et al. (2016) is robust against target selection bias since both close and wide binaries (as defined above) are unresolved in the KIC, the absolute scale of planet formation among binaries would be affected if there are target selection effects with respect to stellar multiplicity.

To investigate potential biases in the Kepler target selection with stellar multiplicity, we used Gaia RUWE values of the stars in the Kepler field of view for both the observed and nonobserved samples. The RUWE value, as discussed in Section 2, is the Gaia RUWE. A RUWE value above 1.2 suggests that the star in question has a binary companion, though there are a small number of cases when a star has an

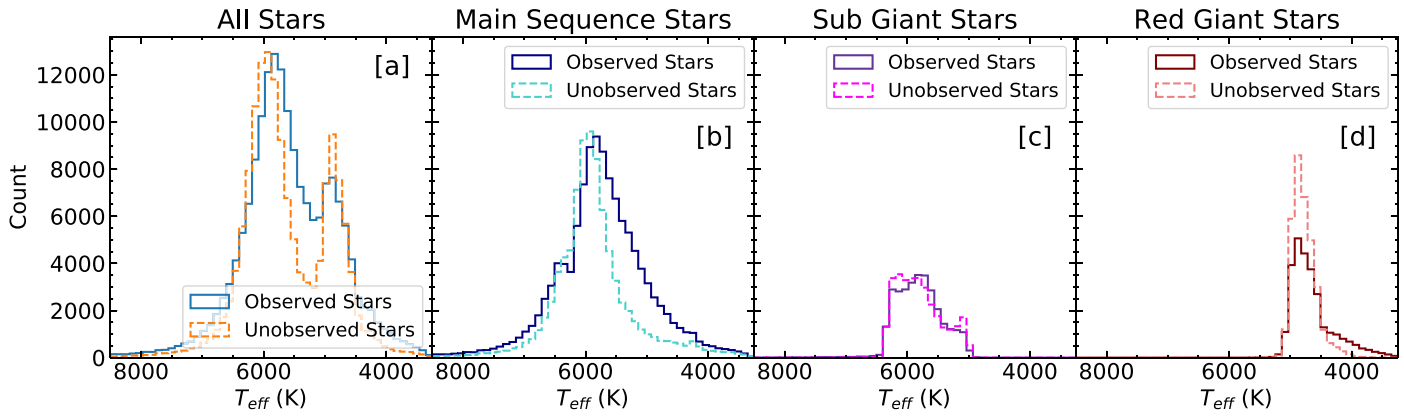

Figure 5. Histogram of the effective temperatures of Kepler stars for different evolutionary states. Solid lines represent the observed sample, and dashed lines represent the nonobserved sample.

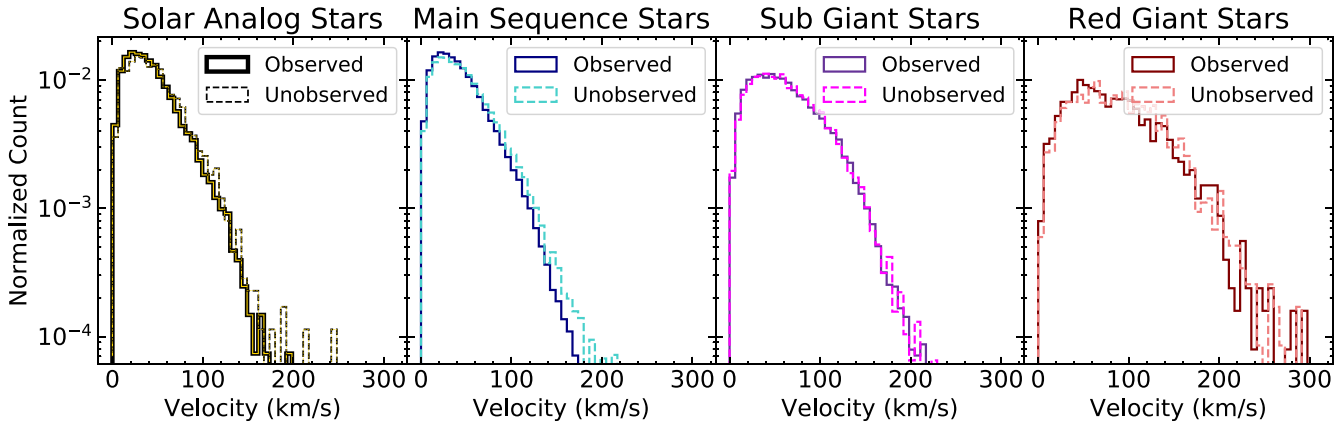

Figure 6. Histogram of the quadrature sum of the proper motions in R.A. and decl. (km s^{-1}) of Kepler stars for different evolutionary states. Solid lines represent the observed sample, and dashed lines represent the nonobserved sample.

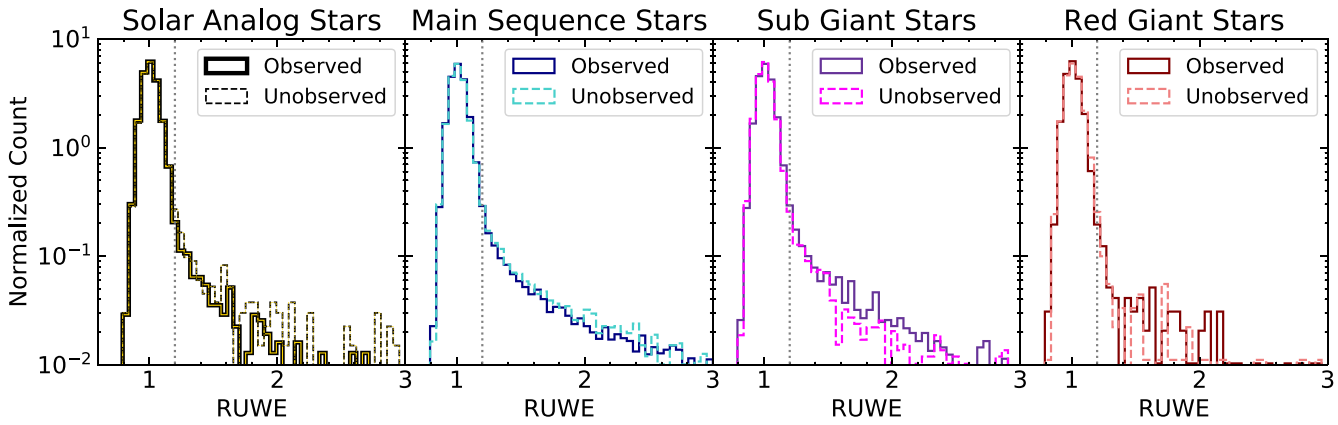

Figure 7. Histogram of RUWE values of Kepler stars for different evolutionary states. Solid lines represent the observed samples, and dashed lines represent the nonobserved samples. The dotted gray line marks the RUWE value beyond which stars are likely to possess companions.

inflated RUWE for reasons other than stellar multiplicity (i.e., stellar variability). Thus, if the Kepler observed and non-observed samples have different distributions for large RUWE we can conclude that the selection function was biased in some way with respect to stellar multiplicity.

Similarly to the analyses of Section 3.4 we first reduced our sample to the stars with $14 < Kp < 16$ mag, both because the selection function was most active in this magnitude range and because the unit-weight error (UWE) has been shown to be inaccurate for Gaia G magnitudes brighter than thirteenth mag (Lindegren 2018). As a next step we matched the nonobserved sample distance distribution to the observed sample distribution individually for each evolutionary state. This is because RUWE values are dependent on the apparent angular separation of binary companions, and the intrinsic physical separations of binary companions will produce different RUWE distributions depending on how far away that system is from Earth. The RUWE distributions are shown below in Figure 7.

RUWE values greater than 1.2 (shown by the dotted gray line in Figure 7) correlate with multiplicity, and hence we focus

our investigation on the stars with large RUWE values. To quantify the fraction of stars with stellar companions, we divided the number of stars with RUWE > 1.2 by the total number of stars. We repeated this process for both the observed and unobserved samples for various evolutionary states.

The red giant observed and nonobserved samples have moderately significant differences and match within 3σ. In contrast, the main sequence, solar analog, and subgiant observed and nonobserved samples differ by $\gtrsim 4\sigma$. While this difference is significant it is important to note that stellar multiplicity, and thus the RUWE values, can also induce significant biases in the evolutionary state classifications themselves by affecting the derived luminosity values. Specifically, the inverse relationship between the samples is likely due to Malmquist bias, which causes main-sequence stars in binary systems to appear as subgiants due to their inflated luminosity values. This would cause the main-sequence observed sample to have a lower fraction of binary stars than the nonobserved sample, because the observed binaries in the main sequence have been misidentified as subgiant stars. Similarly, it would cause the subgiant observed sample to have a greater fraction of binary stars than the nonobserved sample since the observed sample includes main-sequence binary stars as well as subgiant binary stars. We therefore attribute the differences in the RUWE distributions in Figure 7 to effects of binaries on our derived stellar parameters, highlighting the importance of calibrating Malmquist bias when using Gaia to assess the impact of binaries on planet occurrence (A. L. Kraus et al. 2021, in preparation).

4. Host Star Sample

To better control for the biases entering the kinematic and multiplicity analysis for the full sample, we performed a separate analysis focusing only on the sample of stars with transiting planets. We used the Kepler host sample as a basis for our investigations into the differences between the observed and nonobserved samples RUWE and sky-plane velocity distributions. Our host sample consists of 2066 stars from the KOI table of the NASA Exoplanet Archive (Akeson et al. 2013) with either confirmed or candidate planets and Kepler magnitudes fainter than 14 and brighter than 16 mag (Thompson et al. 2018). We randomly selected stars from our observed and nonobserved samples that match the host sample distributions of effective temperature, radius, and distance. By matching our observed and nonobserved samples to the host sample we create two samples that are similar in both evolutionary state and distance. As such, if any differences arise in our comparisons of RUWE and sky-plane velocity values between the matched samples they are due to the Kepler selection function. The comparison between the host-matched observed and nonobserved samples is shown in Figure 8.

Figures 8(a) and (b) confirm that the distributions of effective temperature, radius, and distance are matched between the observed and nonobserved samples. Figure 8(c) shows the RUWE values of the observed and nonobserved samples, and it can be seen that the nonobserved sample is consistently above the observed sample for high RUWE values. $8.2 \pm 0.5\%$ of the observed sample have RUWE values above 1.2 and $11.8 \pm 0.6\%$ of the nonobserved sample have RUWE values above 1.2. This significant 4.8σ difference suggests that Kepler preferentially selected nonbinary stars for observation. This difference remains even when raising the RUWE cutoff to 1.3, 1.4, and 1.5. When the RUWE cutoff is raised to 1.3, $6.9 \pm 0.4\%$ of the observed sample have RUWE values above 1.3 and $9.8 \pm 0.5\%$ of the unobserved sample have RUWE values above 1.3, resulting in a 4.1σ difference. Additionally, when the RUWE cutoff is raised to 1.4, $6.3 \pm 0.4\%$ of the observed sample have RUWE values above 1.4 and $8.5 \pm 0.5\%$ of the unobserved sample have RUWE values above 1.4, resulting in a 3.5σ difference. Finally, when the RUWE cutoff is raised to 1.5, the trend continues as $5.6 \pm 0.4\%$ of the observed sample have RUWE values above 1.5 and $7.7 \pm 0.5\%$ of the unobserved sample have RUWE values above 1.5, resulting in a 3.2σ difference. This result persists when taking into account Gaia EDR3 values.

The difference in elevated RUWE for observed and nonobserved Kepler targets may have implications on studies of planet occurrence as a function of multiplicity. To investigate whether this difference probes close or wide binaries we compared the KIC contamination numbers for both samples, which trace effects of nearby stars that are resolved in the KIC. We found that both samples have KIC contamination numbers that match within 0.5σ, implying that the RUWE difference in Figure 8(c) may be mostly driven by binaries that are unresolved in the KIC. Based on this we speculate that the target selection bias may have been produced by unresolved companions causing broadband colors that deviate from predictions from single-star model atmospheres, which were used to perform the stellar classification in the KIC (Brown et al. 2011). This is further supported by the fact that the difference between the samples is largest in the moderate RUWE regime ($\sim$2–5), which probes higher contrast companions that would remain undetected in the seeing-limited imaging used to construct the KIC. However, we cannot rule out that some fraction of the difference in RUWE values between the observed and unobserved samples can be attributed to wide companions that were intentionally removed.

In contrast to RUWE, the velocity distributions of Figure 8(d) appear very similar, and a Kolmogorv–Smirnov test confirms this conclusion with a p-value of 0.97. This p-value, as well as the similarities in the sky-plane velocity distributions of Figure 8(d), allow us to conclude that Kepler was unbiased with respect to proper motions, confirming the results by McTier & Kipping (2019).

5. Comparison with Gaia DR2 Stellar Parameters

5.1. Revised Radius Comparison

In previous sections of this paper we use the revised stellar parameters calculated with the techniques of Berger et al. (2020) for our analysis. However, since these properties depend on evolutionary models, we also investigated the difference of these parameters to those provided in the Gaia DR2 archive (Andrae et al. 2018). The latter also allows us to investigate the properties of the faintest selected Kepler targets ($Kp > 16$ mag), which were excluded from our classifications.

Figure 9 shows a comparison between the HR diagrams of the Gaia stellar parameters and our revised stellar parameters. The diagrams of Figure 9 look qualitatively similar, confirming our main conclusions of the previous sections. The red giant branch stars show similar color variations along the radius axis, and the subgiant stars are observed at relatively the same rate as

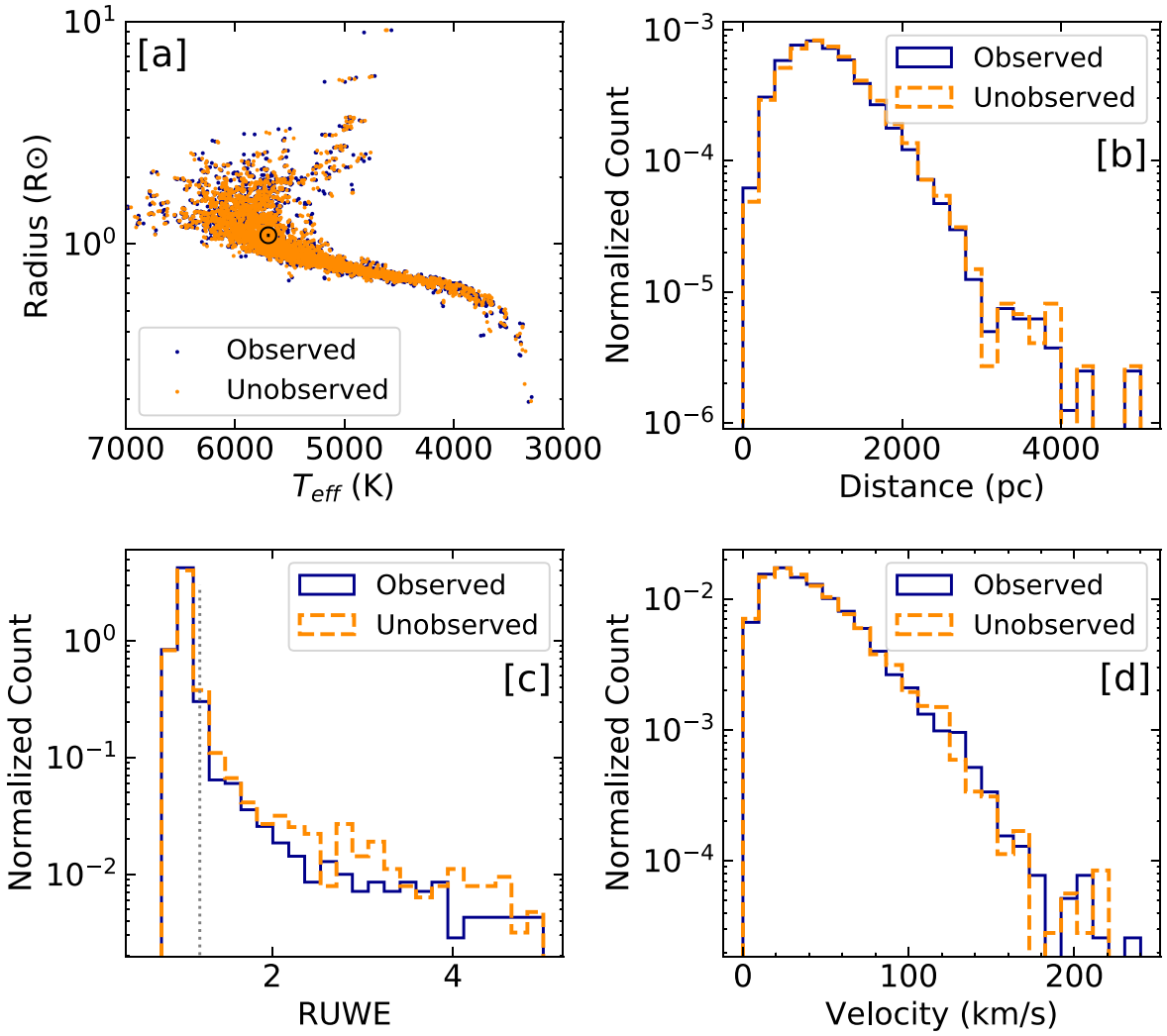

Figure 8. (a) HR diagram of the observed (blue) and nonobserved (orange) samples. (b) Histogram of the distance distributions for the observed and the nonobserved samples. (c) Histogram of the RUWE values for the observed and the nonobserved samples. (d) Histogram of the sky-plane velocity distributions for the observed and the nonobserved samples.

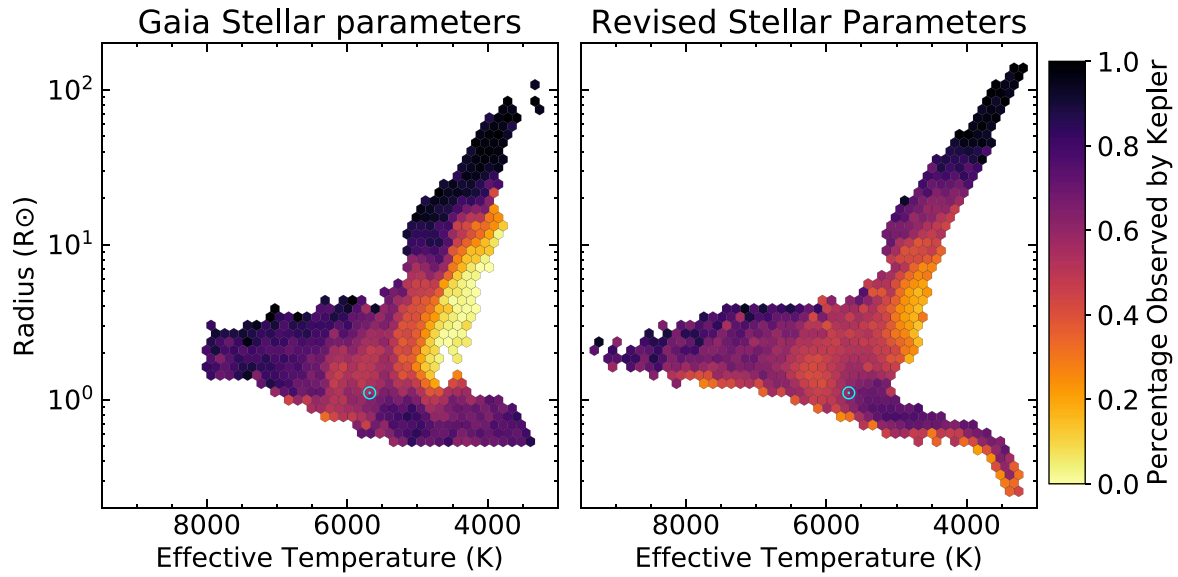

Figure 9. HR diagrams for stars brighter than $Kp = 16$ mag. Figure (a) shows the Gaia radii and effective temperatures. Figure (b) shows our revised radii and effective temperatures. The Sun is shown as the circled dot in each panel.

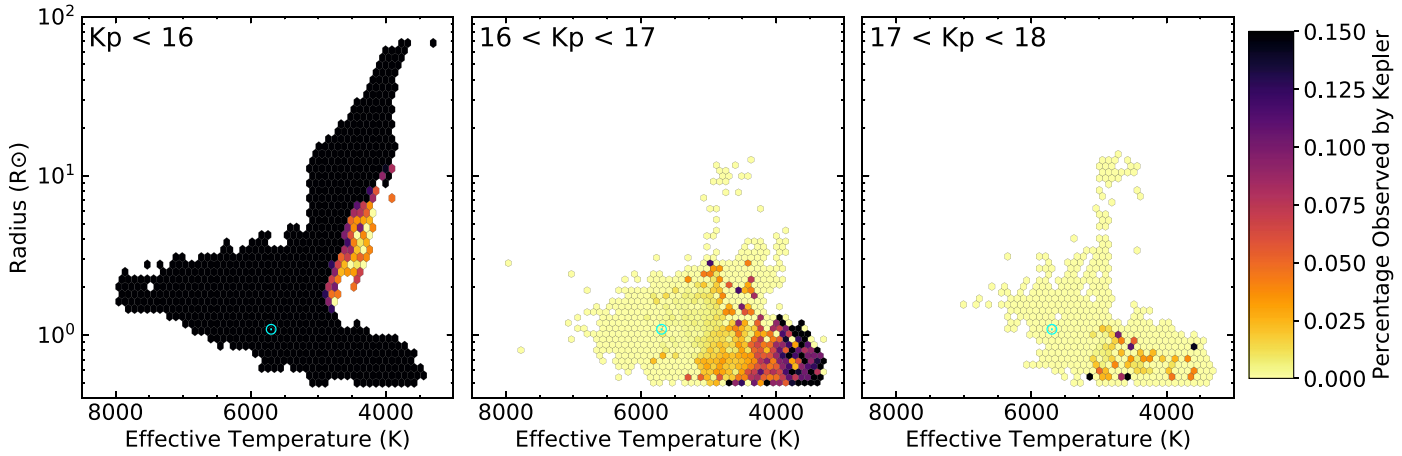

Figure 10. Gaia DR2 radii plotted against $T_{\rm eff}$ for increasing Kepler magnitudes beyond eighteenth magnitude. The color of each bin corresponds to the percentage of points observed by Kepler in that bin. The Sun is shown as the circled dot in each panel. 302,008 stars are plotted in (a), 163,572 stars in (b), and 15,066 stars are plotted in (c).

solar-type stars in both diagrams. However, Gaia DR2 is missing both the hottest and coolest main-sequence stars, due to their cuts on $T_{\rm eff}$ and luminosity (only stars with $3300 < T_{\rm eff} < 8000$ K and $\sigma(L)/L > 0.3$ were given luminosity and $T_{\rm eff}$ values) (Andrae et al. 2018). Some of the differences between these diagrams are due to different treatments of interstellar extinction; however, the overall qualitative agreement confirms that extinction does not have a strong effect on the observed and unobserved percentages.

5.2. Distribution of Faint (Kp > 16 mag) Kepler Stars

One advantage of using the Gaia DR2 stellar parameters is that they provide parameters for stars fainter than $Kp = 16$ mag. These faint stars are displayed in Figure 10, with color-coding showing the percentage of stars observed by Kepler on a scale from 0% to 15%. We observe the that vast majority of targeted stars fainter than $Kp > 16$ mag are cool dwarfs. We suspect that this is due to the fact that toward the end of the Kepler mission it was discovered that M dwarf stars host many planets (Dressing & Charbonneau 2013) and were added to the Kepler target list. We see few stars with $Kp > 17$ mag because of Gaia's inability to derive physical parameters for faint stars (Gaia Collaboration et al. 2018).

5.3. Binary Fraction

Gaia DR2 derived their radii from the Stefan–Boltzmann law. Radii derived this way (rather than using isochrones) for cool dwarfs will form a second main sequence, which is made up of cool dwarf stars that appear more luminous on an H-R diagram than they actually are due to the presence of stellar companions. This "binary main sequence" provides us with another metric to analyze the Kepler selection function's bias with respect to stellar multiplicity.

Figure 11(a) displays the same information as Figure 4(a), except we now plot Gaia DR2 $T_{\rm eff}$ and radii and use the definitions of Berger et al. (2018) to identify cool main-sequence binaries (green line). Figure 11(c) demonstrates that the green line in Figure 11(a) indeed efficiently identifies binaries by comparing the RUWE values of the two cool main-sequence star samples. $39 \pm 2\%$ of the cool main-sequence binary stars have RUWE values greater than 1.2 and $15.2 \pm 0.5\%$ of the cool main-sequence stars have RUWE values greater than 1.2, leading to a difference of 12.1σ.

Figure 11(b) displays the percentage observed of cool main-sequence stars and their neighboring binaries as a function of Kp magnitude. The percentage observed of cool main-sequence stars increases significantly for $Kp > 14$ mag, with ~8% fewer cool main-sequence binaries observed. This qualitatively agrees with the conclusion of Section 4 that Kepler preferentially selected nonbinary stars for observation.

6. Conclusions

In this paper we have analyzed the Kepler mission's target selection function by using Gaia DR2 as the ground truth to characterize the ~500,000 stars that Kepler could have observed, and compared this population to the sample of ~200,000 stars that were selected for observations. Our main results are as follows:

1. We find that the Kepler target selection was efficient at selecting solar-type stars. Specifically, the Kepler target selection is essentially complete for $Kp < 14$ mag, with the main-sequence star selection fraction dropping from 95% to 60% between $14 < Kp < 16$ mag. For stars on the main sequence completeness is best for stars cooler than the Sun and worst for stars hotter than the Sun, with 55% of all solar analogs observed for $Kp < 16$ mag. We find that the observed fraction for subgiant stars is only ~10% lower than the main-sequence stars, implying that many subgiant stars selected for observation were believed to be main-sequence stars.
2. We find that the observed fraction for red giant stars drops from 90% at $Kp = 14$ mag to 45% at $Kp = 16$ mag. Kepler's selection of red giant stars was most strongly biased against cool, low-luminosity giants, with completeness dropping below 30%. This confirms that the KIC was efficient in separating giants from dwarfs, in particular for temperatures between 4000 and 5000 K.
3. We find that the distribution of elevated Gaia RUWE (> 1.2) of the observed and nonobserved main-sequence stars differ at $\sim 5\sigma$ significance, implying a Kepler target selection bias against binaries. This conclusion is robust when taking into account differences in the sample

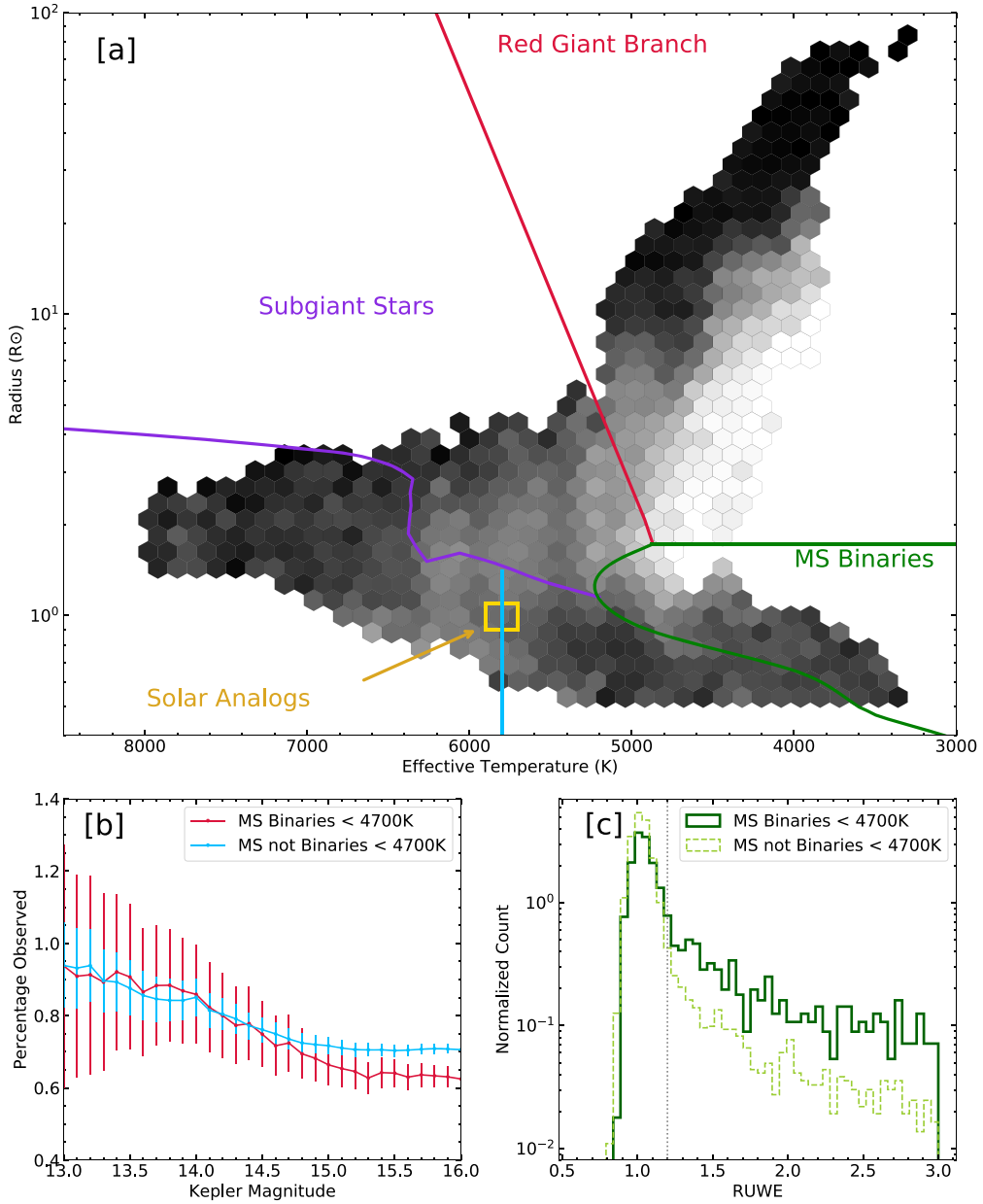

Figure 11. (a): Effective temperature and radius plot of $Kp < 16$ mag with lines defining different evolutionary states. Red separates the giant and subgiants, magenta the subgiant and main-sequence stars, green the suspected main-sequence binaries, blue the stars cooler and hotter than the Sun, and yellow the solar analogs. (b): Percentage observed at each Kp magnitude for main-sequence and main-sequence binary stars with $T_{\rm eff} < 4700$ K. (c): Histogram of the RUWE values for main-sequence and main-sequence binary stars with $T_{\rm eff} < 4700$ K. The dotted gray line marks the RUWE value beyond which stars are likely to possess companions.

properties and supported when using the luminosities of cool main-sequence stars as a proxy for binarity. We find tentative evidence that the RUWE difference may be caused by close binaries that were unresolved in the KIC and speculate that biases in composite broadband colors may have led to this selection bias, but further work will be needed to confirm this conclusion. The difference in elevated RUWE does not affect the previously detected differential planet formation suppression rate for close binaries (Kraus et al. 2016), but highlights the importance of taking into account selection biases for determining the absolute scale of stellar multiplicity effects on planet occurrence.

4. We find that the Kepler target sample is unbiased with respect to galactocentric space velocities compared to the background population of stars Kepler could have selected for observations. This confirms previous results for Kepler exoplanet host stars by McTier & Kipping (2019).
5. We find that the faintest Kepler stars were exclusively selected to be cool dwarfs. The observed M dwarf fraction is $\sim$14% for $16 < Kp < 17$ mag, and falls to $\sim$8% for $17 < Kp < 18$ mag.

Gaia DR2 has enabled the first comprehensive evaluation of the biases and successes of the Kepler selection function, which will be valuable for the study of exoplanet demographics and

stellar populations using Kepler data. For example, the bias against stellar multiplicity identified suggests that future research may require analysis of a control sample of nonobserved, nonhost stars. Future studies combining Gaia RUWE with AO imaging will be required to determine whether the bias identified here is caused by wide or close binaries. Additionally, future Gaia data releases with improved resolution and source completeness will allow more detailed investigations of the selection function bias for Kepler and other space-based transit surveys. We note that Gaia EDR3 was released during the final phases of completing this paper (Lindegren et al. 2020). We have performed basic comparisons of parallaxes, kinematics, and RUWE values for the Kepler sample and confirmed that our main conclusions remain unchanged when using results from EDR3.

We gratefully acknowledge the Gaia and Kepler missions and the people involved for their hard work that has made this paper possible. We thank the anonymous referee for helpful comments on the manuscript. We thank Adam Kraus for discussions on stellar multiplicity bias in the Kepler sample and helpful feedback on the paper. We also thank Robert Jedicke, Aaron Do, Ben Shappee, Michael Bottom, Victoria Catlett, Jingwen Zhang, Casey Brinkman, Ashley Chontos, Vanshree Bhalotia, Nicholas Saunders, Larissa Nofi, Jamie Tayar, and Lauren Weiss for their helpful discussions and feedback.

L.M.W. acknowledges support from Research Experience for Undergraduate program at the Institute for Astronomy, University of Hawaii-Manoa funded through NSF grant 6104374. L.M.W. would like to thank the Institute for Astronomy for their kind hospitality during the course of this project.

T.A.B. and D.H. acknowledge support from a NASA FINESST award (80NSSC19K1424) and the National Science Foundation (AST-1717000). D.H. also acknowledges support from the Alfred P. Sloan Foundation.

This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC; https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. This research has made use of NASA's Astrophysics Data System. This research has made use of the VizieR catalog access tool, CDS, Strasbourg, France. We obtain our Kepler data from MAST, and our Gaia DR2 data from VizieR. This paper has also made use of the open-source TOPCAT application.

Software: pandas (McKinney 2010), SciPy (Virtanen et al. 2020), Matplotlib (Hunter 2007), evolstate.[6]

ORCID iDs

Linnea M. Wolniewicz https://orcid.org/0000-0002-2087-1634
Travis A. Berger https://orcid.org/0000-0002-2580-3614
Daniel Huber https://orcid.org/0000-0001-8832-4488

[6] http://github.com/danxhuber/evolstate

References

Adams, E. R., Ciardi, D. R., Dupree, A. K., et al. 2012, AJ, 144, 42
Akeson, R. L., Chen, X., Ciardi, D., et al. 2013, PASP, 125, 989
Andrae, R., Fouesneau, M., Creevey, O., et al. 2018, A&A, 616, A8
Bányai, E., Kiss, L. L., Bedding, T. R., et al. 2013, MNRAS, 436, 1576
Baranec, C., Ziegler, C., Law, N. M., et al. 2016, AJ, 152, 18
Batalha, N. M., Borucki, W. J., Koch, D. G., et al. 2010, ApJL, 713, L109
Belokurov, V., Penoyre, Z., Oh, S., et al. 2020, MNRAS, 496, 1922
Berger, T. A., Huber, D., Gaidos, E., & van Saders, J. L. 2018, ApJ, 866, 99
Berger, T. A., Huber, D., van Saders, J. L., et al. 2020, arXiv:2001.07737
Borucki, W. J., Koch, D., Basri, G., et al. 2010, Sci, 327, 977
Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A. 2011, AJ, 142, 112
Bryson, S., Coughlin, J., Batalha, N. M., et al. 2020, AJ, 159, 279
Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, ApJ, 809, 8
Dong, S., & Zhu, Z. 2013, ApJ, 778, 53
Dong, S., Zheng, Z., Zhu, Z., et al. 2014, ApJL, 789, L3
Dressing, C. D., Adams, E. R., Dupree, A. K., Kulesa, C., & McCarthy, D. 2014, AJ, 148, 78
Dressing, C. D., & Charbonneau, D. 2013, ApJ, 767, 95
Evans, D. F. 2018, RNAAS, 2, 20
Everett, M. E., Howell, S. B., Silva, D. R., & Szkody, P. 2013, ApJ, 771, 107
Farmer, R., Kolb, U., & Norton, A. J. 2013, MNRAS, 433, 1133
Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, ApJ, 795, 64
Fuhrmann, K. 1998, A&A, 338, 161
Furlan, E., Ciardi, D. R., Everett, M. E., et al. 2017, AJ, 153, 71
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
Gaidos, E., & Mann, A. W. 2013, ApJ, 762, 41
Garrett, D., Savransky, D., & Belikov, R. 2018, PASP, 130, 114403
Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, ApJS, 201, 15
Hsu, D. C., Ford, E. B., Ragozzine, D., & Ashby, K. 2019, AJ, 158, 109
Huber, D. 2017, Isoclassify: V1.2, v1.2, Zenodo, doi:10.5281/zenodo.573372
Huber, D., Silva Aguirre, V., Matthews, J. M., et al. 2014, ApJS, 211, 2
Hunter, J. D. 2007, CSE, 9, 90
Kopparapu, R. K., Hébrard, E., Belikov, R., et al. 2018, ApJ, 856, 122
Kraus, A. L., Ireland, M. J., Huber, D., Mann, A. W., & Dupuy, T. J. 2016, AJ, 152, 8
Kunimoto, M., & Matthews, J. M. 2020, AJ, 159, 248
Law, N. M., Morton, T., Baranec, C., et al. 2014, ApJ, 791, 35
Lillo-Box, J., Barrado, D., & Bouy, H. 2012, A&A, 546, A10
Lindegren, L. 2018, Re-normalising the astrometric chi-square in Gaia DR2 GAIA-C3-TN-LU-LL-124-01, Lund Observatory, http://www.rssd.esa.int/doc_fetch.php?id=3757412
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Lindegren, L., Klioner, S. A., Hernández, J., et al. 2020, arXiv:2012.03380
Mann, A. W., Gaidos, E., Lépine, S., & Hilton, E. J. 2012, ApJ, 753, 90
McKinney, W. 2010, in Proc. 9th Python in Science Conf., ed. S. van der Walt & J. Millman, 56
McTier, M. A. S., & Kipping, D. M. 2019, MNRAS, 489, 2505
Miglio, A., Chiappini, C., Morel, T., et al. 2013, MNRAS, 429, 423
Mulders, G. D., Pascucci, I., Apai, D., & Ciesla, F. J. 2018, AJ, 156, 24
Murphy, S. J., Moe, M., Kurtz, D. W., et al. 2018, MNRAS, 474, 4322
Pascucci, I., Mulders, G. D., & Lopez, E. 2019, ApJL, 883, L15
Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013, PNAS, 110, 19273
Pinsonneault, M. H., Elsworth, Y., Epstein, C., et al. 2014, ApJS, 215, 19
Sharma, S., Stello, D., Bland-Hawthorn, J., Huber, D., & Bedding, T. R. 2016, ApJ, 822, 15
Stello, D., Compton, D. L., Bedding, T. R., et al. 2014, ApJL, 788, L10
Teske, J. K., Ciardi, D. R., Howell, S. B., Hirsch, L. A., & Johnson, R. A. 2018, AJ, 156, 292
Thompson, S. E., Coughlin, J. L., Hoffman, K., et al. 2018, ApJS, 235, 38
Verner, G. A., Chaplin, W. J., Basu, S., et al. 2011, ApJL, 738, L28
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261
Youdin, A. N. 2011, ApJ, 742, 38
Yu, J., Bedding, T. R., Stello, D., et al. 2020, MNRAS, 493, 1388
Ziegler, C., Law, N. M., Baranec, C., et al. 2018, AJ, 155, 161
Zink, J. K., & Hansen, B. M. S. 2019, MNRAS, 487, 246