The Milky Way and its exoplanets

Jesper Nielsen

Lund Observatory
Lund University

2021-EXA177

Degree project of 60 higher education credits (for a degree of Master)
May 2021

Supervisors: Alexander Mustill and Paul McMillan

Lund Observatory
Box 43
SE-221 00 Lund
Sweden

Abstract

The detections of over 4000 exoplanets as of today has shown that they exist in a wide range of different configurations, otherwise known as architectures. Recently, studies have been made trying to link the environment of host stars to the architectures of their planetary systems and have found that the architectures of planetary systems may not necessarily be uniform in the Milky Way. Since effects such as photoevaporation may truncate or dissipate protoplanetary discs and thus affect the formation and evolution of planetary systems, we aim to investigate how planetary system architectures are affected by the galactocentric radius at which their host stars were formed. By determining the ages of stars using a precise Bayesian fitting algorithm and combining these with their metallicities and models of the evolution of the radial metallicity gradient in the Milky Way, the *formation radii* of stars are determined. By binning the stars in formation radius and estimating the occurrence rates in each bin through Markov Chain Monte Carlo simulations of a Poisson process likelihood model which separates different orders of detection, individual multiplicities of planetary systems can be linked with the birth environment of their stars.

Six-planet systems were found to be the most common multiplicity for all formation radii and was also the highest multiplicity considered although due to very large uncertainties and the expectation that high multiplicities are common, these results will need to be further investigated. No statistically significant trend between the occurrence rate of planetary systems and the formation radius of their host stars was found for any of the multiplicities considered but the large uncertainties make the results still inconclusive.

KS tests were performed on the orbital period and planet radii distributions of planets around host stars formed inside and outside a given Galactocentric radius. After debiasing for stellar ages, the null hypothesis that the two subsamples of orbital periods and planet radii were drawn from the same, underlying distribution, could not be rejected for any formation radius. These results hints towards the formation process and evolution of planetary systems being independent of the galactocentric radius at which the system was formed although further investigations are needed.

Popular Science Abstract

Our Solar System is not the only planetary system in the universe. Since the first detected planet orbiting a star other than the Sun, over 4000 *exoplanets* have been detected. These exoplanets live in systems with an incredible diversity. The systems can include single, giant planets orbiting very close to the star, or they can have multiple planets packed tightly together. What is causing these different *architectures* to arise is still unknown to us. What is known is the fact that planets are formed from a disc of gas which orbits around a newly formed star, a *protoplanetary disc*, and the properties of these discs shape the architectures of the system. The properties of the disc is in turn shaped by the properties of its host star. The environment in which stars are formed are therefore incredibly important in shaping the structure of the protoplanetary discs and thus the architectures of planetary systems. For example, stars formed in very chaotic environments with lots of stars surrounding them may experience stellar fly-bys which is when two stars fly very close to each other. The gravity from each star can then cause the planetary systems to become so unstable that planets get ejected from the systems. This means that it is expected that planetary system should look different throughout the galaxy since the environment in our galaxy changes whether or not someone is for example close to the center of it, where there are plenty of stars, or if someone is in the outskirts, where there are much fewer stars. Naively then, one might think that it is possible to observe where in the galaxy planetary systems are, and see if they change. The problem with that method is the fact that *stars move from their birthplace*. If a star is formed, orbiting at some radius R_1 from the center of the galaxy, it can, during its lifetime, move to another orbit at radius R_2. This means that the environments in which stars are observed in today are probably not the environments in which they were born.

All stars are formed from the *Interstellar medium* or ISM, gas floating around between each star system in the Galaxy. The metal contents of the ISM, its *metallicity*, at the specific time and place in which a star is formed is then kept within that star. The ISM is not homogeneous, instead its metallicity is changing throughout the galaxy and evolving over time as stars explode in supernovae, expelling all their formed elements into the gas. This means that the metallicity in the ISM is unique for a specific time and place in the galaxy meaning that the metallicity of a star can be seen as a sort of fingerprint from its birthplace. If both the age and the metallicity of a star is known, it will be possible to estimate where it was formed, its *formation radius*.

In this project, the aim is to estimate where in the galaxy stars which are hosting planets are formed. Then, it will be possible to say something about how the planetary systems look like depending on where they were formed. However, it is important to note that it is not possible to detect all planets which are orbiting around any given star. Most planets today have been detected using the so-called transit method where light from a star is recorded and if a planet is passing in front of the star from our point of view, the light dims. By observing how much the light dims, it is possible to determine how large

a planet is. However, since the sizes of planets are incredibly small compared to that of stars, only large planets orbiting very close to their stars are able to be detected. This means that the planets we have detected right now are probably not representative of all of the planets which exists out there in the whole wide universe. To combat this in this project, a statistical model is set up. It is possible to calculate the probabilities for us to detect a planet around any given star and by comparing the number of planets we would expect to detect compared to the number of planets we have detected, it is possible to estimate the *occurrence rates* of planets which is the fraction of stars that are expected to host planets. By comparing the calculated occurrence rates with the formation radii of their host stars, it is finally possible to understand how planetary system change depending on where in the galaxy their host stars were formed which might bring useful insight in what different environmental aspects which might affect the formation and evolution of planetary systems.

Contents

1 Introduction **4**

2 Determining formation radii **9**
2.1 The Data . . . 11
2.1.1 Stellar Population . . . 11
2.1.2 Planet Population . . . 14
2.2 Age determination . . . 15
2.3 Models . . . 20
2.3.1 Minchev et al. 2018 . . . 22
2.3.2 Frankel et al. 2018 . . . 23
2.3.3 Kubryk et al. 2015 . . . 24
2.3.4 Sharma et al. 2020 . . . 25

3 Detection Bias **27**
3.1 Detection Probability . . . 28
3.2 Exoplanet distribution . . . 28
3.3 Sorting Order . . . 29
3.4 Likelihood Function . . . 29

4 Formation radius distributions **34**
4.1 Binning by age . . . 34
4.2 Other bins . . . 37
4.3 Discussion . . . 39
4.3.1 Perfect mixing assumption . . . 39
4.3.2 Binning by age . . . 39
4.3.3 Binning by Stellar Mass and Radius . . . 40
4.3.4 Binning by Stellar Effective Temperature . . . 40

5 Exoplanet Occurrence **42**
5.1 Full population . . . 42
5.2 Varying formation radius . . . 45
5.3 Architectural analysis . . . 49
5.4 Discussion . . . 53

5.4.1 Comparisons with previous work 53
5.4.2 Occurrence rate dependence on formation radius 54
5.4.3 Architectural analysis 57

6 Conclusions & Future work **59**
6.1 Outlook 60

A Model quantities **67**
A.1 Detection Probabilities 67
A.2 Exoplanet distribution 70
A.3 Sorting Order 71

B Posterior distribution of the occurrence rate given no detections **73**

C Normalisation of expected number of detections **74**

D Additional Figures **75**

Acknowledgements

I would like to start by thanking my two supervisors Alexander Mustill and Paul McMillan for their guidance and help during this project. Their discussion have truly been invaluable. I would also like to thank John Wimarsson and Rebecca Forsberg for fruitful discussions and for being thoroughly wonderful human beings who have helped me throughout my studies. Most of all, I would like to thank Alexandra Roslund for being the best support I could ever wish for and for her guidance throughout this work. She always finds a way to believe in me despite me not doing so myself and for that I am eternally grateful.

This thesis was written in the midst of the Covid-19 pandemic and thus from the confines of my own home so I would finally like to thank our pets Rafiki, Lovis, and Olle for keeping me company by the computer each and every day.

Chapter 1

Introduction

Ever since the first planet companion around a main-sequence star was detected (Mayor & Queloz, 1995) more and more detections have followed in its wake. As of writing this thesis, there are about 4000 confirmed exoplanets in the NASA exoplanet archive (Akeson et al., 2013) with most of these being discovered by the Kepler telescope (Borucki et al., 2010). These planets orbit around their host stars in a wide variety of configurations otherwise known as architectures. Planets have been found to orbit with orbital periods as short as a few hours (Barragán et al., 2018) while in our own Solar System, Neptune has the longest orbital period of all known planets with an orbital period of about 165 years. Furthermore, planets have been observed to have a wide range of physical radii with detected planets having radii well below and above that of Earth. Planets are thought to have formed from a protoplanetary disc of gas which orbits around newly formed stars and the structure, mass, and lifetime of the disc play a key role in shaping the planetary system as a whole (Raymond & Morbidelli, 2020). Considering that the properties of protoplanetary discs (e.g. mass) are dependent on the properties of their host stars (Ansdell et al., 2017), it is easy to assume that the initial architectures of planetary systems are affected by the properties of the host star. While the properties of the star can set the initial architectures, it is important to note that planetary systems undergo further evolution after their formation. Effects such as tides from the star on close in planets, atmospheric stripping, and dynamical evolution causing scattering or changes in the planetary orbits all contribute to the evolution of architectures. Due to the evolution of planetary systems, it will be incredibly difficult to be able to predict the architectures of planetary systems formed at a certain location in the galaxy, however, by understanding how architectures are affected by formation radius, it will be possible to gain useful knowledge in how impactful the birth environment is for the formation and evolution of planetary systems.

The metallicities of stars have been shown to have a large effect on the presence of giant planets with these being more common around metal rich stars (Johnson et al., 2010) due to the fact that more metal-rich discs causes a more efficient giant planet formation. There have been made similar investigations on smaller planets and hints towards smaller planets being slightly more common around more metal-rich stars have been found (Lu

et al., 2020). Due to the difficulties in detecting smaller planets, these results are however, still under investigation. Yang et al. (2020) looked at how stellar properties affect the occurrence of planets and found that hot stars ($T_{\text{eff}} > 6500$ K) had an occurrence rate of about 0.35 compared to colder stars ($T_{\text{eff}} < 5000$ K) which had an occurrence rate of approximately 0.75. The occurrence rate is an important quantity because it can help us answer the question of how common planetary systems are throughout the galaxy. There are generally two different definitions of occurrence rates: the average number of planets orbiting around a star and the fraction of stars hosting planets. As Yang et al. (2020) mentions, considering the fraction of stars hosting planets allows one to investigate more in detail how stellar properties affect planetary systems and so that is the definition of occurrence rate that will be used throughout this thesis. Santos et al. (2017) found stellar abundances for different elements in solar neighbourhood stars and used these to infer the planet building blocks in thin disc, thick disc, and halo stars[1] and found that these varied significantly between the different types of stars meaning that the planetary systems orbiting around these are expected to vary significantly.

The environment in which a star is formed is not only important for the formation of the system but also its evolution. Stars are generally formed in giant molecular clouds meaning that they are often formed in regions with a high density of stars (Lada & Lada, 2003) and thus stellar flybys could be fairly common for these stars. Muñoz et al. (2015) investigated the effects of stellar flybys on protoplanetary discs and found that close encounters could truncate the discs and cause the disc to lose mass, results that are supported by previous studies by for example Breslau et al. (2014). Li et al. (2020) investigated how stellar flybys affected already formed planetary systems and found that while close in planets are resistant to stellar flybys, planets on wider orbits are at risk of being ejected or, if the outer planets are more massive, cause an interplanetary instability which might eject planets. Although close in planets, which are the ones that we would be able to observe, are resistant to stellar flybys, the scattering of planets orbiting far out in the system can cause instability further in in the system and thus affect the observable planets as well. A planetary system orbiting around a star in an environment with a high number density of stars can therefore be at risk of becoming unstable due to the higher probability of stellar flybys. Further, Rodet et al. (2021) showed that a large fraction of the Hot Jupiter population[2] can be explained through the effect of stellar flybys which can happen in dense stellar clusters.

Another important consequence of being formed in high stellar density environments is the effect of photoevaporation which is the result of FUV[3] or EUV[4] photons from nearby

[1]For a good review on the structure of the Milky Way, including the the thin disc, thick disc, and halo, see Bland-Hawthorn & Gerhard (2016)

[2]Giant planets on short orbits are termed Hot Jupiters although the mass and period ranges vary throughout the literature

[3]Far UltraViolet, $\approx$ 6-10 eV

[4]Extreme UltraViolet, $\approx$ 10-127 eV

young and hot stars (OB stars) transferring energy to the disc, causing it to heat up significantly. This temperature increase can then be enough to create a large enough thermal pressure which causes gas loss, usually in the outer regions of the disc where the gas is more loosely bound (Guarcello et al., 2016). Photoevaporation can cause significant mass loss of the disc and has been shown to be a dominating effect of disc truncation and mass loss compared to stellar encounters (Winter et al., 2018). Winter et al. (2020) used the current phase space density of stars in the local neighbourhood to separate stars in high density and low density regions. They then investigated the architectures of planetary systems between these two regions and found that systems around host stars found in high density phase space vary significantly compared to stars in low density phase space. While these results might be mostly caused by an age bias due to younger stars being more likely to host Hot Jupiters (Adibekyan et al., 2021; Mustill et al., 2021), Longmore et al. (2021) performed a similar investigation where they investigated how common planets around stars in higher phase space densities versus low phase space density stars are and found that single planet systems are much more common around stars in clusters despite not having a significant amount of Hot Jupiters in their sample. These results show that stellar clustering does have some effect on the architectures of planetary systems although but the details would have to be further investigated.

The interstellar medium from which stars are formed is not uniform which means that the formation of stars does not happen uniformly throughout the galaxy. This can be seen in e.g. the stellar number density in the Milky Way which can be fit by two exponential functions decreasing with Galactocentric radius (Jurić et al., 2008). Due to the production of heavier and heavier elements in stars which then get ejected into the ISM due to supernovae, the metallicity in the ISM also increases over time compared to the original primordial composition from big bang nucleosynthesis. This means that stars can be formed and evolve in vastly different environments depending on where in the galaxy the star is formed. The metallicity of the ISM (and thus the stars formed from it) has also been shown to vary throughout the galaxy with a reported metallicity gradient of -0.076 dex kpc^{-1} (Spina et al., 2021).

Naively, one might think that it would be possible to assume that star's orbits have remained the same throughout their lifetime. It could then be possible to try to relate their current Galactocentric radius with the architectures of their planetary systems. However, as shown by Sellwood & Binney (2002), stars tend to radially migrate in the galaxy meaning that the Galactocentric radius at which stars are observed may not necessarily be the one it was formed on. There are two main processes which cause radial migration. The first has been dubbed "churning" and is caused by a density wave structure, typically the bar or a spiral arm, scattering stars near their co-rotation across their co-rotation radius. This effectively causes a change in angular momentum and thus a change in guiding radius

of the orbit without changing the eccentricity of the orbit.[5] The other effect has been dubbed "blurring" which is caused by the scattering with a molecular cloud away from co-rotation with the cloud (Schönrich & Binney, 2009) and leads to an increased eccentricity while the angular momentum, and thus the guiding radius, remains the same. Radial migration has been estimated to be happening over Gyr timescales (Anders et al., 2020) which means that the birth environment should affect the planetary systems significantly as the formation of planetary systems generally happens on Myr timescales (Raymond & Morbidelli, 2020) and thus planetary systems are formed before the host stars has had any time to migrate away from the birth environment. Radial migration of stars means that in order to fully understand the effects the birth environment of a host star has on planetary systems, it is necessary to in some way estimate the *formation radius* of the star i.e. the Galactocentric radius at which the star was formed instead of the Galactocentric radius at which it is observed. The method we use to estimate the formation radii of the stars is based on that of Minchev et al. (2018). Therefore, in order to estimate the formation radii of stars, it will be necessary to not only have accurate abundance measurements of the star, but also accurate age estimations of it.

The purpose of this thesis has therefore been to investigate how the architectures of planetary systems are affected by the formation radius of the star. Due to the chaotic evolution of planetary systems, it is difficult to directly infer the effect birth environment has on planetary systems without directly observing young planetary systems. However, understanding how occurrence rates change with formation radius can help provide further insight into what different galactic and stellar properties might affect the formation and evolution of planetary systems. The overall process for this thesis work followed these steps:

1. Find a relevant data set of stars with known abundances where a subset hosts planetary systems
2. Determine the ages of the stars
3. Find relevant models for the evolution of the radial metallicity gradient in the Milky Way
4. Combine the ages and abundances with metallicity gradient models to estimate the formation radii of the stars
5. Set up a statistical model to handle detection biases such that occurrence rates and multiplicities can be estimated for any given population of stars
6. Split up the data set of stars into different formation radius bins and estimate the occurrence rate on each formation radius bin

[5] While we use the term eccentricity here, it is not eccentricity in the sense that the star is on a closed orbit. The eccentricity is in reality the epicycle amplitude and an increase simply causes a star to orbit over a wider range of radii.

The thesis is structured as follows: chapter 2 goes over how formation radii of stars are determined, chapter 3 talks about how the detection bias when detecting exoplanets are handled, chapter 2 shows and discusses the results of the formation radii distributions while chapter 5 shows and discusses the occurrence rate results. The thesis is concluded with future prospects in chapter 6.

Chapter 2

Determining formation radii

In this chapter, the different components necessary for finding the formation radius of stars are described. The chapter is structured as follows. First, the method for determining the formation radii of stars is described as well as the different components required to do so. Then the origin of the data sets of both the stars and planets used is described as well as the different cuts that were made on them. Following is the method used to determine the ages of stars and descriptions of the different metallicity gradients of the Milky Way that are used since as mentioned, these two components are essential in order to determine the formation radii of stars.

As mentioned, radial migration processes such as churning can cause the angular momentum of a star to change over time. This ultimately leads to a change in guiding radius of the star without leaving a dynamical trace (Sellwood & Binney, 2002) meaning that it is difficult, if not impossible to infer a formation radius kinematically. The method that will be used to determine the formation radii is the same as that of Minchev et al. (2018) which is based on the fact that a star's photosphere retain the elemental abundances of the star's original birthplace for a long enough period of its lifetime. A necessary assumption is that the ISM is sufficiently mixed meaning that stars born at a given Galactocentric radius at a given time will all have the same metallicity. This assumptions seem to be true for the present day Milky Way (Nieva & Przybilla, 2012) but is not necessarily true for higher redshifts where scatter in for example the oxygen abundance has been found (Sánchez-Menguiano et al., 2018). As discussed by Minchev et al. (2018), this effect is assumed to symmetric around the mean meaning that while it will affect individual formation radii of stars, it will not necessarily affect the total distributions significantly. Consequences of this assumption is further discussed in section 4.3.1. The mixing assumption means that it is possible to consider the abundances of a star as a fingerprint from the ISM at the time and place where the star was formed meaning that the abundances that are observed in the star today will be the same as in the ISM at the place and time of a star's birth. Therefore, if it is possible to know how the abundances of the ISM evolved over time and throughout the galaxy (if it has done so), it is possible to match the abundances of a star of a certain age to that of the ISM at the time of birth of the star and thus infer the

formation radius of the star. Critical to this method is also the assumption that there are no internal processes in the star which would alter the abundance and thus remove the fingerprint. This is true for iron (Fe) and is in fact true for most elements in both dwarfs and giants. Therefore, it is [Fe/H] that will be used to infer the formation radius. Here [Fe/H] is the ratio of the number of Fe and H (hydrogen) atoms in the star on a log scale relative to the same ratio in the Sun. In theory, it would be possible to use other elements as well, but there exist fewer models describing the evolution of elements other than Fe over time and position in the galaxy.

The method also hinges on the existence of models describing the evolution of the metallicity in the ISM over both position and time. The models that are used in this project are described in detail in sections 2.3.1-2.3.4. The metallicity distribution in all models are assumed to be axisymmetric at any given time meaning that only two coordinates are needed: Galactocentric radius and time. In order to be applicable it is necessary for the metallicity to be strictly decreasing (or increasing) over radius for any given time and that the metallicity is strictly increasing (or decreasing) over time for any given radius. These constraints are necessary in order to avoid any degeneracies and to allow the models to be unambiguous. The second constraint is well known, has been shown observationally, and can be explained by an enrichment process.

During the lifetime of a star, it fuses elements in the core. Late in the lifetime of the star, due to convective flows within the star, elements formed in the core experience a "dredge-up" where they get homogeneously mixed in most of the star. The more mass a star has, the higher temperature and pressure its core can reach, enabling the star to fuse heavier and heavier elements. The most massive stars ($> 8\ M_{\odot}$) are massive enough for their cores to reach temperatures and pressure high enough such that they are able to fuse elements up to iron. At the point of their death, they explode into Type II Supernova, leaving only the core. Due to the mixing process, this means that all the fused elements get expelled into the ISM again, enriching it (Prialnik, 2000). This leads to an enrichment over time where the metallicity of the ISM increases over time for a given radius.

The first constraint has also been shown observationally. For example, the present day metallicity gradient is shown to be strictly negative (Anders et al., 2017; Nandakumar et al., 2020) in the galactic plane.[1] Further, Anders et al. (2017) showed, using the Coro-Gee sample of red giant stars, that the metallicity gradient is negative for any given time. The cause of this is not yet fully understood but is probably best explained by inside-out formation of the Milky Way where the inner part of a galaxy gets enriched earlier while the outer parts, which are formed later, experiences a delayed enrichment process. Further constraints on the models are that for each metallicity gradient, the resulting formation ra-

[1]It should be noted that in the innermost radial bin in Nandakumar et al. (2020), the gradient is positive but the authors note that this bin is plagued by low number statistics which might influence the results.

dius distribution needs to have the following characteristics: 1) the youngest age bin needs to have the same median formation radius as the current median galactocentric radius and 2) the formation radius distribution need to peak at subsequently smaller formation radii for older populations. The former requirement stems from the fact that it is not expected for young stars ($\sim$ 1 Gyr) to have experienced much migration (Minchev et al., 2018). As seen in figure 2.1, most stars in our sample are situated around the solar neighbourhood and it is expected then that the youngest of these stars have formed near the solar neighbourhood. The latter requirement stems from the assumed inside-out formation of the Milky Way (see Minchev et al. (2014) or Bird et al. (2013), for example).

2.1 The Data

2.1.1 Stellar Population

In order to determine anything about a planet population, it is necessary to not only have accurate data on the planetary systems themselves but also on the stellar properties of the host stars. While there are plenty of different planet detection missions which have discovered plenty of planets, none has been as successful in finding as many planets as the Kepler mission (Borucki et al., 2010), a planet detection mission using the transit method for detecting planets. The transit mission utilises the fact that when a planet passes in front of a star from the point of view of the observer, the light from the star dims allowing an observer to calculate the radius of the planet. By timing each transit, it is also possible to estimate the orbital period of the planet. The Kepler telescope, launched in 2009 performed a deep sky survey on a single patch of sky during its entire mission lifetime. This mission set-up means that not only is it uncontested in the number of exoplanets confirmed so far, it is also very easy[2] to perform statistical analysis on the entire sample while taking into account detection biases.

First, we select the stars in the Kepler stellar table which have completeness products provided by Mathur et al. (2017). The completeness products of interest are the observation time T_{obs}, the duty cycle f_{duty}, and the combined differential photometric precision for a specific transit duration t (in hours), $CDPP_t$. The duty cycle is the fraction of the light curve which contains valid data and the $CDPP$ is an estimate of the stellar noise. We then cross match this stellar table with the catalogue by Berger et al. (2020) which is a follow-up work from Berger et al. (2018). This stellar catalogue is the results of isochrone fitting of data from broadband photometry together with Gaia parallaxes and spectroscopic metallicities (if possible). The reason for using this specific catalogue is because a) it uses only known stars in the Kepler field and thus will include a subset of stars known to host planets b) it has very precise and homogeneous measurements on both stellar mass and stellar radius which are important parameters when investigating exoplanet systems, with stellar radii being a key parameter for finding the planet radii. In order to remove

[2]At least compared to other surveys.

stars which may have poorly constrained parameters, we remove stars with a Berger et al. (2020) goodness of fit (iso_gof) < 0.99. Further, stars with a re-normalised unit weight error (RUWE) > 1.2 were also removed. The iso_gof measures the quality of the Berger et al. (2020) fit, while the RUWE is a metric which combined several Gaia goodness-of-fit metrics. The cut in RUWE also helps with removing potentially unresolved binaries that might pollute the data. These cuts were made following the instructions by Berger et al. (2020). Finally, we only used stars with surface gravity $\log g > 4$ following Zink et al. (2019). The reason for only using stars with $\log g > 4$ is because the expression for the limb darkening coefficients used by the Kepler pipeline found by Zink et al. (2019) (equation (A.6)) are only valid for these stars. Since the goal is to estimate occurrence rates for stars using the model as described by Zink et al. (2019), it is therefore necessary to remove these stars. This cut also makes it so that any giants in the stellar sample are removed, as can be seen in figure 2.2. Then, we remove any star that didn't contain any value for the parameters needed for the age determination ([Fe/H], T_{eff}, parallax, K_s-magnitude). Finally, we remove stars with $T_{obs} < 2$yr, $CDPP_{7.5} > 1000$ ppm, and $f_{duty} < 0.6$ to remove any possible instrumentation noise as well as to ensure that a large enough portion of the light curve was filled for each star. It should also be noted that it may not be possible to determine the age of all stars, and since age is a crucial component in determining the formation radius, all stars for which no age can be determined was also be removed. The final data set used contained 62,613 stars.

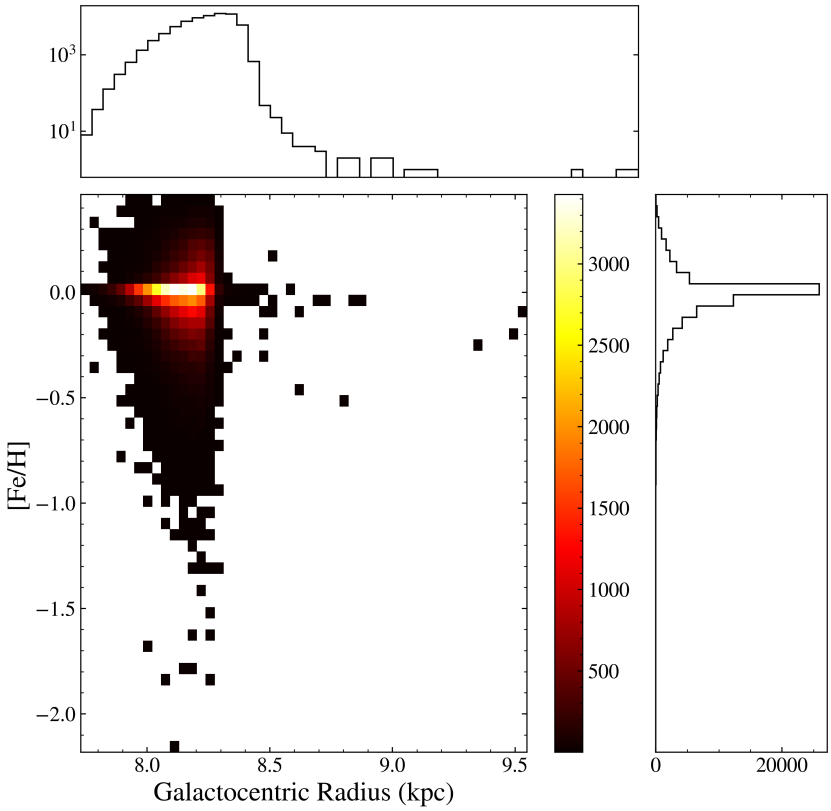

Figure 2.1: Current galactocentric radius versus metallicity as found by Berger et al. (2020). Above of and aside to the 2D histogram are the marginal histograms for each respective quantity. This is the entire stellar population i.e. including stars for which no age could be determined. The median metallicity and median galactocentric radius is -0.002 dex and 8.142 kpc respectively.

Figure 2.1 shows the distribution of the full data (i.e. without considering the age of the stars) with respect their current galactocentric radius and metallicity as found by Berger et al. (2020). Due to the nature of the Kepler mission, most stars lie in the Solar neighbourhood, and have around solar metallicity. Indeed, the median galactocentric radius of the entire sample is 8.142 kpc while the median metallicity of the entire sample is -0.002. Figure 2.2 shows the effective temperature and luminosity in an HR-diagram for the full data and it is clear that most of stars are on the main sequence as expected from the goals of the mission which was to observe Earth-like planets around Solar like stars.

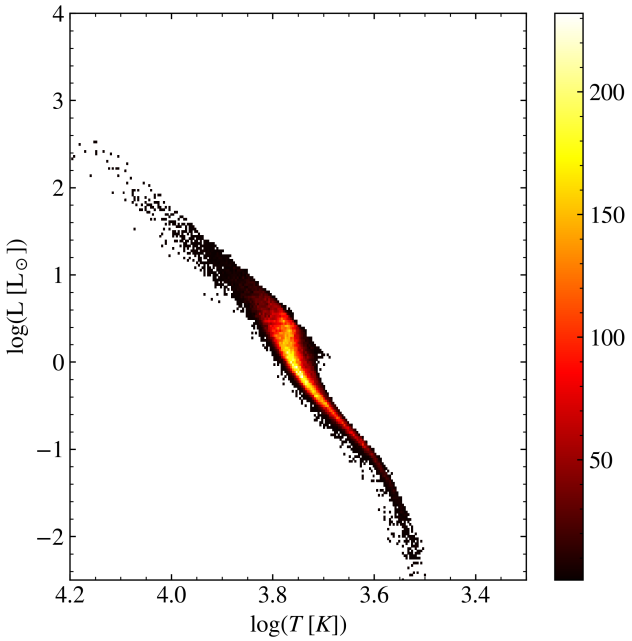

Figure 2.2: HR-diagram showing the effective temperature versus luminosity for the full population. Similar to figure 2.1, this includes those stars for which no age could be determined. Stars with a log $g < 4$ are shown in grey and are faded out in order to show clearly that the giants are the ones removed by this constraint.

2.1.2 Planet Population

The planet catalogue is taken from the Q1-Q17 DR25 Kepler (Thompson et al., 2018) KOI table at exoplanet archive, using only planets flagged as confirmed or candidates. Due to the nature of the transit method, a signal could be a binary star grazing the target star or simply be caused by variability in the light curve. This means that some signals might be false positives and not a planet orbiting around the star. False positives are identified by for example the shape and depth of the light curve. The false positive probability in Kepler is fairly low (up to 10%, see Morton & Johnson, 2011) but in reality it would still necessary to consider and account for false positives in the data (Bryson et al., 2020), however that is outside the scope of this thesis. Including candidates will affect the end result in some way which can be seen in figure 2.3 which shows the detection order distribution for only confirmed planets as well as confirmed planets together with candidates both populations have been cleaned using the cuts as described below. While including candidates does increase the number of systems with planets of higher detection orders by a small margin, the population in the first detection order is increased significantly. This will most likely lead to an increase in occurrence rate estimates for single planets when including candidates.

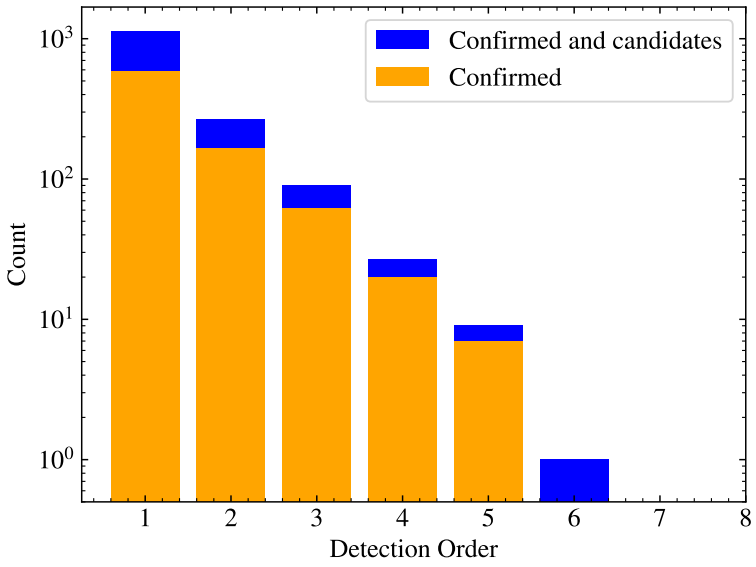

Figure 2.3: The multiplicity distribution of confirmed planets detected by Kepler and confirmed planets together with planet candidates separated by detection order. Including planet candidates does increase the number of higher order systems by a bit but mostly increases the population in lower detection orders.

The planet catalogue was cross-matched with the stellar catalogue to make sure that there existed stellar parameters for all stars with observed planets. Further, we only considered planets with radius R and period P in the ranges $0.5 < R/R_\oplus < 16$, $0.5 < P/\text{day} < 500$ due to the fact that the bias correction model used is only valid in these ranges (see section 3 for more details). We also removed single planet systems with a planet with $R > 6.7R_\oplus$ due to the fact that these systems are most likely a separate population of planets where a giant planet has migrated inwards and scattered the rest of the planets, making it impossible to know which multiplicity these stars were formed with. For details on this population see, for example, Johansen et al. (2012), Steffen et al. (2012), or Mustill et al. (2015). The final planet population can be seen in figure 2.4 with their semi major axes plotted versus their radii together with the Solar System planets. In the planet catalogue, some planets may have no semi major axis given. Therefore, these semi major axes were calculated using Kepler's third law

$$a = \left(\frac{P^2 G M_*}{4\pi^2}\right)^{1/3}. \tag{2.1}$$

2.2 Determining the Ages

The two main components needed in order to infer the formation radius of a star are metallicity and age. Ages of stars are notoriously difficult to determine, mainly due to the

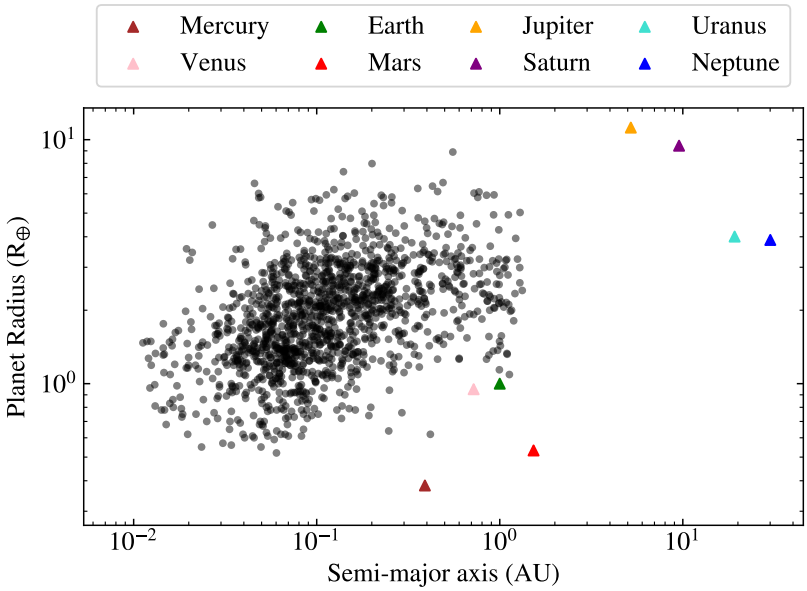

Figure 2.4: Semi major axes and radius of the observed planets in the full planet catalogue. The Solar System planets are also shown to indicate a sense of scale.

fact that the age of a star cannot be determined directly, but rather only through models or empirical methods. There are several methods to determine the ages of stars such as placing the stars on isochrones, gyrochronology, asteroseismology, and others. For a good review, see Soderblom (2010). While the data from Berger et al. (2020) includes ages from their isochrone fitting, for this project, we have determined the ages separately. This is in order to be able to fully capture the probability distribution functions (pdf) of the formation radii. Berger et al. (2020) only report upper and lower uncertainties whereas a separate age determination will produce a full pdf of the age which can then be drawn from.

In order to determine the ages, the method of placing the stars on isochrones will be used. More specifically, a method, which uses a Bayesian fitting algorithm, and was developed by Howes et al. (2019) will be used. This involves calculating the so-called $\mathcal{G}(\tau, [\mathrm{Fe/H}])$-function which is the joint relative likelihood function of both the age and metallicity of a star. $\mathcal{G}$ can be written as

$$\mathcal{G}(\tau, \zeta, \alpha|\mathbf{x}) \propto \int_A \int_\mu \int_{M_*} \xi(M_*)\phi(\mu, A)L(M_*, \tau, \zeta, \alpha, \mu, A|\mathbf{x})d\mu dA dM_*, \tag{2.2}$$

where ϕ is prior density of the extinction A and distance modulus μ, ξ is the mass prior which is the initial mass function of the stellar mass M_* (we use a Salpeter initial mass function), and L is the likelihood function for a multivariate normal distribution of all the parameters. ζ is the theoretical metallicity parameter in the model while α is the theoret-

ical alpha enhancement factor[3] and is usually assumed to be 0. The model parameters are thus: initial stellar mass M_*, age τ, extinction A and metallicity ζ. As K_s magnitudes, which are less affected by extinction, it is be assumed to be negligible. We then marginalise over μ, and M_* assuming a flat prior on μ. The observable stellar data used are magnitudes found using 2MASS K_s photometry, effective temperature $T_{\rm eff}$, and metallicity [Fe/H] as reported by Berger et al. (2020). Further, due to the inclusion of μ, parallax from Gaia DR2 (Gaia Collaboration et al., 2018) as reported by Berger et al. (2020) is included in the calculations. The $\mathcal{G}$- function can be calculated for both the age and metallicity but for this project, the metallicity will be marginalised over and only the age calculation will be used. For the full description of the $\mathcal{G}$-function, see Howes et al. (2019). The isochrones used for the age determination are the MIST isochrones (Dotter, 2016; Choi et al., 2016).[4]

When calculating the $\mathcal{G}$ function, it is important to note that it is not a posterior probability density on its own. Instead the posterior density f for the age can be given as

$$f(\tau|\mathbf{x}) = \frac{\mathcal{G}(\tau|\mathbf{x})\psi(\tau)}{\int_0^\infty \mathcal{G}(\tau'|\mathbf{x})\psi(\tau')d\tau'}, \tag{2.3}$$

as shown in Appendix A of Howes et al. (2019) Here $\psi(\tau)$ is the prior density on the age. A flat prior on the age will be used meaning that the posterior density is proportional to $\mathcal{G}$ and can be found through a normalisation of $\mathcal{G}$. An example of the posterior distribution for the star can be seen in figure 2.5. Finally, it is important to note that there exist a possibility that the posterior density of the age of a given star does not behave nicely i.e. there is no clear peak in the posterior density and a single age cannot be estimated. This is due to the fact that main sequence stars evolve slowly and thus have less distinct isochrone tracks, making isochrone placement a less than ideal method of determining the ages for individual main sequence stars (Soderblom, 2010). This can be seen clearly in figure 2.6 which shows the isochrones for four different metallicities and three different ages for each metallicity.

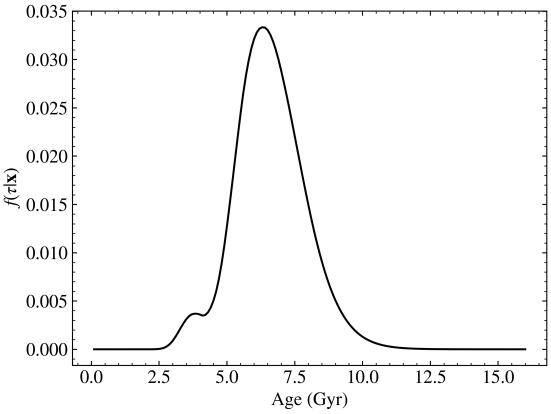

Figure 2.5: Example of the posterior distribution for a star in the sample

[3]The observable counterpart to α is the alpha enhancement [α/Fe]

[4]It should be noted that there are other isochrones sets that could have been used to compare with but since testing different isochrone sets is not the purpose of the project, only MIST was used.

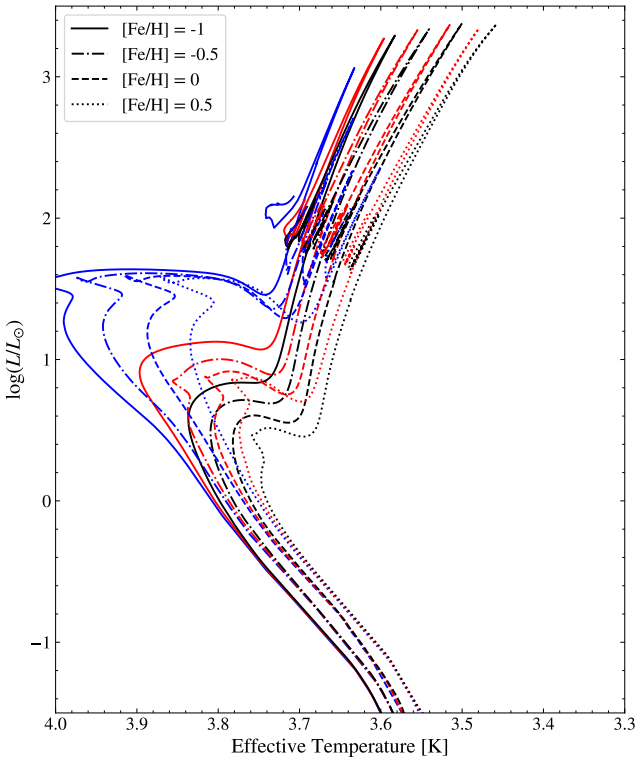

Figure 2.6: Isochrone tracks showing the luminosity versus the effective temperature for four different metallicities. The blue, red, and black tracks for each metallicity represent an age of 1, 3, and 6 Gyr respectively.

In order to test the effect of poor age determination, we estimated the ages of stars which have confirmed planets found by WASP, HAT-P, and HARPS (Pollacco et al., 2006; Hartman et al., 2004; Pepe et al., 2004). The stellar parameters for the WASP and HAT-P stars were extracted from the SweetCat catalogue (Santos et al., 2013) and the V-band magnitude was used instead of the SDSS K_s-band magnitude. Through manual inspection, we found stars which showed a clear peak in the posterior density of the age. The formation radii of stars with distinct ages and the full samples of stars were then compared and the difference between the two samples were found to be very negligible. This means that it is possible to include stars where the pdf does not have clear peaks (some stars may have a completely uniform age pdf for example) without affecting the results significantly.

After determining the ages of all the stars in the final stellar catalogue, we removed those where no values of the $\mathcal{G}$-function could be determined at all. In practice, this meant that the values for the $\mathcal{G}$-function for all age values in the grid was zero. This can happen in

the case when a star lies outside of the isochrones and it is not possible to place the star on any isochrone. This has a minor effect on the final result as can be seen in figure 2.7 which show the same as figures 2.2 and 2.1 but with those stars where no age could be determined was removed. In total, the number of stars removed was 2,349. As can be seen by comparison, there are very minor differences which are not expected to affect the end result.

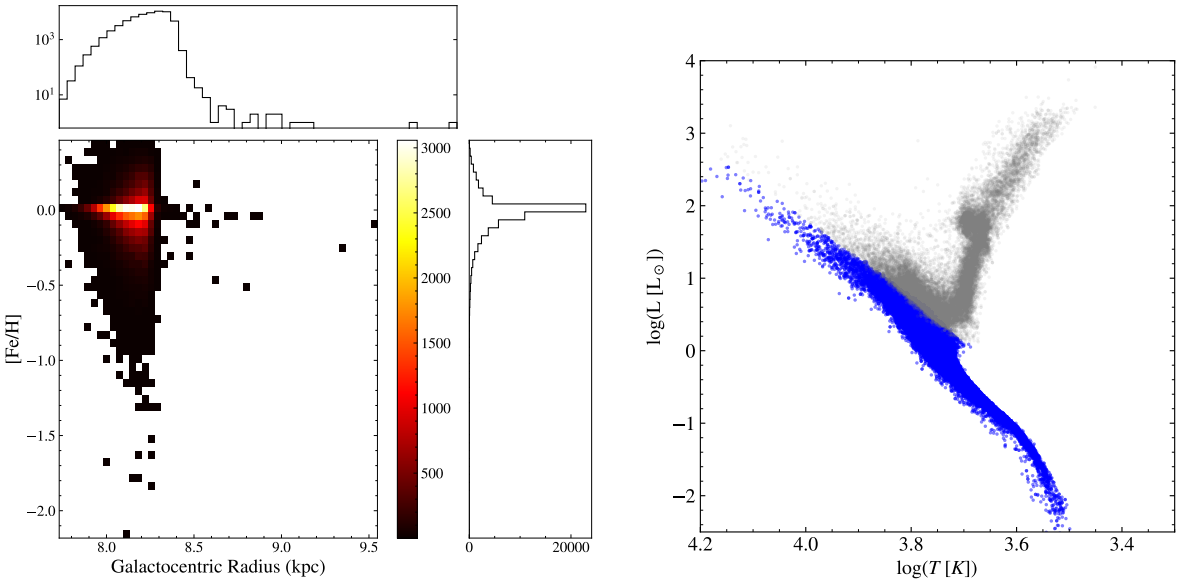

Figure 2.7: Figures similar to figures 2.2 and 2.1 but stars where no age could be determined at all are removed.

It is also useful to compare the ages we found in this project with that of Berger et al. (2020) to check if the results are somewhat similar. Figure 2.8 show histograms of the median ages of the full final stellar sample found in this project as well as the ages of the same stars found by Berger et al. (2020). While the median ages of the sample are similar it is clear that there is a pile-up of very young and very old ages in this project. The pile-up at the old end could simply be caused by the fact that the grid used in this project only extends to 16 Gyr while the grid used by Berger et al. (2020) extends even further. The fractional uncertainty on the ages can be calculated by dividing the maximum of the lower and upper uncertainty (in this case the 32nd and 68th percentile, respectively) with the median.[5] The fractional uncertainty on the ages found in this project is found to be 33% while the fractional uncertainty on the ages found in Berger et al. (2020) is found to be 51%.[6]

[5]This is the same method that is used in Berger et al. (2020)

[6]A slightly lower value than the median fractional uncertainty of their full stellar population which is 56%

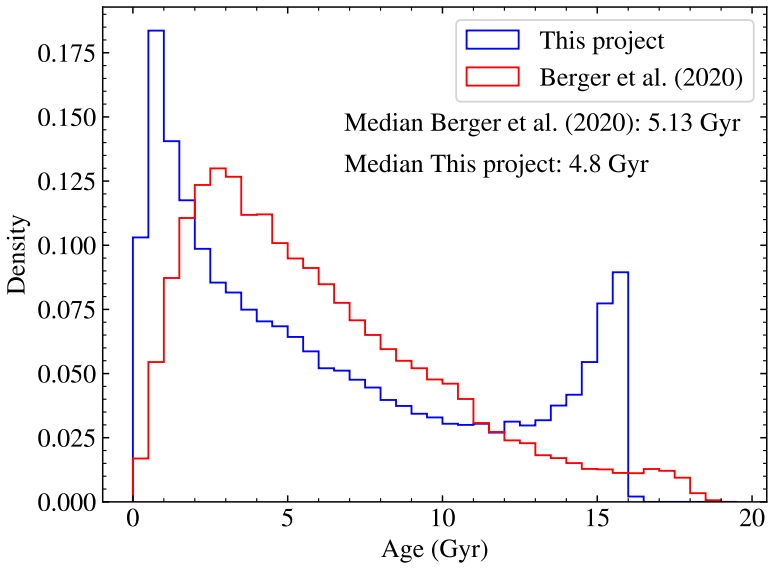

Figure 2.8: Histograms of the median ages of the full final stellar sample calculated in this project and those found by Berger et al. (2020). The median ages found in this project is 4.8 Gyr while the median of the ages found in Berger et al. (2020) is 5.13 Gyr.

2.3 Metallicity gradient models

The radial metallicity gradient models that will be used are the ones developed by Kubryk et al. (2015), Frankel et al. (2018), Sharma et al. (2020), and Minchev et al. (2018). The Minchev, Frankel, and Sharma models have a parameterised description of the metallicity evolution over time while the Kubryk model is based on an N-body+SPH (Smooth Particle Hydrodynamics) simulation of a Milky Way-like galaxy. The Minchev model describes the metallicity as a function of age at the Sun's position in the galaxy as well as the metallicity gradient with respect to the radius as a function of time (figure 2.10). These functions were found using a trial and error method where the two functions were varied until the results matched observations in the solar neighbourhood. The Sharma and Frankel models are a bit more detailed and includes a few more ingredients where the parameters were found using Markov Chain Monte-Carlo (MCMC) simulations.

Figure 2.9 shows the how the metallicity changes with radius for some given ages. As can be seen, they fulfill the requirements that given one of the parameters, the metallicity is either strictly decreasing or increasing with respect to the other parameter. It should be noted that the model by Sharma et al. (2020) has a very rapid metallicity evolution with time for the first few billion years but that there is little to no change in later years. The Minchev, Sharma, and Kubryk models all show a flattening radial metallicity gradient with

time while the Frankel model has the same slope for all times. The validity of a flattening of the metallicity gradient is still debated. As discussed by Anders et al. (2017), it could be possible that the Milky Way formed as a disc with some pre-enriched metallicity where the innermost part evolves faster (due to a higher number density of stars) thus causing a steepening of the metallicity gradient. However, the Milky Way could also have been formed with a initial, steep metallicity gradient and gradually formed more and more stars in the outer parts, flattening the gradient with time. Radial migration of the gas would also be expected to cause a flattening of the gradient with time. This means that a flattening of a gradient would be the more likely alternative meaning that either the Minchev or Kubryk model are the most physically correct models.

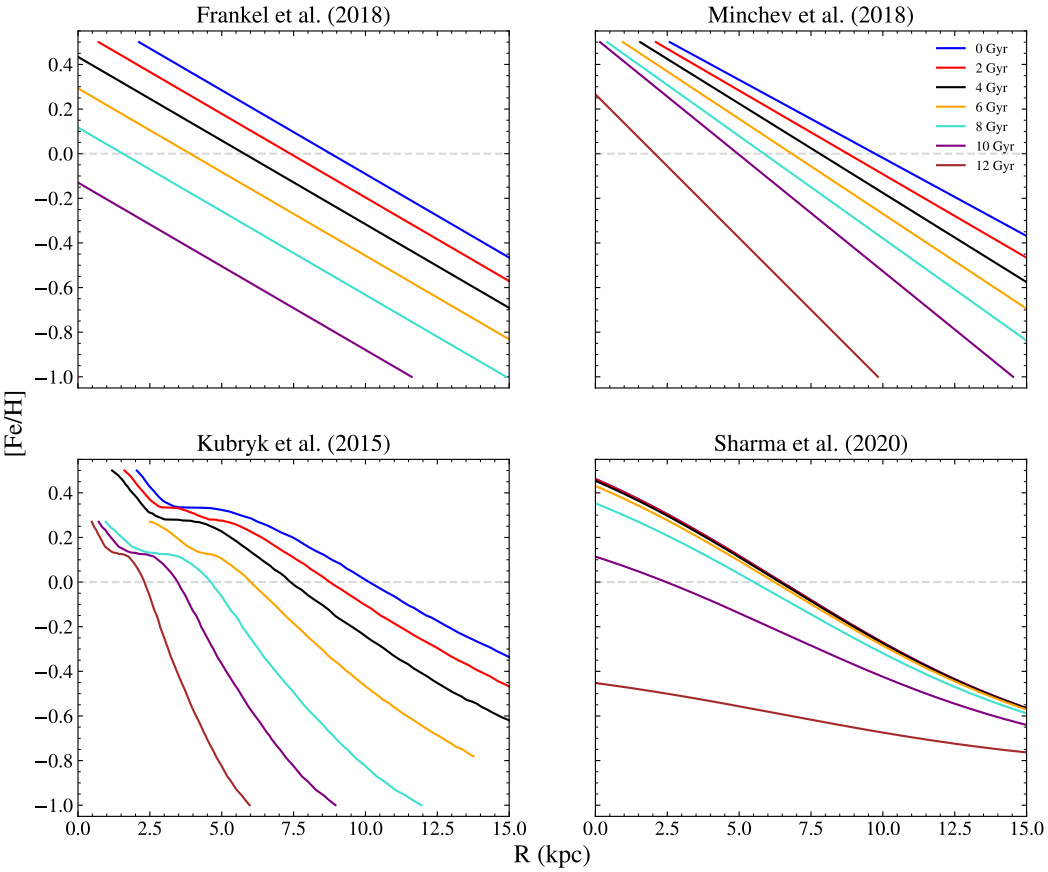

Figure 2.9: The four different metallicity gradient models used. All the models have the assumptions that the metallicity is strictly decreasing with radius for a given age and that for a given radius, the metallicity is increasing with age. The grey dashed line show the median metallicity of the stellar population used.

2.3.1 Minchev et al. 2018

Minchev et al. (2018) developed a model of the metallicity gradient using observational data from a local sample of stars from HARPS-GTO (Delgado Mena et al., 2017) and AMBRE:HARPS (de Laverny et al., 2013). They argue that the youngest stars should be peaked around the solar radii while older stars should successively have formation radii closer to the center of the galaxy as a consequence of inside out galaxy formation. They also argued that the ISM is well mixed at a given radius (i.e. stars are born with the same metallicity at a fixed radius and time) and that the oldest star in the population should be have a formation radius of about 4 kpc to prevent negative formation radii. Through an iterative process, they then apply a metallicity gradient to their observational data in order to produce the desired formation radii distribution. Throughout each iteration, they assumed the present day metallicity gradient to be 0.07 dex kpc^{-1} and strictly decreasing. This is slightly steeper than that of Genovali et al. (2014) who found the present day gradient to be 0.06 dex kpc^{-1}. The final results are shown in figure 2.10 which shows the metallicity [Fe/H]($\text{R}_\odot$) as a function of age τ at the Sun's position in the galaxy and the metallicity *gradient* $\frac{d[\text{Fe/H}]}{dR}$ with respect to the galactocentric radius as a function of τ. Using these two functions it is possible to determine the formation radius of a star given its metallicity [Fe/H] and age τ by linearly interpolating between the metallicity of the ISM in the galactic centre $[\text{Fe/H}]_0$ and at the edge of the galaxy $[\text{Fe/H}]_{max}$ where

$$\begin{aligned} [\text{Fe/H}]_0 &= [\text{Fe/H}](\tau) - \frac{d[\text{Fe/H}]}{dR}(\tau) \cdot R_\odot \\ [\text{Fe/H}]_{max} &= \frac{d[\text{Fe/H}]}{dR}(\tau) \cdot R_{max} + [\text{Fe/H}]_0. \end{aligned} \tag{2.4}$$

Of course, the need for interpolation makes it necessary to make assumptions about R_{max}, something Minchev et al. (2018) did not have to do. For the purpose of this project this is chosen to be 18.5 kpc which is in between the ranges reported for the HI disc found by Nakanishi & Sofue (2016). The distance from the Sun to the galactic center is set to be 8.3 kpc (Gillessen et al., 2017).[7].

[7]In fact, a newer value has been found to be 8.18 kpc (Gravity Collaboration et al., 2019) However, the formation radii were calculated before this was realised. This small change would not lead to a significant enough change in the formation radii so 8.3 kpc is the value used.

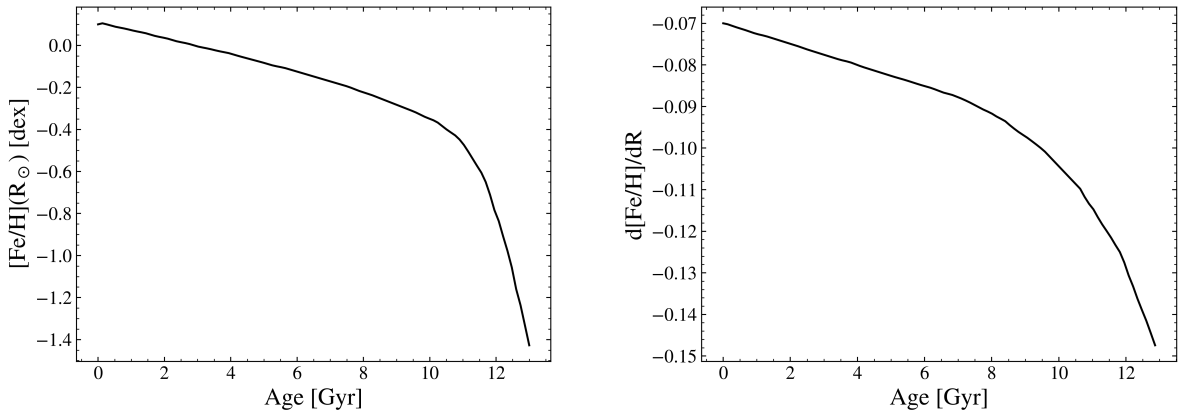

Figure 2.10: ***Left:*** *The metallicity as a function of age at the Sun's position in the galaxy.* ***Right:*** *The rate of change of the metallicity with radius as a function of age. These are both according to the Minchev model. Combining these two allows for the determination of the metallicity gradient as a function of time and galactocentric radius.*

2.3.2 Frankel et al. 2018

The Frankel et al. (2018) model was developed from an attempt to quantify radial migration in the galaxy through a parameterised model. The only assumptions they made are that the metallicity is strictly decreasing with increasing radius for a given age and the metallicity is increasing with time for a given radius but at a decreased rate. They then find an expression for the metallicity as a function of time and galactocentric radius. It is straightforward to invert this expression so that we can get the formation radius as a function of metallicity and age. The expression is then given as

$$R = \frac{[\mathrm{Fe/H}] - F_m + (F_m + \nabla[\mathrm{Fe/H}]R^{now})\left(1 - \frac{\tau}{\tau_m}\right)^{\gamma}}{\nabla[\mathrm{Fe/H}]}, \tag{2.5}$$

where the description of each parameter and their found values can be seen in table 2.1. They fit the values for the parameters using MCMC maximum likelihood estimation on data of red clump giants taken from APOGEE, a red giant survey measuring detailed abundances and radial velocities of stars in the infrared throughout the galaxy (Majewski et al., 2017).

Table 2.1: Descriptions and values for each parameter found by Frankel et al. (2018) in their model

Parameter	Value	Description
F_m	-1 dex	Metallicity of the ISM at the center of the disc at a lookback time of 12 Gyr
τ_m	12 Gyr	Maximum disc age
R^{now}	8.79 kpc	Radius at which the current birth metallicity is solar
γ	0.36	Enrichment factor for the time evolution of the metallicity
∇[Fe/H]	-0.075 dex/kpc	Present day metallicity gradient

2.3.3 Kubryk et al. 2015

In the Kubryk et al. (2015) model, instead of parameterising the metallicity gradient explicitly through observational data, they instead developed a generic model for how a given isotope in their simulated galaxy chemically evolved, which depended on the star formation rate (SFR), isotopic mass release, net infall of the isotope from outside of the galaxy, and net infall of the isotope from adjacent regions within the galaxy. Radial migration was then implemented by considering the probability of a certain stellar mass (of stars that are still alive) migrating to a different zone within the timestep. They then report the metallicity in their galaxy as a function of radius for three different times, 4, 8, and 12 Gyr in simulation time (or 0,4, and 8 Gyr in lookback time which is shown in figure D.3). This requires an interpolation scheme between the lines in order to get a formation radius. Throughout the interpolation, it is assumed that the evolution in time is linear (i.e. fairly similar to that of Frankel et al. (2018)). The interpolation is done by following the conditions

$$R = \begin{cases} L_0 \quad \text{if} \quad \tau = 0 \\ L_0 + (L_4 - L_0)\frac{\tau}{4} \quad \text{if} \quad \tau \leqslant 4 \\ L_4 \quad \text{if} \quad \tau = 4 \\ L_4 + (L_8 - L_4)\frac{\tau - 4}{4} \quad \text{if} \quad 4 < \tau < 8 \\ L_8 \quad \text{if} \quad \tau = 8 \\ L_8 - L_8\frac{\tau - 8}{8} \quad \text{if} \quad \tau \geqslant 8. \end{cases} \tag{2.6}$$

In equation (2.6), L_i is the line given by Kubryk et al. (2015) (and shown in figure D.3) for a lookback time i and τ is the age of the star in Gyr. This will allow for the evaluation of a

formation radius of a star between two lines by considering a linear interpolation between the formation radii on the lines. A drawback with this interpolation is that it is only possible to consider stars which have metallicity within the range of the upper and lower line i.e. it is not possible to evaluate the formation radius of a star with age between 4 and 8 Gyr and a metallicity above ≈ 0.35 dex as the 4 Gyr line is not defined there. This is why the yellow and turquoise seem to have a ceiling in the bottom left figure in figure 2.9. For stars with age $\tau > 8$ Gyr, the lower formation radius limit was set to 0 Gyr as there is not a line to help set a lower limit here.

2.3.4 Sharma et al. 2020

Finally, the Sharma et al. (2020) model was developed in a similar fashion to the one by Frankel et al. (2018) in that it is a parameterised model of the evolution of the metallicity as a function of time and space. In fact, it is mainly an updated version of the model constructed by Sanders & Binney (2015) which provided a lot of the arguments used by Frankel et al. (2018). However, instead of performing a maximum likelihood estimation of observational data, Sanders & Binney (2015) fit a metallicity function to a galaxy simulation performed by Schönrich & Binney (2009) which aimed to match the Milky Way. Sharma et al. (2020) then updated the prescription by adding new prescriptions for the distribution of α-elements and the velocity dispersion of the stars. This then allowed for the formation radius to be expressed as

$$R = \frac{F_{min}}{\nabla[\mathrm{Fe/H}]} \ln\left(-\left(\frac{2F_{min}\tanh\left(\frac{\tau_{max}-\tau}{\tau_F}\right)}{[\mathrm{Fe/H}]-F_{min}}+1\right)\right), \tag{2.7}$$

where the different parameters are described together with their found values in table 2.2. As can be seen in figure 2.9, the Sharma et al. (2020) model has a much weaker time evolution of the gradient at later stages of the galaxy evolution. The radial evolution also isn't a linear relation allowing for a steeper radial profile further into galaxy which might be more physically accurate due to the presence of the bulge in the center of the Milky Way.

Table 2.2: Parameters describing the Sharma et al. (2020) model with their description and values as found by Sharma et al. (2020)

Parameter	Value	Description
F_{min}	-0.85 dex	Birth Metallicity at time $\tau = 0$[a]
τ_F	3.2	ISM metallicity enrichment scale
τ_{max}	13 Gyr	Disc Age today
∇[Fe/H]	-0.08 dex/kpc	Current day metallicity gradient

[a]In Sharma et al. (2020) this parameter is only described as "minimum metallicity" or "birth metallicity". From context, it is assumed to be the ISM metallicity at the time of birth of the galaxy

Chapter 3

Accounting for detection bias in detecting exoplanets

This chapter describes how we account for detection bias using data from the Kepler mission, and thus are able to estimate the occurrence of specific multiplicities. Zink et al. (2019) found that after detecting the first planet in a system, the probability of detection decreases for subsequent planets. Therefore, it is necessary to debias the different multiplicities separately which in turn will have the positive consequence of being able to tell us something about how different multiplicities are affected by the formation radii of their hosts. The statistical method for debiasing as well as all the components required is described in this chapter. The chapter ends with a description of the likelihood function that is calculated in order to estimate the fraction of stars which are hosting planets, given a stellar population.

When aiming to study anything about the exoplanet population as a whole, it is important to take into account completeness or detection bias. It is impossible to draw any useful conclusion without doing so simply due to the fact that any indirect method of detecting planets is heavily favoured towards a certain kind of planet. The radial velocity method requires a large enough pull on the star from the planet which means that more massive planets are more likely to be detected due to them having a stronger signal. Similarly for the transit method, larger planets are more likely to be detected since they have a greater transit depth. Further, planets on shorter orbital periods are also more likely to be detected in a survey because their transit or radial velocity signal occurs more frequently, and for the transit method, planets on smaller orbits are more likely to transit and have larger signals. Indeed, this can be seen in figure 1 of Burke et al. (2015) which shows the completeness model of Kepler for a single star and shows that massive planets closer to the star are much more likely to be observed by Kepler.

With this in mind, it is necessary to develop a statistical model which takes into account the different probabilities for the detection of a planet with certain characteristics. The

model used in this project is based heavily on that of Zink et al. (2019) which is a modified Poisson process likelihood model and an extension of that of Youdin (2011) which is a detection bias model for the Kepler mission. Youdin (2011) developed their model assuming that detections of multiple planets around a given star are independent of each other. Zink et al. (2019) instead included the order of detection in their model and showed that after detecting the first planet, the detection probability decreased significantly enough to affect the occurrence estimate. In the following sections, only the actual quantities which are necessary for the final likelihood function will be written out. For a full explanation of each quantity, see appendix A

3.1 Detection Probability

The full detection probability of a planet with radius R and orbital period P can be written as the product between three different probabilities: the probability P_{rec} that the Kepler pipeline recovers the transit signal, the window probability P_{win} that the planet performs at least three transit during the observation time of Kepler, and the probability P_{tr} that a transit actually occurs from our point of view

$$P_{det,m} = P_{tr,m}(P)P_{win}(P)P_{rec,m}(R,P). \tag{3.1}$$

The recovery probability is ultimately dependent on the Multiple Event Statistic (MES) which is an internal metric within the Kepler pipeline and is analogous to a signal-to-noise ratio. The MES is in turn dependent on the transit depth and the number of transits where the main contributions are from the stellar radius, stellar noise, planet period, and planet radius. Christiansen (2017) found that the recovery probability is well fitted by a gamma function. Zink et al. (2019) expanded upon this and found that the fitted parameters for the gamma function varies significantly with detection order m.

The window probability takes into the account the probability that enough transits can occur within the observation time and is thus only dependent on the period of the planet, the duty cycle, and the observation time. The reason for expanding the window probability to account for three transits is due to the Kepler pipeline detection requirement of at least three transits.

Finally, the geometric transit probability is found through a Monte-Carlo simulation performed by Zink et al. (2019) which looked at 10^6 different lines of sight and calculated the probability for all planets transiting given a detection order. The transit probabilities can be seen in A.1.

3.2 Exoplanet distribution

As well as the detection probability, it is also necessary to model the exoplanet population. Previous studies have done extensive work in characterising this population (see Youdin

(2011), Burke et al. (2015), for example) and it is also included in the Zink et al. (2019) model. It is therefore reasonable to adopt the model as found by Zink et al. (2019) since their model is the one used for the project. The total population of exoplanets are described by two separated functions of orbital period and planetary radii

$$\frac{d^2N}{dRdP} = fg(R)q(P), \tag{3.2}$$

where f is an entangled occurrence factor which will be described more in detail later on while g and q are the power laws describing the exoplanet population with respect to planet radius and orbital period respectively.

3.3 Sorting Order

When the Kepler pipeline searches the light curves for planet candidates, it finds them in order of decreasing MES. The MES is the Multiple Event Statistic and is measured for each transit. It is analogous to a S/N for a planet transit. As can be seen from equation (A.3) in appendix A.1, the MES scales $\propto R^2/P^{1/3}$ meaning that large planets on short period orbits are usually found earlier. The planet distribution over all orders will then be skewed such that for $m = 1$, most of the planets found will be large planets closer to the star. Zink et al. (2019) accounted for this skew by implementing a joint distribution model with the skew being described by a function $O_m(R, P)$ which in turn is dependent on the CDF of $g(R)$ and $q(P)$. This function was found through sorting a uniform distribution of orbital periods and planetary radii using a similar sorting as the Kepler pipeline and then fitting the resulting distribution to a beta distribution. Should the detection with respect to different orders be independent of each other, the resulting distribution would be uniform as well i.e. the parameters in the beta distribution are both equal to 1. For a full description of the function and found parameters, see appendix A.3.

3.4 Likelihood Function

With all the necessary components set up it is now possible to set up the full likelihood function. Handling the detection of each order m as a Poisson process, the likelihood function can be written as

$$\mathcal{L} = \prod_{m=1}^{m_{max}} \left[\prod_{i=1}^{N_{det,m}} f_m P_{det,m}(P_i, R_i) g(P_i) q(R_i) \right] e^{-N_{exp,m}}, \tag{3.3}$$

where $N_{det,m}$ is the number of detected planets and $N_{exp,m}$ is the expected number of detected planets of the order m. m_{max} is the maximum detection order considered. $N_{exp,m}$ is defined as

$$N_{exp,m} = N_* f_m \int_{P_{min}}^{P_{max}} \int_{R_{min}}^{R_{max}} P_{det,m}(P, R) g(P) q(R) O_m(R, P) dP dR, \tag{3.4}$$

where P_{min}, P_{max}, R_{min}, and R_{max} are the limits of each quantity. Following Zink et al. (2019) these limits are set to 0.5 days, 500 days, 0.5 $R_\oplus$, and 16 $R_\oplus$ respectively. Taking the logarithm of the likelihood, we can then get

$$\ln \mathcal{L} = \sum_{m=1}^{m_{max}} \left(\sum_{i=1}^{N_{det,m}} \ln[f_m P_{det,m}(P_i, R_i) g(P_i) q(R_i)] - N_{exp,m} \right). \tag{3.5}$$

The occurrence factor f_m is independent of the detected planets, and for a given m, $N_{exp,m}$ is simply proportional to f_m, which means that the expression can be simplified further

$$\ln \mathcal{L} = \sum_{m=1}^{m_{max}} N_{det,m} \ln f_m + K_{1,m} - f_m K_{2,m}, \tag{3.6}$$

where $K_{1,m}$ is the sum over all detected planets given order m and $K_{2,m} = N_{exp,m}/f_m$. This can only be done because the parameters describing O, g, and q that were found by Zink et al. (2019) are taken to be true. It should be noted that Zink et al. (2019) had a maximum order of seven while the highest order in the full planet population used in this project is six as can be seen in figure 3.1. As the likelihood function is based on a Poisson process, it can be shown that the likelihood of a certain value of f_m being true given zero detections ($P(f_m|N_{det} = 0)$) is proportional to $\exp(-N_{exp})$.[1] This holds true in this modified Poisson process by Zink et al. (2019) as well, since for $N_{det,m} = 0$, $K_{1,m} = 0$ meaning that the only contributing factor for these orders is $-f_m K_{2,m} = -N_{exp,m}$.

[1]The derivation of this is shown in appendix B

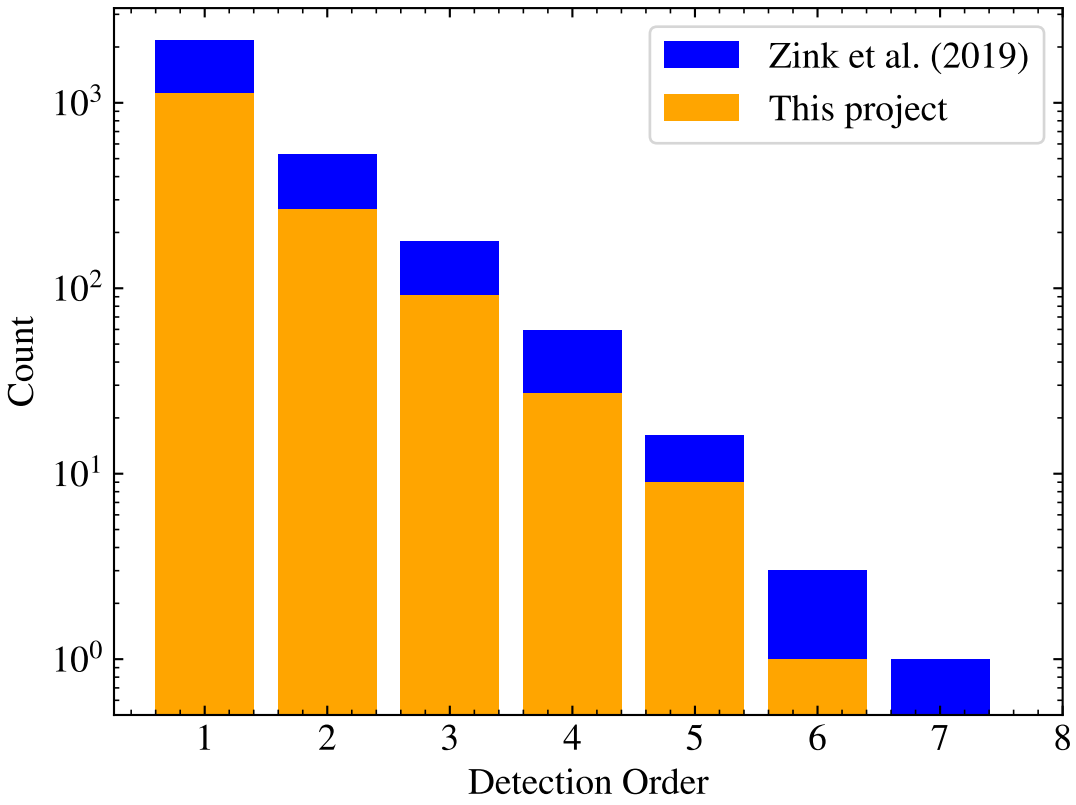

Figure 3.1: Multiplicity distribution of the planet population used in this project overlapping that of Zink et al. (2019) (i.e. the bottom of the bar for both populations is zero). Due to the log-scale on the y-axis, it is clear to see that the differences between the systems with lower multiplicities is significant.

The double integral in equation (3.4) should also be normalised which can be done by considering the case of a detection probability of unity for all orders. Then, the number of expected detected exoplanets at order m should be equal to $N_* f_m$ which means that it is possible to introduce a normalisation factor $C_{norm,m}$ which ensures

$$C_{norm,m} \int_{P_{min}}^{P_{max}} \int_{R_{min}}^{R_{max}} g(P) q(R) O_m(R, P) dP dR = 1. \tag{3.7}$$

Through the separability of this expression, this can be found to be

$$C_{norm,m} = \left(\frac{\Gamma(a_{m,r})\Gamma(b_{m,r})}{\Gamma(a_{m,r} + b_{m,r})} \frac{\Gamma(a_{m,p})\Gamma(b_{m,p})}{\Gamma(a_{m,p} + b_{m,p})} \right)^{-1}, \tag{3.8}$$

Where Γ is the gamma function. The full derivation can be seen in Appendix C. The quantity f_m is an occurrence factor and not the *true* occurrence rate. In order to get the true

occurrence rate, these need to be disentangled. The reason for this is because a multi-planet system might only have one or two transiting planets observed, meaning that f-values for lower m will get a larger contribution than the underlying, true occurrence. Instead, it is necessary to consider F_m which instead is the true occurrence rate and describe the fraction of stars with *at least* m planets. Using the same simulation as for the transit probability, Zink et al. (2019) found the following expression for disentangling the occurrence rates

$$f_m = F_m + \sum_{n=m+1}^{7} F_n \frac{P(n|\overline{(m:n-1)})}{P(m)}, \tag{3.9}$$

where $P(n|\overline{(m:n-1)})$ describes the probability that planet n is detected given that planets $(m:n-1)$ are not detected and $P(m)$ is the probability of detecting planet m. Should all mutual inclinations be completely isotropic and the detection of each planet is independent of each other, the ratio in the sum would be equal to one. The found values for these mixture probabilities can be seen in table 3.1. Finally, it might be useful to instead of considering the fraction of stars with at least m planets, to consider the fraction of stars G_m with exactly m planets which can be defined as

$$G_m = F_m - F_{m+1}. \tag{3.10}$$

Here, it is important to consider the fact that the maximum possible number of planets in a system in this model is seven. It could be possible to assume that the maximum number of planets in any system is not higher than this limit which would mean that $G_7 = F_7$, however this would not be accurate as at least one eight planet system has been found to exist (Shallue & Vanderburg, 2018). Therefore, F_7 will only be used to find the value for G_6 meaning that the range of orders considered for G is between zero and six where G_0 can be found from the fact that $F_0 \equiv 1$.

Table 3.1: The mixture probabilities as found by Zink et al. (2019).

Fraction	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
$\frac{P(2\|\overline{(m:1)})}{P(m)}$	0.67	-	-	-	-	-
$\frac{P(3\|\overline{(m:2)})}{P(m)}$	0.68	0.5	-	-	-	-
$\frac{P(4\|\overline{(m:3)})}{P(m)}$	0.53	1.05	0.50	-	-	-
$\frac{P(5\|\overline{(m:4)})}{P(m)}$	0.53	1.12	1.52	0.46	-	-
$\frac{P(6\|\overline{(m:5)})}{P(m)}$	0.37	1.07	1.85	1.69	1.22	-
$\frac{P(7\|\overline{(m:6)})}{P(m)}$	0.33	0.71	1.64	1.90	1.25	1.22

To calculate the maximum likelihood estimate for the occurrence rates `emcee` (Foreman-Mackey et al., 2013), an affine-invariant ensemble sampler (Goodman & Weare, 2010) was used. The walkers in the MCMC were thus allowed to explore the full range of distribution

for all the parameters F_m which is the output given by `emcee`. A uniform prior is used for each of the parameters F_m. F_1 is allowed to range between 0 and 1 while, in order to maintain the order of the occurrence rates F_m is allowed to range between 0 and F_{m-1}. The physical motivation of this is due to the fact that F_m is the fraction of stars hosting at least m planets meaning that it should be impossible for e.g F_2 to be larger than F_1 as a star can't host at least two planets without hosting at least one. After experimenting with different step numbers, a total of 10000 steps with 200 walkers per MCMC was deemed appropriate to allow the MCMC to converge[2]. After the run, 8000 steps were discarded and the chain for each dimension was thinned by the half of the mean of the auto correlation times calculated using the built in function in `emcee`.

[2]An example of the convergence of the walkers can be seen in figure D.4

Chapter 4

Formation radius distributions

In this chapter, the formation radius distributions of the data are shown. In particular the distributions are shown when binned by stellar age, stellar mass, stellar radius, and effective stellar temperature. The formation radius distributions are all shown to behave as expected. When binning by age, we see that young stars have not had time to migrate very far while older stars are formed further in in the galaxy. The results reflect an inside out formation of the galaxy which is expected. When binning by stellar radius and stellar mass, no difference are found due to the connection between radius and mass for main sequence stars. Cool, less massive, smaller stars are seen to be formed further in in the galaxy due to the fact that older stars are generally smaller, less massive and thus cooler.

In order to fully capture the details of the posterior densities of the age, and thus the formation radius, all stars in the sample were iterated over and 10000 values were drawn from their posterior densities of the age and metallicity where the uncertainty in metallicity is assumed to behave as a Gaussian distribution. Berger et al. (2020) report an upper and lower uncertainty in all their stellar parameters so the mean between these were taken to be the uncertainty. The relative values of the upper and lower uncertainties were determined to not be significantly different so this approximation should not affect the end results in any major way.

4.1 Binning by age

The full formation radius distribution for the four models can be seen in figure 4.1 which show the density for the formation radius. The distributions were smoothed through a kernel density estimation. The three colours represent three different age intervals: $0 < \tau/\text{Gyr} < 4$, $4 < \tau/\text{Gyr} < 8$, and $8 < \tau/\text{Gyr} < 12$ and as can be seen, all models satisfy the two criteria previously laid out: the youngest age bin should peak near the median current galactocentric radius in the entire sample and older age bins should peak at subsequently lower formation radii. Not all models are constructed in a way that does

not yield any negative formation radii which is why there exists negative formation radii in some of the models. Interestingly, the Kubryk et al. (2015) model shows a slight bump for all age bins with a second, small peak at around 1 kpc. By observing the Kubryk model in detail, we can see that in order for young stars to be formed close to the center of the galaxy, they have to be very metal rich and for very old stars they have to be at around solar metallicity. Most stars in the population considered are around solar metallicity meaning that the peak in the oldest bin could be explained by the pile up of older ages that exists in the population as a lot of those stars would be expected to be of solar metallicity. The origins of this bump are discussed more in detail in chapter 4.3.

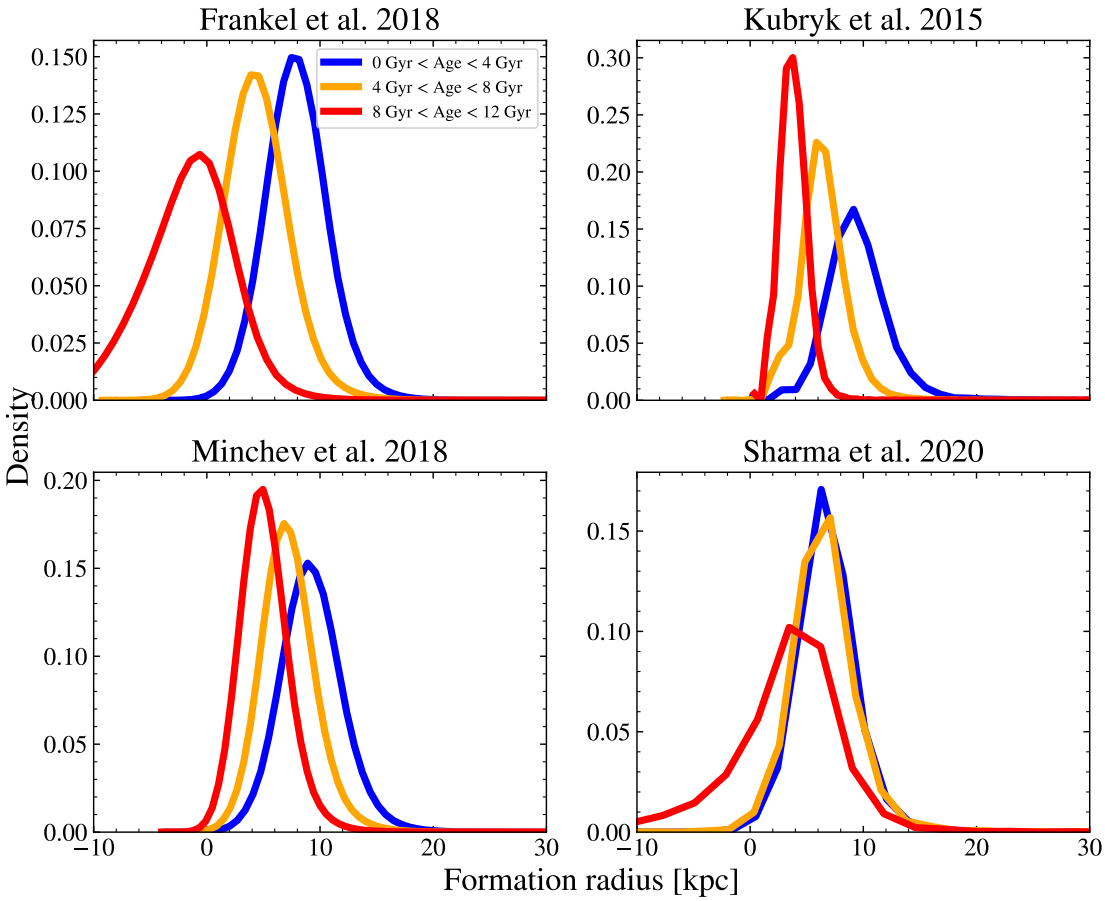

Figure 4.1: Found formation radii distributions for the entire stellar population. The distributions were divided into three age bins in order to check for signs of inside-out galaxy formation. Some models do not restrict the formation radius found to be negative which means that during the interpolation process, some negative formation radii might be found.

Figure 4.2 is the same plot as figure 4.1 but with only stars which are hosts to planetary systems detected by the Kepler mission. Comparing the two figures, we note that it has the same characteristics as the full stellar sample. The host star sample has peaks around similar formation radii as the full sample and show the same inside out galaxy formation. These similarities hold for all four models. This means that those stars which have detected planets are representative of the sample as a whole and are not biased towards a certain formation radius in the galaxy. This result becomes incredibly important when investigating the dependency of formation radius on the occurrence rate. Should most stars with detected planets be formed at a very specific location in the galaxy, then the occurrence rates would be biased towards that location. This result shows that this is not the case.

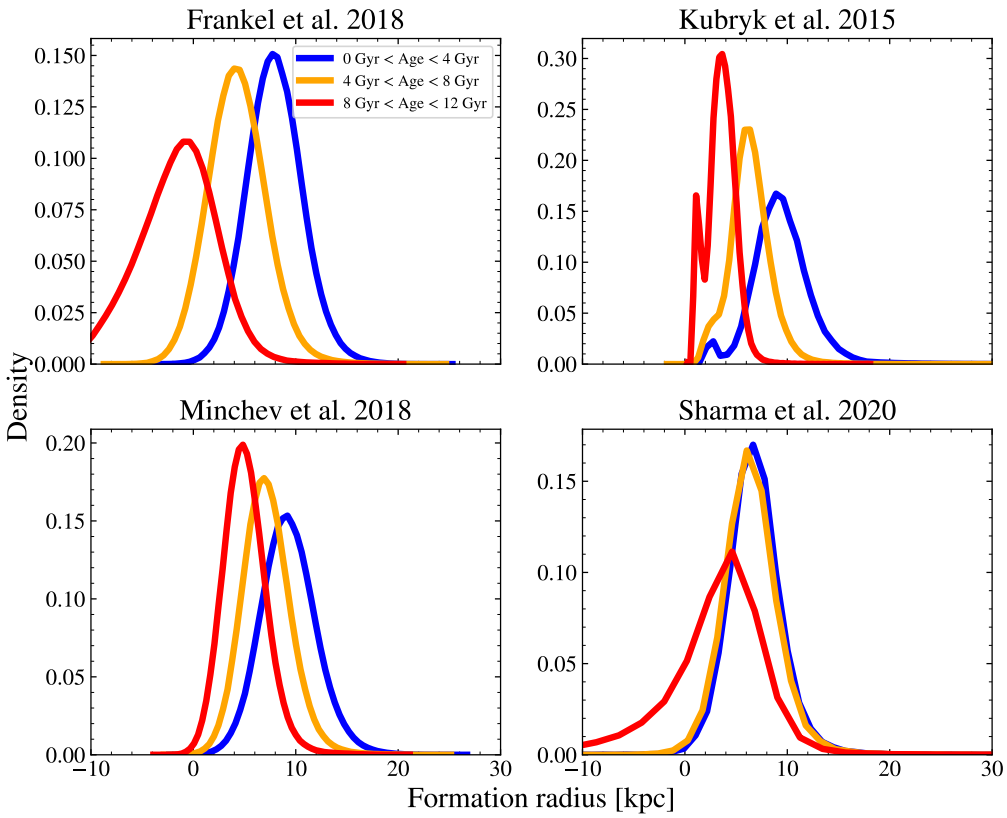

Figure 4.2: Found formation radii distribution for only stars which are hosts to planetary systems.

4.2 Binning by other parameters

Figure 4.3, similarly as figure 4.1 show the formation radius distributions of the full stellar population but with the stars binned by stellar radius, effective temperature, and mass respectively instead of age. Looking at the stellar radius bins, we see that the smaller stars are formed further in in the galaxy compared to larger stars. Similarly, less massive stars are formed further in in the galaxy compared to more massive stars. This effect is most clearly seen in the Frankel and Minchev models but is generally present in all four models. As stars with different radii are roughly similarly distributed throughout the galaxy with the exception of the smallest stars, the detection probability can be thought of as being approximately similar for all formation radii as the stellar parameter considering that the largest effect on the detection probability is stellar radius. Binning stars by stellar temperature shows that the hottest stars (> 7000 K) have not migrated significantly while cooler stars have have been formed further in in the galaxy and migrated outwards. This is due to the significantly shorter lifetimes of hot and massive stars which means that they do not live long enough to migrate far. When using the Kubryk model we can see that the difference between the coolest temperature bin and the rest are much greater compared to when we are using the other three models. Yang et al. (2020) investigated the effect stellar temperature had on occurrence rate and found that the occurrence rate decreased significantly around stars with an effective temperature $> 5500K$. Therefore, when using the Kubryk model, this will have to be taken into account as the occurrence rates for low formation radii might be higher simply due to the temperatures of these stars. This effect is not expected to be as large when using the three other models as the difference between the coolest bin and the rest is significantly smaller.

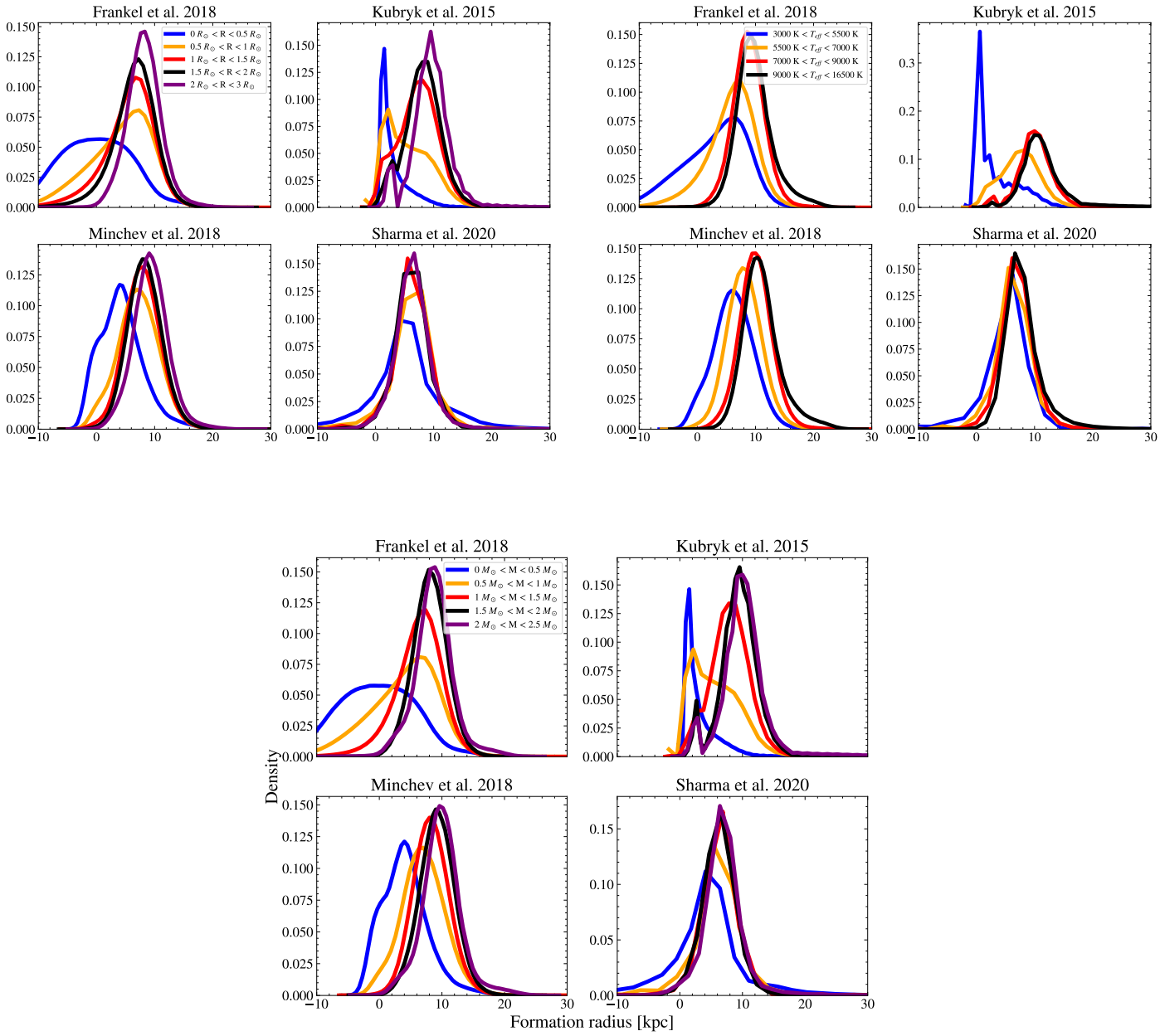

Figure 4.3: Same as figure 4.1 but binned by different stellar parameters. Top left shows the distribution when binned by stellar radius, top right shows effective stellar temperature and bottom shows stellar mass.

4.3 Discussion

4.3.1 Perfect mixing assumption

When estimating the formation radii, we made the assumption that ISM is sufficiently mixed and thus all stars formed at specific radius at a specific time share the same metallicity. While this seem to be true for the present day Milky Way (Nieva & Przybilla, 2012), it might not necessarily be true for higher redshifts, where scatter in the abundance of oxygen has been observed (Sánchez-Menguiano et al., 2018). This scatter means that stars formed on an orbit with the same radius may not share metallicities. This can happen due to the gas not being homogeneous and is instead clumpy which is expected for old, high redshift galaxies (Elmegreen et al., 2009). As discussed by Minchev et al. (2018), this effect seem to be symmetric around the mean meaning that i) the formation radius distributions should be significantly affected and ii) the different populations of stars in the different formation radius bins should not be significantly affected by this assumption. Further, the uncertainty in formation radius due to the mixing assumption is estimated to be less significant than the uncertainty in age which is estimated to be much larger.

4.3.2 Binning by age

The trends in formation radius with age that are seen are important to consider as it is expected that the galaxy has experienced an inside out formation and the youngest stars are not expected to have had time to migrate yet. The distribution of the age bins are well separated, with the exception of the Sharma model, as in this model, the evolution of the metallicity with age has effectively stopped and flattened out after 6-8 Gyr according their model. The idea that the metallicity enrichment of the ISM slows down over time is not unheard of and has been implemented in earlier models as well (e.g., Schönrich & Binney (2009)). The reasoning behind this is the fact that while Fe is produced by type II supernovae all the time and a delayed enrichment from type Ia supernovae (due to them happening on a longer timescale), the star formation rate in the Milky Way has decreased (Snaith et al., 2015), thus causing a decreased pollution rate. Therefore there is an initial increase in metallicity from type II supernovae, and a second enrichment from type Ia supernovae, after which the metallicity in the ISM is thought to only be related to the star formation rate and thus enrichment declines. All of the models show this saturation of the metallicity enrichment to some degree with Sharma being the most extreme case and Frankel being the other extreme with almost no saturation.

The bump in the formation radius distribution according to the Kubryk model at around 0-3 kpc is not seen in any of the other models. This could probably be explained by the apparent pile-up of both very old stars in the stellar population that is used and the fact that Kubryk model has a slower enrichment of the metallicity with age at all times at smaller formation radii meaning that the metallicity does not increase significantly with age at small formation radii. This means that in other models, older stars with solar or

super-solar metallicities, but slightly different ages, will be expected to be formed at very different galactocentric radii while in the Kubryk model small changes in age won't change formation radii much. Since there is a pile up of older stars in the population, it is therefore natural that there is a bump at low formation radii. The distance between the innermost part of the galaxy and the solar neighbourhood is large, meaning that might be unlikely for any given star in the centre of the galaxy to migrate far enough to be observed in the local neighbourhood. However as previously mentioned, the number density of stars in the innermost part of the galaxy is very high, and could be high enough for a non-zero probability of stars migrating to the solar neighbourhood.

As the formation radii distributions for host stars are almost identical to that of the full stellar population, there is no need to worry about host stars being formed at a specific location in the galaxy and skewing the occurrence rate towards there. This result is not as surprising knowing that the formation distributions for different stellar radii (top left in figure 4.3) are almost identical. The stellar parameter with the largest effect on the detection probability of any given planet is stellar radii meaning that if the formation radius distributions of stars of different sizes are identical then planet detections around stars should be approximately equal no matter where the star was formed.

4.3.3 Binning by Stellar Mass and Radius

Less massive stars burn their hydrogen at a much slower rate than massive stars meaning that they are expected to live much longer on the main sequence meaning that older stars are generally less massive. As it is expected for older stars to be formed further in in the galaxy, due to inside out formation, it is therefore natural to see more of the less massive stars we observe being formed further in in the galaxy. Further, due to stars not moving much on the main sequence, low mass stars generally have more uncertain ages which means that the resulting formation radius distribution becomes broader. This effect is seen most clearly in the Frankel model in the bottom of figure 4.3 where the distributions for the least massive stars are broader compared to the distributions for more massive stars.

Binning instead by stellar radius showed almost exactly the same distribution as for stellar mass. This is due to the fact that most of the giant stars have been removed as a result of the log g cuts that were made on the population. The mass and radius of main sequence stars are correlated Prialnik (2000) where larger stars are expected to be more massive so the final formation radius distribution for different radial bins should also be correlated.

4.3.4 Binning by Stellar Effective Temperature

Binning by effective temperature resulted in the formation radius distributions being similar for the different bins except for the coolest bin, meaning that stars with an effective temperature > 5500 K are not formed at a specific galactocentric radius. Cooler stars are seen to have been formed very slightly further in in the galaxy than the hottest stars with

the Kubryk model resulting in the greatest difference out of all four models. This is not unexpected as hot stars are generally very massive and are thus expected to be younger than cooler stars since more massive stars have shorter lifetimes. Cooler stars would then have a larger spread in ages meaning that some are expected to have migrated due to either inside out migration or simply due to their longer lifetimes. When using the Kubryk model, it will be important to note that the stars in the coolest temperature bin are formed significantly further in in the galaxy. Stars with these temperatures have been shown to have a higher occurrence rate of planets as shown by Yang et al. (2020) meaning that the occurrence rates might be biased towards these formation radii when using the Kubryk model specifically.

Chapter 5

Exoplanet Occurrence

In this chapter, the occurrence rates found for the entire data set as well as for different formation radius bins are shown as well as the practicalities of the MCMC procedure. We also look for differences in the orbital period and planet radius distributions of stellar samples of hosts with different formation radii using KS tests.

The KS tests showed that, after debiasing for the ages of stars, no difference between subsamples of stars formed inside or outside any of the tested Galactocentric radii was found. The occurrence rates for all multiplicities showed no statistically significant correlation with formation radius but due to large uncertainties, they remain inconclusive. Further, due to an apparent correlation with the number of stars in each formation radius bin, the final results may be unreliable.

5.1 Full population

Since the stellar population used in this project is similar to that of Zink et al. (2019), it is useful to calculate the occurrence rates of the full sample and compare with that of Zink et al. (2019). The number of stars in their stellar population is 86,605 compared to my population which includes 62,613. This difference most likely comes from the fact that we remove stars where we can't find an age and the data set we use to get stellar parameters is that of Berger et al. (2020) while Zink et al. (2019) use the data set from Berger et al. (2018). Further, the fraction of stars hosting planets for both this study and that of Zink et al. (2019) can be seen in table 5.1. As can be seen it is clear that the population of Zink et al. (2019) host a larger fraction planets for each detection order which means that it is expected that their occurrence rates is slightly larger than mine.

Table 5.1: Fraction of planetary systems for each detection order in this project compared to that of Zink et al. (2019)

Order	Me	Zink et al. (2019)
1	$1.79 \cdot 10^{-2}$	$2.50 \cdot 10^{-2}$
2	$4.25 \cdot 10^{-3}$	$6.05 \cdot 10^{-3}$
3	$1.45 \cdot 10^{-3}$	$2.06 \cdot 10^{-3}$
4	$4.31 \cdot 10^{-4}$	$6.81 \cdot 10^{-4}$
5	$1.43 \cdot 10^{-4}$	$1.85 \cdot 10^{-4}$
6	$1.60 \cdot 10^{-5}$	$3.46 \cdot 10^{-5}$
7	0	$1.15 \cdot 10^{-5}$

The resulting occurrence rates can be seen in figure 5.1 which shows a corner plot of the distributions for all the parameters. The median values for each parameter and their respective uncertainties can be seen in table 5.2 together with the values reported by Zink et al. (2019). The uncertainties reported are given by the 16th and 84th percentile respectively. The general trend from F_1 to F_7 is slightly flatter than that of Zink et al. (2019) but more importantly, there is a striking systematic difference that the occurrence rates found in this project are significantly lower than that of Zink et al. (2019). This is not unexpected as the planet populations used are different and the way the detection probability was handled is slightly different as well. This is discussed in detail in section 5.4.1. Another interesting point is the fact that for all cases except for F_1, F_{m+1} is distributed very close to its upper limit (namely F_m). This effect is slightly weaker for F_7. There are two explanations for this effect. The prior that is imposed might be too strong and could affect the accuracy of the result meaning that $F_{m\geqslant 1}$ should be larger than presented here. The other explanation is that most stars are expected to have high multiplicities and thus it is expected that $F_m \approx F_{m+1}$ for low m. Based on the fact that this effect seem to be slightly weaker for F_7 and the fact that Zink et al. (2019) fitted a survival function to these results and found the average multiplicity to be 8.4 ± 0.31 most stars are expected to have high multiplicities, meaning that the prior should not impose too strong constraints on the results. It would be nice to then try to fit values for even higher orders to confirm this theory but this would require fitting the entire parameter space i.e. all individual parameters in the likelihood model as well as running Monte-Carlo simulations to estimate the transit probabilities for higher orders and mixture probabilities. This is unfortunately outside the scope of this project.

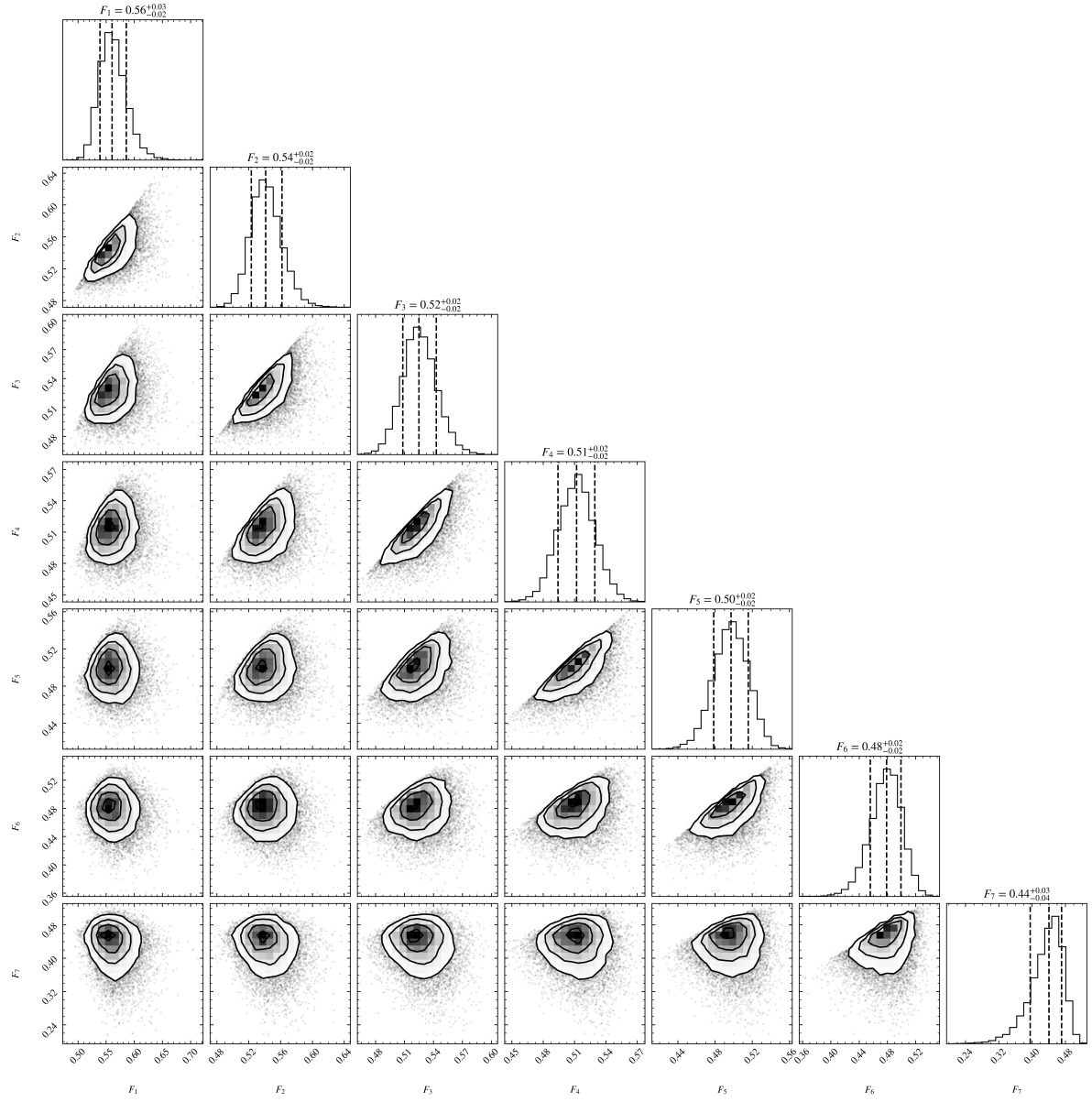

Figure 5.1: Corner plot showing the fractions of stars with at least m planets for the entire stellar population. The plot was made using corner.py (Foreman-Mackey, 2016)

Table 5.2: Occurrence rate estimates for the full stellar population used in this project and that of Zink et al. (2019). The uncertainties shown are the 16th and 84th percentile, the same as reported in Zink et al. (2019).

Occurrence Rate m	Me	Zink et al. (2019)
F_1	$0.56^{+0.03}_{-0.02}$	$0.72^{+0.04}_{-0.03}$
F_2	$0.54^{+0.02}_{-0.02}$	$0.68^{+0.03}_{-0.03}$
F_3	$0.52^{+0.02}_{-0.02}$	$0.66^{+0.03}_{-0.03}$
F_4	$0.51^{+0.02}_{-0.02}$	$0.63^{+0.03}_{-0.03}$
F_5	$0.50^{+0.02}_{-0.02}$	$0.60^{+0.04}_{-0.04}$
F_6	$0.48^{+0.02}_{-0.02}$	$0.54^{+0.04}_{-0.05}$
F_7	$0.44^{+0.03}_{-0.04}$	$0.40^{+0.07}_{-0.09}$

5.2 Varying formation radius

Due to the inherent uncertainty in the age determination and thus the formation radius estimation it is not reasonable to assign a single formation radius value to each star. Thus, in order to capture the shapes of the formation radii posterior probability distribution for each star, a formation radius value was drawn from the distribution for each star. These values were then put in bins between 0 and 20 kpc with the bin width set to be 2 kpc, except for radii larger than 10 kpc where the bin width was set to be 5 kpc to prevent low number statistics. An MCMC simulation was then run, generating a set of walkers corresponding to the probability density of F_m for that bin. The positions of these walkers were then saved and the process was repeated 50 times for each bin. The full walker set was then used to estimate the posterior distribution of the occurrence rates. The same priors and MCMC procedure as for the full stellar population as described in the beginning of the chapter was used. The occurrence rate results for all four models can be seen in figure 5.2 where the occurrence rates (G_m-values, i.e. the fraction of stars hosting exactly m planets, this time) are shown for each model and each formation radius bin. The errorbars show the 1σ (i.e. the 16th and 84th percentiles) confidence interval for the full walker set. Due to the low number of stars formed between 15 and 20 kpc (on the order of 100 stars, about 2-3 orders of magnitude less than the other bins), the results for that bin might be inaccurate and are therefore not as important.

Most notable is the fact that there is no statistically significant trend with formation radius for any multiplicity. Out of all multiplicities, the median values for G_6 has consistently the largest occurrence rate but the large uncertainties make it difficult to say anything concretely about the results. It does show however that high multiplicity system are common throughout the galaxy with a median occurrence rate of $\approx 10\%$ for all formation radii in all models. Instead, it might be of interest to look at G_0 since it is the fraction of stars hosting no planets. Here, the different models predict different results. The Frankel model show that the fraction of stars hosting no planets is largest for the smallest formation ra-

dius bin with a roughly linear decrease with formation radius meaning that the fraction of stars hosting planets increases with formation radius. In the Minchev model G_0 does not have any statistically significant change with formation radius so the same can be said for the fraction of stars with planets. In both the Kubryk model and the Sharma model, G_0 increases initially and then either decreases (Sharma) or stays roughly constant (Kubryk). In both the Sharma and Kubryk model G_0 reaches a maximum between 6-8 kpc meaning that the fraction of stars hosting no planets is at their maximum here.

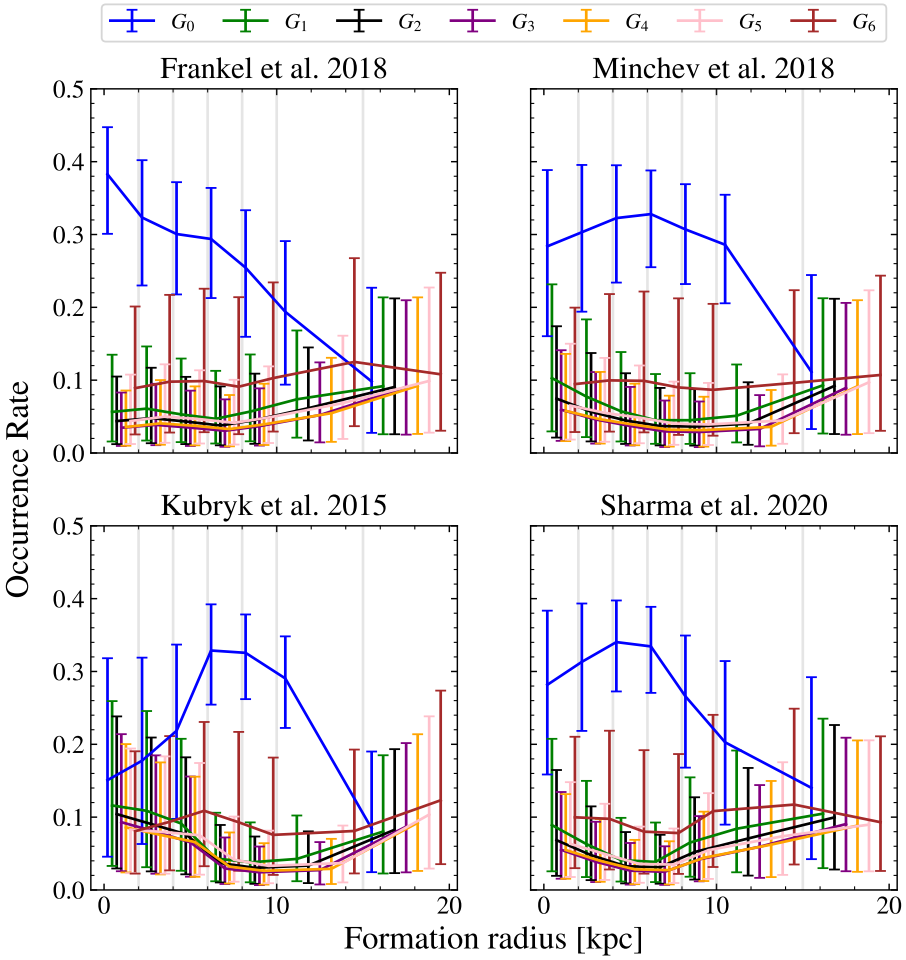

Figure 5.2: Occurrence rates versus formation radii for all four models. As a reminder, G_m is the fraction of stars hosting exactly m planets. The orders are all spread out in each bin for clarification, the grey lines show the different bins used. While $G_7 = F_7$ is included in the calculations, it is not shown as it might cause confusion since it is not a G-value. It is required that the sum of all G_m would be equal to unity which is the case when including G_7.

Despite accounting for detection bias, it is very possible that other relations are being picked up on due to some other biases. As mentioned in the introduction, giant planet occurrence has been shown to be dependent on the stellar metallicity, particularly on short period orbits (Petigura et al., 2018). Further, the occurrence of Hot Jupiters are thought to be highly dependent on the age of their host stars due to their orbital decay caused by tidal effects from the star (Winn, 2019). Young systems would therefore still retain their Hot Jupiters while in older systems, the orbits of these planets might have decayed far enough to collide with the star. In an attempt to remove this metallicity dependence (which is degenerate with formation radius), as well as the age dependence for Hot Jupiters, planets larger than 5.6 $R_\oplus \approx 0.5 R_{jup}$ were removed from the sample and the maximum limit considered in the likelihood calculation were lowered to this value. Further, as M dwarfs are thought to host higher multiplicity planetary systems than other stars (Yang et al., 2020) it is possible that they contribute to the high occurrence rates at low formation radii, especially in the Kubryk model as mentioned earlier, considering that there are a larger fraction of M dwarfs being formed here according to all models, as seen in the bottom of figure 4.3.

In order to remove this possible bias, stars with effective temperatures outside the range $4200K < T_{\text{eff}} < 6100K$ i.e. all stars which are not solar-like GK dwarfs were also removed. As can be seen in the top right of figure 4.3, most temperature bins overlap meaning that there is little to no chance of introducing further bias by removing stars based on their temperatures i.e. it is not expected that stars with a certain formation radius are over-abundant after performing this temperature cut. The results after removing giant planets and only including solar like GK dwarfs is seen in figure 5.3. However, comparing this with previous results makes it clear that removing these planets and stars had no statistically significant effect on the occurrence rates for any order. Finally, an important parameter to consider is the number of stars in each bin. Figure 5.4 shows the mean number of stars in each formation radius bin over all draws. As can be seen, there seem to be a trend where more stars in a bin equals a larger fraction of stars hosting no planets. This could imply that the G_0 results are only a reflection of the number of stars in each bin. The consequences of this is discussed in section 5.4.2.

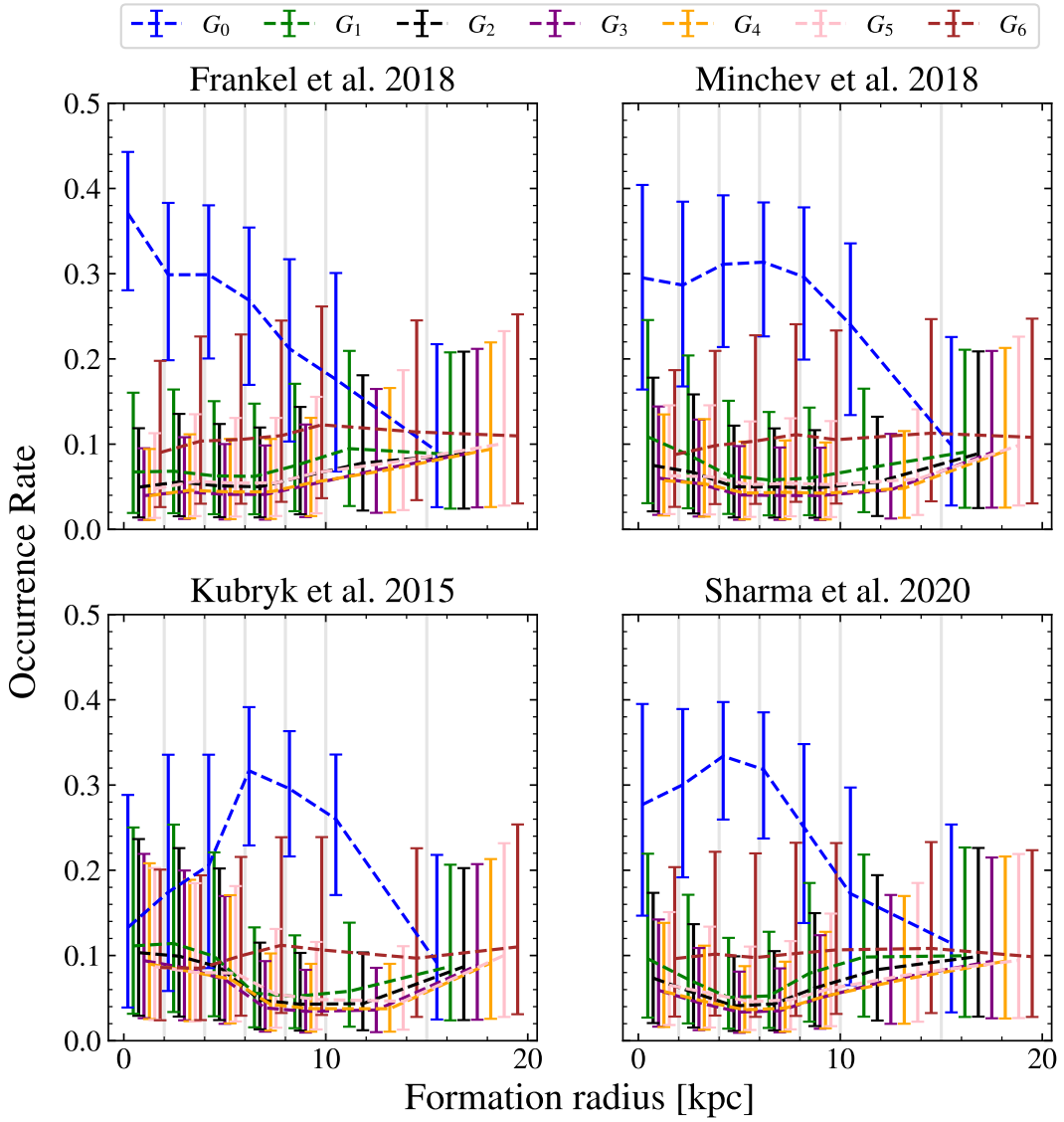

Figure 5.3: same as figure 5.2 but with giant planets and stars with a temperature outside of $4200K < T_{\text{eff}} < 6100K$ *removed. Despite removing these planets and stars, the trend remain almost unchanged.*

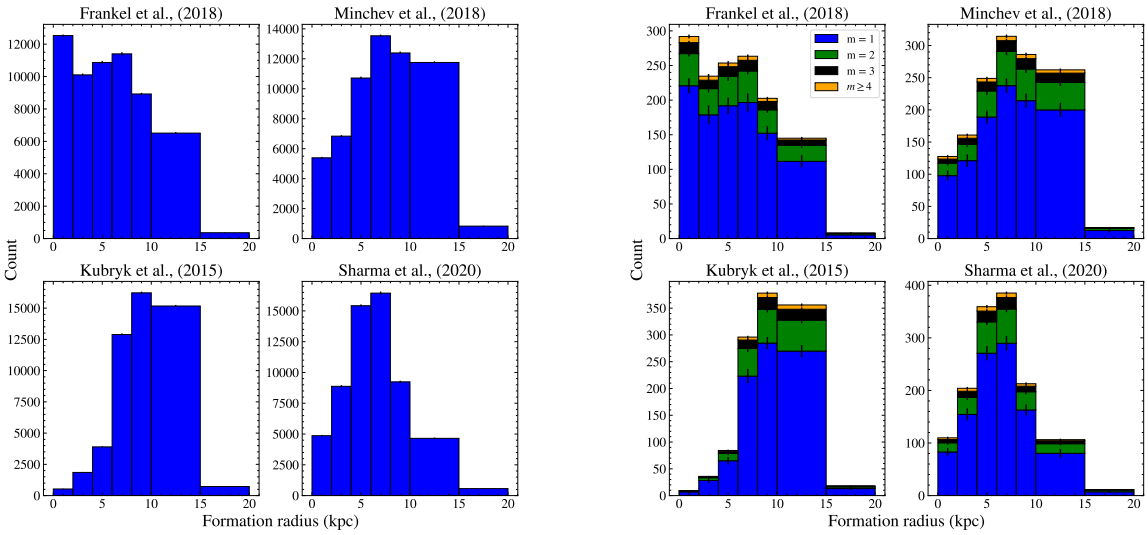

Figure 5.4: ***Left:*** *The mean number of stars in each formation radius bin for each model. The standard deviation is also shown as an errorbar but the standard deviation is significantly smaller than the mean meaning that the number of stars in each bin does not change significantly between draws.* ***Right:*** *Same as the left figure but with the number of host stars shown. The different colours represent the different detection orders which are stacked on top of each other.*

5.3 Architectural analysis

Since the full formation radius pdf's are known, it is possible to create different subsamples of the stars hosting planets based on where they were formed. For a given *critical formation radius* R_{crit}, the planet host sample was split up into two subsamples: stars with at least 68% probability (P_{in}) of being formed inside R_{crit}, and stars with at least 68% probability (P_{out}) of being formed outside R_{crit}. P_{in} was found simply by integrating the formation radius pdf's between 0 and R_{crit} and P_{out} can then be found as $P_{out} = 1 - P_{in}$. KS-tests between the subsamples on the cdf's of orbital period and planet radii was then performed for the following values of R_{crit}: 2, 4, 6, 8, 10, 15, and 20 kpc. A KS-test tests the null hypothesis that two samples could be drawn from the same distribution. The result of a KS-test can be summarised by a p-value which when below a certain significance level, allows for the possible rejection of the null hypothesis. The p-values resulting from the KS tests for planet radii and orbital period can be seen in the left and right panels respectively in figure 5.5. Each point has two numbers showing the number of stars in each subsample with the upper value showing the number of stars formed outside of R_{crit} and the lower value showing the number of stars formed inside R_{crit}. Values for R_{crit} which are missing points did not have any stars in one of the subsamples. A dashed line showing a p-value

of 0.05, which a commonly used significance level, is also shown in the figures and will be the significance value used for the hypothesis testing. A significance level of 0.05 indicates either that the probability of rejecting the null hypothesis given that it is assumed to be true or vice versa.. For planetary radii, in both the Kubryk and Sharma model, p-values for all R_{crit} are above 0.05 meaning that it is not possible to reject the null hypothesis i.e. that the subsamples are drawn from the same, underlying distribution. In the Frankel model, the p-value for $R_{crit} = 6$ kpc is below 0.05 which could mean that it would be possible to reject the null hypothesis while in the Minchev model, the same can be said for an R_{crit} of 10 kpc and 15 kpc.

In the Frankel model, both with a critical radius of 6 kpc and 8 kpc, the KS-test resulted in a p-value well below the threshold with a significant enough sample size in both cases. As most of the stars observed by Kepler are currently in the solar neighbourhood ($R \approx 8$ kpc), an R_{crit} of 8 kpc separates the entire sample into stars who have migrated inwards and stars who have migrated outwards. Stars formed further in are thought to be older meaning that it might be possible that these two samples are biased due to the presence of Hot Jupiters. Hot Jupiter would be more abundant in the stars formed further in as they are younger and could thus bias the KS-testing. However, even when removing all planets with radius larger than 0.5 Jupiter radii, the same results as in the right planel of figure 5.5 are found. In fact, removing the giant planets did not affect the results of the KS tests for the planet radii either.

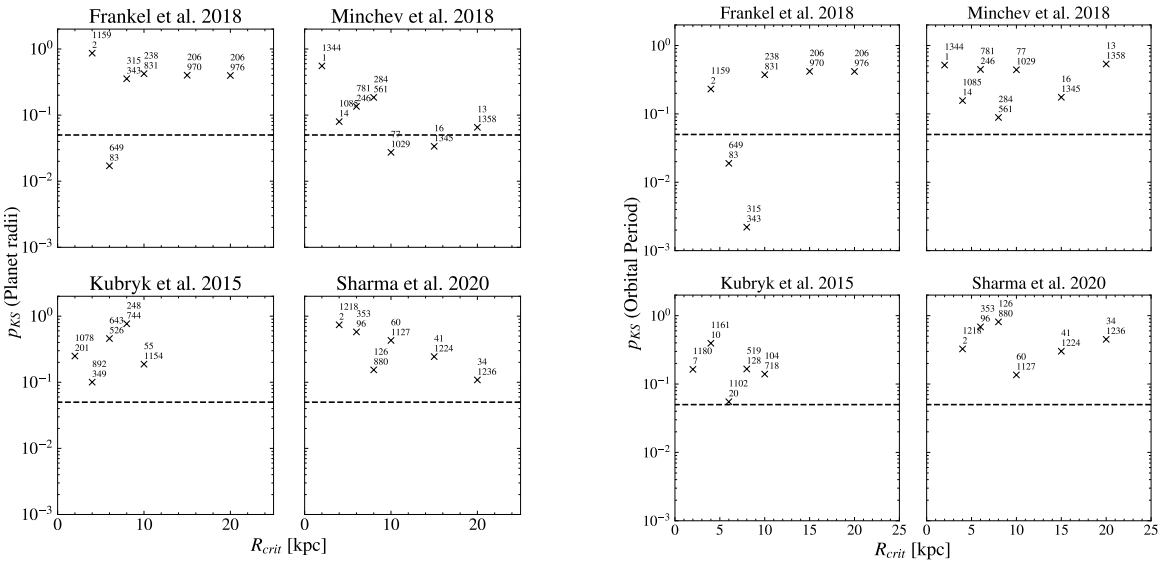

Figure 5.5: p-values for different values of R_{crit}. R_{crit} is the formation radius which separates the two different distributions tested. Stars with an 68% probability of being formed inside of R_{crit} were compared with stars with an 68% probability of being formed outside of R_{crit} through a KS-test. The annotations on each point show the number of stars in the outer and inner distribution respectively. The dashed line shows a p-value of 0.05. The left panel show the p-values for the distributions of planet radii while the right panel show those for the distributions of orbital period

The difference between the two samples in the Frankel model can be best understood through the median ages of all stars in the two subsamples which is shown in figure 5.6. In this model, stars formed inside of 8 kpc (or 6 kpc) are most likely very old stars or very young stars while stars formed outside of 8 kpc (or 6 kpc) have more intermediate ages.[1] Due to the fact that the planetary system is forming together with the star, it may be possible that the orbital periods of planets around younger stars are different than that of older stars, meaning that the difference observed between these two samples may just be caused by an age bias. Figure 5.7 shows the results from the same KS-testing as figures 5.5 but with all stars with a median age less than 2.5 Gyr removed. As seen, the p-values are well above 0.05 meaning that the previously perceived difference was most likely simple caused by an age difference in the stars. This means that the differences in the samples that have been seen are simply caused by the ages of the stars and no other intrinsic property. Interestingly, removing young stars also increased all low p-values for planet radii to be above 0.05 as well. All of these results are further discussed in section 5.4.3.

[1] A similar trend can be seen in the ages for the Minchev model and an R_{crit} of 10 kpc which have the other low p-value for planet radii and the explanation for these results are thus the same no matter which model is discussed

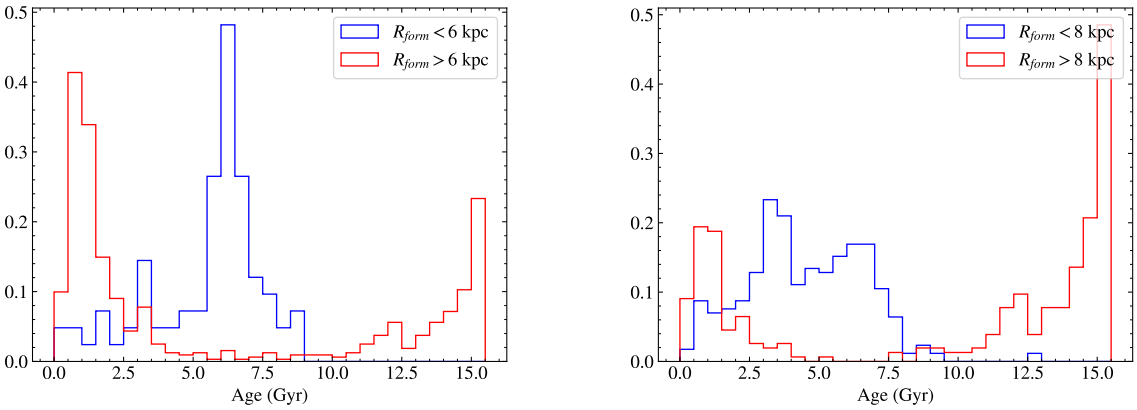

Figure 5.6: ***Left:*** *Density histograms of the ages in the two subsamples formed inside and outside of 6 kpc according to the Frankel model.* ***Right:*** *Same as the left figure but with an R_{crit} of 8 kpc instead.*

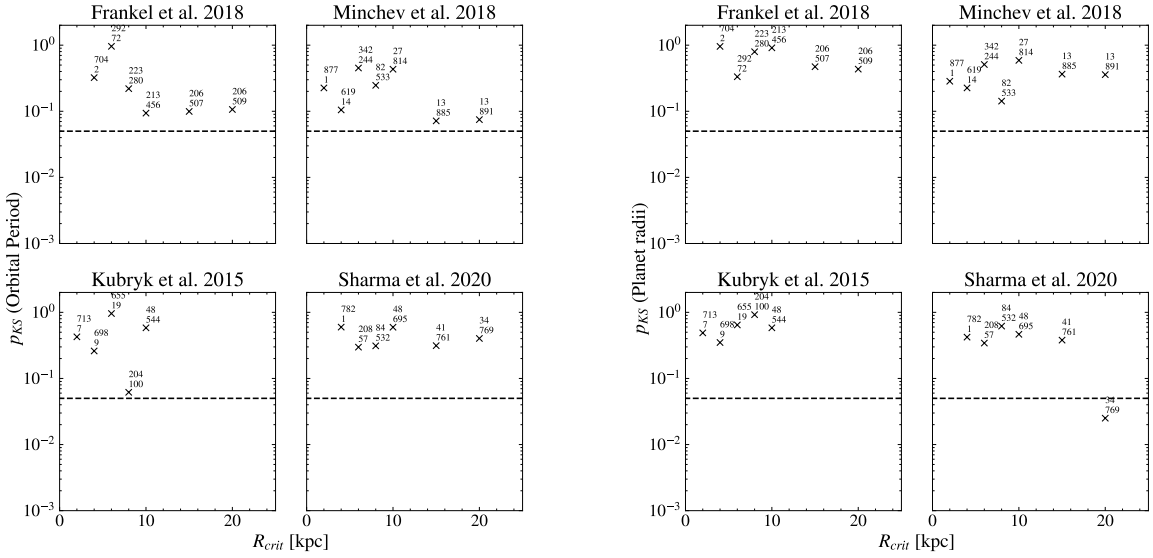

Figure 5.7: ***Left:*** *KS tests for distribution orbital periods after splitting up planet hosts by R_{crit} but with all stars with a median age of < 2.5 Gyr removed.* ***Right:*** *Same as the left figure but for planet radii instead.*

5.4 Discussion

5.4.1 Comparisons with previous work

As mentioned, there is a systematic difference between the occurrence rate values found in this project and those of Zink et al. (2019). This difference in occurrence rates can best be explained by the multiplicity distribution of the two planet populations used in each project (see figure 3.1 and table 5.1). It is clear that Zink et al. (2019) include more planets in their population across all multiplicities and that this difference is larger for lower multiplicities i.e. for 1- or 2-planet systems. We suspect that this difference in the number of planets comes from the fact that Zink et al. (2019) seem to include planets which have host stars that we have removed as a result of the cuts described in section 2.1.1. By investigating equation (3.6) it is clear that increasing the number of planets in lower multiplicity systems while keeping the number of stars constant would increase the estimated value for these occurrence rates. Due to the cascading nature of the occurrence rates it is therefore possible that the occurrence rates for higher orders could be allowed to increase as well. This would also explain the flatter distribution from lower to higher orders as low multiplicity systems are the most affected. Further, as Zink et al. (2019) average their detection probabilities over the entire stellar population it is possible that they overestimate the detection probabilities compared to the values that would have been achieved if individual detection probabilities would have been taken into account. He et al. (2021) investigated the occurrence rates around stars of different spectral types and found that the fraction of solar type (G2V) dwarfs with planets[2] is approximately 0.57 which is comparable to F_1 found in this project using the full data as seen in table 5.2. While it is not possible for a one-to-one comparison due to the fact that the stellar population considered here is a mix of different spectral types, a large fraction of the stars considered are G dwarfs meaning that their contribution is the most significant to the overall occurrence rate.

The separability of the orbital period and planet radii distributions is a major assumption when estimating both the sorting order but also the detection probability itself. Weiss et al. (2018) argue that neighbouring planets are of similar sizes, and that orbital period ratios of adjacent pairs are correlated. Should this be true it is possible to argue that during the formation of planets, they are aware of each other and the probability distributions the radius and period of one planet is a function of the orbital periods and radii of the other planets (Murchikova & Tremaine, 2020) meaning that the distribution functions are not necessarily separable. However, as demonstrated by Murchikova & Tremaine (2020), the same apparent correlations as found by Weiss et al. (2018) can be found even with a universal, intrinsic radius distribution in which the planets do not care about each other when forming. Further, Zhu (2020) also investigated these results and found that the findings of Weiss et al. (2018) can be explained by selection effects due to the specific S/N cuts made. For low detection probabilities (e.g. large stellar radii or short transit

[2]denoted f_{swpa} in He et al. (2021)

duration), only planets with a large radii survive the S/N cuts while with high detection probabilities, both small and large planets survive the cuts. However due to the shape of the radius distribution, most planets in a multiplanet system are expected to be small meaning that in either case, using the specific S/N cuts of Weiss et al. (2018), correlations arise. Despite all this, it is still likely that the sizes of the planets are correlated in some way as shown by Murchikova & Tremaine (2020) who set up four distinct distributions in planetary radii and was able to reproduce the observed distributions and size correlation together with the Pearson/Spearman correlations as a function of minimum radius. This strongly hints that the planetary radii distributions are aware of which system they are in and varies from system to system, but not necessarily that the planets are aware of each other. In such a case, the radius and period distributions are still separable which means that the assumption of separability can be justified.

5.4.2 Occurrence rate dependence on formation radius

When analysing the results of how the occurrence rate changes with formation radius it is important to think about what types of stars correspond to a certain formation radius. As shown in figure 5.8, large formation radii generally correspond to metal poor, young stars while small formation radii are generally either old, solar metallicity stars or metal rich stars with any given age. The Kubryk model generally yields smaller formation radii for a wider range of ages and metallicities compared to the other three while the opposite is true for the Frankel model. Figure 5.8 show the resulting formation radii for a range of different ages and metallicities. The colorbar limits were set to 0 and 20 kpc respectively in order to capture the full range considered. In some areas of the phase space shown however, the formation radius values would exceed these values meaning that the colors are saturated at these points. In practice, this means that some black regions in figure 5.8 have values below 0 kpc and some white regions have values above 20 kpc.

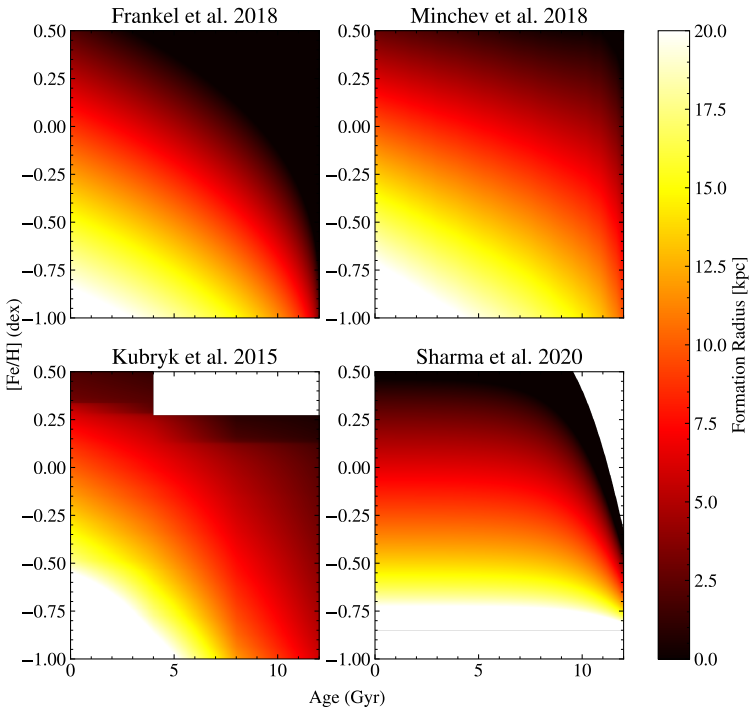

Figure 5.8: The formation radii the four different models output for some metallicities and ages. Both the Kubryk and Sharma models have some regions where a formation radius cannot be inferred. In the Sharma model, the lower metallicity limit comes from the fact that they have set the minimum possible metallicity to -0.85 meaning that no stars have a metallicity below this. In the Kubryk model this corresponds to the upper left region in the bottom left panel of figure 2.9 where the lines as reported by Kubryk et al. (2015) doesn't extend further.

For all multiplicities, there is no statistically significant change with formation radius. It is possible that this result is simply due to the large uncertainties in G_m and that they hide any underlying differences. It would be natural to assume that the fraction of stars hosting no planets would decrease with formation radius as an increased number density of stars would increase the risk for star-star interactions such as fly-bys which might destabilise planetary systems. Further, a larger number density of stars also increases the potential risk of photoevaporation of planetary discs which could cause truncation and thus decreased planet formation efficiency. This is seen when we use the Frankel model while the opposite is seen when using the Kubryk and Sharma models where the fraction of stars with no planets increases up until $\approx$ 6-8 kpc where it reaches a maximum. In the Minchev model, this fraction stays roughly constant for all formation radii up until $\approx 8 - 10$ kpc after which it decreases.

No matter which model is used, the fraction of stars hosting no planets is very low at

high formation radii which correspond to very young, metal-poor stars in all models (high formation radii). When we use the Minchev, Kubryk, and Sharma models, G_0 peak between 6 and 8 kpc while when we use the Frankel model, G_0 peak in the smallest formation radius bin. These peaks do not correspond to the same ages and metallicity combinations which could otherwise explain the differences. This means that the ages and metallicities alone cannot explain the differences between each model. Instead, the more important contribution to the results could be the number of stars in each bin.

Throughout the project, the parameters for the exoplanet distributions as well as the order distribution as found by Zink et al. (2019) have been assumed to be true. This is not expected to affect the results for the full, non-binned case as roughly the same stellar selections as Zink et al. (2019) are used. However, it might be argued that the orbital period and exoplanet radii distributions could be different for different formation radii for the same reasons that occurrence might be expected to be different for different formation radii. Modelling the exoplanet distributions as power laws was generalised for Kepler by Youdin (2011) and the work by Zink et al. (2019) is based on this work. The parameters of interest for the exoplanet distributions are the break values where the power law slope breaks and the slopes for the two different regimes (i.e. $\alpha_1, \alpha_2, \beta_1, \beta_2, R_{br}$, and P_{br} in equation (A.9)). These parameters have been found empirically by modelling planet occurrence rates after taking into account different selection effects such as detection probabilities (Burke et al., 2015; Youdin, 2011, for example). While the parameters have been found empirically and might be found to be different by only considering certain types of stars or stars formed in very specific environments, they are expected to have arisen from intrinsic properties in the formation of the planets. The KS testing in section 5.3 shows that, after debiasing for the ages of stars, the period and radius distributions for all the detected planets seem to be drawn from the same underlying distribution no matter at which galactocentric radius the star was formed. It is still possible that the underlying distributions in orbital period and planet radii, which would be affected by any planets that are not detected yet, is different from the distribution that have been assumed. In fact, fixing the radius and period distributions means that in equation (3.6), $K_{1,m}$ becomes a constant and does not affect the likelihood estimate. Further, when drawing a number of randomly selected stars, the double integral in equation (3.4) changes very little no matter how many stars are drawn or which stars are drawn. Thus, the main parameters which affects the likelihood estimate is the number of detected planets of a certain order $N_{det,m}$ and the number of stars considered. Since $N_* \gg N_{det}$, varying N_* will have a very large effect on the final result compared to the number of planets detected in each order. By allowing for variations in the orbital period and planet radii distributions as well as the sorting order distribution, the detection of individual planets would also affect the likelihood as $K_{1,m}$ would not be constant and the double integral in equation (3.4) would vary more between each bin, thus reducing the effect the number of stars have on the final results.

When comparing figure 5.3 and the left panel in figure 5.4, G_0 follows then number of stars in each bin very closely where a large number of stars corresponds to a larger G_0.

An explanation for this could be the age uncertainties. The median fractional uncertainty in the age of host stars is approximately 50% compared to the entire data set where the median fractional uncertainty is 33 %. This means that the uncertainty in formation radius is generally larger for host stars than for most other stars. Thus, when redrawing formation radii, the mean number of planets could be the same in each bin which would mean that the number of stars in each bin is the only significant factor when estimating the occurrence rate. However, as seen from the right panel in figure 5.4, the mean number of host stars follow the total number of stars meaning that the larger age uncertainties should not affect the final results.

5.4.3 Architectural analysis

When we performed the KS-tests, a significance level was set to $\alpha = 0.05$ in order to assist the analysis which means that, given that the null hypothesis is true, it is expected that 1 in 20 hypothesis tests rejects the null hypothesis, leading to an incorrect rejection. Our KS-testing involves 28 hypothesis tests for both the case for the full data set and the case when very young host stars are removed, meaning that with a significance level of 0.05, it is expected that around one test in each case should reject the null hypothesis, which we see. It is possible that this rejection is simply due to the fact that multiple tests are made instead of the null hypothesis actually being false in these cases. When testing multiple hypotheses, a more appropriate significance level could be found by a Bonferroni correction where the tests would be made at a significance level of α/N_{tests} where N_{tests} are the number of tests made. In reality, a Bonferroni correction would most likely yield a too conservative significance level as the different tests made are not made on independent data sets, meaning that the true significance level most likely lies between the chosen one and the Bonferroni corrected one. Nevertheless, a significance level of 0.05 is still used in the analysis and as we will show later, will not affect the final conclusion.

The KS-testing of the orbital periods distributions showed that when using the Frankel model, it might be possible to reject the null hypothesis that the subsamples of stars born inside and outside of 6 and 8 kpc respectively are drawn from the same distribution. The same can be said of planet radii at an R_{crit} of 6 kpc. However, removing stars with a median age of less than 2.5 Gyr increased the p-values for both orbital periods and planet radii to the point where the null hypothesis could not be rejected anymore. Figure D.1 show cumulative distributions functions of the orbital period of all the planets considered split by stars younger than 2.5 Gyr and stars older than 2.5 Gyr. Here it can be seen, albeit with difficulty, that the distribution of the younger stars host a larger fraction of planets on very short orbital periods which is most likely the cause of the difference in the CDF's. As has been mentioned earlier, planets on short period orbits are expected to experience orbital decay due to tidal effects from the star.[3] This means that it is most likely the tidal decay which is causing the difference between the samples.

[3]Although this effect is stronger for more massive planets.

Regarding the planet radii, the difference cannot be explained as simply. In fact, despite the p-value increasing when removing stars with a median age of less than 2.5 Gyr, the CDF's of the planet radii are not different between stars younger than 2.5 Gyr and stars older than 2.5 Gyr. Figure D.2 show the CDF's of the planet radii for the two sub-samples split by an R_{crit} of 6 kpc according to the Frankel model. As can be seen, the difference lies mainly around 1-2 $R_{\oplus}$ where the planet radii distributions around stars formed further in in the galaxy are skewed towards these sizes whereas the planet radii distributions for host stars formed further out are skewed towards larger planets. It could be so that this result is a consequence of the Fulton gap, which is an underdensity of planets with sizes of less than $2R_{\oplus}$ (Fulton et al., 2017). This gap is thought to arise due to the photoevaporation of atmospheres around planet cores on short period orbits. Planets larger than a certain size are massive enough to retain their atmospheres with their surface gravity while smaller planets gets their atmosphere evaporated such that only the bare core is left meaning that a valley in the radius distribution is created. It could be that younger stars are more efficient at stripping planetary atmospheres, meaning that they are less likely to host planets of sizes of about 1-2 $R_{\oplus}$. Younger stars typically have a higher chromospheric activity (Zhang et al., 2019) which could also cause the star to be more noisy in Kepler observations which in turn causes the MES of a potential planet to decrease. Therefore it might be possible for the detection probability to be larger for younger stars. However, upon inspection, we found that there is no direct correlation between noise and median stellar age meaning that the noise should not affect the final result.

Chapter 6

Conclusions & Future work

The aim of this project was to i) estimate the formation radii of stars and ii) find if there are any connections between occurrence rates of planets and formation radius. The formation radii of stars were determined by utilising the method explained by Minchev et al. (2018) where stars with a certain age and metallicity were assigned a formation radius by modelling the time dependence of the radial metallicity gradient of the Milky Way. By using precise stellar parameters from Berger et al. (2020), the ages of stars observed by the Kepler telescope were determined. The determined ages, combined with the metallicities of the stars, allowed for precise estimations of their formation radii with four different metallicity gradient models. Using the detection bias model set up by Zink et al. (2019), which takes into account mutual inclinations in multiple planetary systems for both geometric transit probabilities and recovery probabilities within the Kepler pipeline, it was possible to estimate the debiased occurrence rates of stars for individual multiplicities formed at different formation radii. It was found that in all four metallicity gradient models, for any of the multiplicities considered, the fraction of stars hosting planets did not show a statistically significant change with formation radii although, due to the large uncertainties, these results are inconclusive. The fraction of stars hosting no planets at all was also considered and while there were variations between each model and formation radius bin, these are most likely explained by the number of stars in each formation radius bin.

KS-testing of the observed population of planets was also performed. In particular, the orbital period and planet radii distributions for stars formed inside or outside a given formation radius was considered. It was found that after debiasing for ages, the null hypothesis for both orbital periods and planet radii (i.e. that the subsamples are drawn from different distributions) could not be rejected for any formation radius or any model. Host star properties, e.g. metallicity, affect the architecture of planetary systems, meaning that the planets know about the system they are in. However, in the case that the orbital period and planet radius distributions are independent of formation radius, the formation and evolution process of planetary systems could be independent of where in the galaxy the host star is formed. This means that the planets are in general not aware of the environment in which the host star was formed. This could seem contradictory to the findings

that photoevaporation from nearby massive stars cause significant mass-loss of protoplanetary discs but Parker et al. (2021) found no correlation between decreased disc-mass and distance from ionising stars. Instead discs are rapidly uniformly destroyed by photoevaporation suggesting an all or nothing effect. This could mean that, while photoevaporation will destroy discs, it might be rare and thus not affect the occurrence rates significantly.

6.1 Outlook and future work

The core idea of this project hinges on the fact that it is possible to precisely estimate the formation radii of stars. While efforts to reduce the statistical uncertainties have been made, there is still a large uncertainty in the metallicity evolution of the Milky Way as evident by the differences in each model. With the recently published third data release of GALAH (Buder et al., 2021) as well as the upcoming 4MOST survey (de Jong et al., 2012), precise and accurate abundances of a large number of stars will be available and thus the evolution of the metallicity in both time and space will be much better constrained thus reducing the potential systematic uncertainties when estimating the formation radii of stars.

Unfortunately, the occurrence rate results presented in this thesis are largely inconclusive. The number of stars has a too large effect on the final results to be able to draw any valuable conclusions from them. One way to minimise the effects the number of stars has on the results would be to allow more parameters in the likelihood model to vary. Throughout this work, parameters for the period and radius distribution of planets, as well as for the sorting order distributions were fixed in order to simplify calculations. However, these assumptions are not necessarily true and would lessen the effect of the number of stars. By allowing these parameters to vary, it would also be possible to expand the work of Zink et al. (2019) to further orders since higher multiplicity systems other than the Solar System has been observed.

Another way would be to instead use the approach of Yang et al. (2020) who performed a similar investigation as this project: bin the stars according to some parameter of interest and perform detection bias corrections in each bin through an MCMC procedure. However, instead of including the detections of planets, they generated planetary systems from debiased, observed distributions and estimated the probability of detection for each planet to estimate how many planets they would expect to observe. This expected number of observations could then be used to compare with the actual number of observed planets in a Poisson likelihood model. As Zink et al. (2019) showed, the detection probability for planets detected after the first one decreases significantly so these would have to still be taken into account. Combining the careful consideration of different detection orders by Zink et al. (2019) with the overall method of Yang et al. (2020) could be a way to properly estimate the effects formation radius has on occurrence rates. Further, with accurate occurrence rates, it would also be possible to generate planetary systems from the found

multiplicities and estimate the fraction of stars hosting habitable planets by making sure that the generated planetary systems are stable, and carefully setting up certain habitability requirements. An example of such work can be found in Zink & Hansen (2019). It would then be possible to estimate how the fraction of stars hosting potentially habitable systems change with formation radius, ultimately leading us to an answer to the question of where in the galaxy life as we know it on Earth is most suitable.

References

Adibekyan, V., Santos, N. C., Demangeon, O. D. S., et al. 2021, arXiv e-prints, arXiv:2102.12346

Akeson, R. L., Chen, X., Ciardi, D., et al. 2013, PASP, 125, 989

Anders, F., Buck, T., Frankel, N., & Minchev, I. 2020, in Contributions to the XIV.0 Scientific Meeting (virtual) of the Spanish Astronomical Society, 115

Anders, F., Chiappini, C., Minchev, I., et al. 2017, A&A, 600, A70

Ansdell, M., Williams, J. P., Manara, C. F., et al. 2017, AJ, 153, 240

Barragán, O., Gandolfi, D., Dai, F., et al. 2018, A&A, 612, A95

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, AJ, 159, 280

Bird, J. C., Kazantzidis, S., Weinberg, D. H., et al. 2013, ApJ, 773, 43

Bland-Hawthorn, J. & Gerhard, O. 2016, ARA&A, 54, 529

Borucki, W. J., Koch, D., Basri, G., et al. 2010, Science, 327, 977

Brakensiek, J. & Ragozzine, D. 2016, ApJ, 821, 47

Breslau, A., Steinhausen, M., Vincke, K., & Pfalzner, S. 2014, A&A, 565, A130

Bryson, S., Coughlin, J. L., Kunimoto, M., & Mullally, S. E. 2020, arXiv e-prints, arXiv:2006.15719

Buder, S., Sharma, S., Kos, J., et al. 2021, MNRAS[eprint[arXiv]2011.02505]

Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, ApJ, 809, 8

Choi, J., Dotter, A., Conroy, C., et al. 2016, ApJ, 823, 102

Christiansen, J. L. 2017, Planet Detection Metrics: Pixel-Level Transit Injection Tests of Pipeline Detection Efficiency for Data Release 25, Kepler Science Document KSCI-19110-001

de Jong, R. S., Bellido-Tirado, O., Chiappini, C., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV, ed. I. S. McLean, S. K. Ramsay, & H. Takami, 84460T

de Laverny, P., Recio-Blanco, A., Worley, C. C., et al. 2013, The Messenger, 153, 18

Delgado Mena, E., Tsantaki, M., Adibekyan, V. Z., et al. 2017, A&A, 606, A94

Dotter, A. 2016, ApJS, 222, 8

Elmegreen, B. G., Elmegreen, D. M., Fernandez, M. X., & Lemonias, J. J. 2009, ApJ, 692, 12

Fang, J. & Margot, J.-L. 2012, ApJ, 761, 92

Foreman-Mackey, D. 2016, The Journal of Open Source Software, 1, 24

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Frankel, N., Rix, H.-W., Ting, Y.-S., Ness, M., & Hogg, D. W. 2018, ApJ, 865, 96

Fulton, B. J., Petigura, E. A., Howard, A. W., et al. 2017, AJ, 154, 109

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

Genovali, K., Lemasle, B., Bono, G., et al. 2014, A&A, 566, A37

Gillessen, S., Plewa, P. M., Eisenhauer, F., et al. 2017, ApJ, 837, 30

Goodman, J. & Weare, J. 2010, Communications in Applied Mathematics and Computational Science, 5, 65

Gravity Collaboration, Abuter, R., Amorim, A., et al. 2019, A&A, 625, L10

Guarcello, M. G., Drake, J. J., Wright, N. J., et al. 2016, arXiv e-prints, arXiv:1605.01773

Hartman, J. D., Bakos, G., Stanek, K. Z., & Noyes, R. W. 2004, AJ, 128, 1761

He, M. Y., Ford, E. B., & Ragozzine, D. 2021, AJ, 161, 16

Howes, L. M., Lindegren, L., Feltzing, S., Church, R. P., & Bensby, T. 2019, A&A, 622, A27

Johansen, A., Davies, M. B., Church, R. P., & Holmelin, V. 2012, ApJ, 758, 39

Johnson, J. A., Aller, K. M., Howard, A. W., & Crepp, J. R. 2010, PASP, 122, 905

Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864

Kubryk, M., Prantzos, N., & Athanassoula, E. 2015, A&A, 580, A126

Lada, C. J. & Lada, E. A. 2003, ARA&A, 41, 57

Li, D., Mustill, A. J., & Davies, M. B. 2020, MNRAS, 496, 1149

Longmore, S. N., Chevance, M., & Kruijssen, J. M. D. 2021, arXiv e-prints, arXiv:2103.01974

Lu, C. X., Schlaufman, K. C., & Cheng, S. 2020, arXiv e-prints, arXiv:2009.06638

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94

Mathur, S., Huber, D., Batalha, N. M., et al. 2017, VizieR Online Data Catalog, J/ApJS/229/30

Mayor, M. & Queloz, D. 1995, Nature, 378, 355

Minchev, I., Anders, F., Recio-Blanco, A., et al. 2018, MNRAS, 481, 1645

Minchev, I., Chiappini, C., & Martig, M. 2014, A&A, 572, A92

Morton, T. D. & Johnson, J. A. 2011, ApJ, 738, 170

Muñoz, D. J., Kratter, K., Vogelsberger, M., Hernquist, L., & Springel, V. 2015, MNRAS, 446, 2010

Murchikova, L. & Tremaine, S. 2020, AJ, 160, 160

Mustill, A. J., Davies, M. B., & Johansen, A. 2015, ApJ, 808, 14

Mustill, A. J., Lambrechts, M., & Davies, M. B. 2021, arXiv e-prints, arXiv:2103.15823

Nakanishi, H. & Sofue, Y. 2016, PASJ, 68, 5

Nandakumar, G., Hayden, M. R., Sharma, S., et al. 2020, arXiv e-prints, arXiv:2011.02783

Nieva, M. F. & Przybilla, N. 2012, A&A, 539, A143

Parker, R. J., Alcock, H. L., Nicholson, R. B., Panić, O., & Goodwin, S. P. 2021, arXiv e-prints, arXiv:2104.03973

Pater, I. D. & Lissauer, J. J. 2015, Planetary sciences, 2nd edn. (Cambridge University Press), 494

Pepe, F., Mayor, M., Queloz, D., et al. 2004, A&A, 423, 385

Petigura, E. A., Marcy, G. W., Winn, J. N., et al. 2018, AJ, 155, 89

Pollacco, D. L., Skillen, I., Collier Cameron, A., et al. 2006, PASP, 118, 1407

Prialnik, D. 2000, An introduction to the theory of stellar structure and evolution, 2nd edn. (Cambridge University Press)

Raymond, S. N. & Morbidelli, A. 2020, arXiv e-prints, arXiv:2002.05756

Rodet, L., Su, Y., & Lai, D. 2021, arXiv e-prints, arXiv:2102.07898

Sánchez-Menguiano, L., Sánchez, S. F., Pérez, I., et al. 2018, A&A, 609, A119

Sanders, J. L. & Binney, J. 2015, MNRAS, 449, 3479

Santos, N. C., Adibekyan, V., Dorn, C., et al. 2017, A&A, 608, A94

Santos, N. C., Sousa, S. G., Mortier, A., et al. 2013, A&A, 556, A150

Schönrich, R. & Binney, J. 2009, MNRAS, 396, 203

Sellwood, J. A. & Binney, J. J. 2002, MNRAS, 336, 785

Shallue, C. J. & Vanderburg, A. 2018, AJ, 155, 94

Sharma, S., Hayden, M. R., & Bland-Hawthorn, J. 2020, arXiv e-prints, arXiv:2005.03646

Snaith, O., Haywood, M., Di Matteo, P., et al. 2015, A&A, 578, A87

Soderblom, D. R. 2010, ARA&A, 48, 581

Spina, L., Ting, Y.-S., De Silva, G. M., et al. 2021, MNRAS

Steffen, J. H., Ragozzine, D., Fabrycky, D. C., et al. 2012, Proceedings of the National Academy of Science, 109, 7982

Thompson, S. E., Coughlin, J. L., Hoffman, K., et al. 2018, ApJS, 235, 38

Van Eylen, V., Albrecht, S., Huang, X., et al. 2019, AJ, 157, 61

Weiss, L. M., Marcy, G. W., Petigura, E. A., et al. 2018, AJ, 155, 48

Winn, J. 2019, in AAS/Division for Extreme Solar Systems Abstracts, Vol. 51, AAS/Division for Extreme Solar Systems Abstracts, 201.04

Winter, A. J., Clarke, C. J., Rosotti, G., et al. 2018, MNRAS, 478, 2700

Winter, A. J., Kruijssen, J. M. D., Longmore, S. N., & Chevance, M. 2020, Nature, 586, 528

Yang, J.-Y., Xie, J.-W., & Zhou, J.-L. 2020, AJ, 159, 164

Youdin, A. N. 2011, ApJ, 742, 38

Zhang, J., Zhao, J., Oswalt, T. D., et al. 2019, ApJ, 887, 84

Zhu, W. 2020, AJ, 159, 188

Zink, J. K., Christiansen, J. L., & Hansen, B. M. S. 2019, MNRAS, 483, 4479

Zink, J. K. & Hansen, B. M. S. 2019, MNRAS, 487, 246

Appendix A

Description of individual quantities in the model

A.1 Detection Probabilities

The window probability is estimated by Burke et al. (2015) to be

$$P_{win} = 1 - (1 - f_{duty})^M - M f_{duty}(1 - f_{duty})^{M-1} - \frac{1}{2}M(M-1)f_{duty}^2(1 - f_{duty})^{M-2}, \quad \text{(A.1)}$$

where f_{duty} is the duty cycle of the star i.e. the fraction of active data intake and $M = t_{obs}/P$ is the fraction between the full observation time of the Kepler telescope and the orbital period of the planet. The window probability is independent with respect to the detection order.

The recovery probability has been characterised by Christiansen (2017) through injecting artificial planets into each Kepler star and then put the light curves through the Kepler pipeline. They did not however, consider detection order and assumed that the recovery probability uniform over all periods. Zink et al. (2019) used the same data as Christiansen (2017) but instead split the data into $m = 1$ injections and $m \geqslant 2$ injections as well as in two period ranges: $0.5 < P/\text{days} < 200$ and $200 < P/\text{days} < 500$ and fit the recovery probability to a Γ_{CDF}

$$P_{rec}(\text{MES}) = \frac{c}{b^a(a-1)} \int_0^{\text{MES}} (x - x_0)^{(a-1)} \exp\left(\frac{-(x - x_0)}{b}\right) dx, \quad \text{(A.2)}$$

where the parameters a, b, c, x_0 are fit. Their found values can be found in table A.1. The MES is the multiple event statistic and is analogous to the S/N and is found to be

$$\text{MES} = 1.003\sqrt{N_{tr}}\frac{\Delta}{\sigma_{cdpp}}, \quad \text{(A.3)}$$

with N_{tr} being the number of transits, σ_{cdpp} being the combined differential photometric precision, and Δ being the transit depth. The σ_{cdpp} can be seen as the stellar noise. The

number of transit is straightforwardly defined as

$$N_{tr} = \frac{t_{obs}}{P}, \tag{A.4}$$

while Zink et al. (2019) found the transit depth to be expressed as

$$\begin{aligned} k &= \frac{R}{R_*} \\ c_0 &= 1 - (u_1 + u_2) \\ \omega &= \frac{c_0}{4} + \frac{u_1 + 2u_2}{6} - \frac{u_2}{8} \\ \Delta &= 1 - \frac{1}{\omega}\left(\frac{c_0}{4} + \frac{(u_1 + 2u_2)(1-k^2)^{3/2}}{6} - \frac{u_2(1-k^2)}{8}\right), \end{aligned} \tag{A.5}$$

where u_1, u_2 are the limb darkening coefficients used to fit the transits within the Kepler pipeline. They are found by Zink et al. (2019) to be related to the effective temperature T_{eff}

$$\begin{aligned} u_1 &= -1.93 \cdot 10^{-4} T_{\text{eff}} + 1.5169 \\ u_2 &= 1.25 \cdot 10^{-4} T_{\text{eff}} - 0.4601. \end{aligned} \tag{A.6}$$

Table A.1: Parameters for the recovery probability and their values as found by the Zink et al. (2019) injections

Period range	a	b	c	x_0
m = 1				
$0.5 < P/\text{days} < 200$	29.3363	0.2856	0.9845	0.0102
$200 < P/\text{days} < 500$	18.4119	0.3959	0.9051	1.0984
m = 2				
$0.5 < P/\text{days} < 200$	21.3265	0.4203	0.9276	0.093
$200 < P/\text{days} < 500$	5.5213	1.2307	0.7456	2.9774

The transit probability for a single planet system can be shown to be (Pater & Lissauer, 2015)

$$P_{tr} = \frac{R_p + R_*}{a(1-e^2)} \approx \frac{R_*}{a}, \tag{A.7}$$

where the orbits are assumed to be circular ($e = 0$). Circular orbits is a reasonable enough assumption as the eccentricities of single planet have been found to be small enough to not affect the period calculation significantly (Van Eylen et al., 2019). For multiple planet systems, the transit probability becomes more complicated and it becomes necessary to look at semi-analytical models and perform Monte-Carlo simulations to investigate the probability of transits (Brakensiek & Ragozzine, 2016). Zink et al. (2019) did exactly this

for planetary systems containing between 2 and 7 planets and for different values of a/R_* (note the inversion). The first planet was found through the different values of a/R_* and the rest of the planets were drawn from their Kepler period sample. The mutual inclinations were made to follow the distribution found by Fang & Margot (2012). Then, the stability of the system was checked for by investigating the planet separations and if any separation is less than 10 percent of its outer neighbour, the entire system is resampled. For a transit to be observed, one of the following equations must be satisfied

$$\begin{aligned} \cos(i)\cos(\omega) - \sin(i)\sin(\omega) &\geqslant \frac{R_*}{a} \\ \sin(i)\sin(\omega) - \cos(i)\cos(\omega) &\leqslant \frac{R_*}{a}, \end{aligned} \tag{A.8}$$

where i is the inclination; uniformly sampled over $\sin(i)$, and ω is the ascending node; also uniformly sampled over $\sin(\omega)$. The results of this sampling can be found in figure A.1 for all different values of m. This is the same as figure 1 in Zink et al. (2019). Given a certain semi major axis, stellar radius, and order m, the transit probability was found through linear interpolation between these found values.

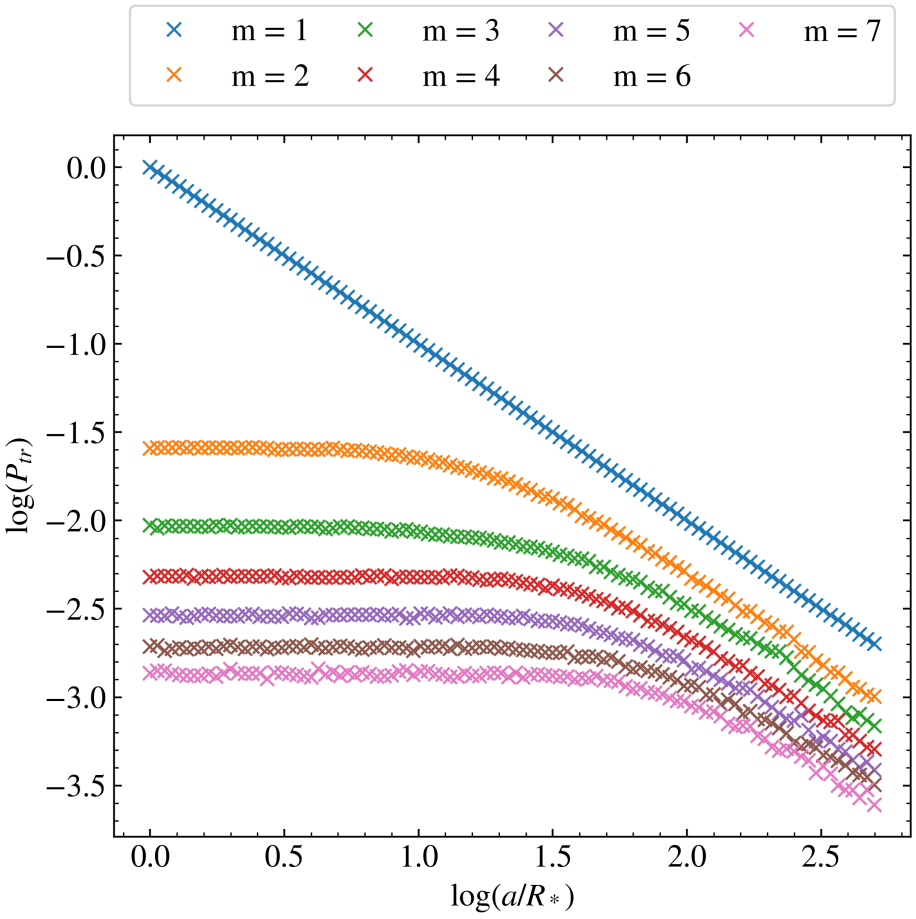

Figure A.1: Found transit probabilities by Zink et al. (2019).

A.2 Exoplanet distribution

The exoplanet distributions for the planet radii and orbital period g and q are most commonly written as power laws with a break point corresponding to R_{br} and P_{br}

$$g(R) = \begin{cases} A_R R^{\alpha_1} & \text{if} \quad R < R_{br} \\ B_R R^{\alpha_2} & \text{if} \quad R \geqslant R_{br} \end{cases}$$
$$q(P) = \begin{cases} A_P P^{\beta_1} & \text{if} \quad P < P_{br} \\ B_P P^{\beta_2} & \text{if} \quad P \geqslant P_{br}, \end{cases} \tag{A.9}$$

where α_1, α_2, β_1, β_2, P_{br}, R_{br} are all fit parameters. Since Zink et al. (2019) found these parameters in their study using the same model that will be used during the project and considering the fact that this model is supposed to be describing the exoplanet population

as a whole, the values found by Zink et al. (2019) will be adopted in this project. The values for the parameters can be found in table A.2. A_R, B_R, A_P, and B_P are coefficients which are determined in order to satisfy normalisation as well as continuity at the breakpoints of the two functions. Specifically, they are determined by solving the following set of linear equations

$$\begin{aligned} &\int_{P_{min}}^{P_{br}} A_P P^{\beta_1} + \int_{P_{br}}^{P_{max}} B_P P^{\beta_2} = 1 \\ &A_P P_{brk}^{\beta_1} = B_P P_{brk}^{\beta_2} \\ &\int_{R_{min}}^{R_{br}} A_R R^{\alpha_1} + \int_{R_{br}}^{R_{max}} B_R R^{\alpha_2} = 1 \\ &A_R R_{brk}^{\alpha_1} = B_R R_{brk}^{\alpha_2}, \end{aligned} \tag{A.10}$$

Table A.2: Parameters for the exoplanet population and their values as found by Zink et al. (2019)

Parameter	Value
α_1	-1.65
α_2	-4.35
β_1	0.76
β_2	-0.64
P_{br}	7.09 days
R_{br}	2.66 $M_\oplus$

A.3 Sorting Order

The joint distribution model is described as the following, given a quantity x

$$P_m(x) \propto P_0(x) C_0(x)^{a_m - 1} (1 - C_0(x))^{b_m - 1}, \tag{A.11}$$

where P_0 is the true probability distribution while C_0 is the CDF. Due to the separability of the exoplanet distribution with respect to radius and period, the skewed portion of the distribution can then be written as

$$O_m(P, R) = C_r(R)^{a_{m,r}-1}[1 - C_r(R)]^{b_{m,r}-1} \cdot C_p(P)^{a_{m,p}-1}[1 - C_p(P)]^{b_{m,p}-1}, \tag{A.12}$$

where again, the parameters $a_{m,r}$, $b_{m,r}$, $a_{m,p}$, and $b_{m,p}$ will be adopted from Zink et al. (2019) and their values can be found in table A.3. The CDF's can be described through

the following expressions

$$C_R(R) = \begin{cases} \int_{R_{min}}^{R} A_R R'^{\alpha_1} dR' & \text{if} \quad R < R_{brk} \\ \int_{R_{min}}^{R_{brk}} A_R R'^{\alpha_1} dR' + \int_{R_{brk}}^{R} B_R R'^{\alpha_2} dR' & \text{if} \quad R \geqslant R_{brk} \end{cases} \tag{A.13}$$

$$C_P(P) = \begin{cases} \int_{P_{min}}^{P} A_P P'^{\beta_1} dP' & \text{if} \quad P < P_{brk} \\ \int_{P_{min}}^{P_{brk}} A_P P'^{\beta_1} dP' + \int_{P_{brk}}^{P} B_P P'^{\beta_2} dP' & \text{if} \quad P \geqslant P_{brk}. \end{cases}$$

Table A.3: Parameters for the sorting order skew as found by Zink et al. (2019)

Parameter	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$
a_r	1.095	1.030	1.028	1.013	0.998	1.065	0.951
b_r	0.923	1.470	2.206	3.063	4.013	4.898	6.614
a_p	0.957	1.152	1.172	1.184	1.183	1.166	1.234
b_p	1.004	1.010	1.000	0.999	0.997	1.006	0.994

Appendix B

Posterior distribution of the occurrence rate given no detections

The goal is to estimate the posterior distribution of f_m, given that there are no detected m-planet systems ($N_{det,m} = 0$). From Bayes theorem:

$$P(f_m|N_{det}) = \frac{P(N_{det}|f_m)P(f_m)}{P(N_{det})} \propto P(N_{det}|f_m), \tag{B.1}$$

where the proportionality comes from the fact that we use a uniform prior for f_m. Since the detection of planets is modeled as a Poisson process,

$$P_m(N_{det}|f_m) \sim \text{Pois}(N_{det}; N_{exp}) = \frac{N_{exp}^{N_{det}}}{N_{det}!} \exp(-N_{exp}), \tag{B.2}$$

where N_{exp} is the expected number of detections

$$N_{exp} = N_* f_m \int_{P_{min}}^{P_{max}} \int_{R_{min}}^{R_{max}} P(R, P) O_m(R, P) q(P) g(R) dR dP. \tag{B.3}$$

The probability to detect zero planets, given a certain value for f_m is then

$$P(0|f_m) = \exp(-N_{exp}) \tag{B.4}$$

Appendix C

Normalisation of expected number of detections

We want to calculate the value for C_{norm} which can be expressed through

$$C_{norm} \int_{P_{min}}^{P_{max}} \int_{R_{min}}^{R_{max}} g(P)q(R)O_m(R,P)dPdR = 1. \tag{C.1}$$

As

$$O_m(p,r) = C_r(R)^{a_{m,r}-1}[1-C_r(R)]^{b_{m,r}-1} \cdot C_p(P)^{a_{m,p}-1}[1-C_p(P)]^{b_{m,p}-1}, \tag{C.2}$$

it is possible to separate the double integral into a period term and radius term

$$\int_{P_{min}}^{P_{max}} \int_{R_{min}}^{R_{max}} g(P)q(R)O_m(R,P)dPdR = \\ \int_{P_{min}}^{P_{max}} C_p(P)^{a_{m,p}-1}[1-C_p(P)]^{b_{m,p}-1}q(P)dP \int_{R_{min}}^{R_{max}} C_r(R)^{a_{m,r}-1}[1-C_r(R)]^{b_{m,r}-1}g(R)dR. \tag{C.3}$$

Through the definition of the CDF, it is possible to perform the following variable substitution $\frac{dC_r}{dR} = g(R) \Rightarrow dR = \frac{1}{g(R)}dC_r$. The integral limits can then be changed as $C_r(R_{max}) = 1$, $C_r(R_{min}) = 0$. These operations are equally true for the period part of the expression. This means that

$$C_{norm} \int_0^1 C_p(p)^{a_{m,p}-1}[1-C_p(p)]^{b_{m,p}-1}dC_p \int_0^1 C_r(R)^{a_{m,r}-1}[1-C_r(R)]^{b_{m,r}-1}dC_r = 1. \tag{C.4}$$

These integrals have analytical solutions expressed by the Γ-function which means that the normalisation factor can be written as

$$C_{norm,m} = \left(\frac{\Gamma(a_{m,r})\Gamma(b_{m,r})}{\Gamma(a_{m,r}+b_{m,r})} \frac{\Gamma(a_{m,p})\Gamma(b_{m,p})}{\Gamma(a_{m,p}+b_{m,p})} \right)^{-1}. \tag{C.5}$$

Appendix D

Additional Figures

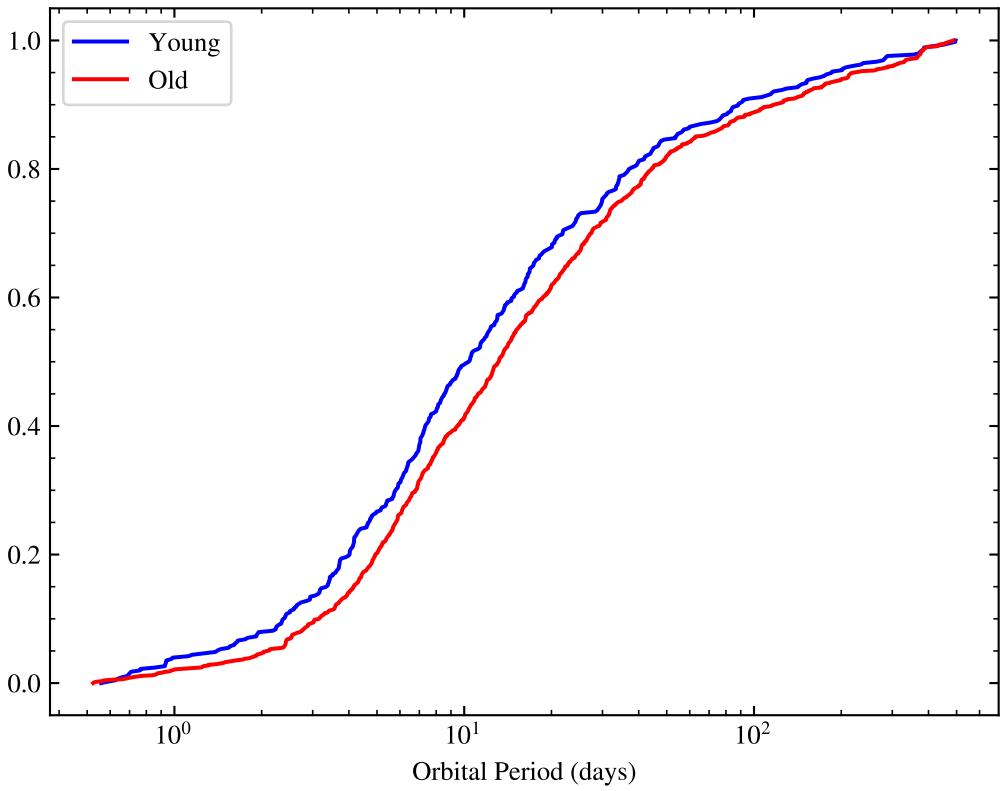

Figure D.1: Cumulative distribution functions of orbital periods of the planet population. The two colors show stars older than 2.5 Gyr and those younger than 2.5 Gyr.

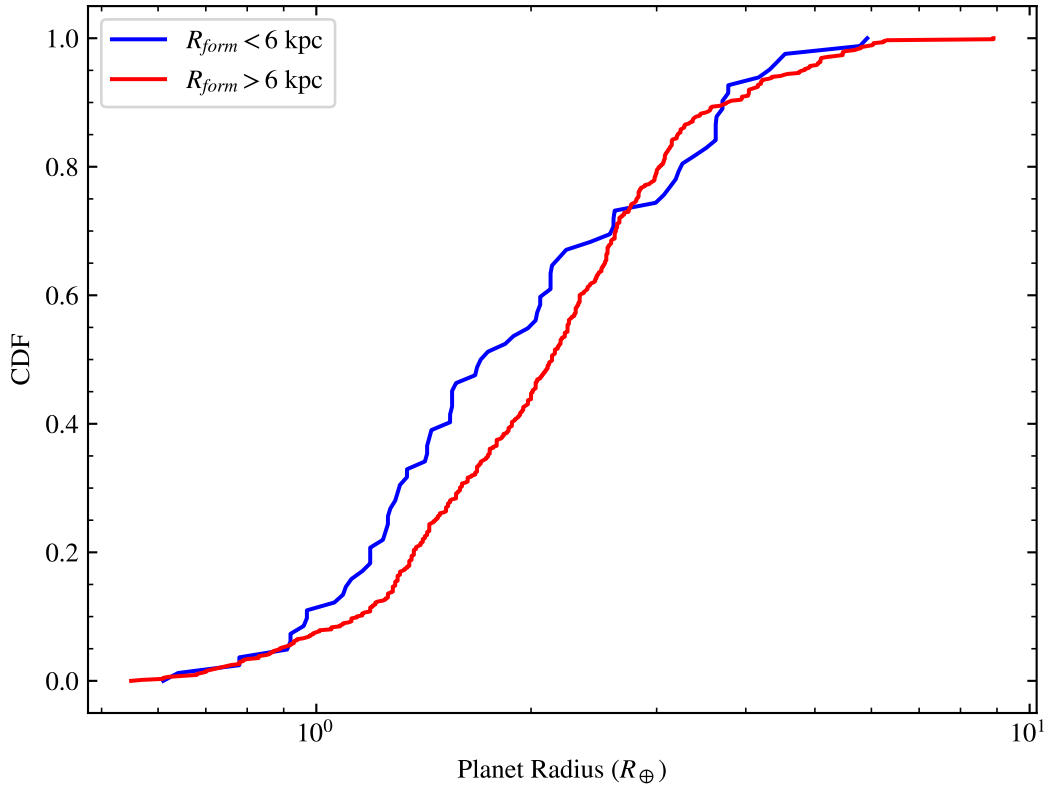

Figure D.2: Cumulative distribution functions of planet radii with host stars formed inside 6 kpc compared to host stars formed outside of 6 kpc according to the Frankel model.

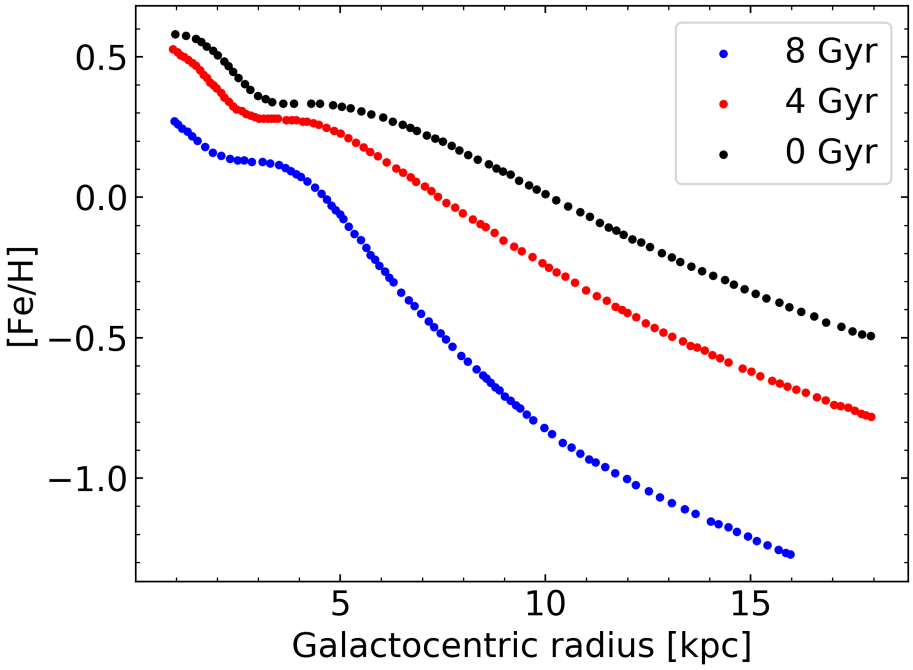

Figure D.3: The reproduced metallicity gradients given by Kubryk et al. (2015). It is important to note that the ages given here is lookback time (or stellar age) and not simulation time which is reported in Kubryk et al. (2015)

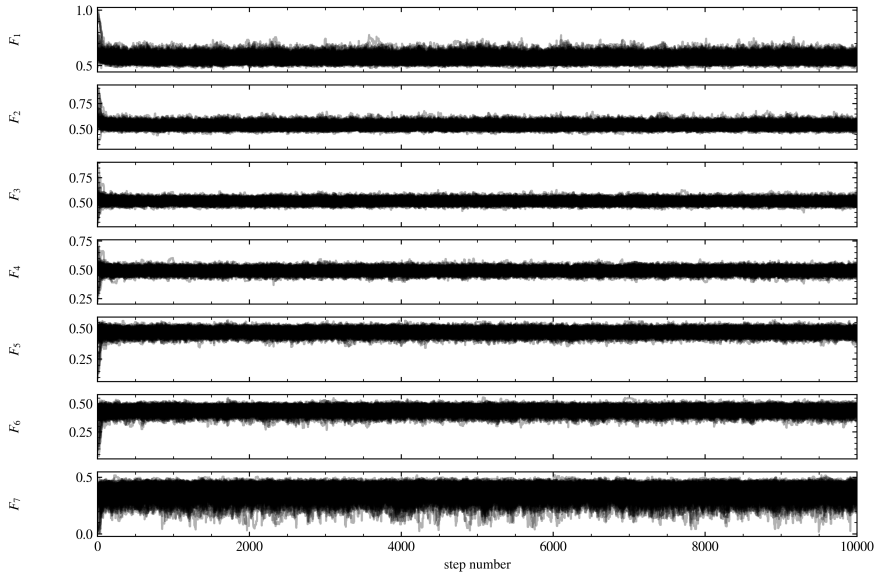

Figure D.4: An example of a single walker chain. In the full procedure, these were generated 50 times for each formation radius bin and then combined in order to produce the final distributions of F_m and G_m.