Composition of terrestrial planets orbiting M dwarfs

Bárbara Maria Teixeira Basto Soares
Mestrado em Astronomia e Astrofísica
Departamento de Física e Astronomia
2021

Orientador
Dr. Vardan Adibekyan, Instituto de Astrofísica e Ciências do Espaço; Faculdade de Ciências da Universidade do Porto

Coorientadores
Prof. Dr. Nuno Santos, Faculdade de Ciências da Universidade do Porto; Instituto de Astrofísica e Ciências do Espaço
Dra. Elisa Delgado Mena, Instituto de Astrofísica e Ciências do Espaço
Dr. Sérgio Sousa, Instituto de Astrofísica e Ciências do Espaço

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

UNIVERSIDADE DO PORTO

MASTERS THESIS

Composition of terrestrial planets orbiting M dwarfs

Author:
Bárbara SOARES

Supervisor:
Vardan ADIBEKYAN

Co-supervisors:
Nuno SANTOS, Sérgio SOUSA,
Elisa DELGADO MENA

A thesis submitted in fulfilment of the requirements
for the degree of MSc. Astronomy and Astrophysics

at the

Faculdade de Ciências da Universidade do Porto
Departamento de Física e Astronomia

December 5, 2021

" Don't stop. Grit your teeth and look straight ahead. Even if you stop and crouch down, Time won't wait for you, or snuggle you and grieve along with you. "

Rengoku Kyoujurou in *Kimetsu no Yaiba*, written by Koyoharu Gotouge

Acknowledgements

This whole thesis project wouldn't be possible without the help and support of a lot of people, especially during this last year, and so I shall try and address them all. I would like to start by thanking my supervisors. Vardan, I thank you for your constant guidance, for sharing so readily your knowledge on so many different areas, for always answering even the smallest of questions and making sure I stayed on track and didn't stray too much on some discussions and for always replying my emails, not minding the ridiculously late (early?) hours they were usually sent. Above all, I sincerely thank you for your unwavering support, even with your new little one, I can only imagine it must not have been easy. Nuno, thank you so much for always being available for on-topic and off-topic discussions and in prompt meetings, for your immense knowledge in exoplanets which always taught me something new whenever we discussed and for showing me (even if it was not your intent) how to better communicate in science. Your classes in the first year were a major part of it and cemented my certainty that I wanted to stay in the Planetary Sciences area. Elisa, I want to thank you for your constant quiet support, for always making the interesting questions and pinpointing the most curious scenarios and what-ifs, as they always got me thinking and questioning my actions, leading to a better understanding of what I was doing. Sérgio, I thank you for always giving me insightful feedback, for pointing out the smallest but important details and for always suggesting different approaches when I was uncertain about some result. I must also not forget to thank you and Vardan for always forwarding me any opportunities you thought could be useful to me. I know this was a project a bit done on the fly, but I would like to express my sincere gratitude to the four of you that made this thesis possible for me.

A special thanks to my friends who made this year easier. To Mariana and Alexandra who, although busy with their own work, always made a bit of time for a meet up, lunch and/or a small chat and for checking in on me from time to time. To Pedro, for answering my machine learning questions, making useful comments and giving valuable suggestions to try. To Nuno, for always cheering me on, keeping me motivated and for letting me crash at his place a whole day when I needed a change. To Francisco, who was constantly nagging me everyday, regardless of getting a reply or not, stubbornly checking in on me and making sure I went to sleep at decent times. To João(zinho) for always tempting me to go eat out. To Carolina, Miguel, Haru and Becca, thank you for your support and patience. I wouldn't have made it here without any of you. I also thank D. Susana and Sr. José Pinto of J.F.L, for promptly offering me a place to work without worries. If I forgot anyone, please know it was not on purpose.

Last but not least, my gratitude goes to my family, and this last paragraph is dedicated to them. Papá, obrigada pela constante disponibilidade para mudar os planos à última da hora, por me deixares fazer as coisas ao meu ritmo (6:20), mostrando sempre o teu apoio constante. Manis, devo-te tudo: muito obrigada pelo apoio, paciência e compreensão infinitos, por me teres encorajado a continuar e andar para a frente quando estava empancada e por me teres ajudado com as decisões difíceis. Hugo e Filipe, obrigada por não me deixarem esquecer que trabalho e empenho dão sempre resultados, independentemente do resto. Ricardo, obrigada pela companhia constante e por prontamente me passares o testemunho, seja a que horas for. Patrícia, obrigada por ouvires os meus devaneios e ofereceres sempre uma crítica ou opinião sincera. Sarinha e Joaninha, obrigada por me relembrarem que a vida não é só trabalho, mas quando for para trabalhar a sério, temos de fazer contar. Raquel, obrigada pelo apoio constante, acho que foste uma das que me acompanhou mais de perto. Por muito ridículas ou rebuscadas que tenham sido *aquelas* trocas de ideias, ajudaram sempre a desanuviar a cabeça ou quando estava bloqueada, e por isso te agradeço. Vovó Nieta, obrigada pela preocupação constante pelo meu bem estar, seja a nível académico ou pessoal, e por sempre me incentivares a não desistir. Agradeço-vos a todos vós o vosso apoio incondicional e inabalável. Sem vocês não teria conseguido e quero que saibam que vos sou grata. Muito obrigada do fundo do coração!

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto

Departamento de Física e Astronomia

MSc. Astronomy and Astrophysics

Composition of terrestrial planets orbiting M dwarfs

by Bárbara SOARES

Most stars in our galaxy are M dwarfs, stars smaller and cooler than Sun-like stars. Because of the detection efficiency, many low-mass/small-sized planets are discovered around low-mass stars, like M dwarfs. To better understand planetary properties, such as their internal structure and their atmosphere, we need to know the characteristics of the planet host star. However, M dwarfs are intrinsically faint in the optical, which makes their observations with most of ground-based state-of-the-art spectrographs challenging. Due to their cooler atmospheres, their spectra contains millions of molecular lines which makes it very difficult to identify and measure the spectral lines corresponding to different chemical elements. By assuming that the Galactic chemical evolution is a 'unigalactic' process, we can estimate M dwarfs chemical composition based on their FGK counterparts, for which the chemical composition can be determined with high precision. We use machine learning techniques, taking metallicities and galactic velocities for a large sample of FGK dwarfs as input parameters, to establish a new method to predict the chemical abundances for M stars.

The first goal of this work is to develop calibration relationships between the stellar parameters (effective temperature, metallicity, and surface gravity) and chemical composition based on FGKM stars from large spectroscopic surveys, and to build a model able to provide abundance estimates for M dwarfs. The second main objective of this thesis is to characterise low-mass planets orbiting M dwarfs. We use the calibrations obtained through machine learning to predict the abundances of major rock forming elements: C, O, Mg, and Si. We then inserted these abundances into a stoichiometric model to determine the iron-to-silicate mass fraction (as it is correlated with the planet's core size), and water-mass fraction (which is important for the habitability of terrestrial planets) of the protoplanetary disk. We then tested whether there is a correlation between the composition of the planetary disk and the densities (which depends on their composition) of the planets.

From the exoplanet databases we selected most precisely characterised low-mass ($M < 10\ M_{Earth}$) and small-sized ($R < 2\ R_{Earth}$) planets (precision in R and M better than 30%) orbiting around M dwarf and for these 10 stars estimated the average iron-to-silicate mass fraction ($\sim 32 \pm 1\%$), water mass fraction ($\sim 53 \pm 5\%$) and total heavy metal content ($\sim 1.5 \pm 1\%$) which appeared to be similar to the Solar System values. We were not able to observe a clear trend between the iron-to-silicate mass fraction of the host star (which we assumed to reflect the composition of the protoplanetary disk) and the density of the planets, possibly due to not enough precision in the estimation of the abundances. Despite this, our ML model provides relatively good estimates for the abundances of the aforementioned elements which can be used to characterise planetary interiors with constraints from the host composition.

— Key-words: Low mass stars; exoplanets; planetary systems; rocky planets; chemical abundances; machine learning; habitable zone.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto

Departamento de Física e Astronomia

Mestrado em Astronomia e Astrofísica

Composição de planetas terrestes que orbitam anãs vermelhas

por Bárbara SOARES

A maioria das estrelas da nossa galáxia são anãs vermelhas, de tipo espectral M, com temperatura e tamanho inferiores a estrelas tipo-Sol. Devido à eficiência da deteção, vários planetas de baixa massa/pequenas dimensões são descobertos em torno destas estrelas. Para melhor compreender as propriedades planetárias, como a sua estrutura interna e a sua atmosfera, é necessário conhecer as características da estrela hospedeira. Contudo, as anãs vermelhas são intrisecamente ténues na banda do visível, o que torna desafiantes a sua observação mesmo com os mais recentes espetrógrafos terrestres. Devido às suas atmosferas mais frias, os seus espectros contêm milhões de linhas moleculares, sendo muito difícil identificar e medir as linhas espectrais correspondentes a diferentes elementos químicos. Assumindo que a evolução química Galáctica é um processo 'unigaláctico', podemos estimar a composição química de anãs vermelhas baseada em estrelas do tipo espectral FGK, para as quais a composição química pode ser determinada com elevada precisão. Nesta tese utilizamos técnicas de *machine learning*, tendo as metalicidades e velocidades galácticas já existentes para uma larga amostra de estrelas tipo FGK como parâmetros de entrada, a fim de estabelecer um novo método para prever as abundâncias químicas de estrelas tipo espectral M.

O primeiro objectivo deste trabalho consiste em desenvolver relações de calibração entre parâmetros estelares (temperatura efectiva, metalicidade e gravidade à superfície) e composição química, baseado em estrelas FGKM de grandes levantamentos espectroscópicos, bem como criar um modelo capaz de estimar abundâncias para anãs vermelhas. O segundo objectivo foca-se em caracterizar os planetas de baixa massa que orbitam anãs vermelhas. Utilizamos as calibrações obtidas através de *machine learning* para prever as abundâncias dos principais elementos constituintes das rochas: C, O, Mg e Si. A seguir, inserimos estas abundâncias num modelo estequiométrico para determinar a fração de massa de ferro-silicatos (correlacionada com o tamanho do núcleo do planeta) e a fracção

de massa de água (importante para a habitabilidade de planetas terrestres) do disco protoplanetário. Seguidamente, testámos se há uma relação química entre a composição do disco protoplanetário e as densidades dos planetas (que dependem da composição deles).

De bases de dados de exoplanetas selecionámos planetas de baixa massa ($M < 10\ M_{Earth}$) e pequenas dimensões ($R < 2\ R_{Earth}$) de melhor precisão (precisão do R e M acima de 30%), a orbitar anãs vermelhas. Para estas 10 estrelas estimámos as frações de massa de ferro-silicatos ($\sim 32 \pm 1\%$), de água ($\sim 53 \pm 5\%$) e de metais pesados ($\sim 1.5 \pm 1\%$), que aparentam ser semelhantes aos valores do Sistema Solar. Não foi possível observar uma relação clara entre a fração de massa de ferro-silicatos da estrela (que assumimos refletir a composição do disco protoplanetário) e a densidade dos planetas, possivelmente devido a insuficiente precisão na determinação das abundâncias. Apesar disto, o nosso modelo de ML fornece relativamente boas estimativas para as abundâncias dos referidos elementos, que podem ser usados para caracterizar interiores planetários com condições vindas da composição da estrela hospedeira.

Contents

Acknowledgements v

Abstract vii

Resumo ix

Contents xii

List of Figures xvii

Glossary xxiii

1 Introduction 1
- 1.1 Detecting exoplanets . . . 3
  - 1.1.1 The radial velocity method . . . 3
  - 1.1.2 The transit method . . . 4
  - 1.1.3 Other methods . . . 6
- 1.2 The star-planet connection . . . 6
- 1.3 M dwarfs . . . 8
  - 1.3.1 M dwarfs as host of small rocky planets . . . 11
- 1.4 Planet interior models . . . 14
- 1.5 Spectroscopic surveys . . . 16
  - 1.5.1 APOGEE . . . 17
  - 1.5.2 GALAH . . . 18
  - 1.5.3 HARPS . . . 18
- 1.6 This work . . . 19

2 Predicting stellar abundances with Machine Learning 21
- 2.1 Machine Learning and Random Forests . . . 22
- 2.2 The training data . . . 25
  - 2.2.1 APOGEE . . . 26
  - 2.2.2 GALAH . . . 36
  - 2.2.3 HARPS . . . 39
- 2.3 Galactic velocities . . . 42
- 2.4 Predicting abundances with RF . . . 43
- 2.5 Testing the models . . . 47

3 Iron-to-silicate mass fraction in planet host stars **53**
3.1 From stellar abundances to iron-to-silicate mass fraction . . . 53
3.2 Determining the composition . . . 55
3.2.1 HARPS data . . . 55
3.3 Additional tests . . . 62

4 Composition of low-mass planets and their host stars **67**
4.1 Selecting the sample . . . 67

5 Conclusions and future work **77**

A The log g problem **81**

B Chemical abundances depending on stellar parameters (and S/N) **85**
B.1 GALAH total sample . . . 85
B.1.1 Temperature dependency . . . 85
B.1.2 Log g dependency . . . 87
B.1.3 Metallicity dependency . . . 88
B.2 Solar Analogues stars . . . 90
B.2.1 Temperature dependency . . . 90
B.2.2 Log g dependency . . . 91
B.2.3 Metallicity dependency . . . 92
B.3 HARPS sample . . . 94
B.3.1 Dependency of chemical abundances with S/N . . . 94
B.3.2 Temperature dependency . . . 95
B.3.3 Surface gravity dependency . . . 96
B.3.4 Metallicity dependency . . . 98
B.4 A brief check on the Galactic chemical evolution . . . 99

C Random Forest results **101**
C.1 APOGEE . . . 101
C.1.1 Temperature . . . 101
C.1.2 Log g . . . 102
C.2 GALAH . . . 104
C.2.1 Temperature . . . 104
C.2.2 Log g . . . 105
C.3 HARPS . . . 107
C.3.1 Temperature . . . 107
C.3.2 Log g . . . 108

D Uncertainties of predicted abundances **109**
D.1 Precision of abundance values . . . 109
D.2 Determining the systematic error . . . 109
D.3 Final error of predicted abundances . . . 110

E Testing the models with external data **111**
E.1 Testing the models with data from Adibekyan et al. (2021) . . . 111
E.2 Testing the models with data from Maldonado et al. (2020) . . . 114

E.3 Testing the models with data from González Hernández et al. (2010) 116

F Additional tests to the ML models **119**
F.1 SWEET-Cat database 119
F.2 Predicting the [Mg/Si] abundance ratio 123
F.3 Predicting the C and O abundances 126
F.4 Compare HARPS and Nissen (2015) data 126
F.5 Predicting Mg and Si abundances with HARPS 1111 stars 128
F.6 Predicting only [Mg/H] and [Si/H] 130

Bibliography **133**

List of Figures

1.1 Confirmed exoplanets with measured masses (in Earth mass units) and known semi-major axis (in astronomical units, AU). The dashed lines represent the value for the Jupiter mass (top line) and the Earth mass (bottom line). Note that the masses discovered solely by the RV method are the minimum masses. Data taken from The Extrasolar Planets Encyclopaedia, found in `http://exoplanet.eu` and accessed in 02/09/2021. 2

1.2 Mass-radius relationships for low mass rocky planets orbiting solar-type stars, with mass measurements better than 20% precision. The solid lines are theoretical mass–radius curves for planets with the different compositions. The shaded grey region in the lower right indicates planets with iron content exceeding the maximum value predicted from models of collisional stripping. Taken from Dressing et al. (2015). 8

1.3 Spectra of GJ 180 (M1.0, top panel), Ross 128 (M4.0, middle panel) and HD 180617 (M8.0, bottom panel) in black, compared with the best-fit BT-Settl model (red). Figure taken from Figure 1 of Rajpurohit et al. (2018). 10

1.4 Confirmed exoplanets with measured masses (in Earth mass units), per discovery year. The dashed lines represent the value for the Jupiter mass (top line) and the Earth mass (bottom line). Note that the masses discovered solely by the RV method are the minimum masses. Data taken from The Extrasolar Planets Encyclopaedia, found in `http://exoplanet.eu` and accessed in 02/09/2021. 12

2.1 Sketch illustrating the general working of reinforcement learning algorithms. Image taken from Géron (2019). 24

2.2 [X/Fe] and its uncertainties dependent on S/N, for X = Mg, Si, C, O. The bottom panels correspond to [Mg/Si] and [C/O] and respective errors. The orange line corresponds to the linear fit adjusted for the data. 28

2.3 HR diagram for different values of S/N. The N on the title represents the sample size. 29

2.4 $\log g$ errors depending on T_{eff}. The colour refers to T_{eff} and is just to evidence the stars disposition: hot stars on the left and cooler stars on the right. 29

2.5 $\log g$ errors depending on metallicity. The colour refers to $\log g$ values. The top group corresponds to Main Sequence stars, the middle on to Red Clump stars and the bottom one to Red Giant Branch stars. 30

2.6 Abundance ratios as a function of [Fe/H] for MS dwarf stars. Colour scheme represents the density of points (stars). 33

2.7 Abundance ratios as a function of [Fe/H] for RGB stars. Colour scheme represents the density of points (stars). 34

2.8 Abundance ratios as a function of [Fe/H] for RC stars. Colour scheme represents the density of points (stars). 35
2.9 Correlations between the different parameters, for the APOGEE sample of RGB stars. 36
2.10 Correlation matrix for the parameters of GALAH data set. 38
2.11 Correlation matrix for the parameters of Solar Analogue stars sample. . . . 40
2.12 Correlation matrix for the parameters of the HARPS sample. 41
2.13 The X axis is the values of T_{eff} (top) and $\log g$ (bottom) and the Y axis is the *Predictions – True value*, for the chemical abundances of [Mg/H] (left) and [Mg/Fe] (right). The black line marks values of predictions equal to the true data. The title includes the abundance ratio we are predicting, the mean difference of *Predictions – True value*, the standard deviation for the predictions and the standard deviation for the true values. 45
2.14 Predictions of Nissen(2015)'s data, using the RFA, RFG and RFH models. Y axis is the difference between the predicted values and the true values, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. Title includes standard deviation of true data and predictions, mean difference of Predictions - True data values, and correlation between prediction and true data, for all samples. The black line marks values of predictions equal to the true data. 49

3.1 Composition for the HARPS sample (left column), RFA (middle column) and RFH predictions (right column). Top row is for the iron-to-silicate mass fraction, middle row for the water mass fraction and bottom row for the mass percentage of all heavy elements. Stars are distributed in galactic populations. 56
3.2 Composition for the HARPS sample (left column), RFA (middle column) and RFH predictions (right column). Top row is for the iron-to-silicate mass fraction, middle row for the water mass fraction and bottom row for the mass percentage of all heavy elements. Stars from the thin disk are divided in two groups. 57
3.3 Cumulative distribution function of the w_f for the three set of results: the true (original) data and the predictions for the RFA and RFH models. Left panel is the w_f value of thin disk stars and right panel is the w_f for thick disk stars. 60
3.4 Predictions of the HARPS data using the APOGEE based random forest, RFA. Y axis is the difference between predicted and true data while on the X axis are the metallicity values. Title includes information on the mean difference and standard deviation of true data and predictions. 60
3.5 Distribution of f_{iron} (top), w_f (middle) and Z (bottom) values, for 500 random stars from APOGEE and predictions of RFA and RFH models. 65

4.1 f_{iron} (top), w_f (middle), and Z (bottom) mass fractions distributions for the 36 stars from the NEA sample. 70
4.2 f_{iron} (top), w_f (middle), and Z (bottom) mass fractions distributions for the 10 M dwarf stars from the NEA sample. 72

4.3 Densities of rocky planets as a function of iron-to-silicate mass fraction. The measured density of each planet was normalised by the expected density of an Earth-like composition for the same mass. The iron-to-silicate mass fraction was calculated from the individual abundances of the host star, using RFA (top panel) and RFH (bottom panel) predictions. Red dots represent planets orbiting M dwarfs and blue dots represent planets around FGK stars. 75

A.1 Histogram for the flags in the $\log g$ parameter. 82
A.2 Histogram of the $\log g$ absolute uncertainties (left) and histogram of the relative uncertainties of $\log g$ (right). 82
A.3 Log g uncertainties depending on the $\log g$ values, for APOGEE stars. . . . 83
A.4 Comparison of HR diagram for uncalibrated parameters, using the same colour scheme: black for RGB, grey for red clump stars, orange for dwarfs and blue for the mix between RGB and dwarfs. On the left we have the diagram from Jönsson et al. (2020); on the right we have the plot obtained in this work. 84
A.5 Same as Figure A.4, but for calibrated parameters. 84

B.1 Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on T_{eff}. 86
B.2 Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on $\log g$. 88
B.3 Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on metallicity ([Fe/H]). 89
B.4 Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on T_{eff}. 91
B.5 Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on $\log g$. 92
B.6 Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on metallicity ([Fe/H]). 93
B.7 Plots of chemical abundances depending on S/N, for [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. 95
B.8 Plots of chemical abundances depending on T_{eff}, for [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. 96
B.9 Plots of chemical abundances depending on $\log g$, for [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. 97
B.10 Plots of chemical abundances depending on metallicity, for [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. 99
B.11 Distribution of [C/Fe] (left) and [O/Fe] (right) abundance ratios with metallicity, for total GALAH sample (top), for the Solar Analogue stars (middle) and for the HARPS sample (bottom). 100

C.1 Difference between the predicted values and the true values, depending on temperature, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data. 102

C.2 Difference between the predicted values and the true values, depending on $\log g$, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data. 103
C.3 Difference between the predicted values and the true values, depending on temperature, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data. 105
C.4 Difference between the predicted values and the true values, depending on $\log g$, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data. 106
C.5 Difference between the predicted values and the true values, depending on temperature, for the chemical abundances of [C/H], [O/H], [Si/H], and abundance ratios of [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data. 107
C.6 Difference between the predicted values and the true values, depending on $\log g$, for the chemical abundances of [C/H], [O/H], [Si/H], and abundance ratios of [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data. 108

E.1 Predictions of Adibekyan et al.(2021)'s data, using the RFA, RFG and RFH models. Y axis is the difference between the predicted values and the true values, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The title includes the standard deviation of true data and predictions, mean difference of $Predictions - True\ data$ values, and correlation between prediction and true data, for all samples. The black line marks values of predictions equal to the true data. 112
E.2 Predictions of Maldonado et al. (2020)'s data, using the RFA, RFG and RFH models. Y axis is the difference between the predicted values and the true values, for the chemical abundances of [C/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe] and [Mg/Si]. The title includes the standard deviation of true data and predictions, mean difference of $Predictions - True\ data$ values, and correlation between prediction and true data, for all samples. The black line marks values of predictions equal to the true data. 114
E.3 Predictions of González Hernández et al. (2010)'s data, using the RFA, RFG and RFH models. Y axis is the difference between the predicted values and the true values, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The title includes the standard deviation of true data and predictions, mean difference of $Predictions - True\ data$ values, and correlation between prediction and true data, for all samples. The black line marks values of predictions equal to the true data. 117

F.1 Abundance predictions of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O], for 211 stars with $T_{eff} < 4000$ K from SWEET-Cat dataset, using the APOGEE, GALAH and HARPS random forest model. . . 121
F.2 Abundance predictions of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O], for all stars from SWEET-Cat dataset, using the APOGEE, GALAH and HARPS random forest model. 122

F.3 [Mg/Si] determined from the difference of Mg and Si abundances depending on the predicted values of [Mg/Si] from Nissen (2015)'s data, using RFA (top), RFG (middle) and RFH (bottom) models. 124
F.4 Same as Figure F.3, but for Adibekyan et al. (2021)'s data. 125
F.5 Comparing the predictions of [C/H] and [O/H] abundances using the RFA, RFG and RFH models and the predictions of Adibekyan et al. (2021). Y axis is the difference between our predicted values and those of the author. X axis represents our values obtained using the RF models. Title includes information about the mean value of $Predictions - Predictions_{Adibekyan}$. . . 126
F.6 Comparing metallicity and individual abundances of common stars between Nissen (2015)'s data and the HARPS sample. 127
F.7 Comparison of difference between predictions and true data, depending on effective temperature, for both HARPS RF models. Both models were applied to the total HARPS sample to predict [Mg/H] abundances (left) and [Si/H] abundances (right). Title includes information on the mean difference between predicted and true values and standard deviation of predictions and true data. 128
F.8 Comparison of difference between predictions and true data, depending on effective temperature, for both HARPS RF models. Both models were applied to Adibekyan et al. (2021)'s data to predict [Mg/H] abundances (left) and [Si/H] abundances (right). Title includes information on the mean difference between predicted and true values and standard deviation of predictions and true data. 129
F.9 Distributions of [Mg/H] and [Si/H] abundances, for the HARPS samples with 527 stars (red) and 1111 stars (blue). 129
F.10 Comparing predictions from RF with individual abundances and abundance ratios as outputs and from RF with only [Mg/H] and [Si/H] as outputs, for the APOGEE (top), GALAH (middle) and HARPS (bottom) samples. Left column are the results for the Mg abundances and right column are for the Si abundances. Title includes information on the mean difference and standard deviation of both predictions and true data. 131

Glossary

AU	Astronomical units
exoplanet.eu	The Extrasolar Planets Encyclopaedia
$\mathbf{f_{iron}}$	iron-to-silicate mass fraction
HZ	Habitable zone
LSR	Local standard of rest
mas	milli-arsecond
mas/yr	milli-arsecond per year
ML	Machine Learning
MS	Main Sequence
NEA	NASA Exoplanet Archive
pc	parsec
RC	Red Clump
RF	Random Forest
RFA	Random Forest trained on APOGEE sample
RFG	Random Forest trained on GALAH sample
RFH	Random Forest trained on HARPS sample
RGB	Red Giant Branch
RV	Radial Velocity
SAS	Solar Analogue stars
SDSS	Sloan Digital Sky Survey
SN Ia	Supernovae type Ia

SN II	Supernovae type II
TTV	Transit-time-variations
$\mathbf{w_f}$	water mass fraction
$\boldsymbol{Z}$	summed mass percentage of all heavy elements

To my parents, my brothers and sisters, an unwavering and bright presence in my life. To my grandma, who always reminded me to keep my heart in the Stars but my head down to Earth.
To my other grandma, who always dreamed about the stars.

Chapter 1

Introduction

For thousands of years, the human kind has been looking at the sky and wondering if there are other worlds out there and what they may be like. This curiosity led to the development of the Planetary Sciences field, which has been slowly growing along the knowledge acquired about extra-solar planets (exoplanets). However, in the last decades we have witnessed a turning point in exoplanetary science with the development of new and better technology. It was in 1992 that Wolszczan and Frail (1992) discovered the first exoplanet-mass object around a millisecond pulsar using a radiotelescope. Three years later, high-precision radial velocity measurements led to the detection of 51 Peg b, the first exoplanet discovered around a solar-type star by Mayor and Queloz (1995). As of today more than 4800 exoplanets have been discovered and confirmed.[1]

More than big, Jupiter-like planets, there has been an increasingly interest in smaller, rocky planets, as they are the main candidates for harbouring life as we know it.

Exoplanets orbiting main-sequence stars are discovered through six main techniques, most of them being indirect methods, like the radial velocity (RV) method or the transit method, where the focus of observations relies on the host star. When two or more of these techniques are used and combined, we can extract more information about the exoplanets (Santos et al. 2020). Figure 1.1 shows the detected and confirmed 2395 planets with measured masses, separated by detection method. The transit method is responsible for 1027 planets, the RV method for 964 planets, and the microlensing and imaging methods enabled the discovery of 153 and 148 planets, respectively. The Transit-Time-Variations (TTV) and the astrometry techniques were the least successful, with 20 and 16

[1] According to The Extrasolar Planets Encyclopaedia on http://exoplanet.eu (Accessed on 21/08/2021)

planets discovered each. However, the number for exoplanets discovered with astrometry might change once the analysis of Gaia data is carried out (Sahlmann 2016). Finally, 67 exoplanets were discovered through other techniques than those already mentioned. Figure 1.1 also includes Earth and Jupiter mass reference as a comparison to the discovered planets. We note that different methods have different sensitivities for a given population of planets. For example, the imaging method is better to detect more massive planets with orbits farther form their star, but the transits method is better to find planets orbiting closer to their star, regardless of their mass. These techniques will be explained in more detail bellow.

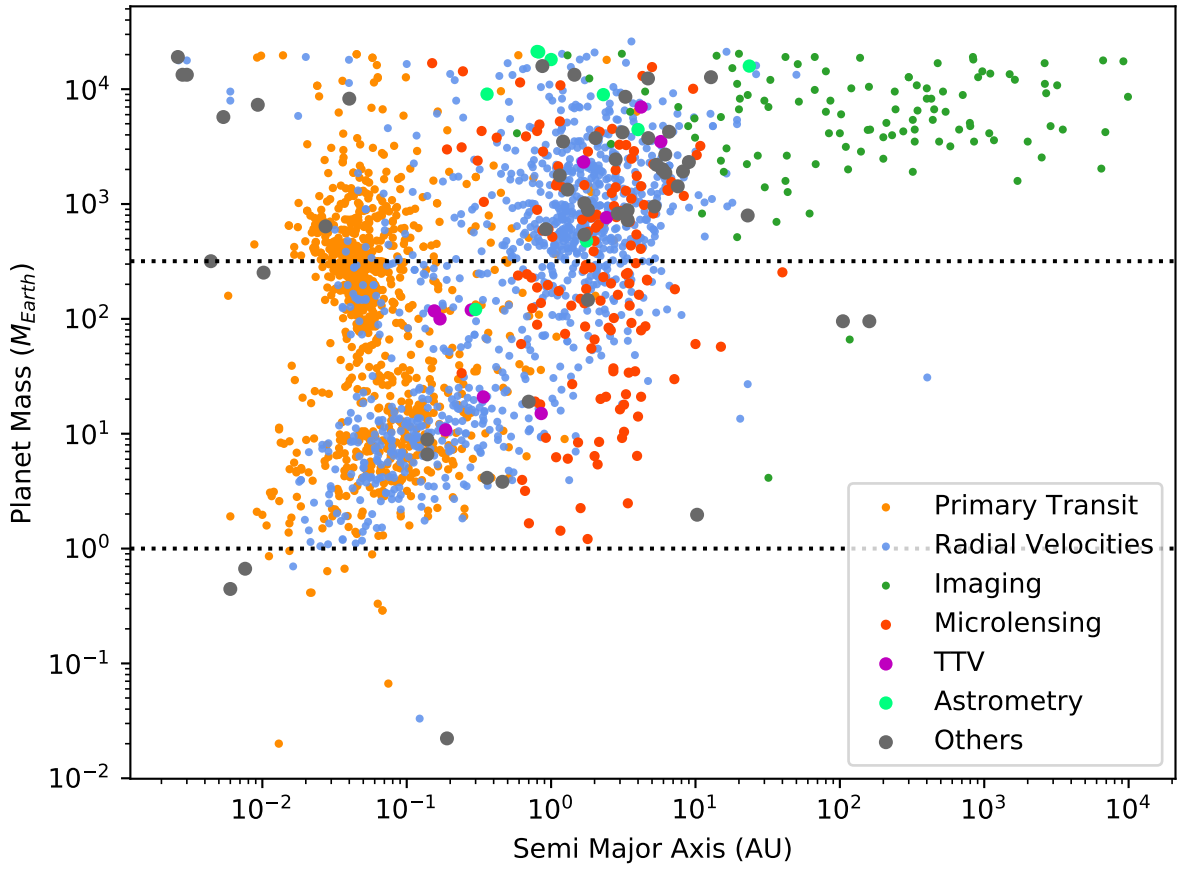


FIGURE 1.1: Confirmed exoplanets with measured masses (in Earth mass units) and known semi-major axis (in astronomical units, AU). The dashed lines represent the value for the Jupiter mass (top line) and the Earth mass (bottom line). Note that the masses discovered solely by the RV method are the minimum masses. Data taken from The Extrasolar Planets Encyclopaedia, found in http://exoplanet.eu and accessed in 02/09/2021.

1.1 Detecting exoplanets

1.1.1 The radial velocity method

In 1952, Struve (1952) proposed the RV method, a technique based on the detection of small variations on the RV of the host star due to the gravitational pull from its planet as it moves through its orbit. When an exoplanet orbits a star, they both orbit their centre of mass, which is closer to the star than to the planet because the star is the more massive body. The orbiting planet causes a Doppler effect on the RV of the star, for the observer. If the star is moving towards the Earth, its light will be blueshifted, while if the star is moving away, its light will be redshifted. This effect enables the deduction of the RV of the star from the displacement of the spectral lines, through the Doppler equation (Rosen and Gothard 2010):

$$\frac{\Delta\lambda}{\lambda} = \frac{\nu}{c} \tag{1.1}$$

where λ is the reference wavelength, $\Delta\lambda$ is the observed wavelength shift, ν is the RV and c is the speed of light.

Equation 1.2 shows how to measure the semi-amplitude of the RV periodic signal, K, created by the exoplanet on the host star (Cumming et al. 1999). This semi-amplitude depends on the ratio between the mass of the star, $M_\star$ and the mass of the planet, M_p.

$$K = \left(\frac{2\pi G}{P}\right)^{1/3} \frac{M_p sin(i)}{(M_\star + M_p)^{2/3}} \frac{1}{\sqrt{1-e^2}} \quad \text{[km/s]} \tag{1.2}$$

G is the gravitational constant, P is the orbital period in days, i is the orbital inclination with respect to the line of sight and e is the orbital eccentricity of the planet.

In general, a planet produces a signal that follows a keplerian function (Hilditch 2001, Wright and Howard 2009, Lissauer and de Pater 2013):

$$RV(t) = \gamma + K[e\cos(\omega) + \cos(\omega + \nu(t))] \tag{1.3}$$

where K is the RV amplitude, e the eccentricity, ω the longitude of periastron, γ the RV of the centre of mass of the binary, and $\nu(t)$ the true anomaly (see Section 2.7 of Hilditch (2001) to a deeper discussion of this topic). By fitting the RV signal through time to a keplerian orbit function, the parameters like the period and eccentricity of the orbit can

be measured. However, there is a degeneracy between the mass of the planet and its orbital inclination, since the method only detects the changes in the motion of the star in the line-of-sight of the observer. As such, it is not possible to determine the true mass of the planet and its orbital inclination, and it is only possible to derive the minimum mass of the planet, $M_p sin(i)$. However, if aided by other detection method, it is possible to solve the degeneracy. One example of this is the use of the transits method when a planet is observed to transit in front of its star, as it makes it possible to get the inclination of the system and therefore, estimate the true mass of the planet.

It is important to note that different phenomena like the stellar activity might mimic planetary signals. The presence of spots on the stellar atmosphere might produce a signal resembling that of a planetary companion, leading to a false positive (see e.g. Queloz et al. 2001).

As can be seen from Equation 1.2, the RV method is more sensitive to short-period, more massive planets, as their gravitational influence will cause larger variations in the star's RV. In order to find Earth-like planets using this method it is necessary a high precision to detect the RV amplitude signal. In recent years, precise RV spectrographs such as HARPS (ESO, Mayor et al. 2003) and HARPS-N (La Palma, Cosentino et al. 2012) were developed and were finally able to achieve long term precision below 1 m/s (see e.g. Pepe et al. 2013a). More recently, it was also developed the ESPRESSO spectrograph (Pepe et al. 2013b, 2021), which can read a precision of about 10 cm/s, a precision required to detect an Earth-like planet orbiting around a solar-like star. ESPRESSO combines the efficiency of a high resolution spectrograph with extreme radial velocity and spectroscopic precision with a larger telescope aperture (than HARPS), since it is mounted at the Very Large Telescope (VLT). The VLT has a large light collecting area allowing to search for planets orbiting around distant or faint stars.

1.1.2 The transit method

Based on the numbers of detected exoplanets, the transits method could be said to be the most successful. Henry et al. (1999) and Charbonneau et al. (2000) were the first to observe a transiting exoplanet, while photometrically monitoring the system HD 209458. The transit method is based on a planet transiting in front of its star, while aligned with the line of sight from the observer. The observed flux from the system will decrease as the exoplanet blocks some of the light from the star (since the planet does not emit as

much light as the star at the observed wavelength). A larger planet will block more light, so it will have a deeper transit. Through the transit method it is possible to measure the ratio between the star radius $R_\star$ and the planet radius R_p from the brightness variation, as shown by Equation 1.4 (Lissauer and de Pater 2013).

$$\frac{\Delta L}{L} \propto \left(\frac{R_p}{R_\star}\right)^2 \tag{1.4}$$

L corresponds to the luminosity of the star for the given observed wavelength, and ΔL its maximum variation.

The probability of detecting exoplanets depends on the geometry of the system, as it has to be aligned with the line of sight from the observer. Therefore, the transit method is most sensitive to exoplanets that orbit closer to their star, as these planets are more likely to be aligned with the observer and its host star.

The CoRoT (Convection, Rotation and planetary Transits) mission (Baglin et al. 2006, 2007) was the first space mission specifically designed to find and study exoplanets. Launched in 2006 by ESA, CoRoT observed 12 000 stars and was pioneer on the discovery and characterisation of rocky planets, smaller than gas giants, a type of planet that had not been observed at the time yet (Léger et al. 2011, Barros et al. 2014). The CoroT mission also focused on detecting and studying stellar vibrations, leading to a development of the asteroseismology field. After 6 years and two missions extensions, the mission ended in June 2014. The *Kepler* telescope (Borucki et al. 2010) launched by NASA in 2009 was a turning point in the discovery of exoplanets through the transits method, having found more than 2600 exoplanets for the duration of both of *Kepler* and *K2* missions. Moreover, researchers are still finding exoplanets in *Kepler*'s data, as more than 150 000 stars were observed (see Vanderburg et al. 2020). The *Kepler* telescope was decommissioned in 2018. In the same year, NASA launched the Transiting Exoplanet Survey Satellite (TESS). TESS's mission is still ongoing and it will observe 200 000 of the brightest stars near the Sun to search for transiting exoplanets (Ricker et al. 2009). ESA plans on initiate the PLATO (PLAnetary Transits and Oscillations of stars) mission (Rauer et al. 2014, 2016) in 2026, whose aim is to detect and characterise terrestrial exoplanets around bright solar-type stars, focusing on planets orbiting in the habitable zone of their star. Being a photometric mission, PLATO will also provide values of planetary radii. Follow up RV and spectroscopic observations will be able to confirm the planetary nature of the targets and determine stellar masses, age and confirm planetary radii values (Catala and Plato Team 2006, Roxburgh et al. 2007).

1.1.3 Other methods

One example of another method used to detect exoplanets is the microlensing method, which detects planets that are located in the observed planetary system and a background distant star. The magnitude of this background star will be magnified as the gravity of the star-planet system will bend the light from the background star, acting as a lens. The amount of increase in magnitude is related to the mass of the exoplanet and its configuration in the orbit around its host star. A higher planetary mass will lead to a higher effect, however it is also possible to detect low-mass exoplanets with this technique (Christie 2006).

Direct imaging method is another example of a technique used to find exoplanets (e.g. Kalas 2011). By carefully blocking the light of a star, it is possible to directly image and characterise the planet around it. However, it can be quite complicated to find the planets, as the residual light from the host star are orders of magnitudes brighter than the exoplanets (Pueyo 2018). On the other hand, this technique can be used to characterise young planetary systems and is sensitive to planets with wide orbits that still emit in the infrared due to the heat of planetary formation (Traub and Oppenheimer 2010). Gauza et al. (2015) discussed how this technique has been particularly successful in finding exoplanets orbiting M dwarfs. The establishment of SPHERE at the VLT allowed the discovery of several exoplanets with this technique (see e.g. Vigan et al. 2016, Bonnefoy et al. 2016, Zurlo et al. 2016).

Another technique not very usual is astrometry, which consists in the measurement of the position of a star on the sky, in other words its coordinates, and its distance. Due to its relative motion, a star's coordinate may change over time. Furthermore, if an exoplanet is orbiting the star, it might cause tiny changes in the position, as its presence will reflect on the star's motion around the centre of mass. The Gaia mission (Gaia Collaboration et al. 2016) is expected to detect many exoplanets with astrometry (see Perryman et al. 2014), including planets around M dwarfs (Sozzetti et al. 2011).

1.2 The star-planet connection

In order to precisely and accurately characterise an exoplanet and their planetary properties such as internal composition or atmosphere, we need to know the characteristics of the host star. From Equations 1.2 and 1.4, it is possible to understand that to obtain values

for the mass and radius of a planet we need to have accurate stellar parameters. Knowing the radius and mass of the planets allows to determine their average density. Although this in itself is not enough to characterise the interior composition of the planet, knowing the density allows to distinguish the planets into three different types, if the precision is sufficiently high. Gaseous planets, sometimes called Jupiter-like planets, have the lowest densities; icy planets are sometimes called Neptune-like, due to their similar density. The last one are the rocky planets, which can be Earth-like planets or super-Earths, depending on their mass (e.g. see Thiabaud et al. 2015a). However, gaseous planets have atmospheres and if only 1% of their mass is a H or He atmosphere, they can have any radius value, in other words, any density.

Recently, Russell (2021) proposed a classification system based on mass-radius and mass-density relationships and composition of planetary bodies. Thus, based upon the mass fraction of H-He gas, planets can be classified as Gas Giant, Rock-Ice Giant and Terrestrial planets. Additionally, the average density can serve as a proxy for the planetary composition, as Dressing et al. 2015 showed and can be see in Figure 1.2. This comes to reinforce the idea that accurate measurements of the host star and its stellar parameters are fundamental to better characterise an orbiting planet.

Moreover, the chemical composition of the molecular cloud that originated the protostar and is reflected on the composition of the stellar atmosphere, is related to the interior chemical composition of a planet (Rubie et al. 2015, Aguichine et al. 2020). Yet, different compositions can lead to the same density, which means there is a degeneracy in the planet interior models (Dorn et al. 2015, 2017). This shall be discussed further on.

In this chapter there will be a more detailed introduction about M dwarfs, their characteristics and why are they one of the most favourable targets when looking for small-sized planets and life outside our Solar System in Section 1.3. In Section 1.4 it is discussed planet interior models while Section 1.5 includes general information about the spectroscopic surveys used in this work. In the end of this chapter, a general overview of this work is presented in Section 1.6

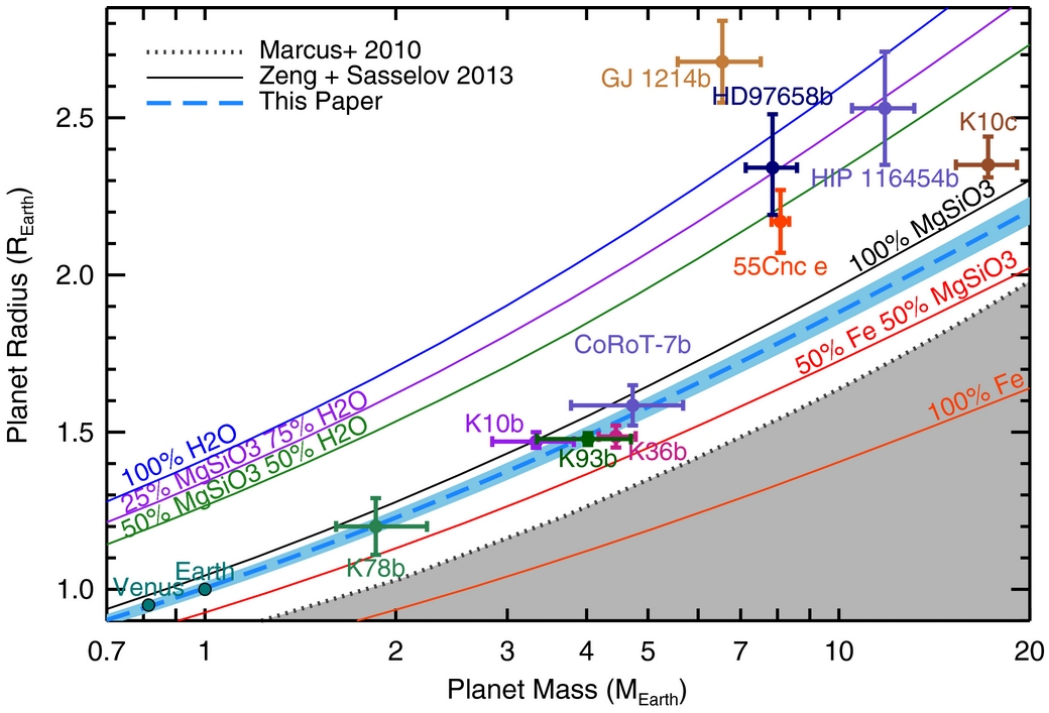


FIGURE 1.2: Mass-radius relationships for low mass rocky planets orbiting solar-type stars, with mass measurements better than 20% precision. The solid lines are theoretical mass–radius curves for planets with the different compositions. The shaded grey region in the lower right indicates planets with iron content exceeding the maximum value predicted from models of collisional stripping. Taken from Dressing et al. (2015).

1.3 M dwarfs

At the end of the 19th century, in the 1880s, Williamina Fleming devised a system to classify stars based on the intensity of hydrogen absorption lines, which causes variation in the wavelengths emanated from stars and results in variation in color appearance. Together with Edward C. Pickering, they created 17 different classes based on the intensity of hydrogen spectral lines, from "A" to "Q". Spectra with the strongest lines were classified as "A" stars, the next as "B," and so on until "O" stars, which have very weak hydrogen lines. "P" spectra was classified as planetary nebulae and "Q" as other types of spectra. The lettering system displayed the gradual decrease in hydrogen absorption in the spectral classes when moving down the alphabet (Oxford-Reference). With this system, they classified over 10,000 stars, published in 1890 in the Draper Catalogue of Stellar Spectra (Pickering 1890).

Some years later, Annie Jump Cannon and Antonia Maury realised that there was a significant overlap between the different types and stars from the same class could have very different colours. Thus they decided to reclassify the stars it in a way that made more

astrophysical sense. Many of the original classes from A to O were dropped, and the order was rearranged in a temperature sequence that has since been used as the primary means of classifying stars, "O B A F G K M", where O stands for the hottest stars and M the coolest ones. This classification method resulted in the Morgan–Keenan system, which was the basis for the Harvard spectral classification (Sobel 2017, Ridpath 2018).

M dwarfs' low temperature is due to being low mass stars, which results in low pressure. This leads to a lower fusion rate of hydrogen to helium, hence the low temperature. A dwarf star refers to a main-sequence star. Therefore, M dwarfs (or red dwarfs, as they are also known) are cool low-mass stars, with masses of $0.08 - 0.55\ M_{\odot}$ and effective temperatures (T_{eff}) of $2500 - 4000$ K (Pecaut and Mamajek 2013). Because they have the lowest mass, M dwarfs are the most abundant stars in the Galaxy and comprise around $75 - 78\%$ of all the stars in the solar neighbourhood (Ledrew 2001, Henry et al. 2006, Winters et al. 2019). Despite this, there is not a single M dwarf that can been seen with the naked eye, as they are very faint.

Like all the others spectral classes, the M spectral type is divided in 10 subclasses, numbered from 0 to 9, and goes from M0 – M9. The spectral types and the subclasses represent a temperature sequence (Payne 1925). The temperature of a star defines their "colour" and surface brightness (Moore 1992).

Because of their cooler atmosphere, M dwarfs' optical and near-infrared spectra has millions of absorption lines due to the presence of diatomic molecular species, mainly oxide molecules (Önehag et al. 2012, Afram and Berdyugina 2019, Kochukhov 2021). In the visible spectra, titanium(II) oxide (TiO) bands are particular relevant and in late M dwarfs vanadium(II) oxide (VO) bands also become more present (Rajpurohit et al. 2018). Figure 1.3 exhibits the spectra (black line) for GJ 180, a M1.0 star (top), Ross 128, a M4.0 star (middle) and HD 180617, a M8.0 star (bottom). It is easy to see how all three spectra are populated by many absorption lines, namely Fe I, V I, Na I and Ti I. It is possible to see that the molecular absorption becomes increasingly more important towards later spectral types, to the extent that almost no atomic line is without molecular blends (Kochukhov 2021). This poses a hindrance when trying to fit a model (red line), as the model cannot recognise the V I lines between the Na I lines.

Usually a star's chemical composition can be known through a spectral analysis since similar to planets, stars also have atmospheres. The photosphere is the atmosphere's lowest and coolest layer and usually the only visible one. The light irradiated by the star

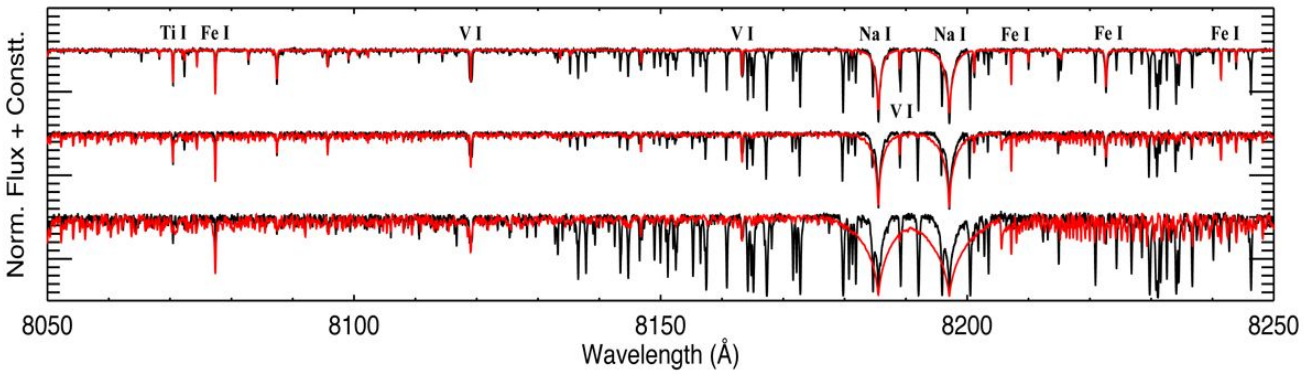


FIGURE 1.3: Spectra of GJ 180 (M1.0, top panel), Ross 128 (M4.0, middle panel) and HD 180617 (M8.0, bottom panel) in black, compared with the best-fit BT-Settl model (red). Figure taken from Figure 1 of Rajpurohit et al. (2018).

stems from this region and corresponds to a continua spectra in the range of wavelengths of the produced radiation. However, when it passes through the higher layers, the elements there present absorb radiation in certain wavelengths, originating an absorption line in the star spectra (see Chapter 7 and Chapter 9 of Rutten 2003). Identifying these absorption lines in a clear way allows us to determine which was the element responsible for it. Therefore, we can deduce the star's composition through its spectra.

However, because of the heavy blending of the spectral lines, it is very challenging to determine their chemical composition. Another drawback is that the presence of so many absorption lines makes it difficult to find the continuum of the spectra (Rajpurohit et al. 2018). Nevertheless, although it is difficult to determine abundances of individual elements in the optical spectra, we still can determine the metallicity (which will be considered as the iron abundance, [Fe/H]) through the strengths of the absorption lines in the spectra (Önehag et al. 2012), using photometric calibrations (Bonfils et al. 2005) as well as spectroscopic calibrations in the K band (Rojas-Ayala et al. 2010), spectrum synthesis (Bean et al. 2006a,b) and studies based on molecular indices (Woolf and Wallerstein 2006).

Since the analysis of M dwarfs optical spectra is quite challenging, few works have attempted to directly determine their composition. Analysing high-resolution H-band spectra from the SDSS–IV APOGEE survey for Kepler-138 and Kepler-186 planet host stars, Souto et al. (2017) determined the chemical abundances of 13 elements (C, O, Na, Mg, Al, Si, K, Ca, Ti, V, Cr, Mn, and Fe). This was done by using a spectrum syntheses which was computed with a line list that was improved by including H_2O and FeH lines. The authors found that both stars are slightly metal-poor and have near-solar C/O ratios.

Kepler-186 exhibits a somewhat enhanced [Si/Fe] ratio and the average of the [α/Fe] relative abundances are consistent with what was expected for a slightly metal-poor star.

Maldonado et al. (2020) determined the stellar abundances of elements other than iron for a sample of 204 M dwarfs from high resolution optical spectra. The data was taken from the HADES radial velocity program sample of M dwarfs, the HARPS GTO M dwarf sample and the California Planet Survey (CPS) for late-K and M-type dwarfs. The authors based their methodology in the use of principal component analysis and sparse Bayesian's methods, trained on a sample of M dwarfs orbiting a FGK companion star with known abundances. Assuming that both companions in a system share the same composition, they can test their method on the M dwarf and validate the results with the FGK companion. Then, they applied their methods on those 204 M dwarfs to determine their stellar metallicity and chemical abundances. They verified that low-mass stars with planetary companions follow the same mass, abundances and metallicity trends than their FGK counterparts.

Besides the aforementioned challenges, a key aspect to accurately characterise these small and low-mass stars consists in studying these stars in the wavelengths where its emission flux reaches the maximum. Therefore, it is important to also analyse the spectra of M dwarfs in the NIR region. Ishikawa et al. (2020) performed a detailed chemical analysis of five M dwarfs in binary systems with G or K-type stars companions. Through a line-by-line analysis on high resolution near-infrared (NIR) spectra, the authors derived the chemical abundances of eight elements (Na, Mg, K, Ca, Ti, Cr, Mn, and Fe), for which the results are concordant with those of the primary star. The authors concluded that one cannot estimate the overall metallicity or any chemical ratio (e.g. [Me/Fe]) without consistently determining the true abundance of individual elements. As such, while the knowledge of the NIR region is not as good as the optical band, it can still help to provide more information about the stellar parameters (see also Lindgren et al. 2016).

1.3.1 M dwarfs as host of small rocky planets

Thanks to the continuous technological advances, different exoplanet detection techniques have been able to find an increasingly number of small, Earth-size planets, and the regime of Earth-mass exoplanets was finally reached, as it can be seen in Figure 1.4.

Currently, M dwarfs are the main targets when searching for Earth-sized or Earth-mass planets with the potential to harbour life outside our Solar System. One of its main

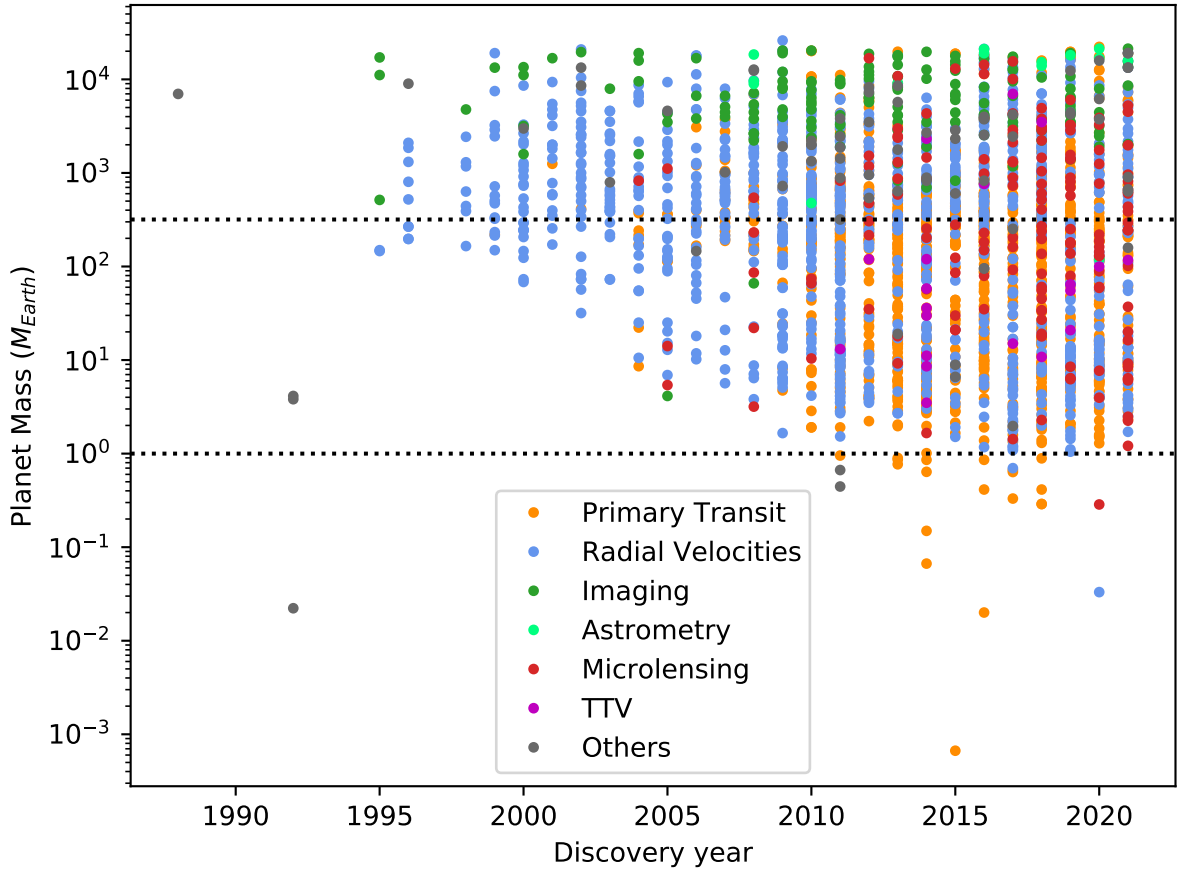


FIGURE 1.4: Confirmed exoplanets with measured masses (in Earth mass units), per discovery year. The dashed lines represent the value for the Jupiter mass (top line) and the Earth mass (bottom line). Note that the masses discovered solely by the RV method are the minimum masses. Data taken from The Extrasolar Planets Encyclopaedia, found in `http://exoplanet.eu` and accessed in 02/09/2021.

distinct feature is its lifetime in the main sequence, which is in the order of $\sim 10^{12}$ years and greatly exceeds that of the Sun, an expected $\sim 10 \times 10^9$ years (Adams and Laughlin 1997). Their lifespan surpasses even the age of the universe, which has an actual age of 13.8×10^9 years (Planck Collaboration et al. 2020). As such, M dwarfs have not yet had time to leave the main sequence since the beginning of the Universe. This long lifetime grants the star some stability which might be beneficial to the development of life in a planetary companion (see Shields et al. 2016, Lingam and Loeb 2017 but also Tarter et al. 2007 and Ballard 2019).

Looking at Equations 1.2 and 1.4, it is evident why M dwarfs are the easy candidates to detect and study small and low-mass Earth-like planets. M dwarfs have lower stellar mass and so the gravitational pull from the planet would result in a increased semi-amplitude value, which makes it easier to detect lower mass planets around these stars with the RV

method. Similarly, when using the transits method, most rocky planets are found around M dwarfs because M dwarfs have small radii. As such, planet-to-star radius ratio will be higher which leads to a higher ratio of the star's surface blocked by the planet, implying a deeper transit and a increased observed signal for smaller host stars.

Another reason is related to the proximity of their habitable zone (HZ), which is a zone where Earth-like planets could retain liquid water on their surface making the environment hospitable to life (Huang 1959, Hart 1978, Kopparapu 2013, Shields et al. 2016). Its boundaries depend on the stellar parameters and the planetary atmosphere. Planets with denser atmospheres would need to be at a larger distance of its host star to harbour liquid water when compared with a planet with a thinner atmosphere (Kopparapu et al. 2014, Kopparapu 2018). Due to their small mass, small size, lower temperature and lower luminosity compared to hotter Sun-like stars, their HZ is located much closer to the star, so planets to be found in this zone would have a shorter orbital period and higher RV amplitudes (Kopparapu 2013). Not only ground-based RV searches but also surveys of transiting exoplanets from space confirm the existence of a substantial population of small planets orbiting M dwarfs (e.g. Affer et al. 2016, Gillon et al. 2016, 2017). The closest Earth-size planets with potential habitability are all in M dwarfs planetary systems (e.g. Quintana et al. 2014, Affer et al. 2016, Anglada-Escudé et al. 2016 Lingam and Loeb 2017, Luque et al. 2019, Damasso 2019).

Looking for Earth-like planets around M dwarfs, besides its advantages also has disadvantages. In particular, the magnetic activity of these stars (Reiners 2012, Feiden and Chaboyer 2013) can have a major impact on planetary atmospheres and compromise the habitability of planets orbiting M dwarfs (see e.g. Mullan and MacDonald 2001, Chabrier et al. 2007, Feiden and Chaboyer 2014, MacDonald and Mullan 2014, Kochukhov 2021). These stars are known to have a large percentage of their surface covered with spots and numerous flares (Schmidt et al. 2014, Yamashiki et al. 2019) that can jeopardise the possibility of life (Yamashiki et al. 2020). The fraction of active M dwarfs rapidly increases towards later spectral types (Reiners 2012, Kochukhov 2021).

Flares and coronal mass ejections (Moschou et al. 2019) lead to short wavelength radiation and recurrent energetic particles emission, and these events are thought to have a significant impact on the atmospheres of potentially habitable rocky planets (Lammer et al. 2007, Khodachenko et al. 2007). Moreover, stars with strong magnetic fields may deposit enough energy in the orbiting planet via induction heating, which results in a

melted interior of rocky planets with closer orbits, increased volcanic activity or in extreme cases a mantle permanently molten (see Kislyakova et al. 2018 for a more general discussion and an example on Kislyakova et al. 2017).

As it was mentioned before, planets in M dwarfs HZ orbit closer to the star, which makes them more vulnerable to X-ray and UV radiation. In extreme cases, the stellar X-ray radiation can evaporate the volatile elements of an orbiting planet (Penz et al. 2008). In addition, planets in closer orbits are more susceptible to be tidally locked, leading to an extreme condition in the opposite sides of the planets: the side facing the star would have very high temperatures and the side facing away would have freezing temperatures.

Nevertheless, there have already been discovered planetary systems with a M dwarf as the host star and with planets with potential for life, as is the case of TRAPPIST-1 (Grimm et al. 2018) or Proxima Centauri (Damasso 2019). As Adibekyan (2017) discusses, knowing the stellar metallicity and chemical abundances of these type of stars can provide a better understanding on the formation and evolution of exoplanets, when one is looking for potentially habitable planets.

1.4 Planet interior models

In order to better understand planetary formation, evolution and internal structure it is necessary to have knowledge about the composition of planets (Adibekyan et al. 2012a, Dorn et al. 2015, Santos et al. 2017). Atmospheres of planets can potentially be directly studied through spectroscopic data from observations of transiting planets. So far, no atmosphere of an Earth-like planet has been detected yet, but the results suggest they are expected to be different (Madhusudhan 2012). The determination of the internal composition of planets is even less direct.

With the increasing number of discovered exoplanets and the search for habitable worlds that could harbour life, there has been a growing interest in planet interior models, as a planet's composition is directly connected with its habitability (Bond et al. 2010a, Thiabaud et al. 2014). Different detection methods provide information on a planet's mass and radius, which allows the determination of its average density. However, just the average density is not enough to correctly constrain interior structure models (Dorn et al. 2015), as different compositions can lead to the same density, which means there is a degeneracy in the models (Rogers and Seager 2010). One assumption usually behind the

calculation has to do with terrestrial planets composition. From solar system observations and planet formation and differentiation models, it was possible to recognise that some interior compositions are more likely than others (Unterborn et al. 2016, Dorn et al. 2017). Thus, rocky planets are generally thought to be differentiated in an iron-rich core, a silicate mantle (where all Mg and Si are assumed to be) and a crust (Dorn et al. 2015). In order to estimate the rocky composition of exoplanets, these models take mainly one of two approaches. In the first one it is assumed that the solid components formed in equilibrium from the material of the solar nebula in order to link the composition and formation process of planets, so that equilibrium chemistry can be used as a constraint for modelling (Elser et al. 2012, Thiabaud et al. 2014). The second one assumes that the planetary elemental ratios Fe/Si and Mg/Si are essentially identical to stellar values (see Thiabaud et al. 2015b), due to the fact that Fe, Si and Mg condense at similar temperatures (Lodders 2003). Dorn et al. (2015) showed that these ratios can be used to model planets interior structure, as they yield tighter constraints on the internal structure and help break the models degeneracy by coupling structure and composition of core and mantle through bulk abundance ratios. The authors also reported how the Fe/Si and Mg/Si stellar ratios in specific are vital in constraining the mantle composition and internal structure of rocky planets (see also Dorn et al. 2017).

Until recently, planet composition models were only considering volatile species, constraints from observations of exoplanet atmospheres and models of ice formation, mainly to evaluate the possibility of presence of liquid water on planetary surfaces (Johnson et al. 2012). Although refractory species are of great importance in the formation of rocky planets, very few (e.g. Elser et al. 2012, Johnson et al. 2012) have included them in the model calculations and how they would affect the planetary formation process (see e.g. Bitsch et al. 2013, 2015a,b, Bitsch and Savvidou 2021). Johnson et al. (2012) tried to combine dynamic and chemical models, based on condensation sequences under equilibrium conditions. There was some success with the element ratios of C/O and Mg/Si, as they reveal a correlation between the composition of the host star and that of the planets (Dorn et al. 2015). With a sample of 451 FGK stars, Delgado Mena et al. (2010) were particularly successful as they showed the C/O and Mg/Si ratios are proxies for the distribution of carbides and silicates in planetary systems (Bond et al. 2010a,b). The authors also conclude that the inherent degeneracy in determining interior structure from mass and radius measurements depends not only on measurement accuracy, but also on the size and density

of the exoplanet. Therefore, precise observations of stellar elemental abundances are crucial to constraint planetary bulk composition and to reduce model degeneracy (see Dorn et al. 2015, Unterborn et al. 2016 for more details and Santos et al. 2015 as an example). Interestingly, this long-held assumption that the compositions of telluric exoplanets and their host stars are identical has very recently been challenged (Adibekyan et al. 2021) on a sample of most precisely characterised planets with uncertainties in mass and radius better than 30%.

1.5 Spectroscopic surveys

As mentioned in Section 1.3, the stellar parameters like the effective temperature, metallicity or surface gravity ($\log g$) can be determined by performing a spectroscopic analysis of the stars. A good precision and accuracy are fundamental for a better characterisation of a host star and its planets (Adibekyan et al. 2018). Moreover, a careful assessment of the stellar properties are key in order to properly understand planetary formation and evolution (Santos et al. 2000, 2004, Adibekyan 2019).

During the last decade, several large surveys have been launched to discover and characterise exoplanets. Some examples of ground-based surveys are the CORALIE survey (ESO - La Silla Observatory; see e.g. Naef et al. 2001, Rickman et al. 2019), the SpHere INfrared Exoplanet (SHINE) survey, part of the SPHERE GTO Program on the ESO VLT (Meyer 2019, Desidera et al. 2021), the already mentioned HARPS and HARPS-N (Pepe et al. 2000, Rupprecht et al. 2004, Cosentino et al. 2012), and ESPRESSO GTO (Pepe et al. 2013b, 2021), mounted at the VLT. For space-based surveys, we have the *Kepler* survey (Borucki et al. 2010, Vanderburg et al. 2020) and its posterior mission K2 (Howell et al. 2014) from ESA and the the ongoing TESS from NASA (Ricker et al. 2009), just to mention a few.

In parallel to the aforementioned exoplanet discovery surveys, there have been several large spectroscopic surveys launched that aimed mostly at precise chemical characterisation of stars in the solar neighbourhood, and beyond. Some examples include APOGEE (Majewski et al. 2017), GALAH Rupprecht et al. 2004, GAIA–ESO (Randich et al. 2013) and RAVE (Steinmetz et al. 2006). These spectroscopic data are of great importance for exoplanet host star characterisation as well. In the current work we made use of data from three different catalogues, which will be explained in the next subsections: APOGEE, GALAH and HARPS.

1.5.1 APOGEE

The Apache Point Observatory Galactic Evolution Experiment 1 (APOGEE-1, Deshpande et al. 2013, Majewski et al. 2017) is one of the programs in the Sloan Digital Sky Survey III (SDSS–III), with results first released in July 2013. It is a large-scale, stellar spectroscopic survey conducted in the near-infrared (H-band) portion of the electromagnetic spectrum. The APOGEE-1 consists of two 300-fiber cryogenic spectrographs that operate on board the 2.5 meter Sloan Foundation Telescope and the 1 meter NMSU Telescope at Apache Point Observatory (APO) in New Mexico, USA and the 2.5 meter Irénée du Pont Telescope of Las Campanas Observatory (LCO). This survey is unique among large spectroscopic surveys because it is able to acquire medium-resolution ($R \sim 22500$), high signal-to-noise ratio (> 100 per pixel), infrared (1.51–1.70 μm) spectra of 146 000 stars across the entire Milky Way (Majewski et al. 2017, Wilson et al. 2019). Beside spectra, the APOGEE-1 survey also provides data for RV, stellar parameters (including effective temperature, surface gravity and metallicity) derived using the APOGEE Stellar Parameters and Chemical Abundance Pipeline (ASPCAP) as well as chemical abundances of 17 elements (C, N, O, Fe, Na, Mg, Al, Si, S, K, Ca, Ti, V, Mn, Ni, Cr and Co) also determined with the Chemical Abundances Pipeline (Shetrone et al. 2015, Holtzman et al. 2015, García Pérez et al. 2016). With RV precision around 0.07 km/s and detailed chemical abundances combined with the Gaia astrometric data, the APOGEE survey provides valuable insights into the dynamical structure and chemical history of the Galaxy (Majewski et al. 2017).

In 2019 it was released new data from the Southern Hemisphere survey of the Apache Point Observatory Galactic Evolution Experiment 2 (APOGEE-2, here on referred to simple as APOGEE, Majewski et al. 2016), the sixteenth data release (DR16) from SDSS–IV, the fourth and penultimate phase (Majewski et al. *in prep.*). The new survey made bright-time observations from 2014 to 2020. All of the data have been re-reduced and re-analyzed using the latest pipelines, resulting in a total of 473,307 spectra of 437,445 stars. Abundances of the neutron-capture element Ce are included for the first time, adding to the abundance determination of a total of 26 species − C, C I, N, O, Na, Mg, Al, Si, P, S, K, Ca, Ti, Ti II, V, Cr, Mn, Fe, Co, Ni, Cu, Ge, Rb, Ce, Nd, and Yb (Jönsson et al. 2020, Ahumada et al. 2020). The high-resolution, signal-to-noise ratio and wavelength bands remain the same as in previous releases, but the RV have an improved precision of ~ 200 m/s.

1.5.2 GALAH

The GALactic Archaeology with HERMES (GALAH, De Silva et al. 2015) survey is a Large Observing Program using the HERMES instrument with the Anglo-Australian Telescope, to observe a million stars in our Galaxy. The survey examines the chemical changes of the Galaxy over time and, using specific observational techniques, studies the different chemical elements in stars of different ages (De Silva et al. 2015).

The High Efficiency and Resolution Multi Element Spectrograph, HERMES, is a four channel fibre-fed spectrograph with high resolution and multi-object capability. It provides a spectral resolution of $R \sim 28\,000$ in standard mode for 392 stars simultaneously over a 2 degree field of view. If using a slit mask, the spectral resolution can increase to $R \sim 50\,000$. HERMES provides simultaneous observations in four fixed bands: blue (471.5 − 490.0 nm), green (564.9 − 587.3 nm), red (647.8 − 673.7 nm) and NIR (758.5 − 788.7 nm). These bandwidths have been optimised to maximise the number of chemical element lines captured. HERMES spectra contain absorption lines from 29 chemical elements, including light proton-capture elements, $\alpha-$elements, odd-Z elements, iron-peak elements and n-capture elements from the light and heavy s-process and the r-process − Li, C, O, Mg, Si, Ca, Ti, Na, Al, K, Sc, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Rb, Sr, Y, Zr, Ba, La, Ru, Ce, Nd and Eu. For each star, GALAH also provides RV accurate to within 0.5 km/s (De Silva et al. 2015, Sheinis et al. 2015).

From November 2013 to February 2019, new observations were made as part of the GALAH survey, leading to the third data release (DR3). This data release provides 678 423 spectra for 588 571 mostly nearby stars, of which 81.2% stars are within < 2 kpc distance from the Sun. DR3 consists of 65% dwarfs, 34% giants and the remaining 1% are unclassified stars (Buder et al. 2021).

1.5.3 HARPS

The High Accuracy Radial velocity Planet Searcher (HARPS, Mayor et al. 2003) at the ESO La Silla 3.6 meter telescope is dedicated to the discovery of extra-solar planets, with accurate measurements of the RVs of the stars.

HARPS is a fibre-fed high resolution ($R \sim 115\,000$) echelle stable spectrograph designed to obtain high accuracy RV measurements. For solar-type stars on solar vicinity, a signal-to-noise ratio of 110 enables to reach an RV accuracy of about 1 m/s, enough to detect Saturn-like and Neptune-like planets orbiting these stars. Under the right conditions,

it might also be able to detect super-Earths with closer orbits (less than 0.1 AU of its host star, see e.g. Suárez Mascareño et al. 2017). With a spectral range between $378 - 691$ nm, the HARPS spectrograph covers nearly the entire visible spectrum and a little of near-ultraviolet. With these characteristics, HARPS provides one of the most precise Doppler-measurements in the world, being second only to ESPRESSO. The HARPS spectroscope is mainly used for the search for exoplanets and in the field of asteroseismology (Pepe et al. 2000, Mayor et al. 2003, Rupprecht et al. 2004).

HARPS high-resolution spectra and precise measurements allows for a more precise characterisation of the observed stars and determination of their stellar parameters and chemical abundances (e.g. Delgado Mena et al. 2017, 2018, 2019, Sousa et al. 2021). Additionally, since many of the HARPS targets are solar analogues (Adibekyan et al. 2012a), it proves to be an ideal sample to use as M dwarfs' abundance calibrators (see Section 2.4).

1.6 This work

This thesis will be focused on determining the composition of low-mass planets and their host stars. This task starts by training a machine learning model on a significantly large sample of stars from large spectroscopic surveys with precise and accurate data (APOGEE, GALAH and HARPS), in order to learn how to determine the stellar composition just from the metallicity and total galactic velocity. This process shall be explained in Chapter 2, starting with a dedicated section about machine learning and the process used. The next section follows by explaining how the best subsamples for each data set (APOGEE, GALAH and HARPS) were constrained from the original sample. The last sections of this chapter explain how the machine learning model is trained on the smaller samples to predict the chemical abundances of stars, based on the metallicity and total galactic velocity parameters, and how this model is validated with external data. With the necessary models trained to determine the chemical composition, Chapter 3 proceeds to describe how through a stoichiometric model it is possible to obtain the planetary building blocks composition, namely the iron-to-silicate mass fraction, when information about certain chemical abundances are known. In Chapter 4 the aforementioned model is applied to a sample of planetary systems of low-mass planets with the most precise mass and radius measurements. Lastly, in Chapter 5 the conclusions will be presented and future work based on the results will be discussed. The supplementary data and information,

supporting the main results and conclusions are presented in the appendixes, followed by the bibliography for this work.

Chapter 2

Predicting stellar abundances with Machine Learning

In the last few decades, the astronomy field has had a major breakthrough. In a matter of a few decades, the humankind has manage to understand more of the universe when compared to the last few centuries. This was mainly possible because of the technology development and the start of large scale missions, which led to a massive amount of new data incoming. Each new mission or each new survey brought with them huge quantities of data, leading the amount of information that needed to be analysed to increase exponentially. However, such amounts of data were impossible to analyse in good time, even with computers. This is where Machine Learning (ML) came to the rescue.

Samuel (1959) defined ML as the field of study that gives computers the ability to learn without being explicitly programmed. It includes many different algorithms that could make a digital computer go through a process that could be compared to learning. ML is particularly useful for problems that are either too complex for the traditional approaches and require a lot of fine-tuning or long lists of constrains, or that have no known algorithm. Furthermore, it can also help to find correlations or new trends in a sample of data, leading to a better understanding of the problem. In particular, applying ML techniques in astronomy is getting increasingly popular, providing insights that were not possible before (Howard 2017, Baron 2019).

2.1 Machine Learning and Random Forests

Machine learning systems can be classified in different types, based on their learning method. Whether or not they can learn progressively on the spot can classify them on online learning and batch learning. They can also be divided if they learn by comparing new data entries to known data entries (instance-based learning), or by detecting patterns in the training data and building a predictive model (model-based learning). Lastly, ML systems can be classified depending on whether they trained with human supervision (Howard 2017). There are four major categories: supervised learning, unsupervised learning, semisupervised learning, and reinforcement learning. The algorithm used in this work falls into the first classification type, so we will briefly explain these four categories (for a more detailed explanation, please refer to Chapter 1 and Chapter 7 of Géron 2019).

Supervised learning

In supervised learning, the training set given to the algorithm includes the set of *features*, also called *predictors*, which are the starting point for the model, and a variable that describes the samples used to train the ML model. The training set includes the desired solutions too, which are called *labels* (Chong et al. 2021). Two of the most common supervised learning tasks are classification and regression. The first one, given a training set, learns how to classify and separate the data into different categories. The second one aims to predict a target numeric value, given a set of features with many examples, including both features and labels (Baron 2019). While classification could be said to be more of a qualitative task, regression is certainly more of a quantitative task (Géron 2019).

Some of the most important supervised learning algorithms include k-Nearest Neighbors (kNN), Linear Regression, Logistic Regression, Support Vector Machines (SVMs), Decision Trees and Random Forests (RF) and some Neural networks (e.g. Baron 2019, Auddy and Lin 2020). In this work we will be using Random Forests to apply a ML model to the data, a process that shall be explained in Section 2.4.

Unsupervised learning

Contrary to supervised learning, in unsupervised learning the given training data is unlabelled, which means there is no distinction between the *predictors* set and the solution set,

and the system tries to learn without a guide (Baron 2019). These can be very useful for a very complex and unlabelled data as they can help identify unsuspected patterns in the data. K-Means, Hierarchical Cluster Analysis (HCA), and Principal Component Analysis (PCA) are some of the most used unsupervised learning algorithms (e.g. see Webb et al. 2020).

Semisupervised learning

As the name suggests, semisupervised learning consists on a mix between supervised and unsupervised learning, as the training data set is only partially labelled. For example, a system may consist of many unsupervised algorithms which are trained sequentially in an unsupervised manner but then are fine-tuned using supervised learning techniques (see discussion in Mey and Loog 2019).

On example of this are the *deep belief networks* (DBNs). Based on unsupervised components called *restricted Boltzmann machines* (RBMs) stacked on top of one another, it lets the RBMs be trained unsupervised, and in the end whole system is adjusted using supervised learning algorithms (e.g. Yang et al. 2021).

Reinforcement learning

Reinforcement learning is a very different category of the previous three. In this context, the *agent*, which corresponds to the learning model, observes the environment and perform actions, getting a reward or penalty in return. Figure 2.1 illustrates the way reinforcement learning works. Based on this system, the *agent* must then understand by itself what is the best strategy, called a *policy*, to get the most rewards over time. This *policy* will define what action the *agent* should choose in a given scenario. Reinforcement learning algorithms are very frequently implemented by robots to learn things (Géron 2019).

Random Forest

As mentioned before, the ML algorithm used in this work is a supervised learning algorithm, more specifically a Random Forest (RF). This technique uses a group of predictors (*ensemble*) and an algorithm that applies this technique is called an *ensemble method*. Usually, an ensemble method has better performance than its constituent predictors, especially if there are constituent predictors with worse precision, since the final predicted

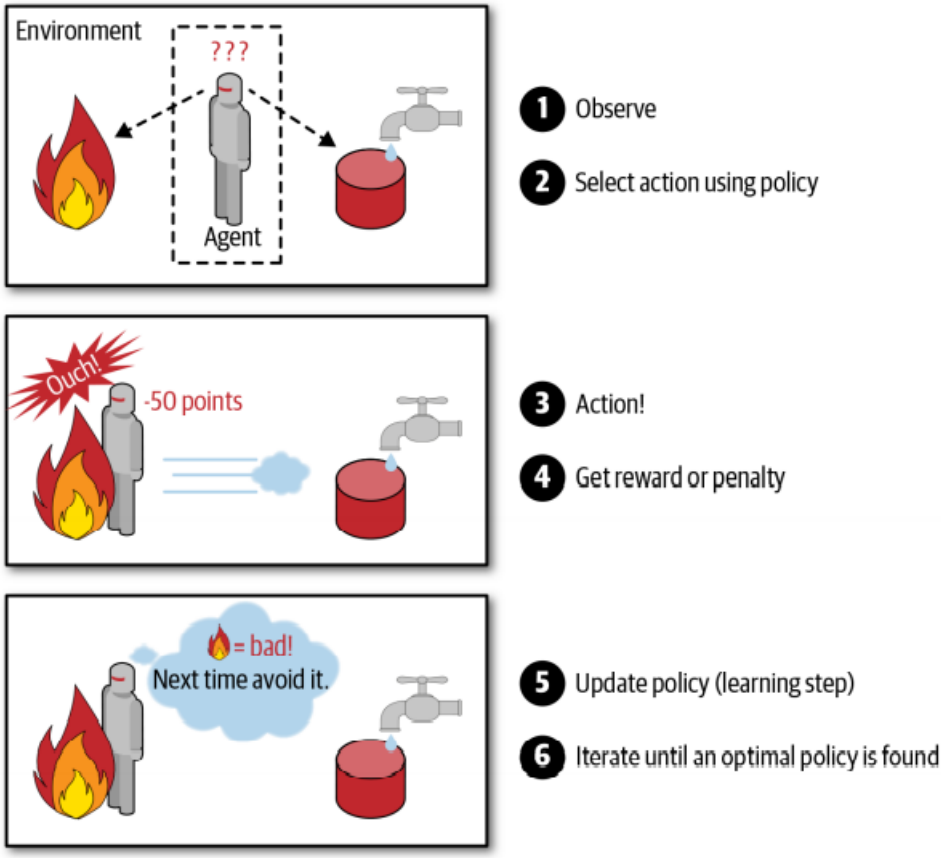


FIGURE 2.1: Sketch illustrating the general working of reinforcement learning algorithms. Image taken from Géron (2019).

value is the average of all individual predictions (Dubath et al. 2011). This way, an extreme prediction would be smoothed out by the average. In this sense, a RF is an ensemble method with more than one predictor and uses the combination of the predictions of the estimators to improve its performance. These estimators are called decision trees (Géron 2019).

A decision tree is an algorithm that classifies an object as a label by verifying several conditions, and is composed of *nodes* (cells where a condition has to be checked) and *leaves* (cells with the predicted value and no conditions to check). A branch begins with the first node of the tree and continues to one of its leaves, where they present their prediction (Ulmer-Moll et al. 2019). An ensemble of decision trees constitutes a random forest. The decision trees have a different number of leaves. Parameters including the number of trees and leaves in a random forest are called hyperparameters. So, in other words, a RF is an ensemble of decision trees diversified by being trained on random subsets of the features samples (Géron 2019). Because of this and despite its simplicity, a RF is one of

the most powerful ML algorithms available (Dubath et al. 2011, Ulmer-Moll et al. 2019, Melton 2020).

Another important and quite useful tool of RF is the fact that it allows to measure the relative *importance* of each feature, by looking at how much on average the impurity of a feature is reduced by tree nodes that use that feature, across all trees within the forest. In other words, the *importance* amounts to the predictive strength of each feature and roughly tells us what would be the decrease in the overall prediction if a feature were to be excluded (Richards et al. 2011). This is a quantitative attribute, as it is the weighted average, with each node's weight corresponding to the number of training samples associated to them. Therefore, RF is highly valuable when dealing with large data sets, as it is also able to provide a quick understanding of what features contain more information needed to better predict the desired parameters.

2.2 The training data

In order to study the internal composition of terrestrial planets, we need to know the stellar composition with high precision. In this regard, abundances of the main rock forming elements (e.g. C, O, Mg, Si, and Fe) are of particular importance.

As Adibekyan (2019) explains, when talking about the abundance of the X element, we are referring to the atomic abundances of stars relative to Sun using the logarithmic abundance (number of atoms) of H, traditionally set to $\log \epsilon(H) \equiv 12$. In the specific case, when the abundance of the element X is provided relative to the abundance of the sun, square brackets are used:

$$[X/H] = \log \left(\frac{N_X}{N_H}\right)_{star} - \log \left(\frac{N_X}{N_H}\right)_{sun} \tag{2.1}$$

with N_X as the number of atoms of element X.

Abundance ratio of two elements, let's say element X and iron (Fe), can be expressed as:

$$[X/Fe] = \log \left(\frac{N_X}{N_{Fe}}\right)_{star} - \log \left(\frac{N_X}{N_{Fe}}\right)_{sun} \tag{2.2}$$

For a metallicity of [Fe/H] (sometimes iron abundance is used as a proxy for overall metallicity) and the abundance of an element ([X/H]), the abundance ratio can be obtained through the following:

$$
\begin{aligned}
&[X/H] - [Fe/H] = \\
&= \log\left(\frac{N_X}{N_H}\right)_{star} - \log\left(\frac{N_X}{N_H}\right)_{sun} - \log\left(\frac{N_{Fe}}{N_H}\right)_{star} + \log\left(\frac{N_{Fe}}{N_H}\right)_{sun} \\
&= \log\left(\frac{N_X}{N_{Fe}}\right)_{star} - \log\left(\frac{N_X}{N_{Fe}}\right)_{sun} \\
&= [X/Fe]
\end{aligned}
\tag{2.3}
$$

Therefore, knowing the metallicity and the abundance of an element, it is always possible to get their abundance ratio respecrive to iron.

In this work, we use data from three different large spectroscopic surveys: APOGEE, GALAH, and HARPS. These data sets are described below.

2.2.1 APOGEE

From the APOGEE website [1], we download the DR16 through the `allStar-r12-l33.fits` file from the Science Archive Server (SAS), which houses parameter summary catalogues, spectra, and all intermediate data products in flat files within a directory structure. This database contains information for 473 307 stars.

We start by excluding stars for which the derived quantities are unreliable, be it from a data point of view, or from the determined stellar parameters point of view (stars with the STAR_BAD flag in the ASPCAPFLAGS column), leaving 428 213 stars. Next, we exclude stars with unpopulated chemical abundances (flagged by "-9999,99") of [X/Fe], where X is the chemical symbol of the element we want to consider, in our case C, O, Mg and Si. For that we set the condition [X/Fe]$\neq$ -9999,99 for all 5 elements mentioned and for metallicity ([Fe/H]$\neq$ -9999,99).

The APOGEE data also has a flag system to warn the user of any problem the data might have or if any special care is needed for a certain column, parameter or subset of data. This flag system is based on bitmasks [2] and each warning corresponds to a number. If a certain star has one or more warnings, the flag columns will have the sum of powers of two of these warnings. For example, if a certain star has warnings number 1 and number 2, on the flag columns we will have the number $6 = 2^1 + 2^2$. If the star has no warning, then the flag columns will have the number 0. As such, we make use of this system to

[1] https://www.sdss.org/dr16/irspec/spectro_data/
[2] https://www.sdss.org/dr16/algorithms/bitmasks/

also exclude stars that might have any warning associated to the five elements we will be using (X_FLAG==0) and consider only the main survey sample, without data from extra missions (EXTRATARG==0), leaving us with 210 237 stars.

In order to have the best data possible, we selected stars with a high signal-to-noise ratio (S/N) spectra. In Figure 2.2 we plot the abundance ratios of [C/Fe], [O/Fe], [Mg/Fe] and [Si/Fe] and their respective uncertainties as a function of S/N. Our goal was to identify the minimum S/N above which the dispersion of stellar abundances is not large. We look at the range of values $100 < S/N < 500$. A similar exercise was performed for the [Mg/Si] and [C/O] radios, as they are important to constrain the internal structure of terrestrial planets. To ensure there is no dependency of the chemical abundances on the S/N, we do a linear fit for all six abundance ratios, as shown in Figure 2.2.

By looking at the slope of the linear fit, we can confirm by its null value there is no dependency of the abundance ratios with the S/N. For all chemical abundance ratios and their uncertainties, the dispersion is higher around $S/N \sim 100$, but it decreases with increasing S/N and stabilises to a smaller value around $S/N \sim 300$. Therefore, data with $S/N \geq 300$ should have less associated errors. To make sure this restriction on the S/N was not excluding any spectral type, we plot all the data in the spectroscopic Hertzsprung-Russel (HR) diagram, i.e. $\log g$ depending on T_{eff}, and then do the same but for constrains on the S/N value, shown in Figure 2.3.

Constraining the sample to a higher S/N (e.g. S/N $>$ 500, or S/N $>$ 1000), would mean a smaller sample size, which would make difficult to build a prediction model, as it requires a large enough sample. With the constrains on S/N $>$ 300, we get a sample of 41 690 stars.

Looking at the top left figure of Figure 2.3, there seems to be two different groups of stars and a clear gap around $\log g \sim 4$. To make sure that the gap is of astrophysical origin and there is not any issue with the $\log g$ or any unknown associated uncertainty, in Figure 2.4 we plot the $\log g$ errors depending on the T_{eff}.

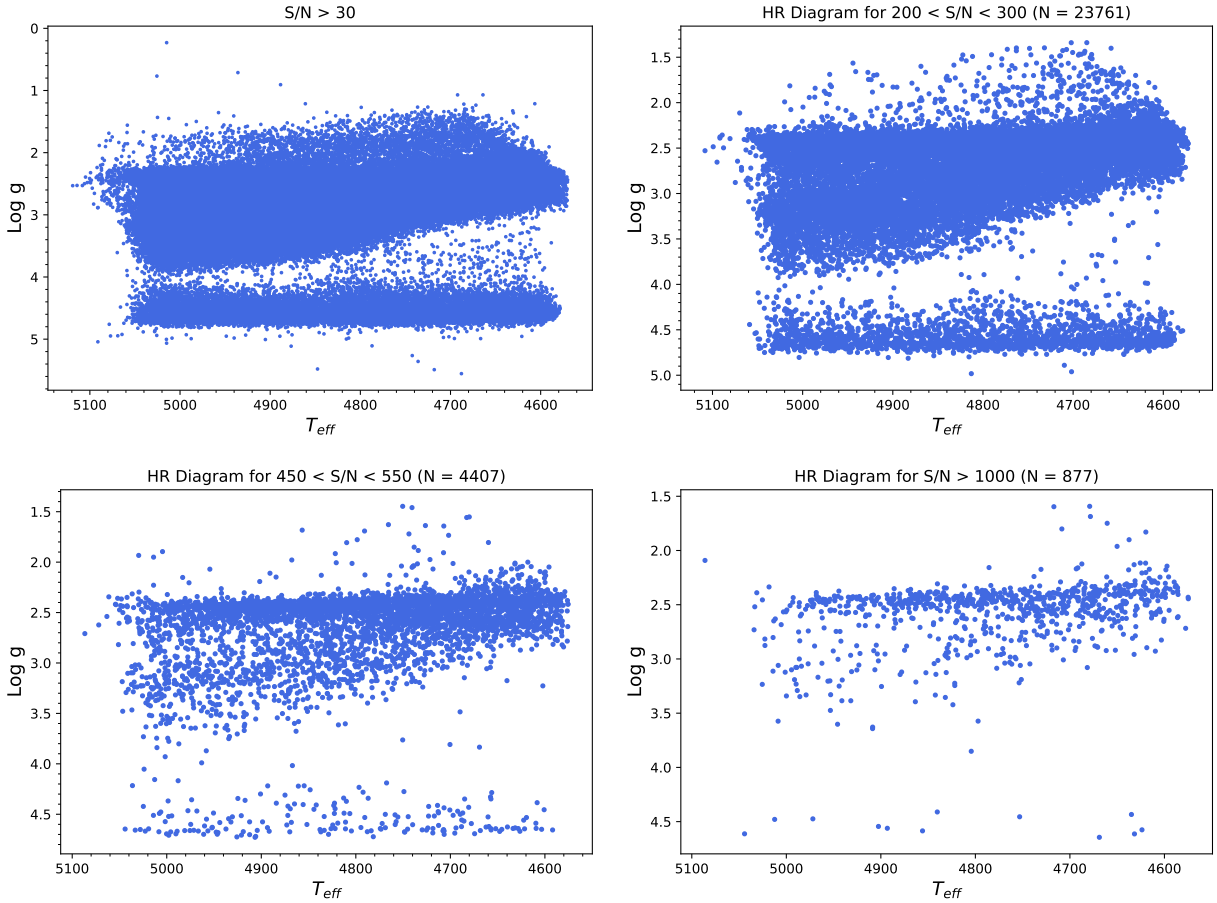


FIGURE 2.3: HR diagram for different values of S/N. The N on the title represents the sample size.

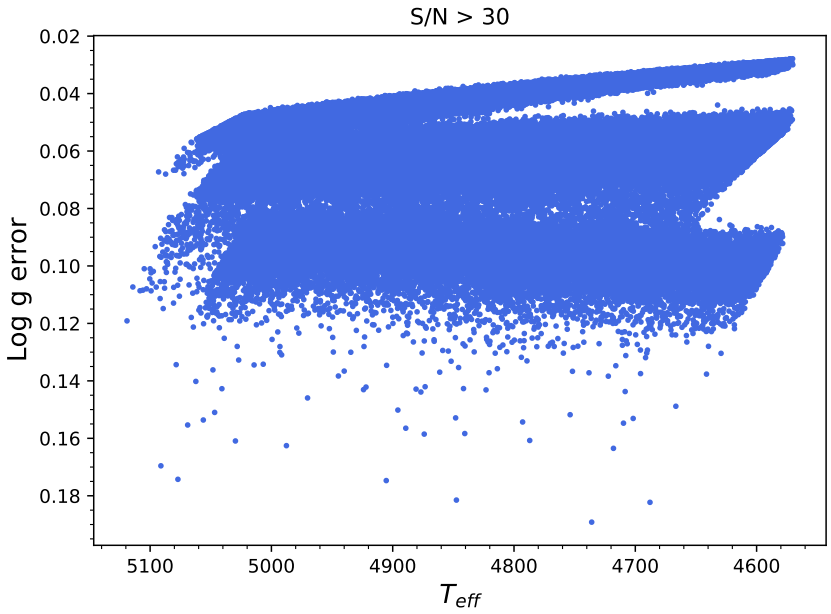


FIGURE 2.4: $\log g$ errors depending on T_{eff}. The colour refers to T_{eff} and is just to evidence the stars disposition: hot stars on the left and cooler stars on the right.

Figure 2.4 shows in fact that the $\log g$ data seems to be separated in three distinct groups instead of just two. We know this is not due to the $\log g$ being a stellar parameter, as the T_{eff} had no problem and showed values for all range. To check if something similar also happened in the metallicity, we plot the $\log g$ errors depending on the stellar metallicity. Figure 2.5 shows that, similar to T_{eff}, the metallicity has no constrains on the range values. Therefore, this data division is related to the $\log g$. We noticed that the three different groups seem to have very clear ranges of $\log g$ values, with the top one having $\log g > 4$, the middle one $1 < \log g < 2.5$ and the bottom one around $2.5 < \log g < 3.5$.

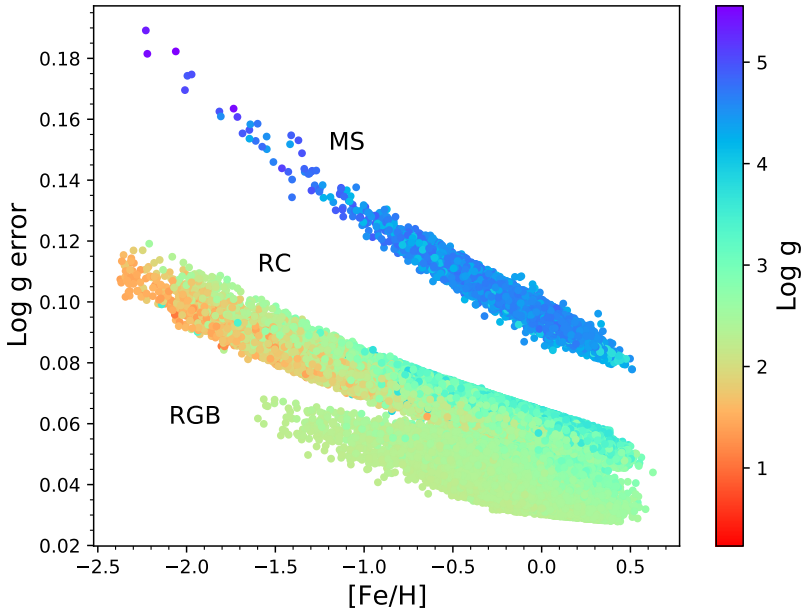


FIGURE 2.5: $\log g$ errors depending on metallicity. The colour refers to $\log g$ values. The top group corresponds to Main Sequence stars, the middle on to Red Clump stars and the bottom one to Red Giant Branch stars.

After exploring the bitmask system[1] used by APOGEE, we were able to understand different calibrations were made to the $\log g$, depending on the type of stars (see Appendix A for a more detailed explanation on this issue, what were the problems and how it was solved). In that sense, we have a calibration for Main-Sequence stars (MS), a calibration for Red Clump stars (RC), a calibration for Red Giant Branch stars (RGB) and a calibration for RGB-MS transition stars. Taking this into account, the top group in Figure 2.5 corresponds to MS stars, the middle one to RC stars and the bottom one corresponds to RGB stars. Similarly in Figure 2.3, the upper group is mostly giants and the lower group are dwarfs, as can be guessed from their logg values. Different calibrations

[1] https://www.sdss.org/dr16/algorithms/bitmasks/#APOGEE_PARAMFLAG

of the $\log g$ forced us to verify whether it is possible to join these three groups into one data set. Below, we explain how we select the best sub-sample based on the expectations from Galactic chemical evolution. The RGB-MS group is the smallest sub-sample and so we decided to exclude it from the beginning.

Most elements were formed through nucleosynthesis occurring during a star's lifetime (McWilliam 1997). Starting from hydrogen, massive stars have increasingly more nucleosynthesis processes in their interior, leading to the production of various elements besides helium and carbon. These elements formed through α-processes are often called α-elements and include species like O, Mg, Si, S, Ca and Ti. At the end of their lives, these stars explode as type II supernovae (SN II). The elements produced in their interiors enrich the interstellar medium and become available to be incorporated in clouds of dust where star formation occurs (see Woosley and Weaver 1984, Giacobbe 2005). Consequently, the new formed stars get a higher [α/Fe] ratio than their ancestors.

Contrary to massive stars that have a short lifetime ($\sim 10^6$ years), low mass stars are longer lived, with the lowest mass ones having a lifespan up to $\sim 10^{12}$ years. In the single degenerate model of SN Ia (Whelan and Iben 1973, Li and van den Heuvel 1997, Lü et al. 2009), we consider a binary system where one of the stars is a low mass star evolved into a white dwarf and the companion is a star with higher mass. If the outermost layers of the companion go beyond its Roche lobe, they stop being gravitationally bound and the white dwarf starts to accrete that mass. This mass accretion continues until the star reaches the maximum limit of mass they can support, after which the white drawf explodes in a type Ia supernovae (SN Ia). For the double degenerate model (Webbink 1984, Nelemans et al. 2001), the process is analogous, but for a binary system of two white dwarfs instead.

SN Ia produces mostly iron peak elements, like Cr, Mn, Co, Ni and, more importantly, Fe. With the explosion, these elements are released into the interstellar medium and will integrate the chemical composition of future stars (Maguire 2017). This means the [α/Fe] will decrease with the increase of the Fe abundance. As such, for the same metallicity value, older stars will have a higher [α/Fe] value due to being enriched with the α-elements produced from SN II, but not Fe because low mass stars did not "have time" yet to explode in a SN Ia.

As an example, the old thick disk population of the Milky Way have high [α/Fe] ratios at low metallicities, which decreases with [Fe/H] (Bensby and Feltzing 2010, Kordopatis et al. 2011, Adibekyan et al. 2012a). On the other hand, younger stars formed after the

SN Ia already have time to explode will have an enhanced iron abundance, consequence of the iron produced in SN Ia, and so lower [α/Fe] ratio. This is verified by the stellar population of the thin disk. Besides stars, the thin disk contains dust and gas and a considerable amount of molecular clouds where star forming is occurring. In fact, stellar formation in the this disk result of gas accretion at the later stages of a galaxy formation. Thus, the stars in the thin disk are younger than those from the thick disk and on average with a higher iron content (Bensby and Feltzing 2010).

On the other hand, the younger population of the Milky Way disk, the thin disk population, is more metal-rich than the thick disk populations. The [α/Fe] ratio of the thin disk stars also decreases with [Fe/H], however, overall the [α/Fe] ratio is lower than for the stars in the Galactic thick disk (e.g. Adibekyan et al. 2012a, Bensby and Feltzing 2010, Haywood et al. 2018).

Therefore, for the APOGEE sample, we expect to see these two populations: the thin and thick disks. So, for the different calibration groups, we analyse the dependency of [C/Fe], [O/Fe], [Mg/Fe] and [Si/Fe] abundance ratios with the metallicity. Figure 2.6 shows the results for the MS population, Figure 2.7 for the RGB stars and Figure 2.8 for RC stars.

As we can see in Figure 2.6, the MS stars group does not clearly reveal these two separate populations for any abundance ratios, as we were expecting, making it impossible to know if we are looking at a group of older stars, younger stars or a mix of both. We attribute this as relatively low quality of the determined abundances and decided to discard this subgroup from our final sample.

Contrary to MS stars, the RGB group in Figure 2.7 clearly shows the two different populations one would expect to see. For all chemical abundances ratio, the upper group corresponds to stars of the thick disk, which are older and have a higher [α/Fe] ratio. The lower group are younger stars from the thin disk, with a higher iron content and lower [α/Fe] ratio.

Similarly to RGB stars, the RC stars of Figure 2.8 also reveal the two separated populations of stars from the thin and thick disk. However, besides there being fewer stars in this calibration group, in the [C/Fe] and [O/Fe] plots the thick disk group is more difficult to identify.

In order to avoid systematic effects within the sample, the best strategy is to restrict the stellar parameter space to only use similar stars. Jönsson et al. (2020) recommend to

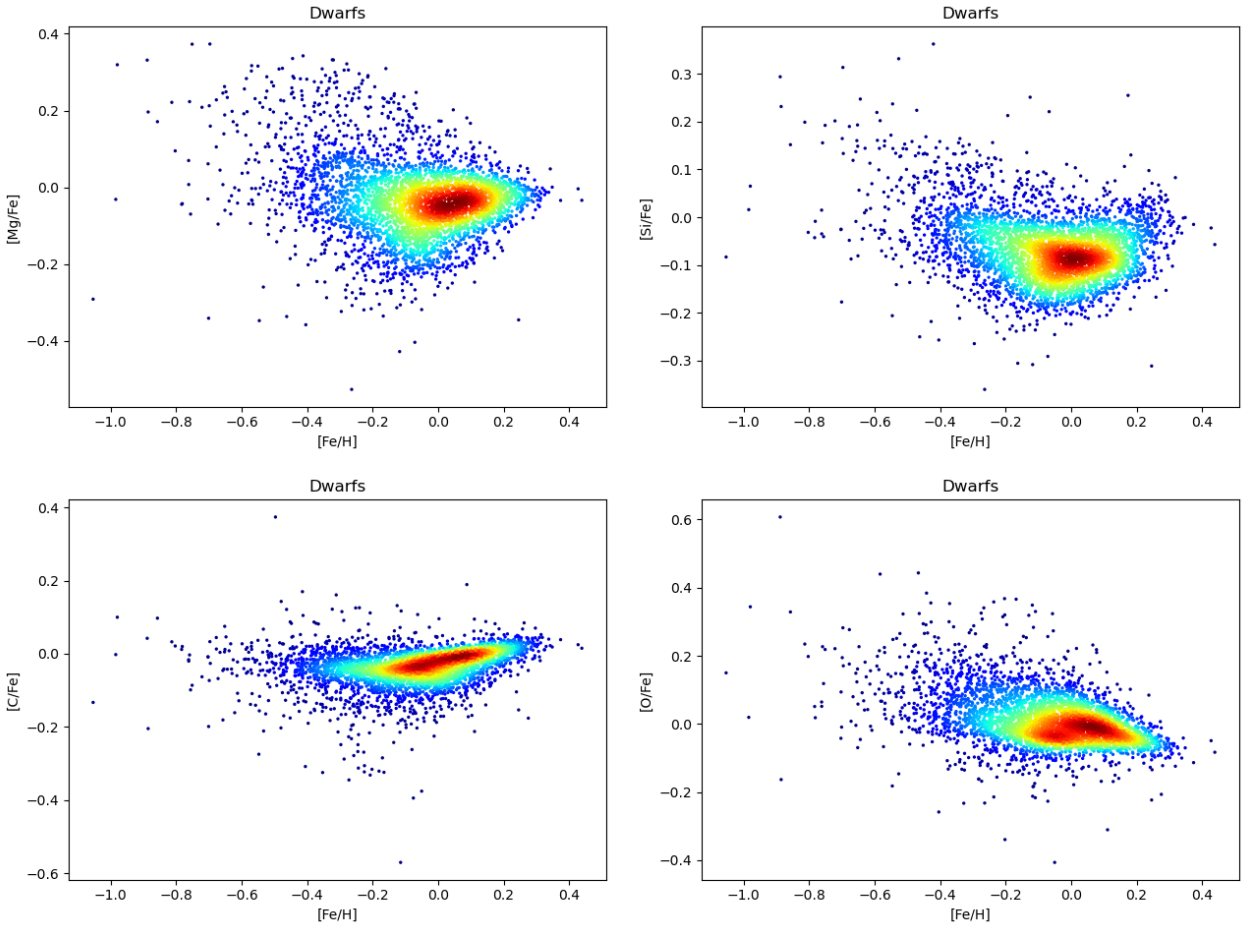


FIGURE 2.6: Abundance ratios as a function of [Fe/H] for MS dwarf stars. Colour scheme represents the density of points (stars).

focus on giants, since they are the group with the best data, but if possibly to avoid the coolest stars with $T_{eff} < 4000$ K (see Section 6.4 of Jönsson et al. 2020). The authors also advise against mixing RGB and RC stars, as these last ones might show some systematic differences (see Section 6.10.2 from Jönsson et al. 2020). As a result, we will be using only the RGB stars for the APOGEE sample (Henrik Jönsson, private communication, January 21, 2021).

The last condition we must ensure is to consider only stars with [Fe/H]≥ -1, as we do not expect to find a significant number of planets for stars with a metallicity value below around -1 (Johnson and Li 2012, Mortier et al. 2012), and we want to study the chemical composition of planet host stars. No planet has been found for stars with [Fe/H] < -1. With a metallicity of [Fe/H] $= -1.00 \pm 0.07$, BD+20 2457 b and BD+20 2457 c are the planets with the lowest metallicity known (Niedzielski et al. 2009).

Adding this to all the constrains made before, we get a sample with 21 788 stars (note

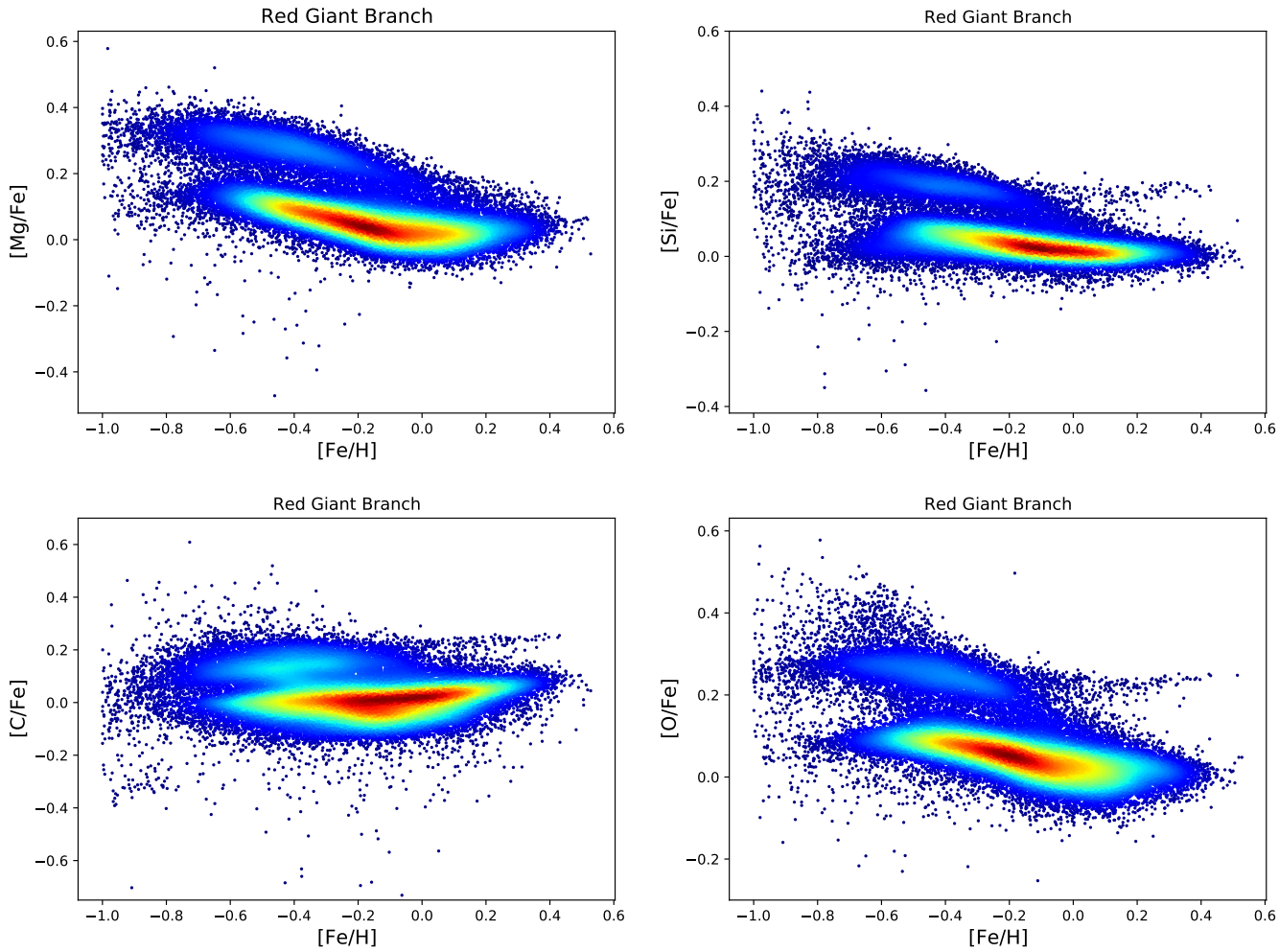


FIGURE 2.7: Abundance ratios as a function of [Fe/H] for RGB stars. Colour scheme represents the density of points (stars).

that only 1148 stars have $T_{eff} < 4000$ K). Table 2.1 summarises the information of the stellar parameters of the stars on this sample.

TABLE 2.1: Range of values of the stellar parameters for the stars in the APOGEE sample.

	Min value	Max value
T_{eff} $[K]$	3425	5056
$\log g$ [dex]	1.07	3.44
[Fe/H] [dex]	– 0.99	0.47

In order to analyse the correlations between the the S/N, stellar parameters and the chemical abundance ratios, we build the Spearman correlation matrix of Figure 2.9 with the APOGEE sample of RGB stars. In the following we will call this sample the APOGEE sample.

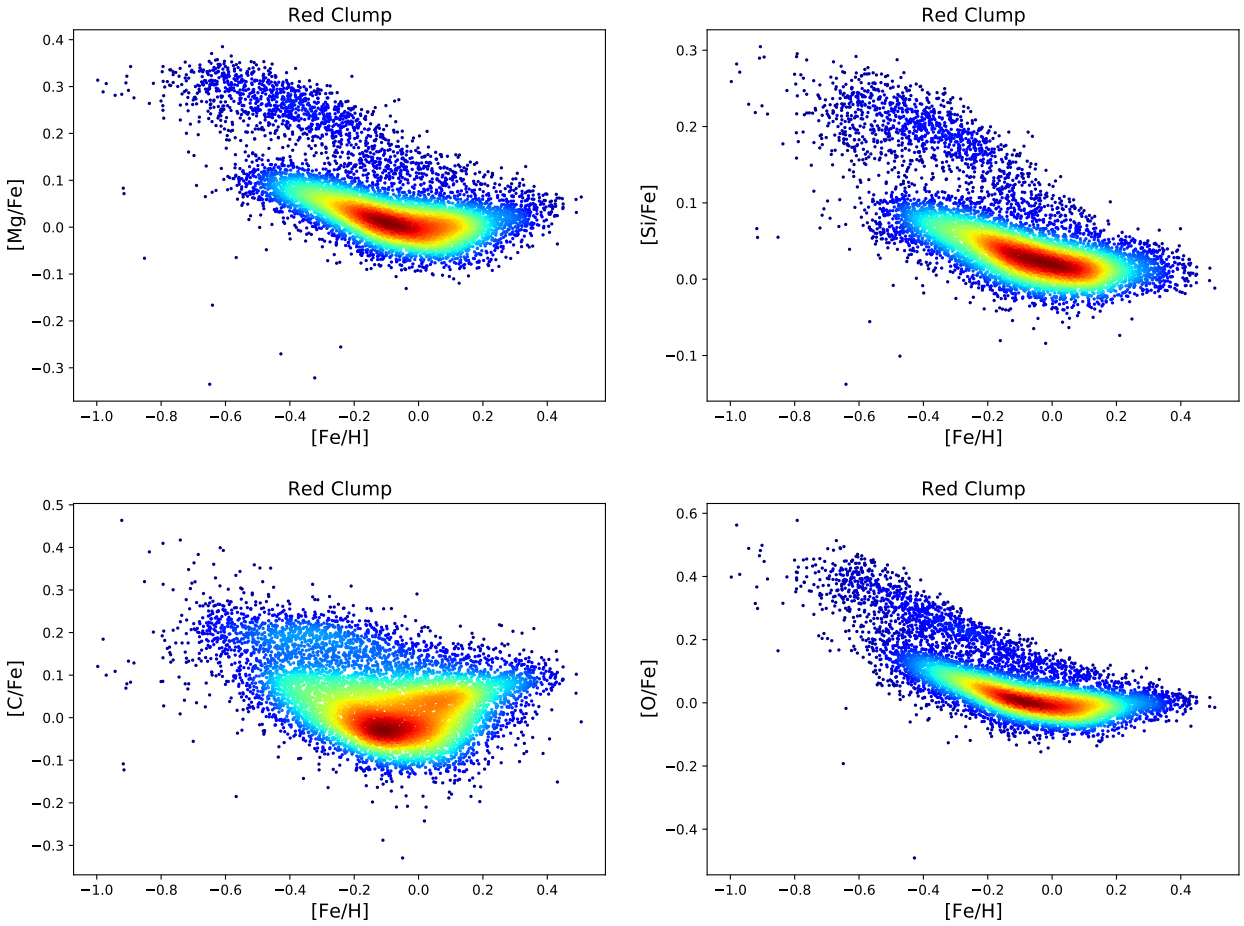


FIGURE 2.8: Abundance ratios as a function of [Fe/H] for RC stars. Colour scheme represents the density of points (stars).

As we had already deduced from Figure 2.2, the S/N does not have much of an impact on the chemical abundances nor does it have on the stellar parameters. The metallicity however, is tightly correlated with the abundances of Mg (0.92), Si (0.97), C (0.94) and O (0.95). This is expected since from the chemical evolution we know that as [Fe/H] increases all the abundances of other elements increase (i.e. as time passes, more elements are created) though not at the same rhythm. While having little relevance to this work, we also note that the [Mg/H], [Si/H], [C/H] and [O/H] abundances are strongly correlated between themselves.

Although not as high, the [C/O] abundance ratio also has a significant correlation with metallicity and those four abundances. The [Mg/Fe] and [Si/Fe] ratios are also strongly correlated between themselves and have some negative correlation with the metallicity, though not as significant. This means the metallicity might be of use to predict all these chemical abundances, be it through a positive or negative correlation.

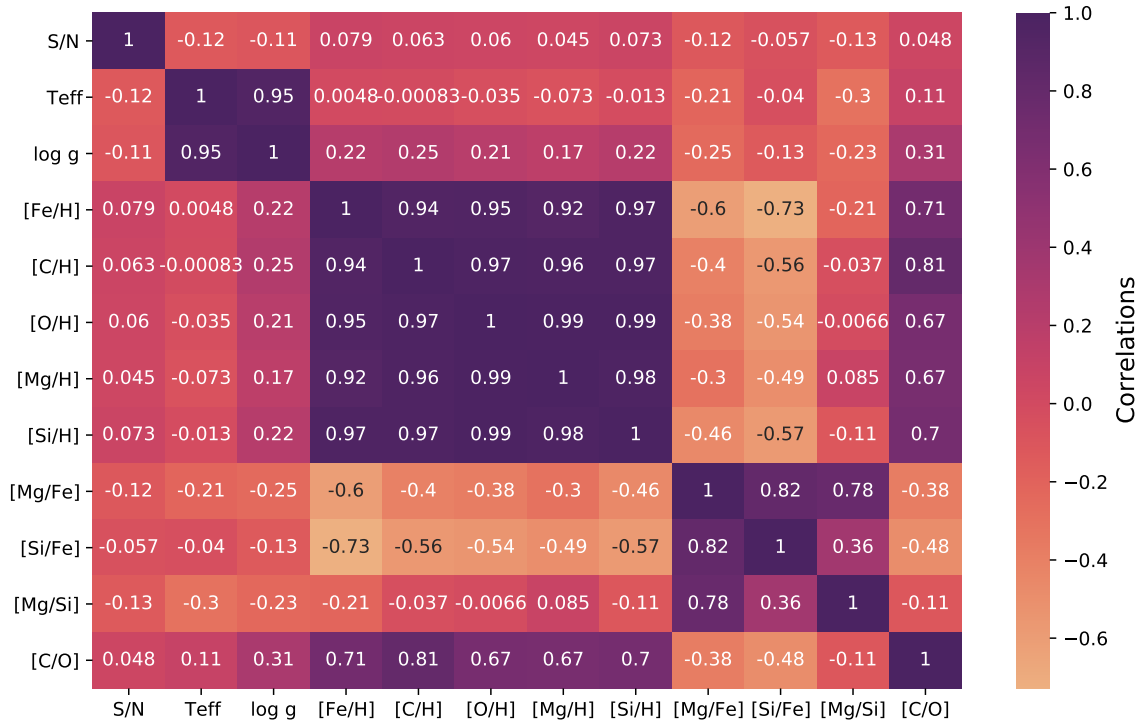


FIGURE 2.9: Correlations between the different parameters, for the APOGEE sample of RGB stars.

Aside from a tight correlation between each other, the T_{eff} and the $\log g$ do not have much of an influence on the chemical abundances (correlations coefficients of ~ 0 and ~ -0.2, respectively), meaning that these two parameters will not be a constrain to predict these quantities.

2.2.2 GALAH

As mentioned before, for the GALAH data we will be using the third data release. From GALAH DR3 website[1] we download the `GALAH_DR3_main_allstar_v2.fits` catalogue, which includes stellar parameters and elemental abundances.

The GALAH sample does not have different calibrations for different types of stars, and so from an initial sample of 588 571 stars, we start by selecting the highest quality data. Similar to APOGEE, GALAH also makes use of a bitmask system, so we can choose the stars with reliable stellar parameters, which refers to T_{eff} and $\log g$ (flag_sp==0), reliable determinations of metallicity (flag_fe_h==0) and chemical abundance ratios of [C/Fe], [O/Fe], [Mg/Fe] and [Si/Fe] with no warnings (flag_C_fe==0, flag_O_fe==0, flag_Mg_fe==0

[1] https://www.galah-survey.org/dr3/overview/

and flag_Si_fe==0). This leaves a sample with 111 037 stars. Table 2.2 shows the range of values for the effective temperature, $\log g$ and metallicity, with and without the aforementioned constrains.

TABLE 2.2: Stellar parameters range for the stars in the GALAH sample.

	Min value	Max value
Without constraints		
T_{eff} (K)	3000	7993.1
Log g (dex)	-0.45	5.18
Metallicity	-4.781	1.0
With constraints		
T_{eff} (K)	4432.3	7979.7
Log g (dex)	0.81	4.79
Metallicity	-1.657	0.996

The metallicity distribution of the constrained sample stars is wider than that of the known exoplanet host stars. Metallicity is one of the main input parameters in the RF model and it is important to guaranty a wide range of the parameter space for the training data. The changes in metallicity between applying or not the constrains do not have an impact on the number of stars since we are only interested in stars with [Fe/H]≥ -1.

From the GALAH survey we are interested in selecting the stars with the highest quality data. As such, we are focused on solar-type MS stars which have T_{eff} and $\log g$ well within the range of the values of the constrained sample.

Similar to the APOGEE data, we analyse the dependency of chemical abundances on stellar parameters. We could not include the S/N analysis as the data set did not have a specific column with that information for each star. Appendix B contains the plots of the chemical abundances depending on the stellar parameters (T_{eff} on B.1.1, $\log g$ on B.1.2 and metallicity on B.1.3).

In order to have a quantitative understanding, in Figure 2.10 we plot the correlation matrix for the GALAH sample.

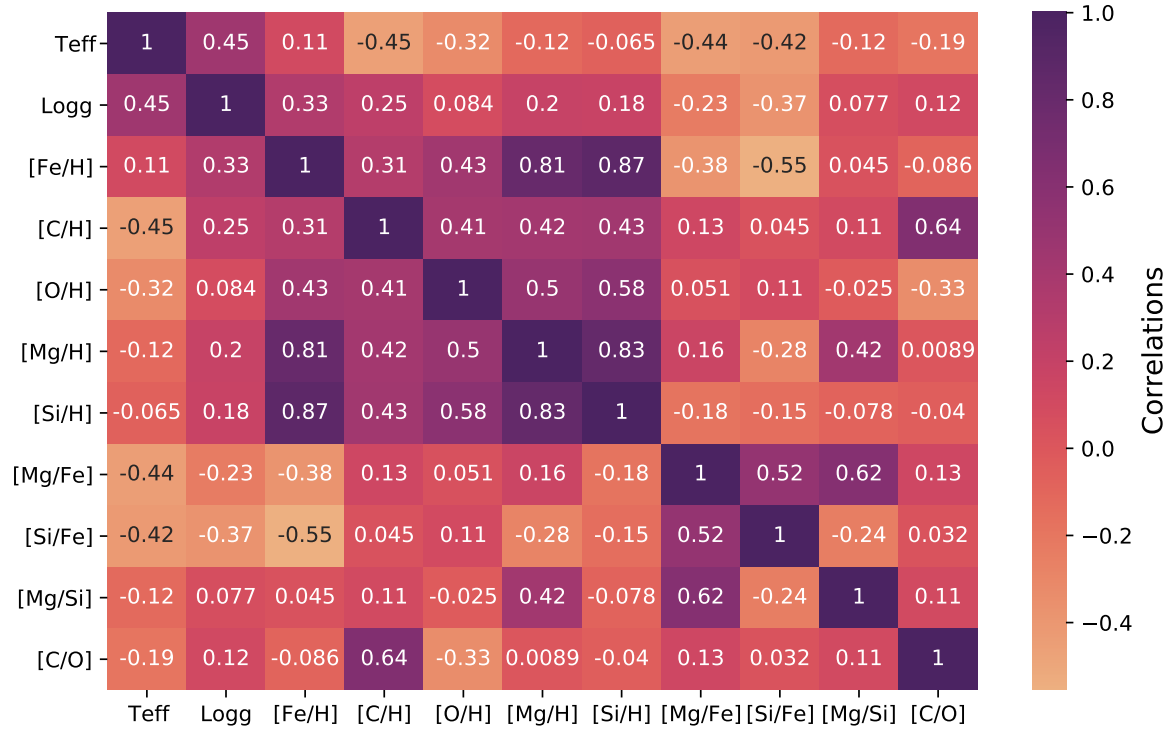


FIGURE 2.10: Correlation matrix for the parameters of GALAH data set.

Overall, the T_{eff} and $\log g$ do not show any significant trend on the data. For the T_{eff}, we can see a weak correlation in the [C/H], [Mg/Fe] and [Si/Fe] abundance ratios. On the $\log g$, it is also possible to notice a weak correlation on the [C/H] and [C/O] abundance ratios. These correlations might have astrophysical origin or could be result of the specific methodology applied to determine the stellar parameters and chemical abundances. For example, hotter stars should on average be also more massive and as such, be overall younger. As such, the abundance - T_{eff} relation can reflect the abundance - age relation which can be a result of the Galactic chemical evolution. Stellar surface gravity is also related with the stellar age for a given spectral type of stars. On the other hand, again it is the metallicity that shows better results. In particular, there is a strong positive correlation between metallicity and the abundances of Mg and Si, and a weaker but still noticeable anti-correlation between metallicity and the [Si/Fe] abundance ratio. The observed correlation between the abundances of different elements with metallicity is a natural consequence of the Galactic chemical evolution.

As discussed earlier, in a high quality data [Mg/Fe] vs [Fe/H] and [Si/Fe] vs [Fe/H] figures should reveal two sequences of stars with high and low content of α elements relative to iron. Figure B.3, however, does not allow to find these two populations. Because

the sample size allows and since the correlations in Figure 2.10 are not very good, we decided to limit the GALAH sample to only Solar Analogues stars (SAS), which we define as stars with the following stellar parameters:

- $T_{eff} = T_{\odot} \pm 300\ K$
- $\log g = \log g_{\odot} \pm 0.2$ dex
- $[\mathrm{Fe/H}] \geq -1$ dex

$T_{\odot}$ stands for the solar T_{eff} and $\log g_{\odot}$ corresponds to the solar $\log g$. We considered $T_{\odot} = 5772$ K and $\log g_{\odot} = 4.44$ dex (Perryman 2018).

The SAS sample is comprised of 21 472 stars. We performed the same exercise of the stellar parameters impact on the chemical abundances as it was done for the full sample. The matrix correlation for the SAS sample is present on Figure 2.11. The figures showing the dependence of the abundances and abundance ratios on T_{eff}, $\log g$, and metallicity are shown in appendices B.2.1, B.2.2 and B.2.3, respectively.

In general, the effective temperature and $\log g$ do not have a strong correlation with any chemical abundance ratios. As opposite to the APOGEE data, in Figure B.6 we cannot see a clear separation between the thin and thick disk populations in any of the [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] vs [Fe/H] plots. Given the low [α/Fe] ratios of most of the stars, it is possible that they belong to the Galactic thin disk population. On the other hand, the metallicity shows a very strong correlation of 0.91 and 0.96 with [Mg/H] and [Si/H], respectively, and a considerable correlation with [C/H] and [O/H] of 0.55 and 0.59. The observed strong correlations with metallicity makes it a good input parameter for the RF to predict the abundances of different elements, which led us to choose the SAS sample, (hereon on called GALAH sample).

2.2.3 HARPS

For the HARPS sample, we followed an approach similar to the work of Santos et al. (2017) to constrain the data. Starting from the work of Adibekyan et al. (2012a), we have a data set based on spectra with high signal-to-noise ratios and high resolution obtained with the HARPS spectrograph with the abundances of Mg, Si, and Fe for 1111 stars.

The oxygen abundance values originally came from Bertran de Lis et al. (2015), but were updated in Delgado Mena et al. (2021), together with carbon abundances. Although

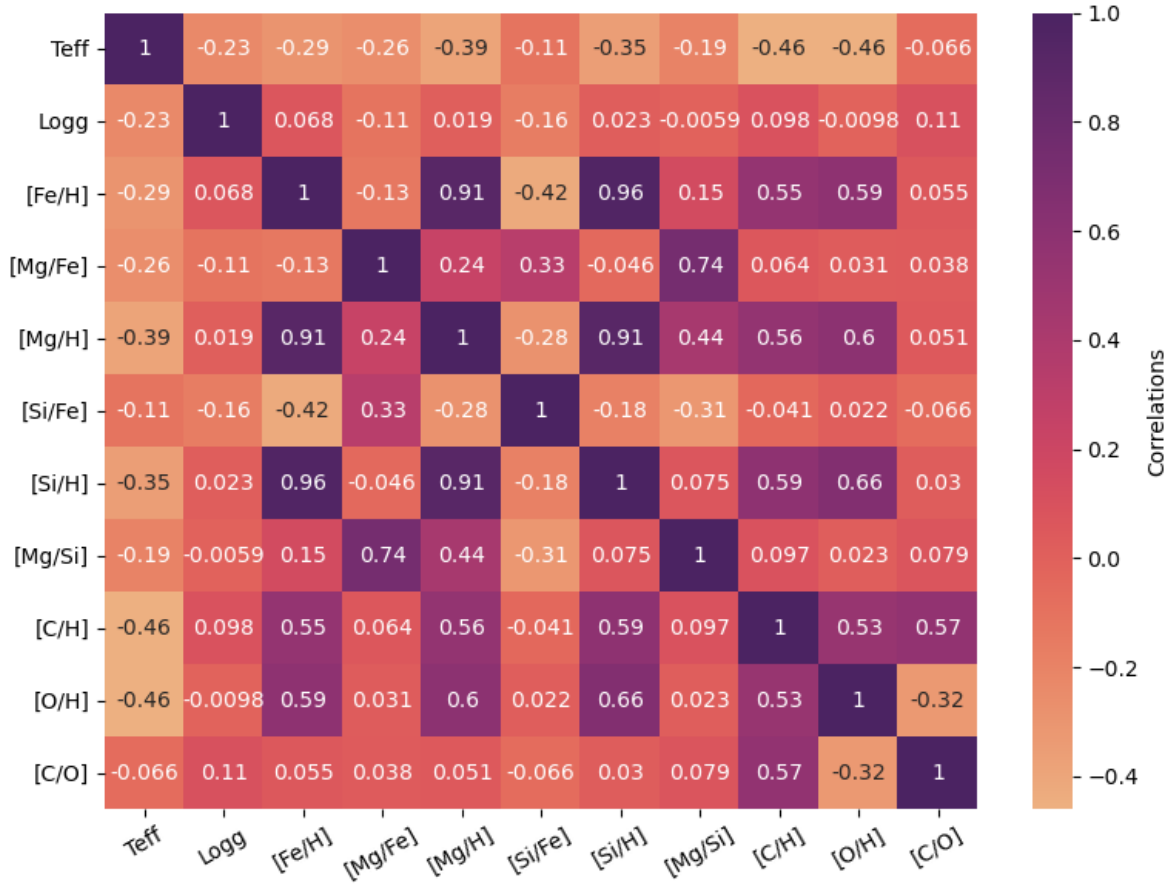


FIGURE 2.11: Correlation matrix for the parameters of Solar Analogue stars sample.

the abundances come from different works, all of them were derived by the same group based on the same set of uniform stellar atmospheric parameters (T_{eff} and $\log g$) from Sousa et al. (2008, 2011a,b). In the end, we have 535 stars with abundances for C, O, Mg, Si, and Fe that are a good representative of the solar neighbourhood.

Similar to what was done for APOGEE, we verify if the chemical abundances depend on the S/N ratio. In Figure B.7 we show the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] versus the S/N. Although there is a slightly larger dispersion for the smaller S/N values, it is possible to see that there is no dependency on the S/N itself. We did not perform any cut on S/N and decided to use the full data set. We do not exclude stars with [Fe/H]< -1 as there are only 2 stars with values of -1.1 and -1.05 each.

We plot the [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] chemical abundance ratios depending on the T_{eff} (Figure B.8), $\log g$ (Figure B.9) and metallicity (Figure B.10), to check if they might depend on the stellar parameters. Like

for APOGEE data, note that in Figure B.10 B.10 we can see a clear separation between the thin and thick disk populations in the [Mg/Fe] vs [Fe/H] and [Si/Fe] vs [Fe/H] plots. The correlation matrix in Figure 2.12 provides a more numerical representation of these dependencies.

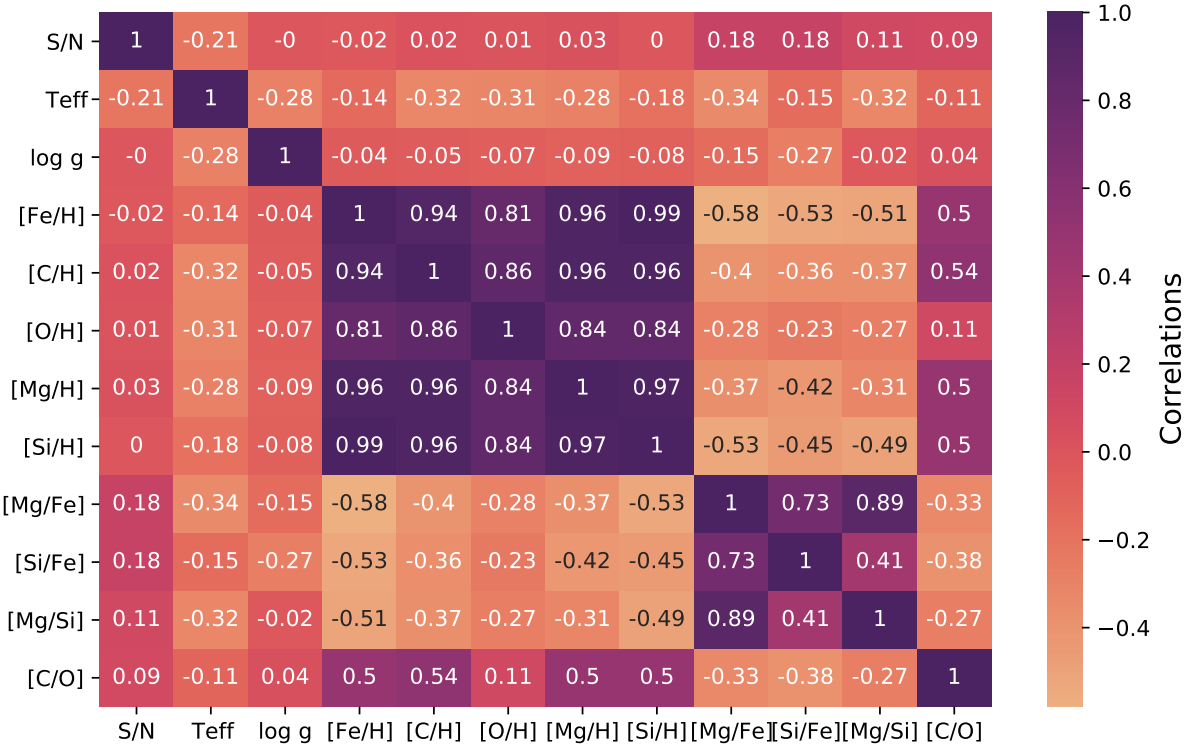


FIGURE 2.12: Correlation matrix for the parameters of the HARPS sample.

The S/N ratio, T_{eff} and $\log g$ do not show any particular correlation with the chemical abundance ratios. The metallicity however, reveals a high correlation with the element abundances and while it is visible on the [O/H] plot, it is particularly strong for the abundances of Mg, Si and C (see referred plots from Figure B.10). This means that with this data set we can also use the metallicity as a proxy to try and predict the abundances, at least for [C/H], [O/H], [Mg/H] and [Si/H].

The correlation matrix shows that the individual abundances correlate better with the metallicity than the abundance ratios. The later ones can be calculated when the individual abundances are precisely predicted. Although not relevant for the current work, we see that the [Mg/H] and [Si/H] abundances also have a very high correlation with the [C/H] and [O/H] abundances. As the Mg and Si abundances are easier to determine from the star spectra, they can be used to know the abundances of [C/H] and [O/H] if it is not possible to determine them from the stellar spectra.

Having ascertained that T_{eff} and $\log g$ do not have much leverage on the chemical abundances and ratios, we can now proceed to trying to make a prediction model for each of these data sets.

2.3 Galactic velocities

As we have clearly seen for the APOGEE and HARPS samples, stars from different galactic populations can have the same metallicity and still have different α-elements abundances, which include O, Mg and Si. Therefore, we need to take into account the different galactic populations. Stars from the thin and thick disk have very distinct kinematics (Bensby and Feltzing 2010, Adibekyan et al. 2012a, just to mention a few works) and we will use this property to better predict abundances.

In that sense, we want to compute the galactic velocity components, U, V and W, where U is the velocity component directed outwards from the centre of the Milky Way, V is aligned with the local direction of the Galactic rotation, and W is positive toward the north Galactic pole (Bashi and Zucker 2019). In order to accomplish this, we follow the approach described in Johnson and Soderblom (1987). To calculate the space velocity components, for each star, we need the right ascension (ra) and declination (dec) in degrees, the proper motion in right ascension (pmra) and in declination (pmdec), both in mas/yr, the distance in pc (or the parallax in mas), and the RV in km/s. Johnson and Soderblom's method however is a heliocentric one, meaning the velocity components are relative to the Sun, which itself is not stationary. Considering $(\mathrm{U,V,W})_\odot = (-8.5, 13.38, 6.49)$ km/s the solar motion relative to the local standard of rest (LSR), it is possible to calculate the velocity components with respect to the LSR. The LSR is defined as the rest frame of a star at the location of the Sun that would be on a circular orbit in the gravitational potential of the Galaxy (Coşkunoğlu et al. 2011).

In order to get the necessary information to compute the galactic velocity components, for each data set we search the stars in the Gaia Archive[1] DR2 and from there get the six parameters needed. From the 21 788 stars in the APOGEE sample, 6803 do not have any Gaia information or are missing some of the parameters, leaving a final sample of 14 985 stars. 3431 out of 21 472 GALAH stars do not have any Gaia information. Of the remaining stars, 20 do not have any RV information. This amounts to 18 021 stars for the

[1] https://gea.esac.esa.int/archive/

GALAH sample. 8 stars from the initial 535 stars on HARPS sample did not have Gaia information. This leaves a final HARPS sample with 527 stars.

2.4 Predicting abundances with RF

In order to create a predictive model, we will resort to the `Scikit-learn` library (Pedregosa et al. 2011) from Python 3. From the `model_selection` module, we use the `train_test_split` function to split the data in training set and test set. Depending on the sample size, a ratio of 70:30 or 60:40 was used to split the data in training and test set. The `random_state` parameter was fixed to ensure the subset for the training data had values covering the whole range and there were no values underrepresented, which could influence the predictions. From the `preprocessing` module, we use the `StandardScaler` function to scale the data by being fitted to the training set. With the module `ensemble` we call the `RandomForestRegressor` function to initiate a regression RF. Depending on the sample size, we choose the number of decision trees in the RF in order to optimise computational time. The `metrics` module provides the `r2_score` and `mean_squared_error` functions, responsible for the evaluation of the model.

As we have three data sets, we will create a model for each one. For APOGEE data, we used the 60:40 training/test ratio to split the data, a `random_state=2` and a forest with 500 decision trees. The `test_size` parameter corresponds to the size of the test sample, i.e the training/test ratio. The `random_state` is a parameter that controls where the division of training/test set is made. We fix this parameter to make sure the training sample was spread along the range and there was no value underrepresented. An example would be if the training set were centred around higher [Fe/H] values. This would lead to worse predictions for stars with sub-solar metallicities, since the model did not see enough examples of these cases. With the GALAH data, we also used the 60:40 training/test ratio to split the data, but a `random_state=6` and a forest with 800 decision trees instead, since the GALAH sample is larger than APOGEE. As the HARPS sample is smaller, we used a 70:30 training/test ratio to do the split, a `random_state=6` and a forest with 800 decision trees, the later ones being the same as for the GALAH conditions. Due to the smaller size of the sample, having 500 or 800 decision trees does not affect much the computational time needed to run the code. Hence we leave it at 800 decision trees. Table 2.3 summarises the characteristics of the three RF models.

TABLE 2.3: Characteristics of the three RF models.

Model	Base dataset	# Stars	Type of stars	Training/test ratio	# Decision trees
RFA	APOGEE	14985	RGB stars	60:40	500
RFG	GALAH	18021	Solar analogues	60:40	800
RFH	HARPS	527	Solar-like stars	70:30	800

To make a model able to predict the chemical abundance of stars based on their metallicity and galactic velocity, we have the metallicity [Fe/H] and total galactic velocity $V_{total} = \sqrt{U^2 + V^2 + W^2}$ as input information, i.e. information we give to the model. Initially we tested the model with the 3 velocity components individually and the sum of them and found that V_{total} contains more information than the individual components. Based on the information of those parameters, we want to predict the abundances of [C/H], [O/H], [Mg/H], [Si/H] and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. For each sample, we train the respective model on the training set by showing it the inputs of various examples and their expected chemical abundances so the model can start to try to find a pattern. Tables 2.4, 2.5 and 2.6 summarise the information and evaluate how the model behaves, for each data set.

On Figure 2.13 we present the plots of the predictions done by the RF of the HARPS sample (RFH), for the [Mg/H] abundance and [Mg/Fe] abundance ratio. Together with the plots of the predictions done by the RF of the APOGEE sample (RFA) and the GALAH sample (RFG), all the other figures can be found in Appendix C.

TABLE 2.4: Parameters of the RFA model. From left to right, the columns are the abundances to predict, R^2 score, mean values of Prediction - True value, standard deviation of the predicted and true data, mean-squared-error (MSE) of the training and test set.

Abundances	R^2	mean diff	std pred	std true	Train MSE	Test MSE
[C/H]	0.90	-0.00082	0.288	0.296	0.0012	0.0089
[O/H]	0.90	-0.00047	0.220	0.227	0.0007	0.0050
[Mg/H]	0.88	-0.00061	0.228	0.239	0.0009	0.0069
[Si/H]	0.95	-0.00069	0.234	0.238	0.0004	0.0029
[Mg/Fe]	0.30	-0.00068	0.071	0.100	0.0009	0.0069
[Si/Fe]	0.46	-0.00076	0.058	0.074	0.0004	0.0029
[Mg/Si]	-0.09	0.00008	0.021	0.044	0.0003	0.0021
[C/O]	0.46	-0.00034	0.075	0.093	0.0006	0.0047

TABLE 2.5: Parameters of the RFG model. Meaning of the columns is the same as for Table 2.4.

Abundances	R^2	mean diff	std pred	std true	Train MSE	Test MSE
[C/H]	0.18	0.00168	0.106	0.163	0.0031	0.0217
[O/H]	0.18	0.00203	0.096	0.148	0.0025	0.0181
[Mg/H]	0.80	-0.00178	0.189	0.204	0.0011	0.0082
[Si/H]	0.90	-0.00004	0.174	0.181	0.0005	0.0034
[Mg/Fe]	-0.06	-0.00179	0.043	0.088	0.0011	0.0082
[Si/Fe]	0.10	-0.00005	0.037	0.062	0.0005	0.0034
[Mg/Si]	-0.12	-0.00174	0.039	0.086	0.0011	0.0083
[C/O]	-0.12	-0.00035	0.074	0.160	0.0041	0.0287

TABLE 2.6: Parameters of the RFH model. Meaning of the columns is the same as for Table 2.4.

Abundances	R^2	mean diff	std pred	std true	Train MSE	Test MSE
[C/H]	0.88	0.0167	0.203	0.220	0.0009	0.0060
[O/H]	0.59	0.0082	0.143	0.175	0.0018	0.0124
[Mg/H]	0.91	0.0102	0.206	0.211	0.0008	0.0040
[Si/H]	0.97	0.0071	0.224	0.234	0.0002	0.0016
[Mg/Fe]	0.52	0.0102	0.080	0.090	0.0007	0.0039
[Si/Fe]	0.58	0.0071	0.052	0.057	0.0002	0.0014
[Mg/Si]	-0.04	0.0031	0.037	0.046	0.0004	0.0022
[C/O]	0.42	0.0085	0.092	0.129	0.0014	0.0095

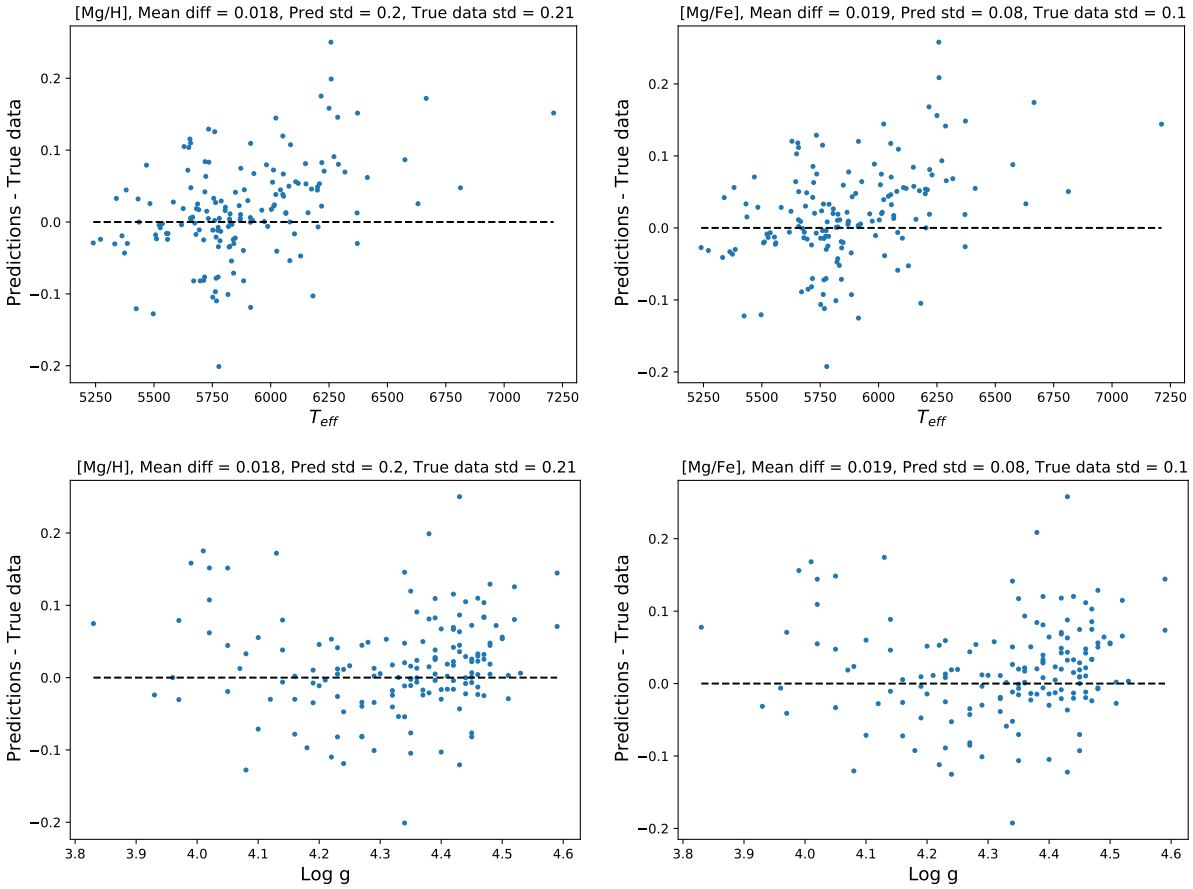


FIGURE 2.13: The X axis is the values of T_{eff} (top) and $\log g$ (bottom) and the Y axis is the *Predictions – True value*, for the chemical abundances of [Mg/H] (left) and [Mg/Fe] (right). The black line marks values of predictions equal to the true data. The title includes the abundance ratio we are predicting, the mean difference of *Predictions – True value*, the standard deviation for the predictions and the standard deviation for the true values.

From Figure 2.13, one can see how the mean difference is always around 0.1. Yet, having a small mean difference does not necessarily signify the predictions are good. If some of them are higher than the expected value and others lower than expected, they cancel each other out, leading to a mean difference close to 0. Therefore, it is also necessary to look at the dispersion (standard deviation), R^2 score and mean-squared-errors (MSE).

R^2 is the coefficient of determination, i.e. the regression score. A perfect model would have the best possible score, 1. The coefficient can also be negative because the model can be arbitrarily worse. A null score means the model always predicts the expected value, disregarding the input features. For all three models and for each abundance and abundance ratio, the standard deviation of the predicted data is never greater than that of the true data, meaning that at least the predictions have less dispersion than the original values, as expected. The MSE measures how well the model predictions match the observed data and is calculated as:

$$MSE = \left(\frac{1}{N}\right) \times \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{2.4}$$

where N is the total number of observations, y_i the expected value of the ith observation and $f(x_i)$ the predicted values of the ith observation. The closer the model predictions are to the observations, the smaller the MSE will be. Therefore, the smaller the test MSE is, the better. The training MSE tells us how closely the model predictions match up with the data that we used to train the model, which does not provide any information on how well (or bad) the model behaves. The training MSE, however, can be a good indicator of overfitting if it is much smaller (around an order of magnitude or more) than the test MSE. In these cases, it would mean the model focused too much on finding patterns in the training data that are caused by random chance, as it was fitted too closely to the training data. Therefore, when applied to new data, the model would perform poorly. In our results, both values are very close to 0, meaning we can exclude the possibility of overfitting any of the models.

Looking at the R^2 column, we note that all three models are capable of at least predicting accurately the Mg and Si abundances. In fact, the RFA model has a high R^2 for all C, O, Mg and Si abundances. The RFG however can only accurately predict [Mg/H] and [Si/H]. The RFH model reveals a good score for all abundances except [O/H]. No model shows good results on the predictions of the [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]

abundance ratios. Yet, knowing the metallicity and the element abundances, these ratios can still be calculated arithmetically.

2.5 Testing the models

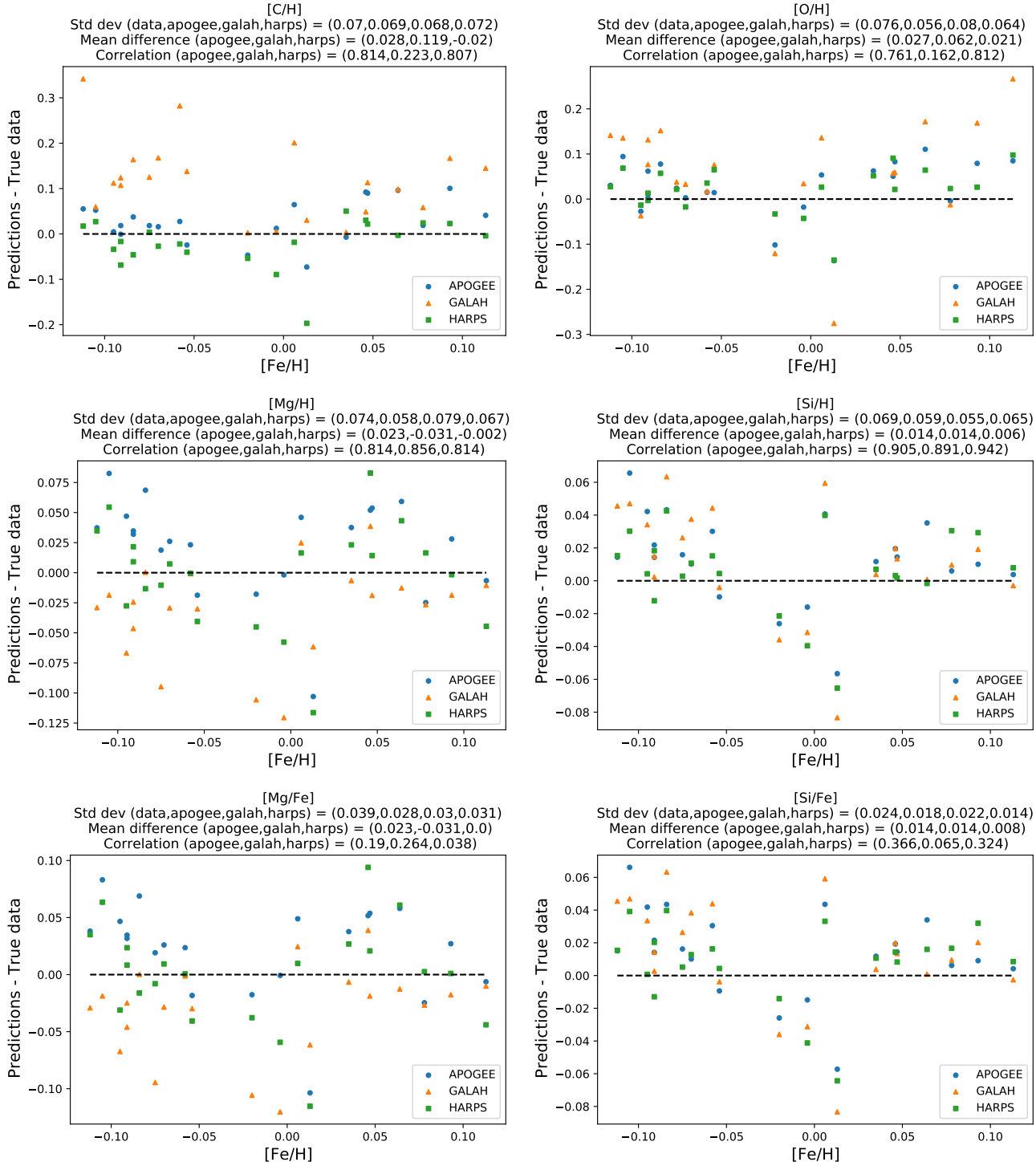
With each model trained and tested in its respective sample, it is necessary to analyse how they behave when applied to different data sets. To further test and validate our predictions we used data sets from four independent works. Apart from the test done with Nissen (2015)'s data, the figures and tables illustrating the results are presented in Appendix E.

The first data set comes from Nissen (2015), which contains a sample of 21 solar twins in the solar neighbourhood. The author has derived the metallicity and abundance ratios of [O/Fe], [C/Fe], [Mg/Fe] and [Si/Fe]. Knowing the metallicity, we determined the element abundances of O, C, Mg and Si through:

$$[X/H] = [X/Fe] + [Fe/H] \tag{2.5}$$

where X is the element of which we want to determine the abundance. The [Mg/Si] abundance ratio can be determined by: [Mg/Si] = [Mg/Fe] − [Si/Fe] = [Mg/H] − [Si/H]. In an analogous manner, we can determine the [C/O] ratio based on the abundances of C and O. Before applying the RF models to this data set, we first determined the Galactic space velocity components as discussed in Section 2.3. From the Gaia archive we retrieved the information on the coordinates, parallax, proper motions and RV (and associated uncertainties) of the stars. Figure 2.14 shows the difference between predictions and true data, depending on metallicity, for all three RF. The figures include information on the standard deviation of true data and three sets of predictions. This is to check whether there is an increase in dispersion on the predictions in relation to the original values. The figure also contains information on the mean difference and correlation between predictions and true data. Tables from 2.7 – 2.10 summarise the performance of the models, including the R^2 score. RFG is the model with predictions farther from the expected value, which can also be seen by its R^2 values on Table 2.9. Although its R^2 score is negative, we can see on Figure 2.14 that, in general, RFH model predictions were closer to RFA, which was the model with best performance. In particular, the predictions for [Mg/H] and [Si/H] by the RFA model were close to the expected value (see Section F.4 for more

information about Nissen(2015)'s data in relation to HARPS). For this sample, RFA was the model with best performance, followed by RFH. The RFG model has the worse predictions.



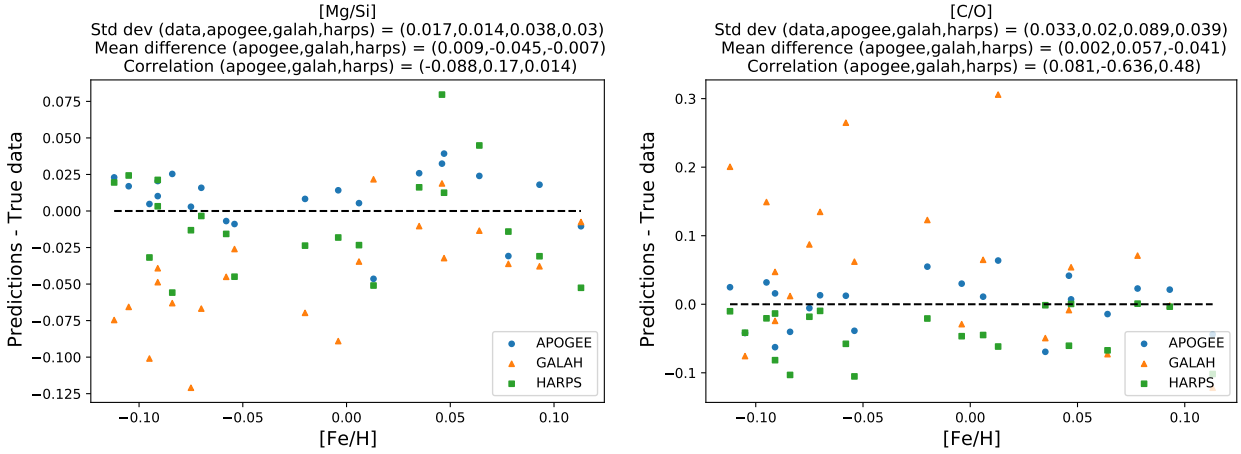


FIGURE 2.14: Predictions of Nissen(2015)'s data, using the RFA, RFG and RFH models. Y axis is the difference between the predicted values and the true values, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. Title includes standard deviation of true data and predictions, mean difference of Predictions - True data values, and correlation between prediction and true data, for all samples. The black line marks values of predictions equal to the true data.

TABLE 2.7: Mean difference values for Nissen(2015)'s data.

	Mean Difference							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.0127	0.0122	0.0224	0.022	0.0128	0.0122	0.0005	0.0004
GALAH	0.1118	0.1739	0.2056	0.1116	-0.0649	-0.0028	-0.0621	0.094
HARPS	0.141	0.1541	0.1476	0.1993	0.0048	0.0179	-0.0131	-0.0517

TABLE 2.8: Standard deviation values for Nissen(2015)'s data.

	Standard deviation							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
Data	0.07392	0.06948	0.06993	0.07589	0.03865	0.02391	0.01661	0.03336
APOGEE	0.06902	0.06639	0.08542	0.06378	0.01709	0.00851	0.01288	0.02638
GALAH	0.08527	0.06242	0.06053	0.05631	0.04444	0.02042	0.04865	0.06189
HARPS	0.04908	0.06649	0.04541	0.04516	0.03896	0.01269	0.0325	0.02403

TABLE 2.9: R^2 score for the predictions of Nissen(2015)'s data.

	R^2 score for the predictions							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.633	0.8318	0.0945	0.0709	-0.3314	-0.4058	-0.7416	-0.1457
GALAH	-1.8444	-5.3885	-8.2909	-1.9555	-4.57	-0.4147	-23.9478	-10.8368
HARPS	-3.1428	-4.0796	-4.158	-6.8817	-0.7365	-0.7351	-4.4251	-3.1432

TABLE 2.10: Correlation between predictions and true data, for Nissen(2015)'s data.

	Correlation between predictions and true data							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.831	0.926	0.752	0.619	0.025	0.114	-0.113	0.373
GALAH	0.687	0.923	0.717	0.544	-0.418	0.176	-0.536	0.107
HARPS	0.68	0.895	0.533	0.305	0.195	0.095	0.07	-0.186

Very recently, Adibekyan et al. (2021) determined the abundances of C, O, Mg, Si and Fe of a sample of 27 solar-type stars. Based on these abundances, we additionally determined the abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O], and total Galactic velocity to test the RFA, RFG and RFH models. As for the Nissen (2015) sample, Figure E.1 shows the difference between predictions and true data, depending on metallicity, for all three RF, including information on the standard deviation of all four samples, the mean difference and correlation between predictions and true data. Tables from E.1 – E.4 summarise the performance of the models, including the R^2 score. Although it was not as poorly as for the Nissen (2015) data, the RFG model is still the one with worse predictions, with the R^2 score being especially low. In general, while RFA is the one with smallest mean difference values, it also has higher dispersion that the predictions from RFH. Both of them have a strong correlation between the expected and predicted values. Based on the R^2 values, the RFA model is the best one, with RFH being the best second. The RFG model has the worst predictions.

From a sample of 7 solar twins and 95 solar analogues, González Hernández et al. (2010) determined the abundance ratios for 24 elements, including the abundance ratios of [C/Fe], [O/Fe], [Mg/Fe], [Si/Fe]. Based on the aforementioned abundance ratios and metallicities ([Fe/H]) of the stars, we calculated the abundances of [C/H], [O/H], [Mg/H]

and [Si/H] abundances. Similar to the previous tests, we also determine the total Galactic velocity.

Figure E.3 shows the predictions for each model and information on the mean difference, standard deviation, R^2 score and correlations are presented in Tables E.9 – E.12. Once again RFG seems to be the model with worse behaviour, which is particularly clear from the [C/H] and [C/O] plots, as we can see that for lower metallicities RFG tends to predict abundances higher than expected. Similarly, the RFG model predicts systematically overestimated values for the [C/O] ratio. Although less clearly visible, for metal-rich stars the RFG model predicts systematically lower [Si/H] abundances. Also on the [C/H] plot, for higher values of [Fe/H], the model RFA tends to predict increasingly higher values than expected. Other than that, the RFA model has a behaviour akin to RFH. Looking to Tables E.9 and E.10, there is no indication that one model is strikingly better or worse than others. However, Table E.11 shows how the RFH model has high values of R^2 for abundances of Mg, Si and C, meaning that these predictions are very close to what is expected. The observed small mean differences in Table E.9 confirms this. The RFH model is followed by the RFA model, also with good R^2 score for the same abundances, although not as high. RFG model also has very good R^2 values for Mg and Si abundances, but falls short on the rest of the abundance ratios. Table E.12 indicates that, in general, RFH and RFA have a good match for Mg, Si, C and O abundances, with high correlation coefficients, while those obtained with the RFG model are smaller for these elements.

Recently, Maldonado et al. (2020) provided a sample of 193 M dwarfs with the derived chemical abundance of individual elements such as C, O, Na, Mg, Al, Si, Ca, Sc, Ti, V, Cr, Mn, Co, Ni, and Zn, as well as the stellar metallicity. Using principal component analysis (PCA) and Bayesian regression methods to reduce the dimensionality of the data and find the variable(s) with higher variance, Maldonado et al. (2020) trained a model with spectra of M dwarfs in binaries orbiting a FGK star to predict the individual chemical abundances. In most cases, their results show an agreement between the derived abundances and those measured in the FGK primary to be better than 0.10 dex.

Knowing the stellar metallicity and the individually chemical abundances, we can compute the [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] abundance ratios. With parameters from the Gaia survey, we determined the total Galactic velocity. Figure E.2 shows the results of applying the three models to the sample. Tables E.5 – E.8 summarise the behaviour of the models by providing information about the mean difference, standard deviation,

R^2 score and correlation between predictions and true data. Maldonado et al. (2020) did not provide abundance of oxygen, as such we could not make a prediction for this element. Aside from the [C/H] abundance where RFG seems to tend to higher abundance predictions, Figure E.2 shows that all predictions are relatively close to each other. Yet the R^2 score on Table E.7 reveals that the RFG is the model with more abundance predictions with negative scores and Table E.8 shows that the predicted and true values for this model have the weakest correlation, while RFA and RFH have similar values.

Additionally, we performed a Kolmogorov–Smirnov (KS) test between the RFA and RFH predictions and the true data, to check whether for M dwarfs the predictions and original data can be drawn from the same distribution, which entails having *p-value* $>$ 0.05. The higher the *p-value*, the higher the probability of both samples being taken from the same parent. We do not include the results from RFG model as it provided the worst results. For [C/H] predictions, the results using the RFH model reveal a higher *p-value* of 0.7, meaning there is a high probability both predicted and original values came from the same distribution i.e. the predicted abundances are trustworthy. The same does not apply for the RFA result, with a *p-value*=0.02, meaning that the distribution of the predicted abundances does not reflect well the distribution of the abundances of the trained data sets. We cannot make conclusions. Likewise, the results for [Mg/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] show small p-values, for both RFA and RFH models. RFA and RFH models have similar results. RFG was the model with worst results.

Overall, all models have better results when predicting individual abundances than when trying to predict the [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] abundance ratios. Since the RFG has consistently worse results, we will be using the RFA and RFH models in our analysis.

Additional tests performed to verify the consistency of the models can be found on Appendix F.

Chapter 3

Iron-to-silicate mass fraction in planet host stars

3.1 From stellar abundances to iron-to-silicate mass fraction

In chapter 2 we explained how to predict the abundances of the main rock forming elements (C, O, Mg, Si and Fe) which together with H and He are the most important to control the species expected from equilibrium condensation models, namely H_2, He, H_2O, CH_4, Fe, $MgSiO_3$, Mg_2SiO_4, and SiO_2 (Lodders 2003). There are more silicate forms, but the aforementioned ones ($MgSiO_3$, Mg_2SiO_4, and SiO_2) are the most relevant (Dorn et al. 2015). As such, the model we will be using includes only the mineral phases of the main rock-forming elements that predominate the crust, the upper and lower mantle, and the core of a terrestrial planet. A simplified model for the expected mass fractions of different compounds using these species is therefore a decent approach.

In this work, we adopted the stoichiometric model from Santos et al. (2017, 2015) to determine the composition of the planetary building blocks where the planets formed. In this model, when the ratio of the number of atoms of Mg and Si is between 1 and 2 ($1 < N_{Mg}/N_{Si} < 2$) Mg and Si are assumed to be equally distributed between olivine (Mg_2SiO_4) and pyroxene ($MgSiO_3$). This is true for most of the stars in the solar neighbourhood. Considering this, the molecular abundances and therefore the mass fraction can be found from the atomic abundances with simple stoichiometry:

$$
\begin{aligned}
N_O &= N_{H_2O} + 3N_{MgSiO_3} + 4N_{Mg_2SiO_4} \\
N_{Mg} &= N_{MgSiO_3} + 2N_{Mg_2SiO_4} \\
N_{Si} &= N_{MgSiO3} + N_{Mg_2SiO_4} \\
N_C &= N_{CH_4}
\end{aligned}
\tag{3.1}
$$

with N_X representing the number of atoms of each species X relative to hydrogen.

When $N_{Mg}/N_{Si} \leq 1$, it is assumed magnesium is incorporated only into pyroxene, leaving the remaining silicon to form mostly SiO_2. For this scenario, the stoichiometric relations would be:

$$
\begin{aligned}
N_O &= N_{H_2O} + 3N_{MgSiO_3} + 2N_{SiO_2} \\
N_{Mg} &= N_{MgSiO_3} \\
N_{Si} &= N_{MgSiO3} + N_{SiO_2} \\
N_C &= N_{CH_4}
\end{aligned}
\tag{3.2}
$$

In the rare case, when $N_{Mg}/N_{Si} \geq 2$, the Si would be embedded in olivine and the remaining Mg would be included in other minerals, mostly oxides. Although the model of Santos et al. (2017, 2015) does not take this case into consideration, it is not a problem for the current work, since the stars analysed in this work always have $N_{Mg}/N_{Si} < 2$.

It is important to mention that Equations 3.1 and 3.2 assume that the composition of H, He, C, N, O, Mg, Si, and Fe dominate the composition of the solid material, a good approximation since these elements constitute the main contributors to the overall composition in the Solar case (Asplund et al. 2009).

Stars with ratios $C/O > 0.8$ are expected to have a mineralogy significantly different than an Earth-like composition, with carbides forming instead of the silicates, leading to planet building blocks strongly enriched in carbon (Bond et al. 2010a) instead of having an Earth-like composition. Only 58 out of 18 021 stars from the GALAH sample were found with a C/O ratio above 0.8, which is consistent with observations that seem to indicate only a small fraction of the stars in the solar neighbourhood have $C/O > 0.8$ (Delgado Mena et al. 2010, Fortney 2012, Brewer and Fischer 2016). As this was the sample whose model performed worst and was later discarded, these stars were kept in the sample.

These simple relations allow to compute the expected mass fractions for the planets building blocks, namely the iron-to-silicate mass fraction (also designed by iron mass fraction, f_{iron}), the water mass fraction (w_f), and the summed mass percentage of all heavy elements (Z), as presented by Santos et al. (2017):

$$f_{iron} = m_{Fe}/(m_{Fe} + m_{MgSiO_3} + m_{Mg_2SiO_4} + m_{SiO_2}) \tag{3.3}$$

$$w_f = m_{H_2O}/(m_{H_2O} + m_{Fe} + m_{MgSiO_3} + m_{Mg_2SiO_4} + m_{SiO_2}) \tag{3.4}$$

$$Z = (m_{CH_4} + m_{H_2O} + m_{Fe} + m_{MgSiO_3} + m_{Mg_2SiO_4} + m_{SiO_2})/M_{tot} \tag{3.5}$$

where $m_X = N_X \times \mu_X X$, $M_{tot} = N_H \times \mu_H + N_{He} \times \mu_{He} + N_C \times \mu_C + N_O \times \mu_O + N_{Mg} \times \mu_{Mg} + N_{Si} \times \mu_{Si} + N_{Fe} \times \mu$ and μ_X stands for the mean molecular weight of an element X.

These quantities offer valuable information on parameters relevant for the planet formation process, such as the mass fraction of all heavy elements, the ice mass fraction in solid planets, or giant planet cores that form beyond the water-ice line (Dorn et al. 2017).

3.2 Determining the composition

As mentioned in Sections 2.5 and F.1, the RFG model is the one with worst performance and therefore, was discarded. As such, we will use only the predictions of RFA and RFH determine the composition of planet building blocks in the protoplanetary disk using the stoichiometric model of Santos et al. (2017, 2015). We further validate these prediction models on data from the HARPS sample (Section 3.2.1 and Section 3.3) and later on data from the APOGEE sample (Section 3.3).

3.2.1 HARPS data

Based on the chemical composition of stars, Adibekyan et al. (2012a) classified HARPS sample stars into different Galactic populations.

Following the chemical boundary addressed in Adibekyan et al. (2012a), the stars from the HARPS sample were classified into different galactic populations. We have 444 stars belonging to the thin disk with low [α/Fe] ratio and 81 stars with enhanced α-element abundances, which is characteristic of thick disk stars. Adibekyan et al. (2011) also identified a possibly new population of alpha-rich metal-rich ([Fe/H]> -0.2 dex) stars. In the literature there is no consensus whether this population is a separate one (Adibekyan

et al. 2011) or just constitutes the metal-rich tail of the Galactic thick disk (Bensby et al. 2014).

In this work, the alpha-rich stellar population is divided in 53 thick disk stars (with [Fe/H]< -0.2 dex) and 28 α-rich metal-rich stars (with [Fe/H]≥ -0.2 dex), hereon called hαmr. Finally, our sample contains 2 halo stars kinematically identified.

We start by applying the RFA and RFH models to the aforementioned. With the original abundances and the two sets of predictions, we apply the stoichiometric model to each of them.

Figure 3.1 presents the distribution of iron-to-silicate mass fraction, f_{iron} (top panels), water mass fraction, w_f (middle panels) and summed mass percent of all heavy elements, Z (bottom panels), derived with the stoichiometric model. Both the original and the predicted parameters show that the protoplanetary disk around the thick disk stars has the lowest amount of f_{iron}, and Z, and at the same time largest amount of w_f. The results also agree on the thin disk population having the largest amount of f_{iron}.

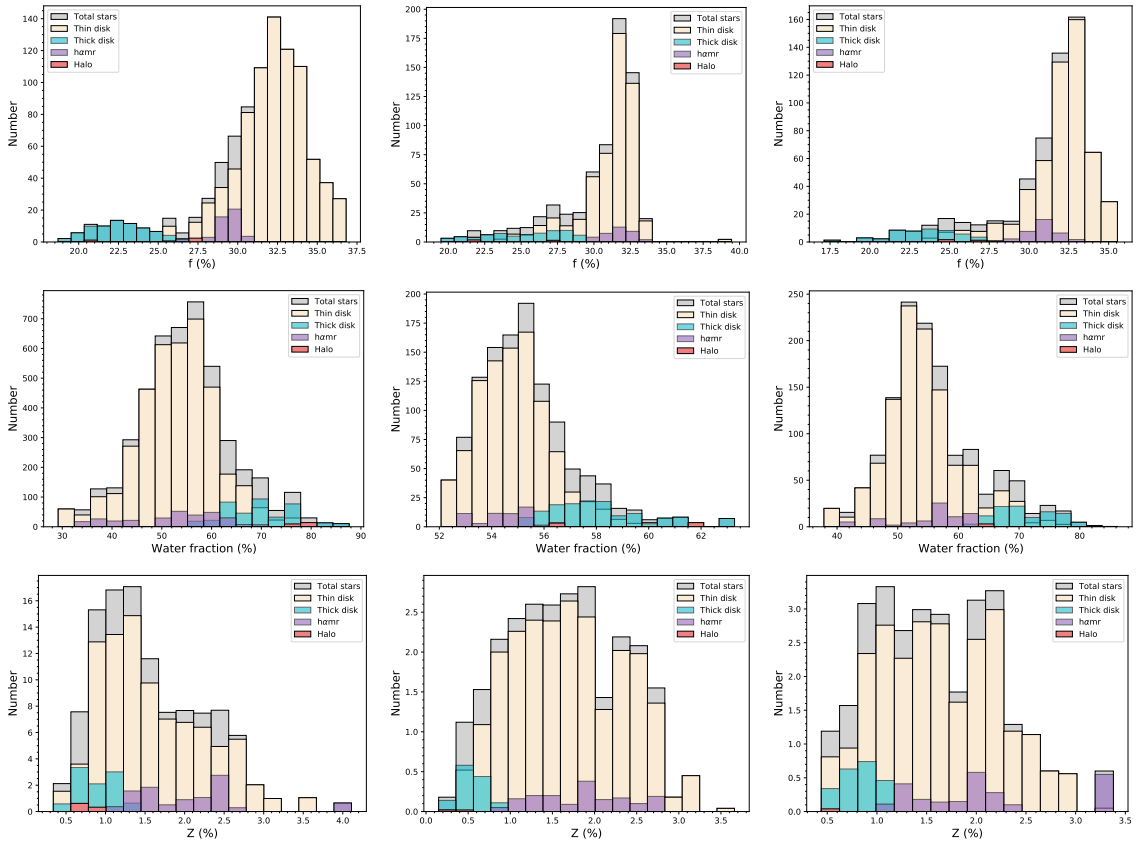


FIGURE 3.1: Composition for the HARPS sample (left column), RFA (middle column) and RFH predictions (right column). Top row is for the iron-to-silicate mass fraction, middle row for the water mass fraction and bottom row for the mass percentage of all heavy elements. Stars are distributed in galactic populations.

The model was applied to the original HARPS sample (left column), the RFA predictions (middle column) and RFH predictions (right column).

As Adibekyan et al. (2012b) discussed, low-metallicity thin-disc stars present an increasing α-element enhancement as [Fe/H] decreases. To test the impact of these α-enhancement on the composition of planet building blocks, in Figure 3.2 we compare the distributions of different parameter for metal-rich and metal-poor thin disk stars.

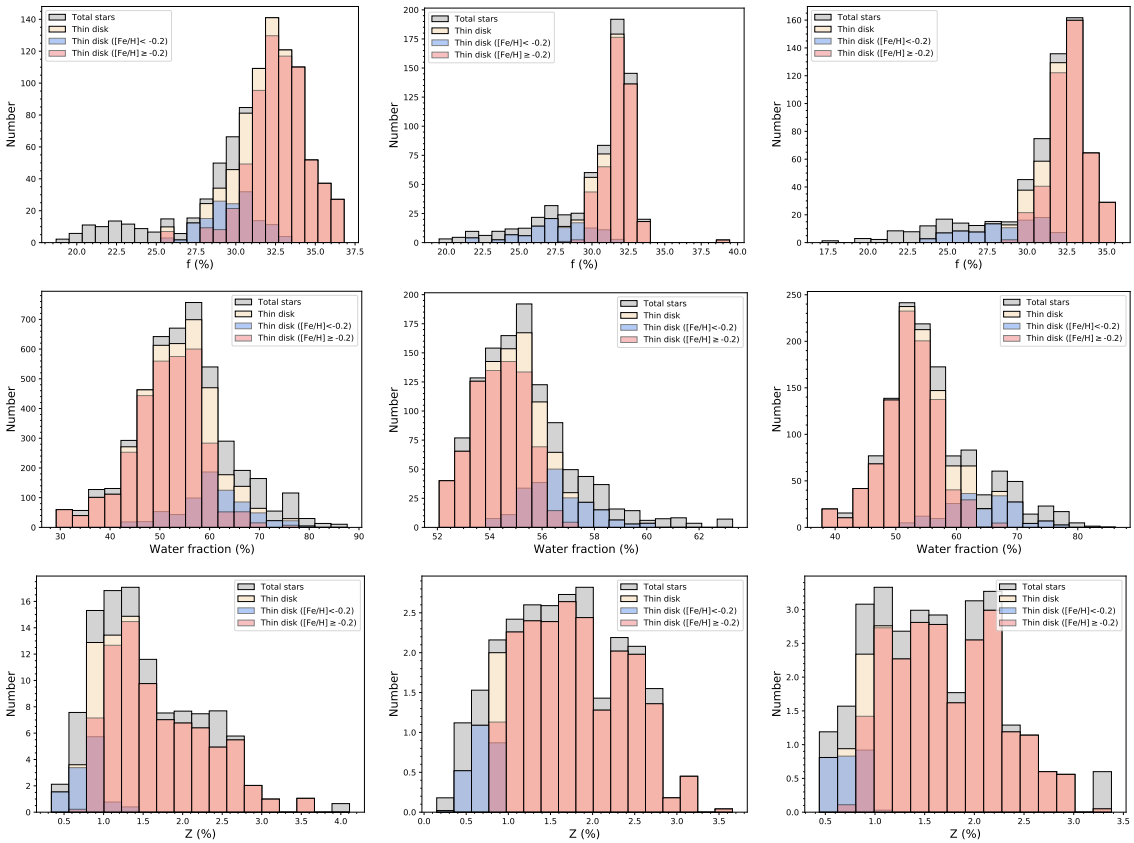


FIGURE 3.2: Composition for the HARPS sample (left column), RFA (middle column) and RFH predictions (right column). Top row is for the iron-to-silicate mass fraction, middle row for the water mass fraction and bottom row for the mass percentage of all heavy elements. Stars from the thin disk are divided in two groups.

For each stellar population, in Table 3.1 we provide the mean values and standard deviation of different parameters. The table shows that the f_{iron} based on the predictions of RFA and RFH results are on average very close to those determined from the abundances of the original HARPS sample. Likewise, the standard deviation values for both predictions and true data are similar. However, when one looks to the respective plots for the different galactic populations or just in the thin disk, the RFA distribution is much narrower than the RFH distribution and even more so than the HARPS data distribution, despite all the distribution peaks being around $\sim$ 32%.

TABLE 3.1: f_{iron} for original 525 HARPS stars (top), predictions done with RFA (left) and predictions done with RFH (right).

Population	Average	Std
f_{iron}		
Thin	32.17	1.94
Thin ([Fe/H]< -0.2)	29.64	1.56
Thin ([Fe/H]≥ -0.2)	32.75	1.50
Thick	22.72	1.78
hamr	29.04	1.22
Halo	24.17	3.08
Total	31.01	3.47

Population	Average	Std
f_{iron}		
Thin	31.02	2.05
Thin ([Fe/H]< -0.2)	27.79	2.30
Thin ([Fe/H]≥ -0.2)	31.77	0.99
Thick	25.28	2.75
hamr	31.54	0.86
Halo	24.58	2.47
Total	30.45	2.75

Population	Average	Std
f_{iron}		
Thin	31.85	2.06
Thin ([Fe/H]< -0.2)	28.75	2.22
Thin ([Fe/H]≥ -0.2)	32.57	1.16
Thick	23.52	2.16
hamr	30.74	1.21
Halo	25.41	1.01
Total	30.93	3.24

As expected, thin disk stars have a higher iron-to-silicate mass fraction, due to them being younger stars, born when the interstellar medium had already been enriched with iron through SN Ia explosions. On the other hand, thick disk stars have higher water mass fraction. This is because the thick disk stars are old and have higher abundances of α-elements (that includes O) and lower metallicity.

For the w_f values, the distributions for the galactic populations all show a peak around $\sim$ 54%, despite the RFA distribution again being narrower than the original or RFH distributions. Furthermore, looking at Table 3.2 we notice that the average and standard deviation values for RFA are smaller than expected in some populations (metal-poor stars of the thin disk, thick disk and halo).

It is interesting to see that the w_f distribution peak for the thick disk and for the metal poor thin disk stars changed on the RFA results (see distribution in Figure 3.1 and average value in Table 3.2). For the original data and RFH results, the thick disk has a distribution peak around $\sim$ 70%, but on the RFA result it is around $\sim$ 58% instead. Similarly, for the metal poor thin disk stars the distributions of HARPS data and RFH sample reaches its maximum values at $\sim$ 66%, but for RFA results this value changes to $\sim$ 56%.

TABLE 3.2: w_f for original 525 HARPS stars (top), predictions done with RFA (left) and predictions done with RFH (right).

Populations	Average	Std
w_f		
Thin	53.84	7.36
Thin ([Fe/H]< -0.2)	61.60	6.79
Thin ([Fe/H]≥ -0.2)	52.04	6.24
Thick	71.55	6.72
hamr	56.15	8.22
Halo	78.78	2.30
Total	55.84	9.17

Population	Average	Std
w_f		
Thin	54.81	1.32
Thin ([Fe/H]< -0.2)	56.50	1.23
Thin ([Fe/H]≥ -0.2)	54.42	0.99
Thick	57.98	1.84
hamr	54.69	0.99
Halo	59.39	2.67
Total	55.14	1.69

Population	Average	Std
w_f		
Thin	55.03	6.03
Thin ([Fe/H]< -0.2)	63.37	5.26
Thin ([Fe/H]≥ -0.2)	53.10	4.31
Thick	71.78	5.76
hamr	56.03	4.66
Halo	73.39	8.40
Total	56.85	7.86

By plotting on Figure 3.3 the cumulative distribution of the w_f of all results, for the thin disk stars (left panel) and for the thick disk stars (right panel), we can observe how regardless of the galactic population, the water mass fraction for RFA predictions is restricted to a narrow range of values. The RFA models expects stars from the thin disk to have only $50 - 60\%$ of water mass fraction and $55 - 64\%$ of water mass fraction for stars in the thick disk. The cumulative function for the predictions from the RFA models has an abrupt increase (around $\sim$ 50% for thin disk and $\sim$ 55% for thick disk) in the likelihood of occurrence, which reinforces the fact that for the RFA model, the w_f distribution is very narrow when compared to those of RFH or the original data. On the other hand, the water mass fraction of RFH predictions follow a more smooth growth, with w_f values across the whole range, a behaviour much more similar to what is expected, based on the true data sample.

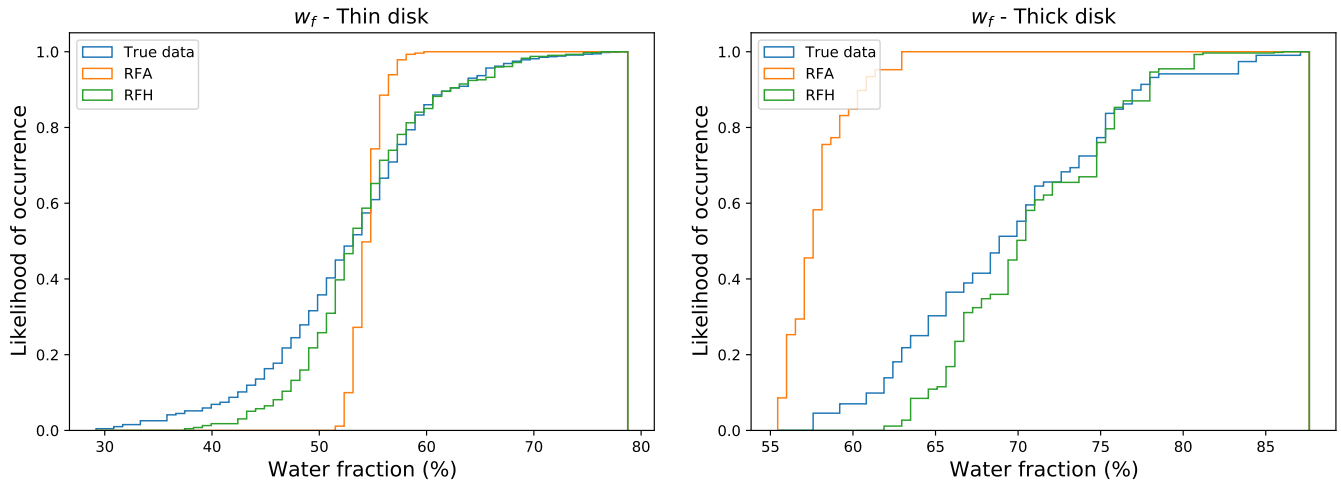


FIGURE 3.3: Cumulative distribution function of the w_f for the three set of results: the true (original) data and the predictions for the RFA and RFH models. Left panel is the w_f value of thin disk stars and right panel is the w_f for thick disk stars.

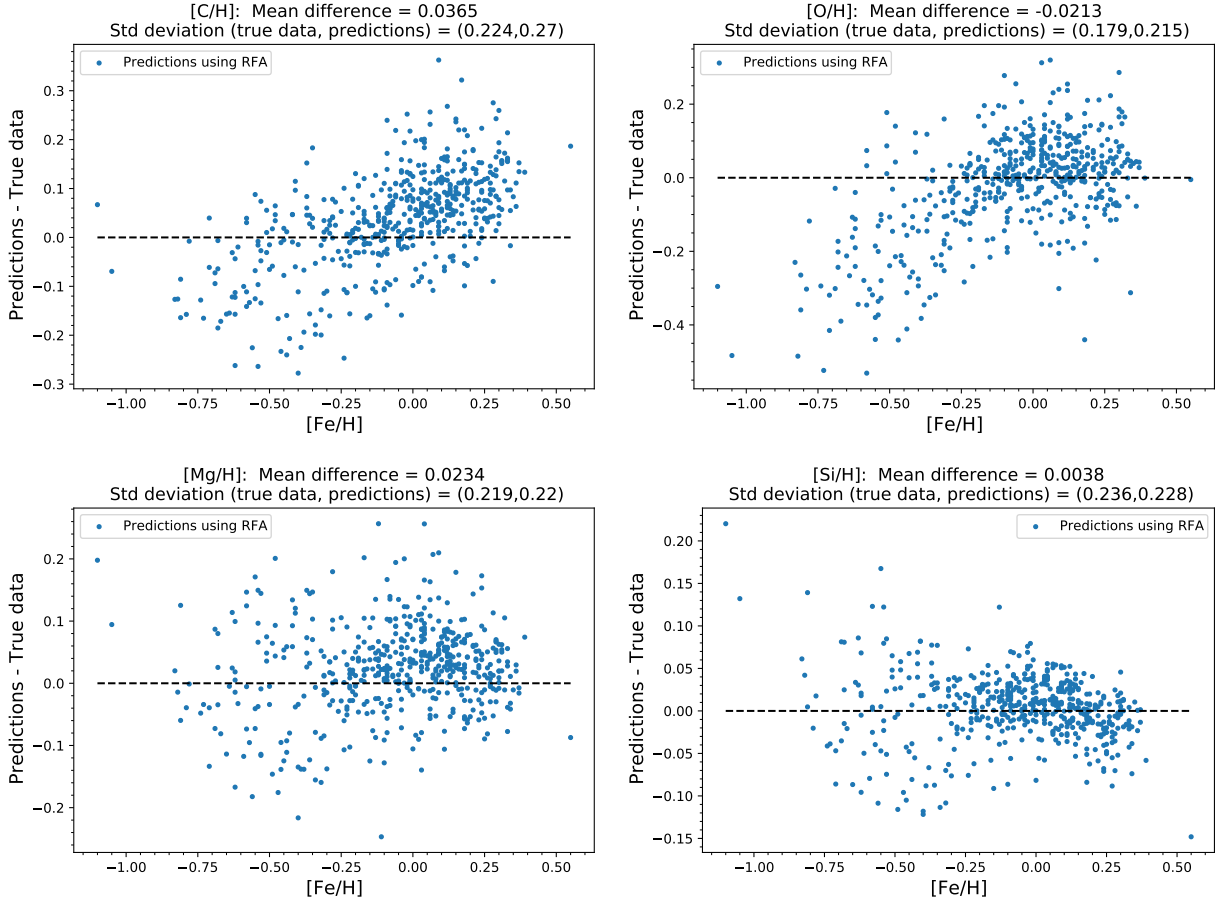


FIGURE 3.4: Predictions of the HARPS data using the APOGEE based random forest, RFA. Y axis is the difference between predicted and true data while on the X axis are the metallicity values. Title includes information on the mean difference and standard deviation of true data and predictions.

TABLE 3.3: Z for original 525 HARPS stars (top), predictions done with RFA (left) and predictions done with RFH (right).

Population	Average	Std
Z		
Thin	1.37	0.57
Thin ([Fe/H]< -0.2)	0.74	0.17
Thin ([Fe/H]≥ -0.2)	1.52	0.53
Thick	0.85	0.24
hamr	1.91	0.62
Halo	0.84	0.07
Total	1.35	0.59

Population	Average	Std
Z		
Thin	1.44	0.62
Thin ([Fe/H]< -0.2)	0.67	0.15
Thin ([Fe/H]≥ -0.2)	1.62	0.55
Thick	0.50	0.15
hamr	1.59	0.49
Halo	0.35	0.03
Total	1.35	0.65

Population	Average	Std
Z		
Thin	1.35	0.52
Thin ([Fe/H]< -0.2)	0.75	0.15
Thin ([Fe/H]≥ -0.2)	1.49	0.48
Thick	0.76	0.19
hamr	1.66	0.50
Halo	0.52	0.04
Total	1.31	0.54

This difference for the RFA values is likely related to the predicted abundances of [O/H]. In Figure 3.4, where we show the difference between the predicted and expected abundances of oxygen, we see that stars with metallicity [Fe/H] < -0.2 tend to have a predicted abundance lower than the true value. Although the number of metal-poor stars is not as large as the number of metal-rich ones, stars with very low metallicity e.g. [Fe/H] < -0.6 dex can have a difference of up to -0.5 dex from what was expected. This explains why the w_f distribution peak for metal poor thin disk and thick disk stars is lower than the peak for the true data. Note that thick disk stars are also metal poor, as their metal rich counterpart comprises the hαmr group.

Table 3.3 and Figures 3.1 and 3.2 clearly suggest that both RFA and RFH models predict results well compatible with the expected values.

Taking into consideration all the results mentioned, we conclude that for the 3 parameters discussed in this section (f_{iron}, w_f and Z) the RFH model works better than the RFA model.

3.3 Additional tests

To complement the tests done in Section 3.2.1, we proceed to do two additional tests.

From the HARPS sample we selected the stars with metallicities between -0.5 and 0.3. The density of stars in this metallicity range is relatively high. Then we reapply the stoichiometric model to this data set. The results summarised in Tables 3.4, 3.5 and 3.6 lead to conclusions consistent with those presented in Section 3.2.1: overall, the predicted parameters based on the RFH model are in better agreement with the original values, when compared with those predicted with the RFA model.

TABLE 3.4: f_{iron} for smaller subset of 455 HARPS stars (top), predictions done with RFA (left) and predictions done with RFH (right).

f_{iron}	Average	Std
Thin	32.33	1.79
Thin ([Fe/H] < -0.2)	30.07	1.31
Thin ([Fe/H] ≥ -0.2)	32.77	1.52
Thick	23.89	1.34
hamr	29.00	1.22
Total	31.71	2.63

f_{iron}	Average	Std
Thin	31.20	1.58
Thin ([Fe/H] < -0.2)	28.58	1.68
Thin ([Fe/H] ≥ -0.2)	31.71	0.91
Thick	27.48	1.44
hamr	31.51	0.86
Total	31.03	1.75

f_{iron}	Average	Std
Thin	32.05	1.77
Thin ([Fe/H] < -0.2)	29.38	1.93
Thin ([Fe/H] ≥ -0.2)	32.56	1.18
Thick	25.08	1.31
hamr	30.71	1.23
Total	31.61	2.31

TABLE 3.5: w_f for smaller subset of 455 HARPS stars (top), predictions done with RFA (left) and predictions done with RFH (right).

w_f	Average	Std
Thin	53.56	6.92
Thin ([Fe/H] < -0.2)	60.54	5.73
Thin ([Fe/H] ≥ -0.2)	52.21	6.29
Thick	67.52	5.26
hamr	55.71	7.84
Total	54.40	7.56

w_f	Average	Std
Thin	54.77	1.12
Thin ([Fe/H] < -0.2)	56.11	0.95
Thin ([Fe/H] ≥ -0.2)	54.51	0.95
Thick	56.82	0.99
hamr	54.76	0.93
Total	54.87	1.19

w_f	Average	Std
Thin	54.69	5.49
Thin ([Fe/H] < -0.2)	62.29	4.30
Thin ([Fe/H] ≥ -0.2)	53.22	4.35
Thick	67.42	3.51
hamr	55.77	4.56
Total	55.40	6.03

TABLE 3.6: Z for smaller subset of 455 HARPS stars (top), predictions done with RFA (left) and predictions done with RFH (right).

Z	Average	Std
Thin	1.35	0.50
Thin ([Fe/H] < -0.2)	0.79	0.14
Thin ([Fe/H] ≥ -0.2)	1.46	0.47
Thick	1.02	0.18
hamr	1.83	0.44
Total	1.36	0.51

Z	Average	Std
Thin	1.41	0.53
Thin ([Fe/H] < -0.2)	0.72	0.12
Thin ([Fe/H] ≥ -0.2)	1.55	0.47
Thick	0.64	0.09
hamr	1.55	0.44
Total	1.38	0.54

Z	Average	Std
Thin	1.32	0.44
Thin ([Fe/H] < -0.2)	0.79	0.12
Thin ([Fe/H] ≥ -0.2)	1.43	0.41
Thick	0.91	0.16
hamr	1.60	0.39
Total	1.32	0.44

A similar analysis was done for APOGEE data. From the whole sample, we randomly selected 500 stars and applied the RFA and RFH models to predict the abundance, which were then inserted in the stoichiometric model to determine the composition of the protoplanetary disk. Then, using the stoichiometric model, we determine the composition. APOGEE data are not divided into different Galactic populations and we used the full data set as a single sample. The results of this test are shown in Figure **??** and are summarised in Table 3.7. They show that there is a good agreement between the average values of predictions and true data, for three parameters (f_{iron}, w_f, and Z), despite the distributions of the predicted parameters having lower standard deviation values when compared with the ones of the original distributions.

TABLE 3.7: Average values and standard deviations for f_{iron}, w_f and Z values, for the original data and predictions made with RFA and RFH models.

	Average	Std	p-value
f_{iron}			
Original	29.25	3.73	—
RFA	29.32	3.12	0.0955
RFH	29.32	3.12	0.0955
w_f			
Original	55.96	2.70	—
RFA	55.89	2.12	0.0416
RFH	55.82	2.14	0.0587
Z			
Original	1.01	0.54	—
RFA	0.997	0.52	0.7189
RFH	0.999	0.52	0.6659

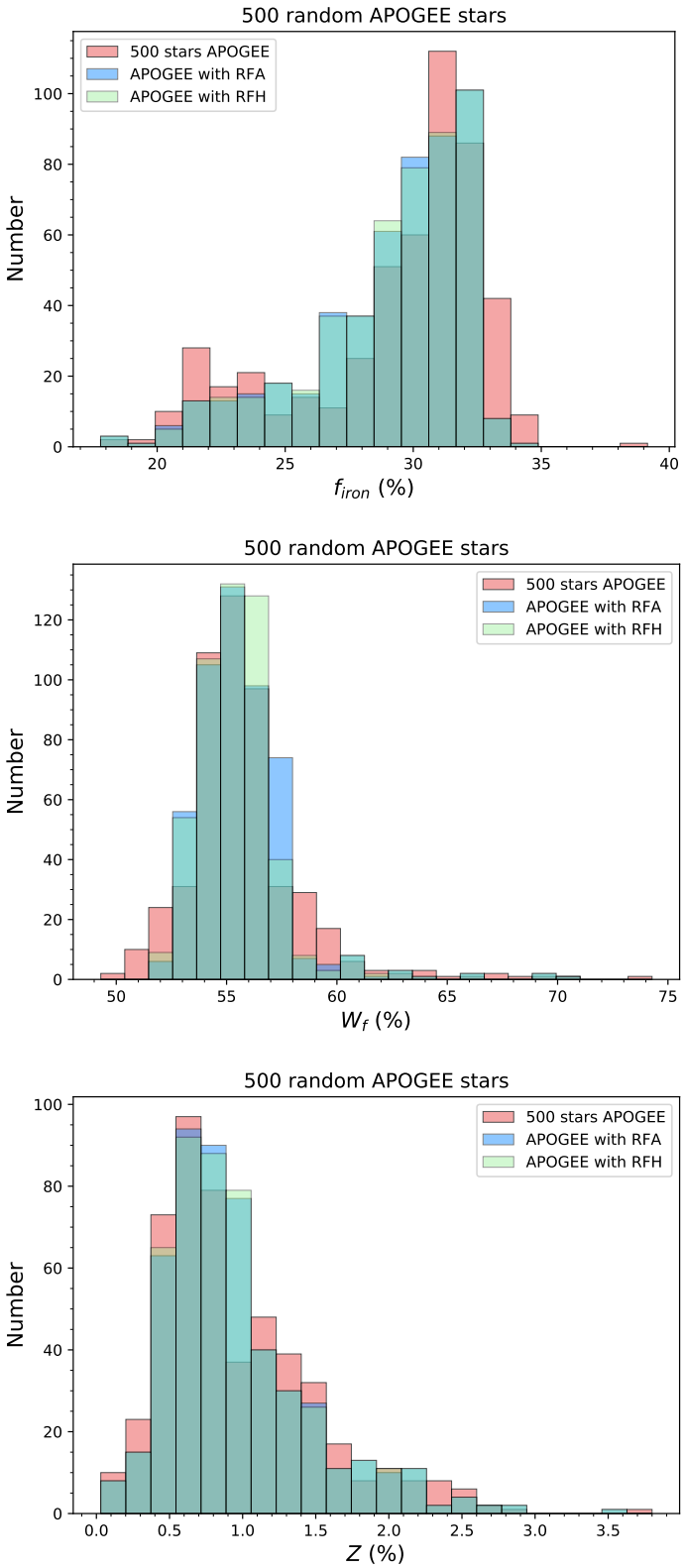


FIGURE 3.5: Distribution of f_{iron} (top), w_f (middle) and Z (bottom) values, for 500 random stars from APOGEE and predictions of RFA and RFH models.

Chapter 4

Composition of low-mass planets and their host stars

4.1 Selecting the sample

After having tested the prediction models and the stoichiometric model on different data from various authors, we now apply them on our sample of stars, among which some are M dwarfs, taken from the NASA Exoplanet Archive[1] (for short, NEA).

The NEA is an online catalogue providing information about the exoplanets and their host stars, namely stellar parameters (like temperatures, surface gravity, positions and magnitudes), exoplanets parameters (such as masses and orbital parameters) and discovery/characterisation data. The catalogue contains invaluable information for the study and characterisation of extra-solar planets and their host stars.

From NEA, we selected planets with mass and radius measurements better than 30%. To ensure we only have terrestrial planets and avoid the sub-Neptune regime, we restrict the sample to planets with $M < 10M_{\oplus}$ and $R < 2R_{\oplus}$ (Van Eylen et al. 2018). Finally, we want to include only RV-based mass measurements by rejecting mass estimations based on mass-radius empirical relations and Transit-Time-Variations-based mass determinations, since Mills and Mazeh (2017) showed that planets with masses from the TTV and RV methods reveal different mass distributions.

However, unlike The Extrasolar Planets Encyclopaedia[2] (for short, exoplanet.eu), the NEA does not have a specific column for the mass determination method. Therefore, from

[1] https://exoplanetarchive.ipac.caltech.edu/
[2] exoplanet.eu

a sample of 20 stars with planets from Adibekyan et al. (2021) taken from exoplanet.eu and with RV-based mass measurements, we analyse each star case by case to try and narrow down the common information between stars with RV-based mass. We pay special attention to the mass related columns: `Planet Mass` provides information on the mass of the planet, `Planet Mass*sin(i)` is the minimum mass of a planet as measured by RV (the *i* stands for inclination). `Planet Mass*sin(i)/sin(i)` is the column indicating the quotient of the lower limit of the measured mass planet (mass*sin(i)) and the sine of its inclination (sin(i)). This column is specific to references where the inclination and minimum mass is provided but the true mass is not reported. The column `Planet Mass or Mass*sin(i)` provides the best planet mass estimation available, in order of preference (Mass, Mass*sin(i)/sin(i) or Mass*sin(i)), depending of availability. The `Planet Mass or Mass*sin(i) Provenance` column gives information on the provenance of the measurement of the best mass. Besides mass related columns, we also look at the `Inclination` column, which is the angle between the plane of the orbit and the plane perpendicular to the line-of-sight of an observer on Earth.

From 20 stars, 19 had the same common information: they had `Planet Mass` values but not `Planet Mass*sin(i)` neither `Planet Mass*sin(i)/sin(i)` (empty columns). The `Planet Mass or Mass*sin(i) Provenance` column displayed the `Mass` option and the stars had information on the `Planet Mass or Mass*sin(i)` and `Inclination`. In this manner, we constrain the selection to stars that satisfy this column arrangement. We note that this is not a foolproof criteria, as there may be stars with RV-based mass that do not fit this condition. Yet, it is still better than simply looking at stars discovered with the RV methods. Regardless, all the stars were crosschecked with exoplanet.eu to make sure their masses were determined with the RV method. This leaves us with a sample of 39 planets orbiting 36 stars, among which 10 are M dwarfs and 26 are FGK stars.

To predict the stellar abundances of the rock forming elements, our RF model requires stellar metallicity and total galactic velocity. The SWEET-Cat catalogue (Santos et al. 2013, Sousa et al. 2021) provides information about the metallicities of the stars. The parameters (parallax, proper motion, coordinates, and radial velocities) required to determine the space velocity components we extracted from NEA.

We inserted the metallicities and total velocities of the stars into the RFA and RFH models to predict the individual abundances. Since these are abundances relative to the Sun, they are transformed in absolute abundances following the solar reference values

given by Asplund et al. (2009) for Fe ($\log\epsilon(Fe) = 7.50$), Mg ($\log\epsilon(Mg) = 7.60$) and Si ($\log\epsilon(Si) = 7.51$). Suárez-Andrés et al. (2017) and Bertran de Lis et al. (2015) provide a solar composition for C ($\log\epsilon(C) = 8.50$) and O ($\log\epsilon(O) = 8.71$), respectively. For He, it was adopted the value of $\log\epsilon(He) = 10.93$ from Lodders (2003). As already mentioned in Section 2.2, the logarithmic number of H atoms is traditionally set to $\log\epsilon(H) = 12$.

With the predicted abundances we proceed to apply the stoichiometric model to determine the composition of the protoplanetary disk, assuming its overall chemical composition is reflected in the composition of the stars formed there.

The uncertainties in the f_{iron} for each star were determined through the Monte Carlo method. Considering a Gaussian distribution with the derived abundance as the mean and the derived uncertainties of the abundances as the standard deviation, we randomly drew 10 000 values of abundances. The uncertainty corresponds to the standard deviation of these randomly drawn abundances.

The distributions of the f_{iron}, w_f and Z are shown on Figure 4.1, and the mean values and standard deviations of these distributions are presented in Table 4.1

TABLE 4.1: Average values and standard deviations for f_{iron}, w_f, and Z mass fractions for the 36 stars from the NEA sample, for both prediction models (RFA and RFH).

	Average	Std
RFA		
f_{iron}	31.1	2.2
w_f	54.3	1.2
Z	1.5	0.7
RFH		
f_{iron}	31.4	3.1
w_f	55.0	7.4
Z	1.4	0.6

Except for the overall amount of solid material (Z), the standard deviations values for the RFA results are significantly lower than those of RFH, pointing to the fact that RFA composition distributions are narrower, particularly for the water mass fraction, which is consistent with the results presented in Section 3.2.1 and 3.3

In fact, when looking at the w_f distribution on Figure 4.1, we see that RFA results have water mass fraction between $\sim 52\% - 57\%$ with an average of $54.3 \pm 1\%$, while RFH results have a wider range ($\sim 42\% - 73\%$) with an average of $55.0 \pm 7\%$. The mean values determined with both models are similar and are slightly lower than the $w_f \sim 67\%$ of the

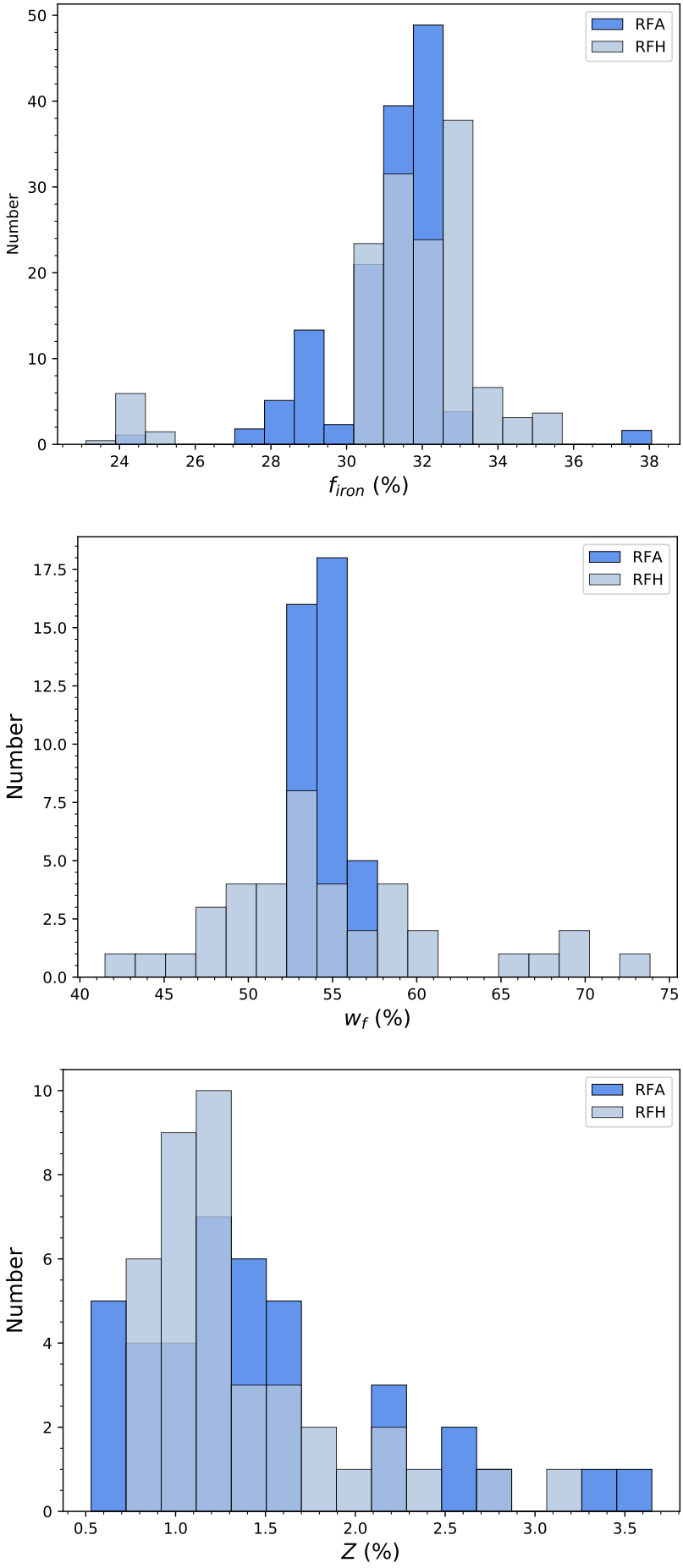


FIGURE 4.1: f_{iron} (top), w_f (middle), and Z (bottom) mass fractions distributions for the 36 stars from the NEA sample.

planet building blocks for the Solar System, derived by Lodders (2003). Using the same stoichiometric model as this work, Santos et al. (2017) predicted $w_f = 60\%$ for the Solar System, a value closer to what we found.

The values estimated for Z with both models are in agreement with each other, with most stars having $Z \sim 1 - 1.5\%$. The average values are $1.5 \pm 0.7\%$ and $1.4 \pm 0.6\%$ for RFA and RFH, respectively, and are consistent with Lodders's value ($Z \sim 1.3\%$) and Santos et al.'s value ($Z = 1.26\%$) for the Solar System

For the f_{iron}, the results are also very similar, with an average of $f_{iron} \sim 31.1 \pm 2\%$ for RFA model and $f_{iron} \sim 31.4 \pm 3\%$ for RFH model. These values are consistent with the iron content of the telluric planets of the Solar System, where $f_{iron}^{Venus} \sim 32\%$ (Aitta 2012), $f_{iron}^{Earth} \sim 32\%$ (McDonough 2003) and $f_{iron}^{Mars} \sim 30\%$ (Khan et al. 2018). The only exception is Mercury, with a $f_{iron}^{Mercury} \sim 70\%$ (Hauck et al. 2013). Santos et al. (2017) predicted a $f_{iron} = 33\%$ for the building blocks of the Solar System, which is also consistent with the values obtained.

As we want to look at the composition of smaller rocky planets orbiting M dwarfs, we need to analyse the results specifically only for the 10 M dwarfs and their 10 planets. Although the sample size is very small, we also performed a Kolmogorov–Smirnov (KS) and Anderson-Darling (AD) tests to check if both predictions, though coming from different models, could have similar distributions, for these M dwarfs.

As briefly explained in Section 2.5, the KS test verifies if two samples could be drawn from the same distribution, which would mean having $p\text{-}value > 0.05$ for a higher probability of both samples being taken from the same parent. The K-sample Anderson-Darling (AD) test verifies if different samples can be modelled as coming from a single population, with no need to specify the distribution function. If the AD statistic is higher than the critical values, for different levels, then the null hypothesis that two samples came from the same distribution can be rejected. Table 4.2 summarises the average value and standard deviations, as well as the p-value for the KS test and the statistic of the AD test. The histograms with the distributions can be seen in Figure 4.2.

For the f_{iron}, we have an average value of $31.9 \pm 2\%$ for predictions from the RFA model and $32.2 \pm 1\%$ for the RFH model. Once again, these values are consistent with the iron content of the rocky planets of the Solar System, namely for Venus ($\sim 32\%$), Earth ($\sim 32\%$) and Mars ($\sim 30\%$).

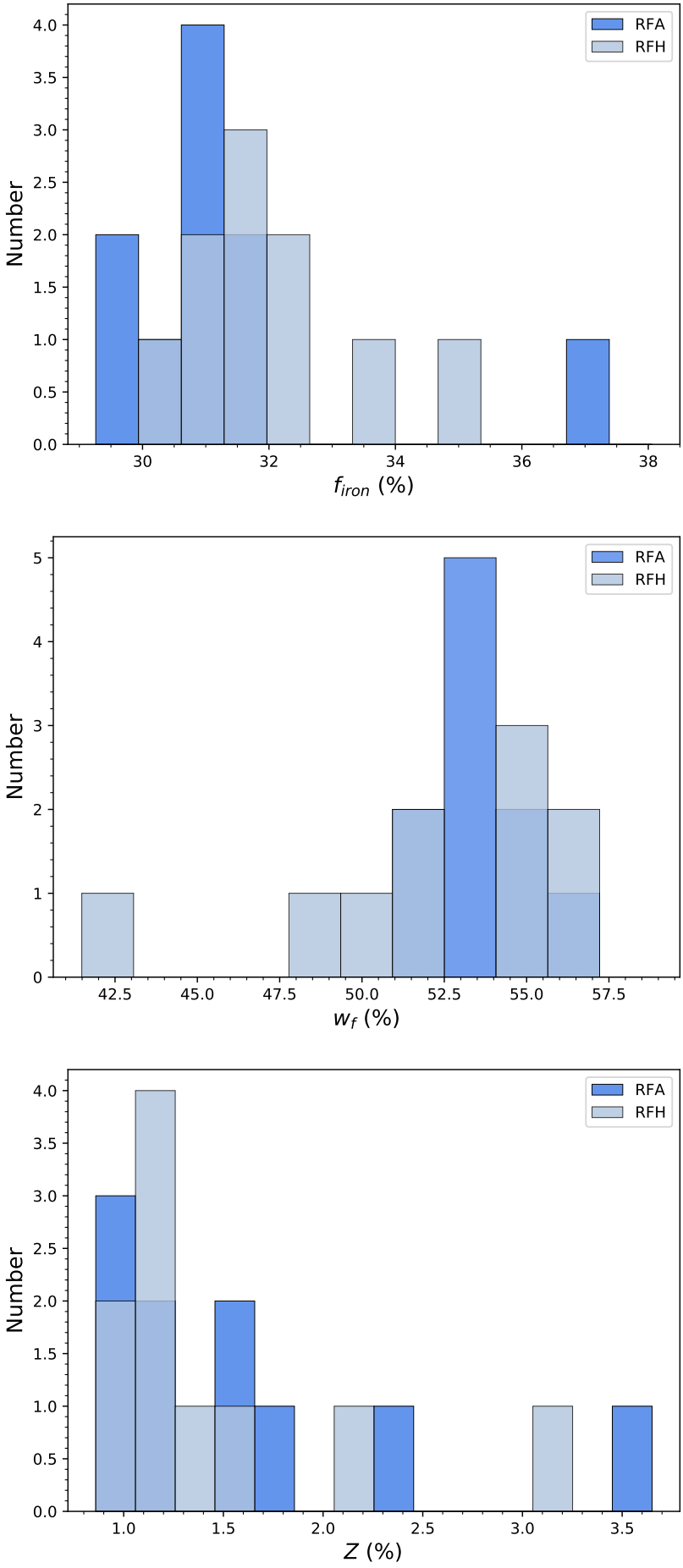


FIGURE 4.2: f_{iron} (top), w_f (middle), and Z (bottom) mass fractions distributions for the 10 M dwarf stars from the NEA sample.

TABLE 4.2: Average values, standard deviations and KS and AD test results, for f_{iron}, w_f, and Z mass fractions of the 10 M dwarfs, for both RFA and RFH prediction models. The levels associated to the critical values are 25%, 10%, 5%, 2.5%, 1%, 0.5% and 0.1%.

	Average	Std	p-value	stats	critical values	significance level
RFA						
f_{iron}	31.9	2.3	0.7869	-0.374	[0.325 1.226 1.961 2.718 3.752 4.592 6.546]	0.25
w_f	53.9	1.3	0.4175	0.668	[0.325 1.226 1.961 2.718 3.752 4.592 6.546]	0.175
Z	1.6	0.8	0.9945	-0.883	[0.325 1.226 1.961 2.718 3.752 4.592 6.546]	0.25
RFH						
f_{iron}	32.2	1.4	0.7869	-0.374	[0.325 1.226 1.961 2.718 3.752 4.592 6.546]	0.25
w_f	53.1	4.7	0.4175	0.668	[0.325 1.226 1.961 2.718 3.752 4.592 6.546]	0.175
Z	1.5	0.7	0.9945	-0.883	[0.325 1.226 1.961 2.718 3.752 4.592 6.546]	0.25

The water mass fraction distribution for the RFA remains narrower, as we can see on the middle panel of Figure 4.2, where RFH results spread between $\sim 41\% - 57\%$ but RFA's are all between $\sim 52\% - 56\%$. The stoichiometric model predicts an average $w_f \sim 53.9 \pm 1\%$ and $\sim 53.1 \pm 5\%$, for the RFA and RFH abundances, respectively.

These values are slightly smaller than that of the Solar System ($\sim$ 67% as determined by Lodders (2003), 60% as Santos et al. (2017) predicted), i.e the planet building blocks of these planetary systems had less water content and thus the rocky planets orbiting M dwarfs might have less water available to be incorporated in their composition. A lower water mass fraction could affect the evolution and habitability of a planet (Noack et al. 2017). Compared to heavier species (like silicates and iron) that constitute the main building blocks of the inner part of the proto-planetary disc, water accumulates in the outer regions during the later condensation of volatile atoms (Ward 1997, Luger et al. 2015, Santos et al. 2017). As such, objects formed in the outer disk are expected to have high water content in their composition. Moreover, the amount of water available affects the amount of condensates beyond the ice line so that it is expected to be related to core or planet formation efficiency (Santos et al. 2017).

The results for the overall amount of solid material (Z) are very similar for both RFA and RFH models, with an average value of $Z \sim 1.6 \pm 0.8\%$ and $Z \sim 1.5 \pm 0.7\%$, respectively. Moreover, these values are also very similar to the $Z \sim 1.3\%$ of the Solar System determined by Lodders (2003). This is also true for $Z = 1.26\%$, as predicted by Santos et al. (2017).

Recent results by Adibekyan et al. (2021) have shown that one can expect a correlation between the planet density and f_{iron} of the planets building blocks, for rocky planets, as inferred from the host star abundances (see Figure 2 from Adibekyan et al. 2021). The

higher the iron-to-silicate mass fraction of the protoplanetary disk, the higher is expected to be the density of the planet, since higher iron content in the planets means a larger core.

In order to check if this correlation is also present in our sample, we plot the planet density ρ depending on f_{iron}. It is, however, important to notice that compression can lead to planets with the same composition but different masses having different densities, since the more massive planet will feel a greater compression effect and hence, have a higher density. To account for this effect, we normalised the densities to the density of a planet with Earth-like composition (Dorn et al. 2015). In Figure 4.3 we show our results for the RFA (top panel) and RFH (bottom panel) abundance predictions by separating the M dwarfs (red dots) from the solar-type stars (blue dots).

Even for the solar-type stars, the planet densities do not show a clear correlation with the estimated f_{iron}. For a given iron-to-silicate mass fraction, there is a range of planet density values. The lack of the observed correlation can be related to the fact that instead of homogeneously determining accurate abundances for the sample stars, we predicted them with RF models. As such, the dispersion in planetary densities observed could be due to not enough precision when estimating the abundances, leading to large uncertainties when determining the composition. As the figure shows, the f_{iron} has large error bars, preventing a possible detection of the expected trend. In fact, Adibekyan et al. (2021) stressed the importance of precise abundance determinations when dealing with the compositions of low-mass planet hosts. Alternatively, this dispersion could have an astrophysical explanation, since it is possible that in protoplanetary disks with similar f_{iron}, planets with different densities can be formed owning to specific processes during the formation of planets. An example of this is the case of Mercury.

The rocky planets of the Solar Systems formed in the same disk, but Mercury is known for having a higher bulk density, which implies that the planet's average composition is more enriched in metallic iron than the other telluric planets of the Solar System. Mercury's massive iron core that represents a core mass fraction of $\sim 69 - 77\%$ (Hauck et al. 2013) and the planet's smaller size suggest that processes like evaporation, mantle-stripping collision between protoplanets or interactions with the young Sun's intense solar wind (Benz et al. 1988, 2007, Asphaug and Reufer 2014, Spalding and Adams 2020) might have occurred during the planetary formation. This led to Mercury losing part of its mantle and consequently having a higher density and a very large core when compared with the rest of the rocky planets of the Solar System (Clement and Chambers 2021).

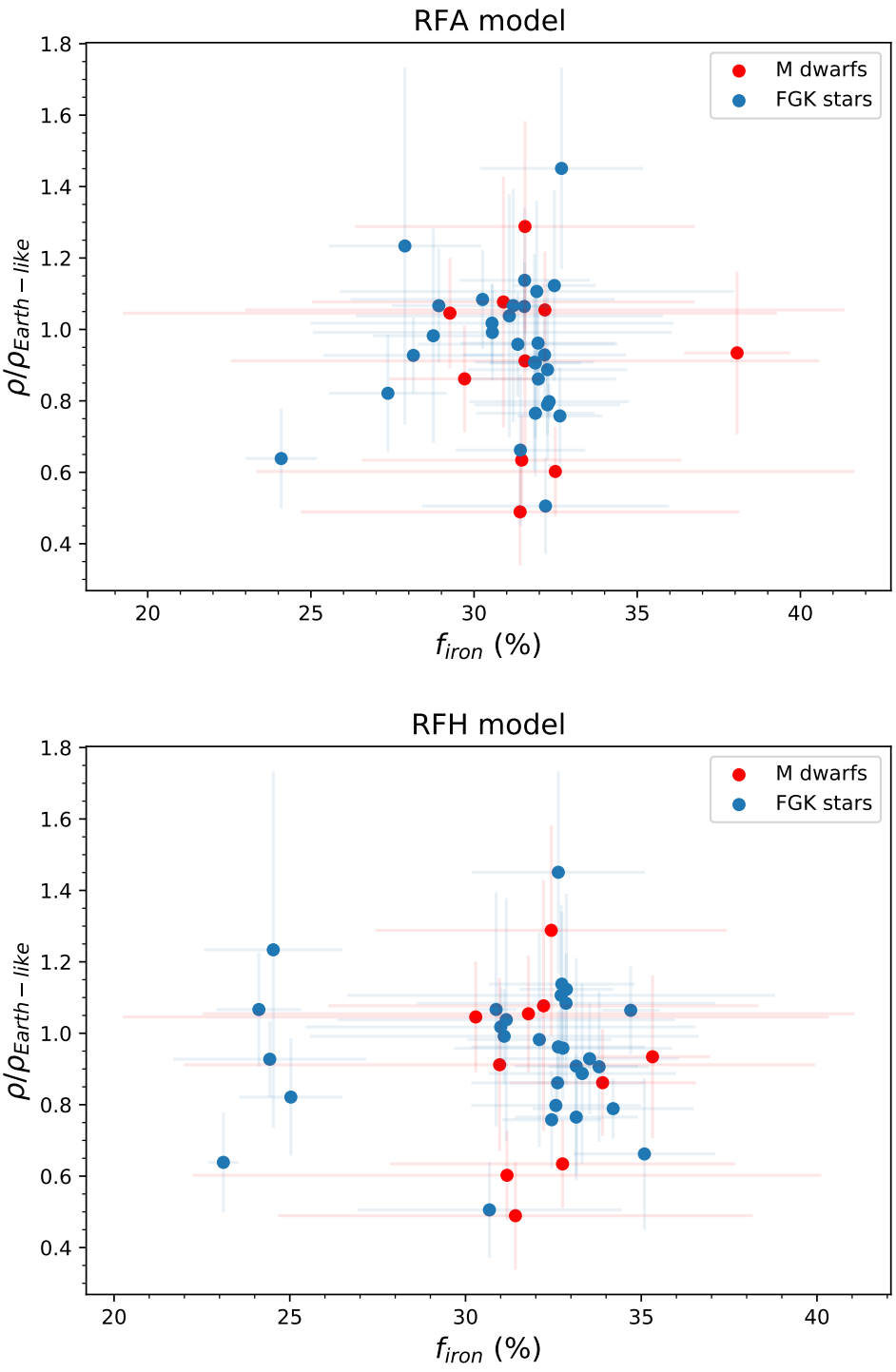


FIGURE 4.3: Densities of rocky planets as a function of iron-to-silicate mass fraction. The measured density of each planet was normalised by the expected density of an Earth-like composition for the same mass. The iron-to-silicate mass fraction was calculated from the individual abundances of the host star, using RFA (top panel) and RFH (bottom panel) predictions. Red dots represent planets orbiting M dwarfs and blue dots represent planets around FGK stars.

Therefore, in order to study exoplanetary systems and to better understand planet formation processes, it is necessary but not sufficient to know the composition of the protoplanetary disk where the star(s) and planet(s) formed. This composition is reflected on stellar atmospheres and planetary composition and has an impact during planet formation.

Chapter 5

Conclusions and future work

In this thesis, we focus on determining the chemical composition of planets orbiting M dwarfs. M dwarfs comprise more than 70% of the stars in the Galaxy and are excellent candidates to search for Earth-like planets in their HZ. In this regard, an accurate characterisation of these stars is crucial to the study of exoplanets orbiting them, particularly in determining their composition and internal structure. Unfortunately, it is very difficult to determine the composition of individual elements for M dwarfs. Because of their cooler atmospheres, their spectra contains millions of molecular lines which makes it very difficult to identify and measure the chemical abundances of the elements present.

To get around this problem, we start by building random forest-based models to predict the chemical abundances of stars from their metallicity and kinematics. We used three different spectroscopic surveys (APOGEE – Deshpande et al. 2013, GALAH – De Silva et al. 2015, and HARPS – Mayor et al. 2003) to select the stars with the highest quality data and then to build corresponding models (RFA, RFG and RFH). The kinematic properties of the stars, in particular the total space velocities, are determined by using the Gaia DR2 astrometric data (Randich et al. 2013).

By testing the models on external data, we concluded that the RFG model yields the worst predictions. By performing different tests, we concluded that the RFH and RFA models are able to predict abundances of the main rock forming elements (C, O, Mg, and Si) with a satisfactory precision.

Subsequently, we used these abundances in the stoichiometric model presented in Santos et al. (2015, 2017) to determine the composition of the planet building blocks. In particular we determined the iron-to-silicate mass fraction - f_{iron}, water mass fraction - w_f and summed mass percent of all heavy elements - Z. By testing the composition

based on the RFA and RFH predictions against published data, we conclude that the RFA model tends to produce a narrower distribution than expected, specially for the water mass fraction. The predictions for Z of both models agree very well with each other and with the published values.

As a next step, from the NASA Exoplanet Archive[1], we select selected potentially rocky planets with mass $M < 10\ M_{\oplus}$ and radius $R < 2\ R_{\oplus}$. Only planets with R and M measurements better than 30% precision were considered. We also require planets only with RV-based mass measurements, as RV-based and TTV-based mass values have different mass distributions (Mills and Mazeh 2017). Our sample consits of 39 planets orbiting around 36 stars, of which 10 are M dwarfs and 26 are FGK-type stars.

The stellar metallicities of the stars are taken from the SWEET-Cat catalogue (Sousa et al. 2021, Santos et al. 2013) and the parameters (coordinates, proper motions, RV, parallax) to determine the Galactic total velocity are taken from the NASA archive. The total velocity and metallicity of the stars were then used to predict the abundances of C, O, Mg, and Si with the RFA and RFH models.

Inserting these abundances into the aforementioned stoichiometric model, we estimated the composition of protoplanetary disks for these 36 stars. For the protoplanetary disks of the 10 M dwarfs, the average values of f_{iron}, w_f, and Z calculated using the RFA and RFH model predictions are, respectively: $f_{iron} \sim 31.9 \pm 2\%$ and $f_{iron} \sim 32.2 \pm 1\%$, $w_f \sim 53.9 \pm 1\%$ and $w_f \sim 53.1 \pm 5\%$ and finally $Z \sim 1.6 \pm 0.8\%$ and $Z \sim 1.5 \pm 0.7\%$. Except for the w_f, the determined compositions are compatible with the Solar System values ($f_{iron}^{Venus} \sim 32\%$ (Aitta 2012), $f_{iron}^{Earth} \sim 32\%$ (McDonough 2003) and $f_{iron}^{Mars} \sim 30\%$ (Khan et al. 2018), $w_f \sim 67\%$ and $Z \sim 1.3\%$ as determined by Lodders 2003). A lower water mass fraction, estimated for some of the protoplanetary disks, not only affects the ice mass fraction in solid planets and the amount of objects formed beyond the water ice-line (Ward 1997, Luger et al. 2015, Santos et al. 2017), but can also affect the habitability of the planet, even if they are in the HZ of the star (Noack et al. 2017).

Planetary sciences is one of the fastest advancing fields in Astronomy, one of the main goals being the detection and characterisation of Earth like planets in the HZ of their stars. In this regard, M dwarfs are favourable targets since i) the detection of low-mass, small-sized planets is easier around low-mass and small-sized stars, and ii) the HZ around M dwarfs is closer to the stars which again makes easier the detection of planets with the

[1] https://exoplanetarchive.ipac.caltech.edu/

current instruments and methodologies. However, the atmospheric temperature of M dwarfs makes it very difficult to determine their chemical composition from their (especially optical) spectra, due to the presence of millions of molecular lines. In that sense, the predictive models developed in this work may prove a valuable tool to go around this problem. Being based on the metallicity and total galactic velocity is an asset, since their kinematics properties can be measured and [Fe/H] is a parameter known to be possible to be determined through M dwarfs spectra.

In the coming years, mission like PLATO (ESA, Rauer et al. 2014, 2016) and ARIEL (ESA, Puig et al. 2016) will be launched with the main objective to detect and characterise exoplanets. PLATO will find and study a large number of exoplanetary systems, focusing on the properties of terrestrial planets in the HZ around solar-like stars. Ariel will study what exoplanets are made of, how they formed and how they evolve, aiming to measure the atmospheric chemical composition and thermal structures of hundreds of transiting exoplanets and proving invaluable data to better understand planetary science far beyond the boundaries of the Solar System. By previously determining the chemical composition of different planetary systems, it is possible to select the most interesting and promising candidates to harbour an Earth-like planet in the HZ of its stars, with the potential for life.

Regarding future work, the next step would be to look at each M dwarf planetary system individually to understand why planets formed in a disk with similar composition would have different densities, and what processes during planetary formation might have played a role.

Based on the available information from different surveys, different versions of the model can be built. Usually, the Mg, Si and Fe abundances can be relatively easily determined for solar-type stars, but the same is not true for the C and O abundances, as their spectral absorption lines might be more difficult to detect. With that in mind, it is also possible to built a predictive model where one would give [Fe/H] and one (or both) of the [Mg/H] and [Si/H] abundances, and predict the remaining individual abundances. Trying the different possibilities would allow to see what would be the best model and understand which individual abundances have more information. As the kinematic properties might not always be available, a model that did not depend on those would also be a good option. This type of tool would help on the analysis of stars with less chemical abundances known and on the study of planetary systems, their chemical composition and formation process.

All these improvements will be implemented and the codes developed during this thesis will be make public after the results are published in a refereed journal.

Appendix A

The log g problem

When selecting the best sample from APOGEE data, it was necessary to ensure we had precise stellar parameters. The PARAMFLAG column has information about stellar parameters, such as T_{eff}, $\log g$, microturbulent velocity $\log(\nu_{mic})$, overall scaled-solar chemical abundance [M/H], carbon ([C/M]), nitrogen ([N/M]), $\alpha-$element abundances [α/M] (O, Mg, Si, S, Ca, Ti, Ti II) relative to the solar abundance ratio, projected rotational velocity $\log(v \sin i)$ and a final entry O currently unused. Each entry is an array with nine elements, corresponding to the nine parameters mentioned but now with flags associated. Each flag equals a warning for that information and is represented by a number, as described in Table 12 from Holtzman et al. (2015). The flag system uses bitmasks, meaning that each warning has an associated number and these numbers are the exponent of powers of base 2. If a parameter has more than one warning, the final number is the sum of these powers of base 2. For example, let's say the effective temperature has associated a warning for grid edge and calibration range. These warnings correspond to the numbers 8 and 9, as described on Table 12 from Holtzman et al. (2015). So, on the first entry of the array correspondent to that star on the PARAMFLAG column, we will have a number which will be $2^8 + 2^9 = 256 + 512 = 768$.

When analysing the flags for the $\log g$ column, we see on the histogram from Figure A.1 that the main warnings may come from single flagged warnings, as is the case for the first bar, where $16777216 = 2^{24}$ which is a currently unused flag. However, the warnings can also be a combination of warnings, as is the case of the second bar where $25165861 = 2^{24} + 2^{23} + 2^5 + 2^2 + 2^0$. This is a mix between "other error condition" flag (flag number 2) and "parameter within 1/8 grid spacing of grid edge" flag (flag number 0). However, the problem lies in flags number 24, 23 and 5 which are marked as unused flags (see bitmask description[1]), although they are clearly being used. Around half of the stars has warning number 24, so this cannot be simply a mistake.

[1] https://www.sdss.org/dr14/algorithms/bitmasks/#APOGEE_PARAMFLAG

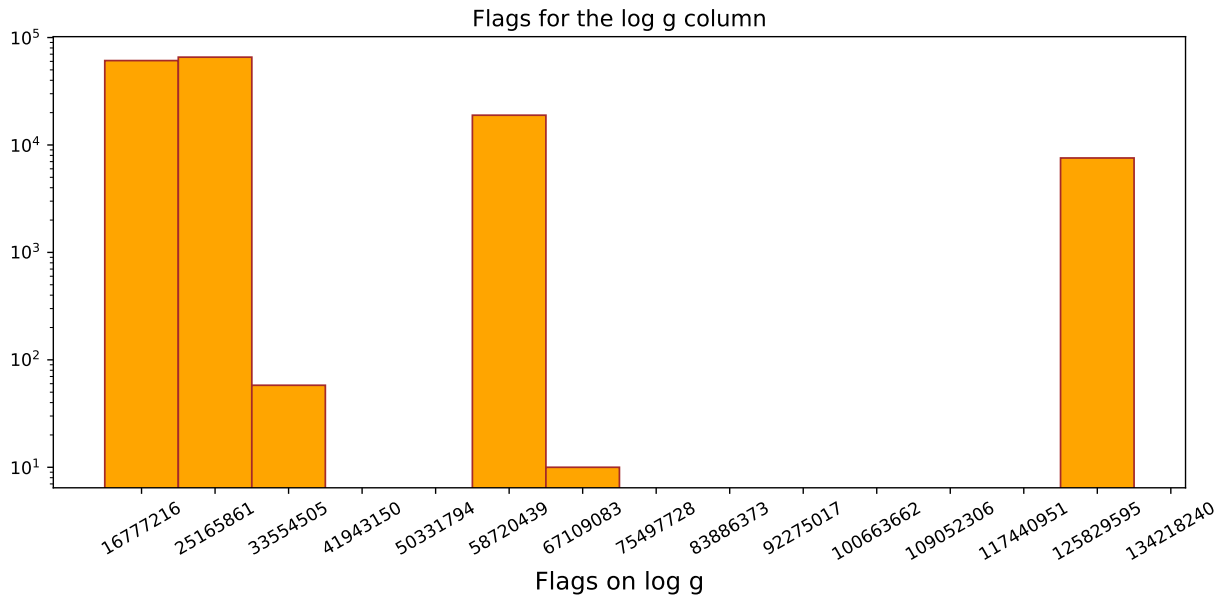


FIGURE A.1: Histogram for the flags in the log g parameter.

In order to see if there was any data set with particularly high uncertainty that could explain the unknown flags, we do an histogram with the errors from log g (absolute and relative), present in Figure A.2. From the left plot, we can see that the distribution seems to have two peaks, one around 0.055 and another around 0.09. Moreover, as we can see from the right plot, the majority of the data has 1% – 3% of relative uncertainty, which is good as it is not very high. However, this does not explain the high number of unknown warnings.

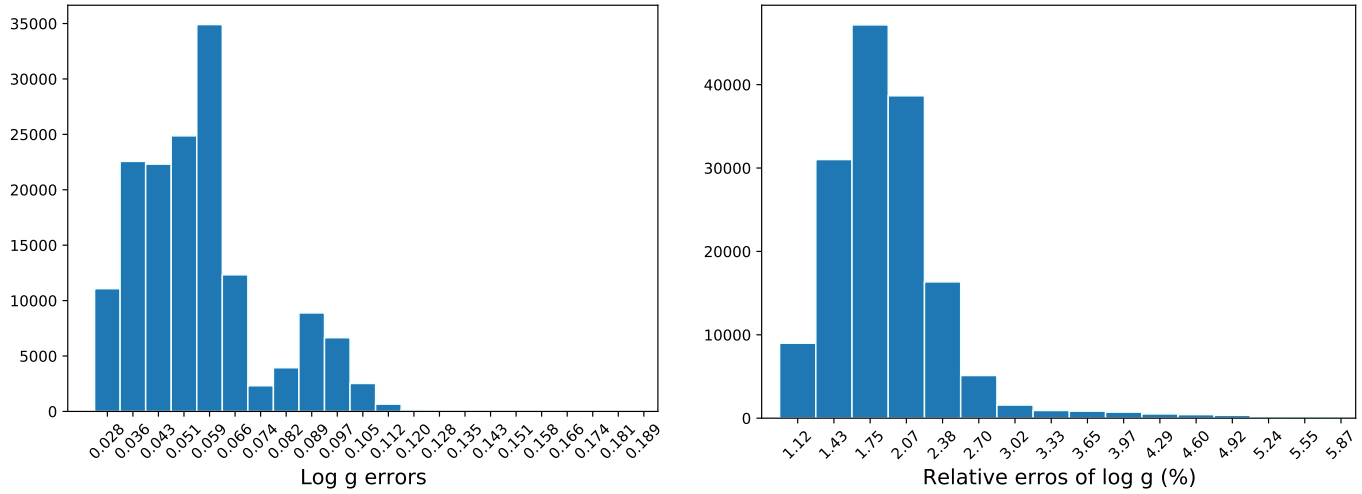


FIGURE A.2: Histogram of the log g absolute uncertainties (left) and histogram of the relative uncertainties of log g (right).

In another effort to understand what was happening, we directly plotted instead the log g uncertainties depending on its log g value, as shown by Figure A.3. We can see that the uncertainties are certainly higher when log $g > 4$ but there is a wider range of uncertainties values when log < 4.

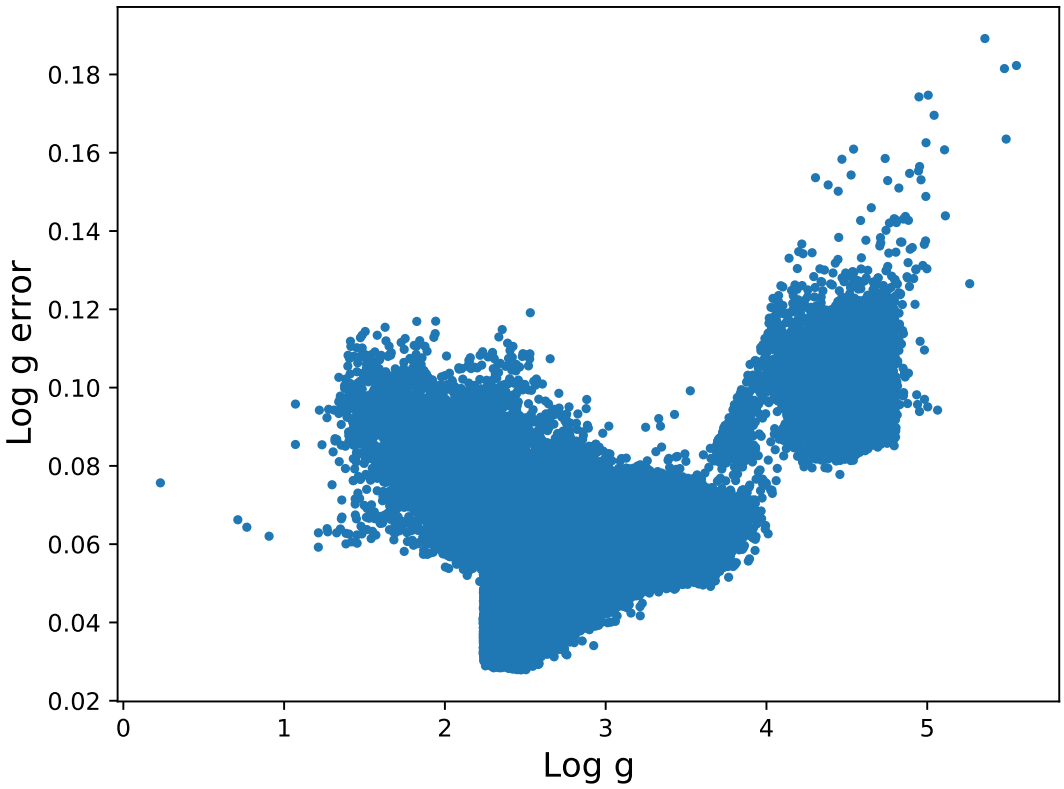


FIGURE A.3: Log g uncertainties depending on the log g values, for APOGEE stars.

Apart from the fact that there appears to be two distinct groups on the data, the maximum relative error is no larger than 6% and we find no indication that there is anything wrong with the log g data or if there is any special care we need to have when analysing it. Therefore, we decided to directly contact the first author of the APOGEE DR16 paper (Jönsson et al. 2020). After exchanging emails with Dr. Henrik Jönsson, we were able to realise what was going on. As described in section 5.2.2 from (Jönsson et al. 2020), the stars were divided in 4 calibration groups: dwargs, red clump stars, red giant branch stars (RGB) and a RGB-dwarfs transition stars. The division was made through a series of conditions, mainly on T_{eff} and log g. To account for this change, new bits were added to PARAMFLAG column (log g entry) providing information about which log g calibration was applied: LOGG_CAL_RC (red clump, RC), LOGG_CAL_RGB (red giant branch, RGB), LOGG_CA_MS (main sequence, MS), or LOGG_CAL_RGB_MS (RGB-MS transition). This was not however described further in the web documentation, as was their intention when writing the paper. Therefore, on the bitmask description webpages, warnings 24–27 should have been updated to:

- 24: LOGG_CAL_RC – log g calibration of red clumps has been applied
- 25: LOGG_CAL_RGB – log g calibration of RGB stars has been applied
- 26: LOGG_CAL_MS – log g calibration of dwarfs has been applied
- 27: LOGG_CAL_RGB_MS – log g calibration of RGB-dwarf transition stars has been applied

In order to verify this, we used the calibration flags to try and reproduce the HR diagram of Figure 5 from Jönsson et al. (2020), where the authors classified the stars in the $\log g$ "calibration-categories". This was done for spectroscopic stellar parameters (Figure A.4) and for the parameters after the calibration (Figure A.5). In both cases the obtained plot is roughly equal to what was expected.

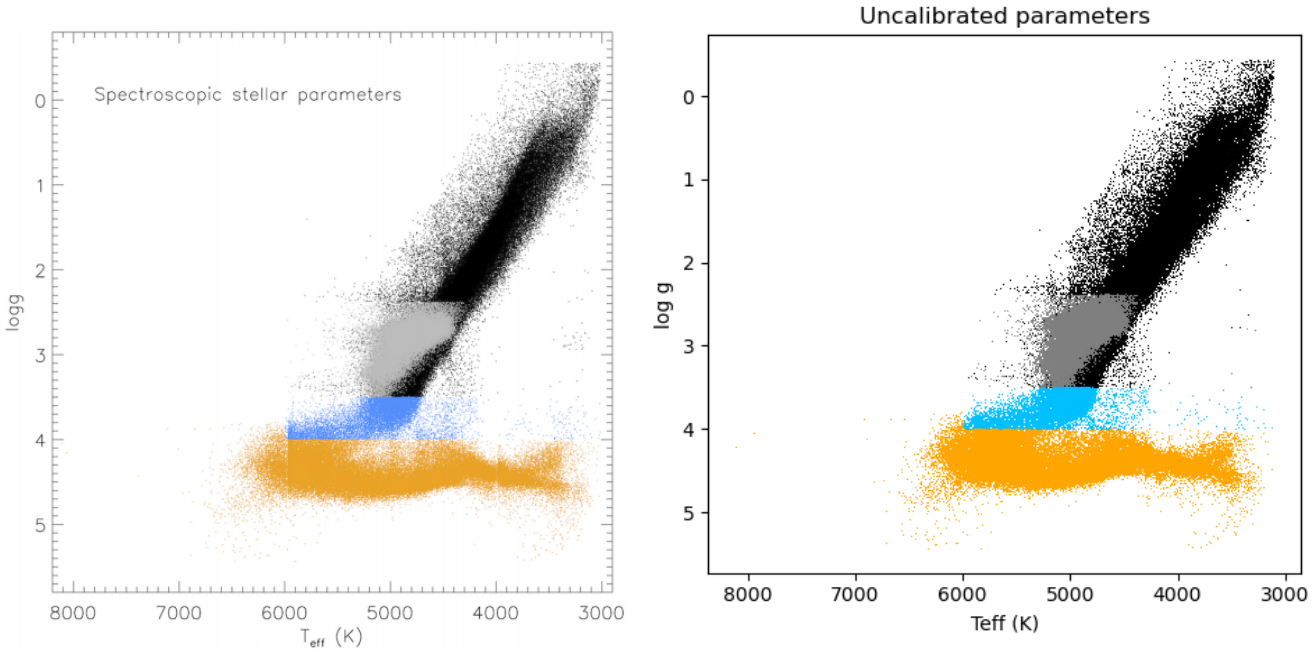


FIGURE A.4: Comparison of HR diagram for uncalibrated parameters, using the same colour scheme: black for RGB, grey for red clump stars, orange for dwarfs and blue for the mix between RGB and dwarfs. On the left we have the diagram from Jönsson et al. (2020); on the right we have the plot obtained in this work.

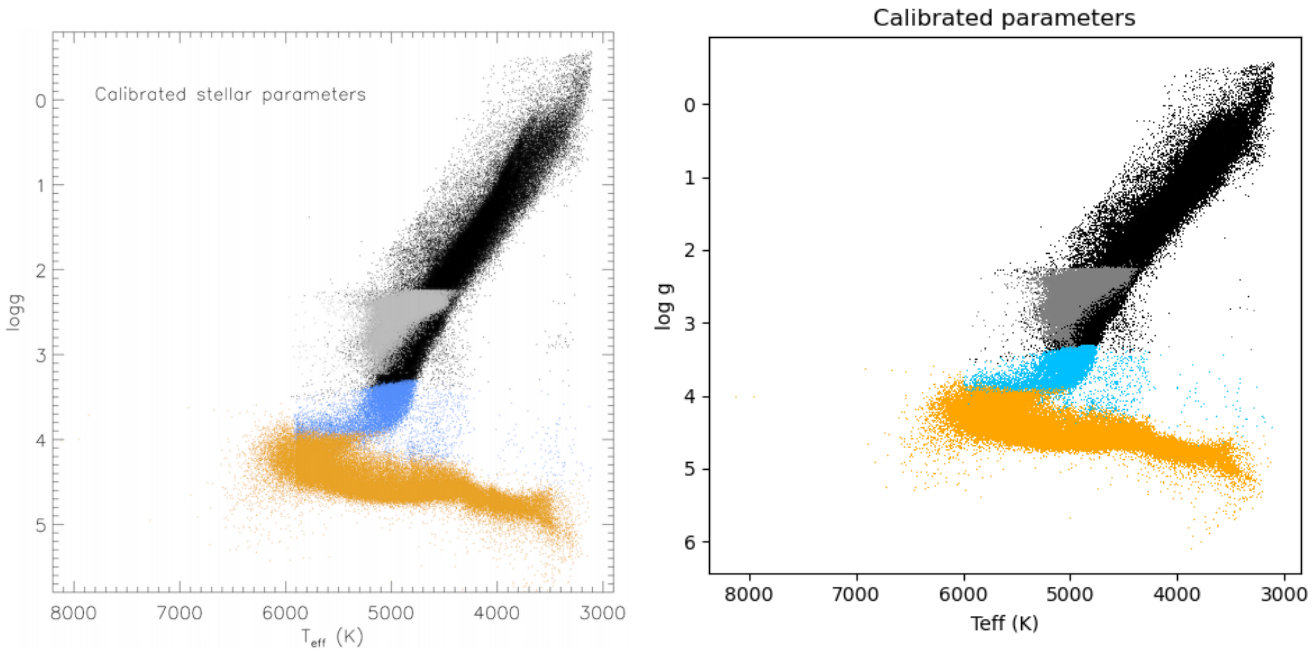


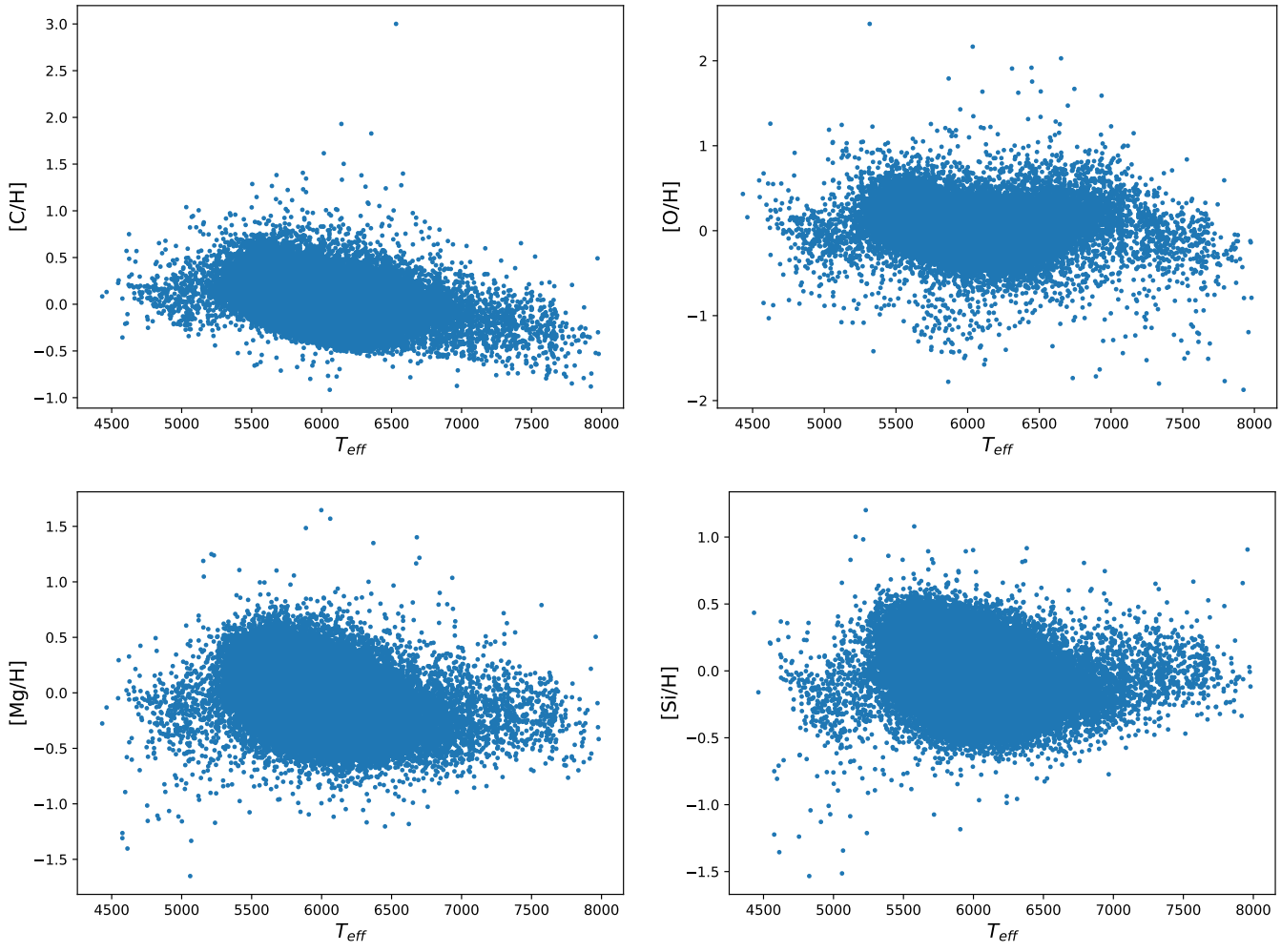
FIGURE A.5: Same as Figure A.4, but for calibrated parameters.

Appendix B

Chemical abundances depending on stellar parameters (and S/N)

B.1 GALAH total sample

B.1.1 Temperature dependency



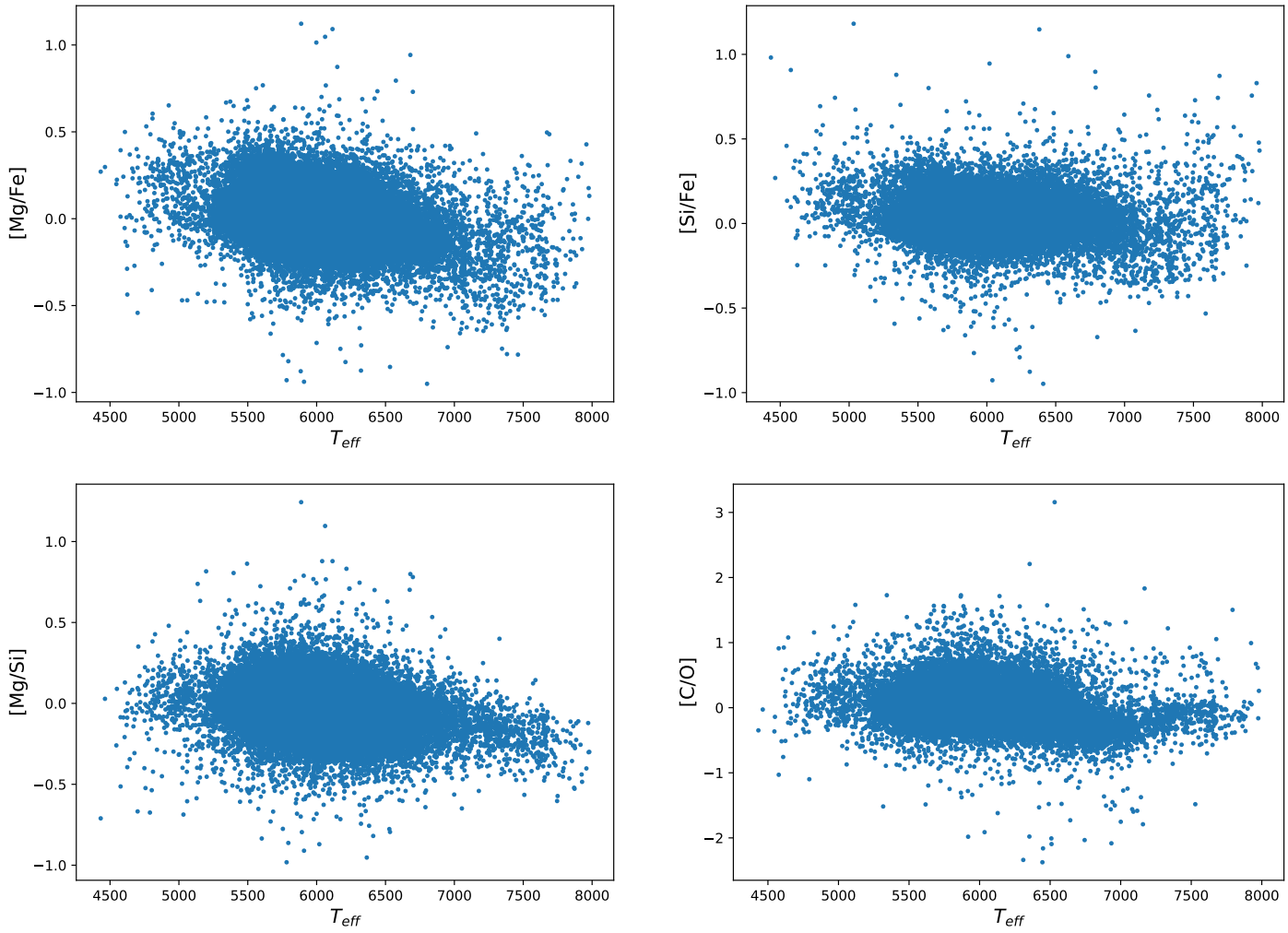
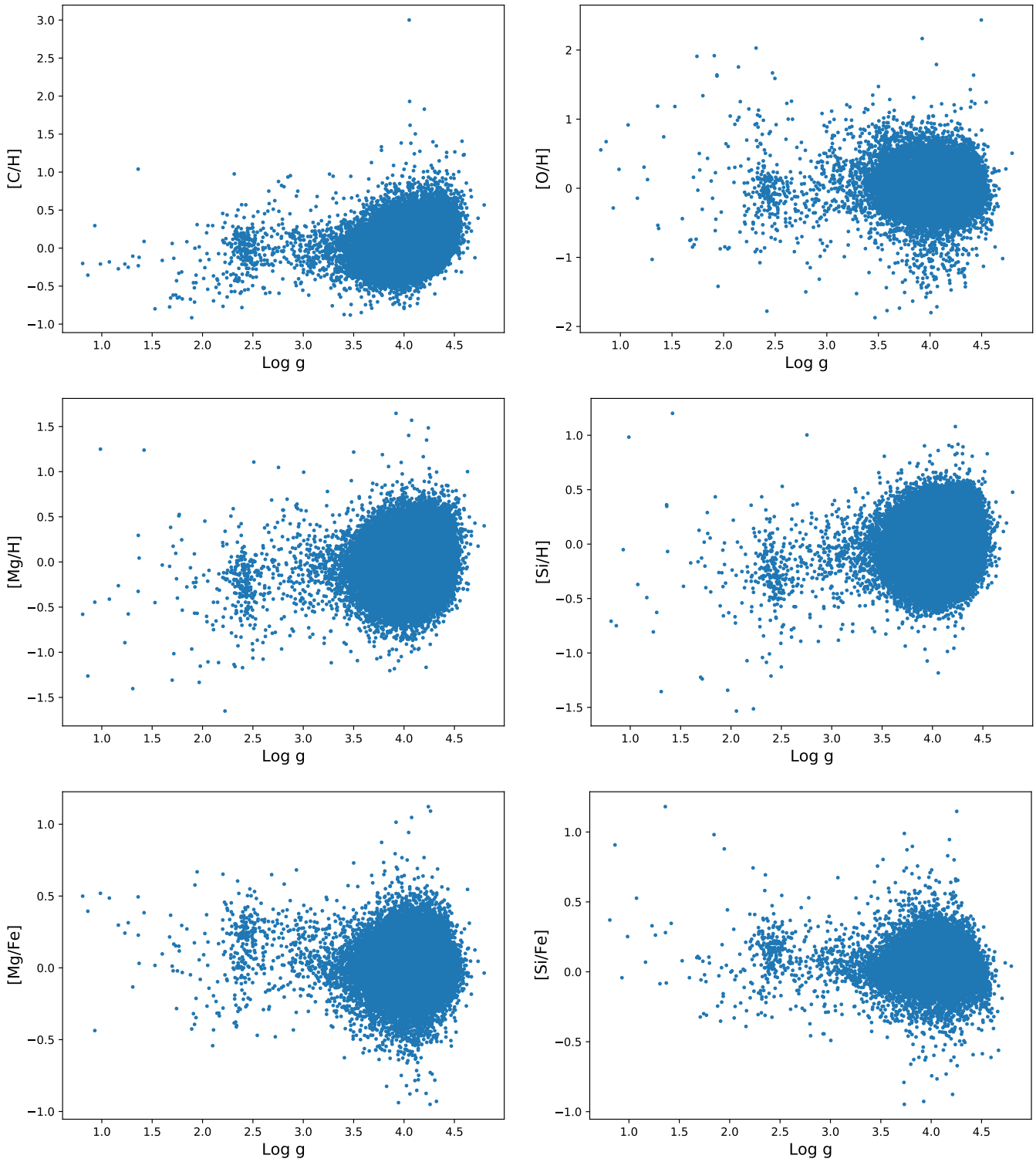


FIGURE B.1: Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on T_{eff}.

B.1.2 Log g dependency



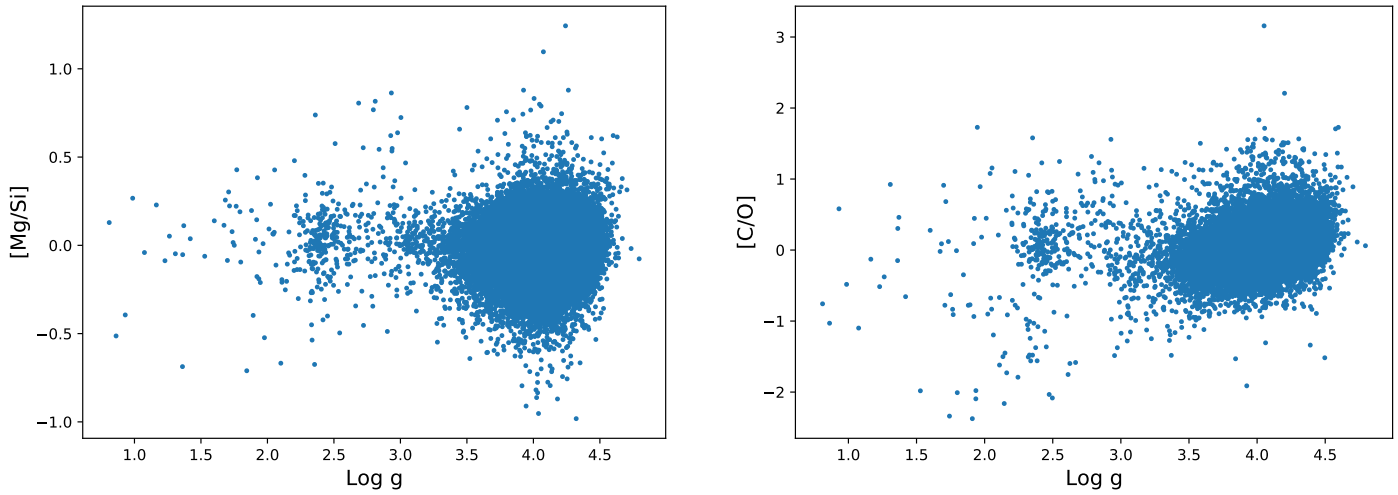
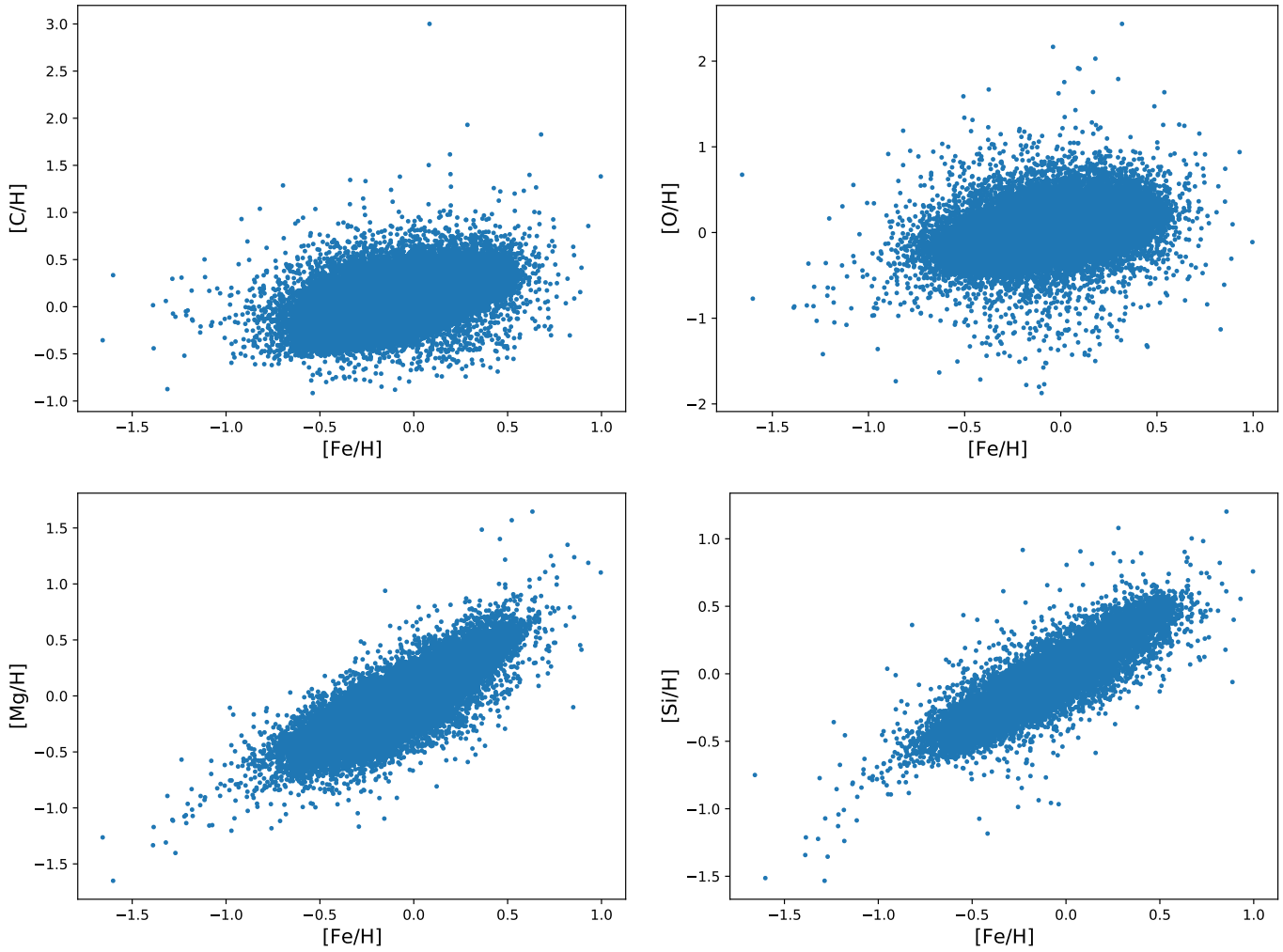


FIGURE B.2: Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on $\log g$.

B.1.3 Metallicity dependency



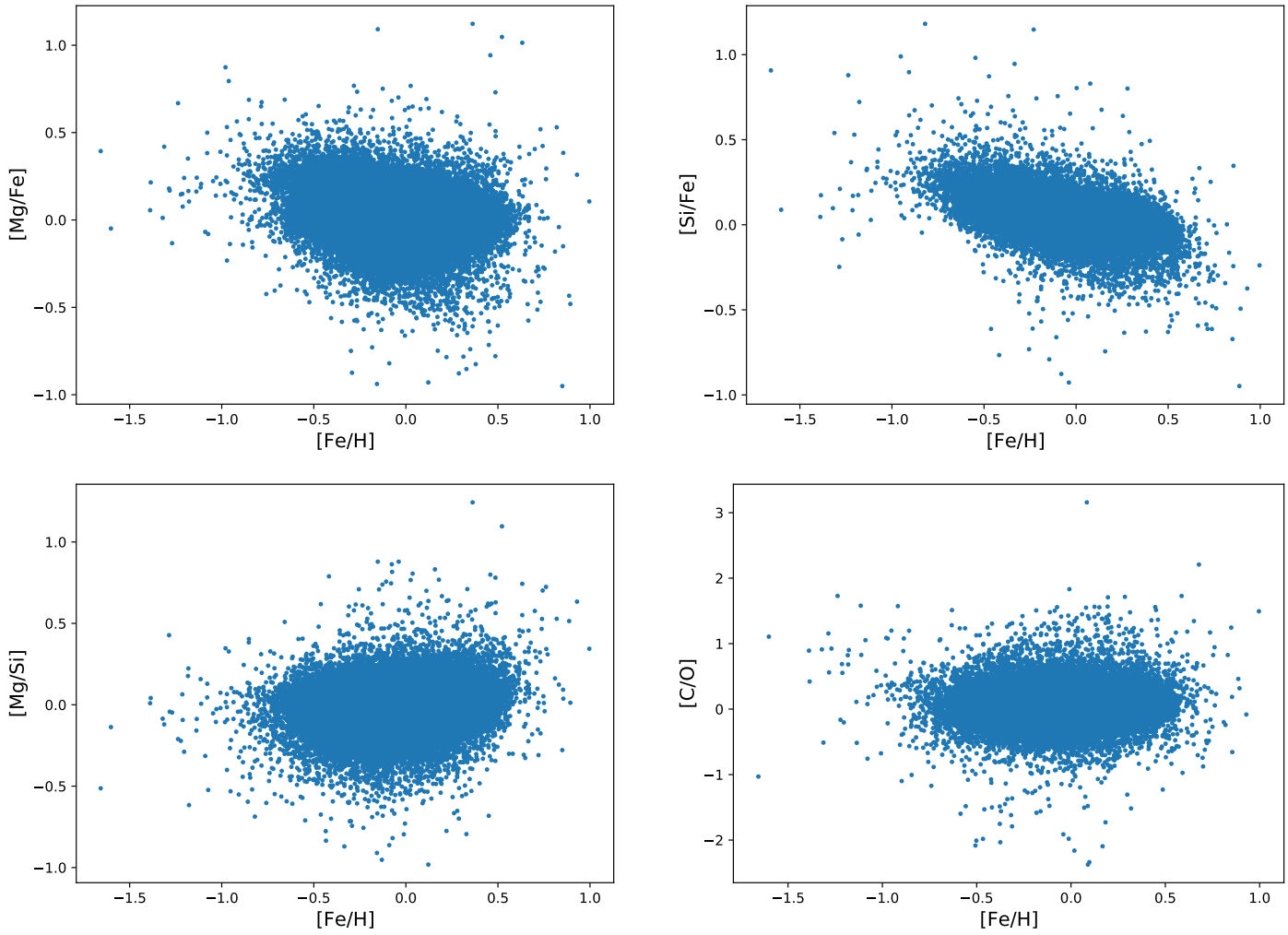
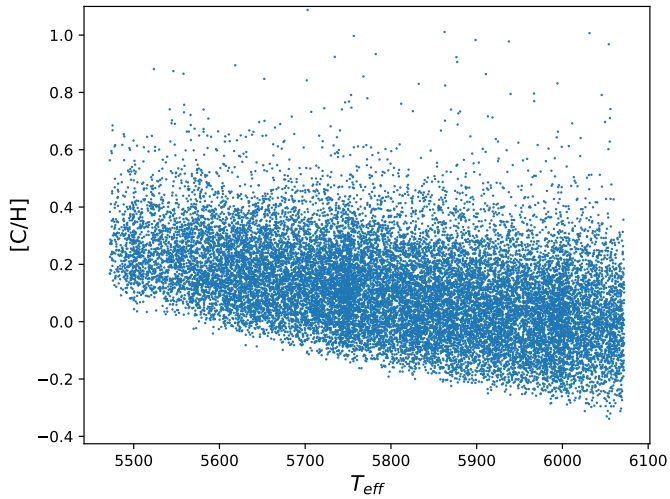
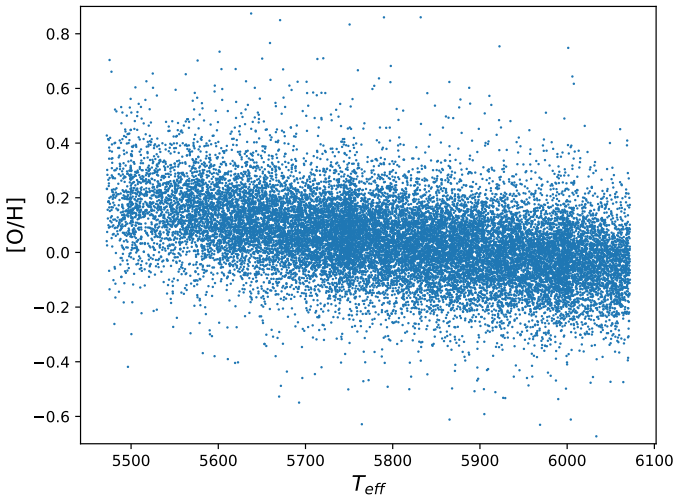


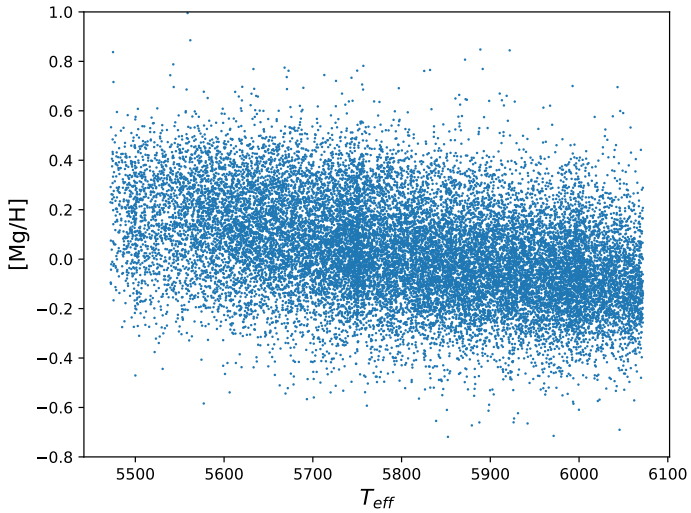
FIGURE B.3: Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on metallicity ([Fe/H]).

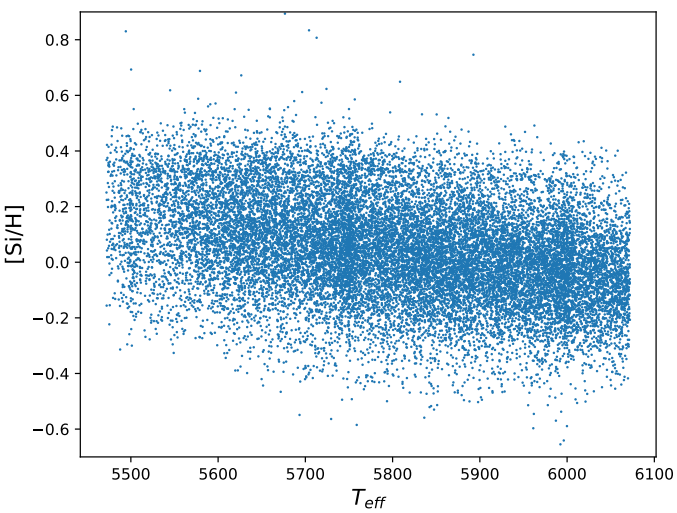
B.2 Solar Analogues stars

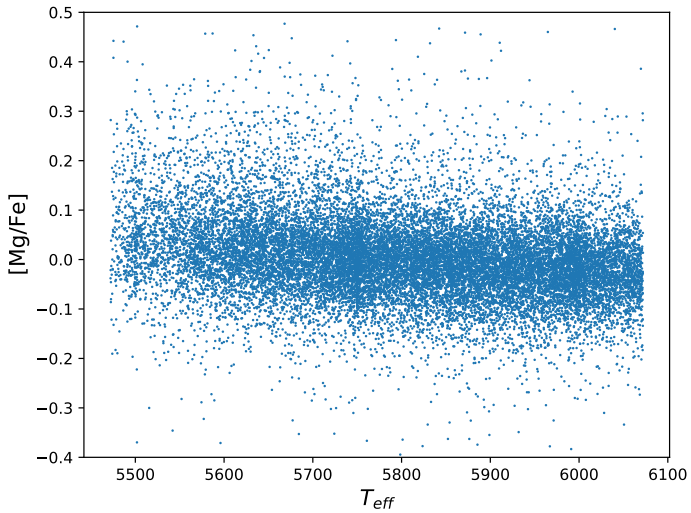
B.2.1 Temperature dependency

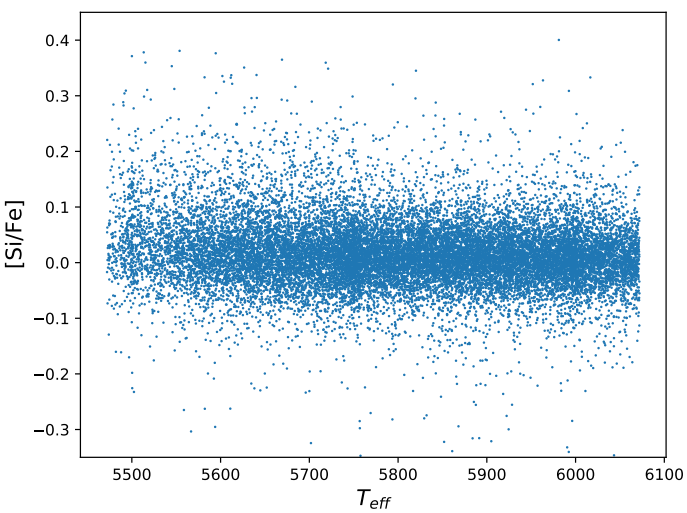












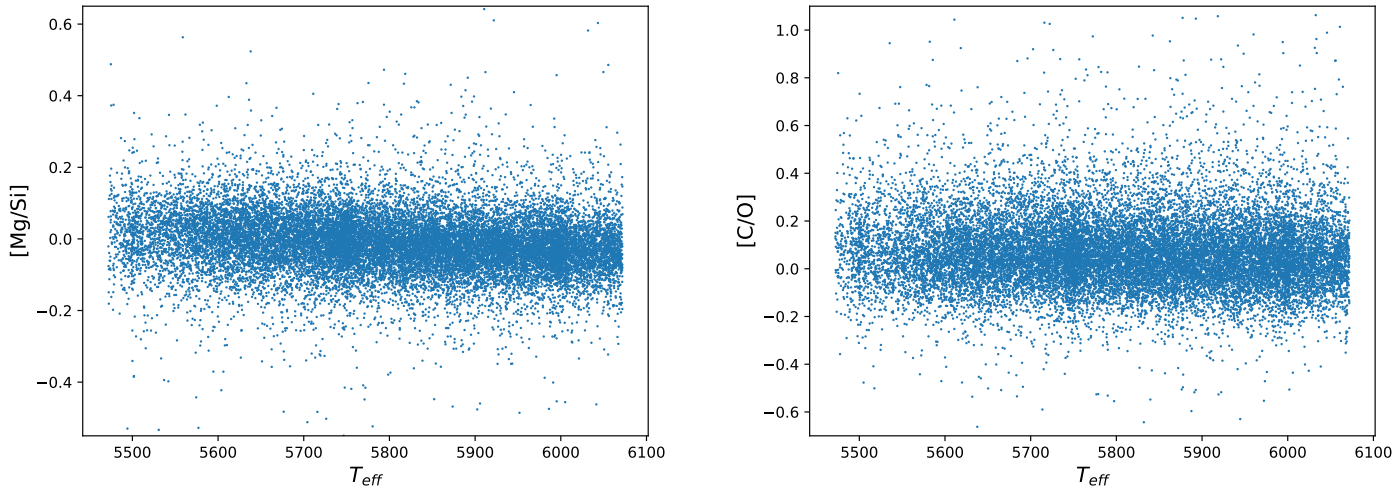
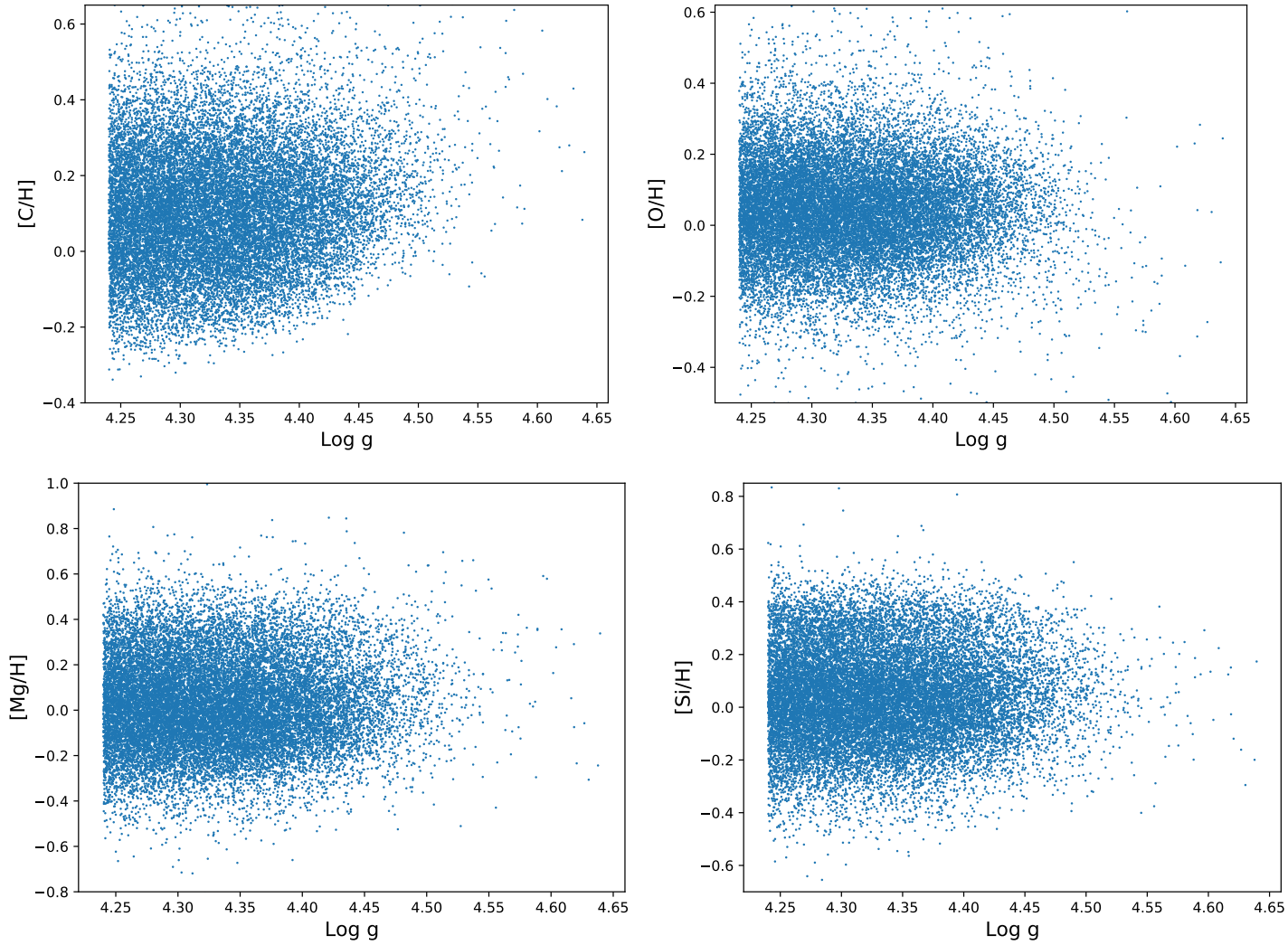


FIGURE B.4: Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on T_{eff}.

B.2.2 Log g dependency



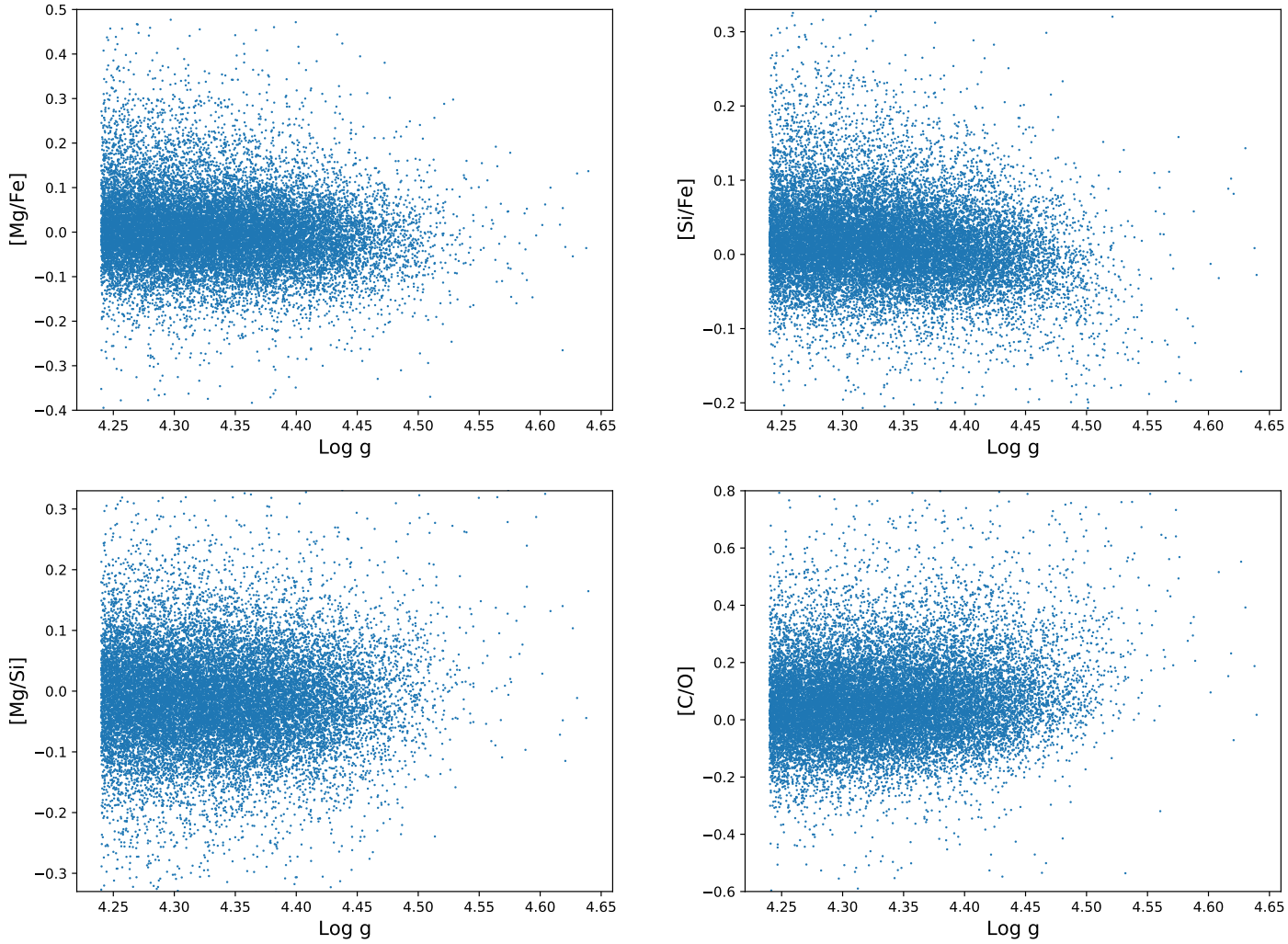
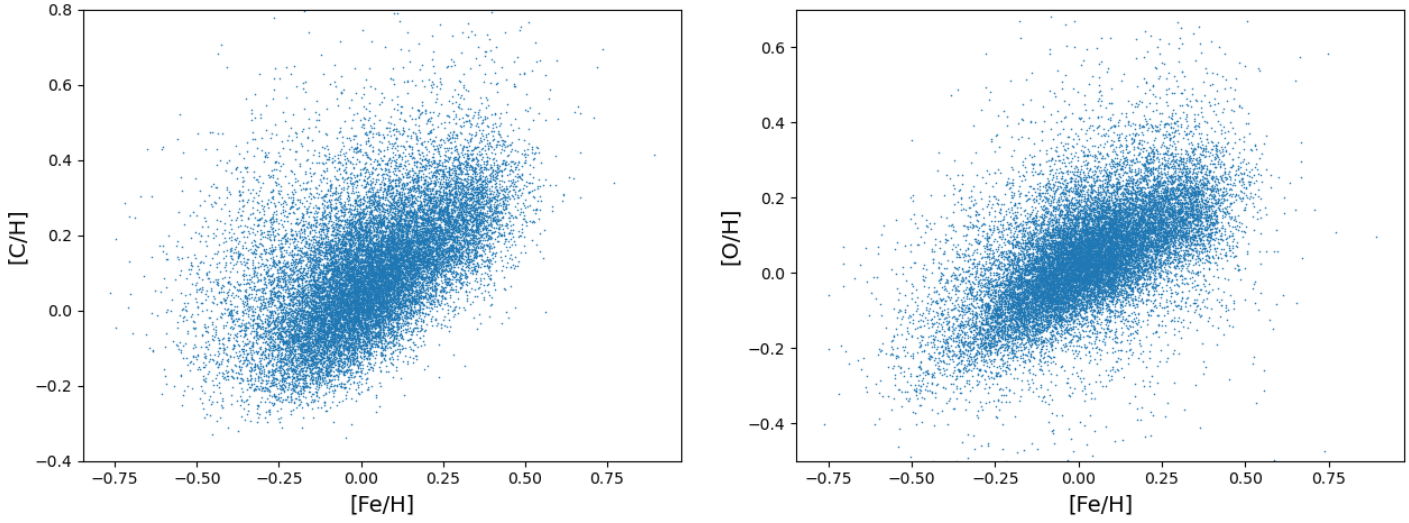


FIGURE B.5: Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on $\log g$.

B.2.3 Metallicity dependency



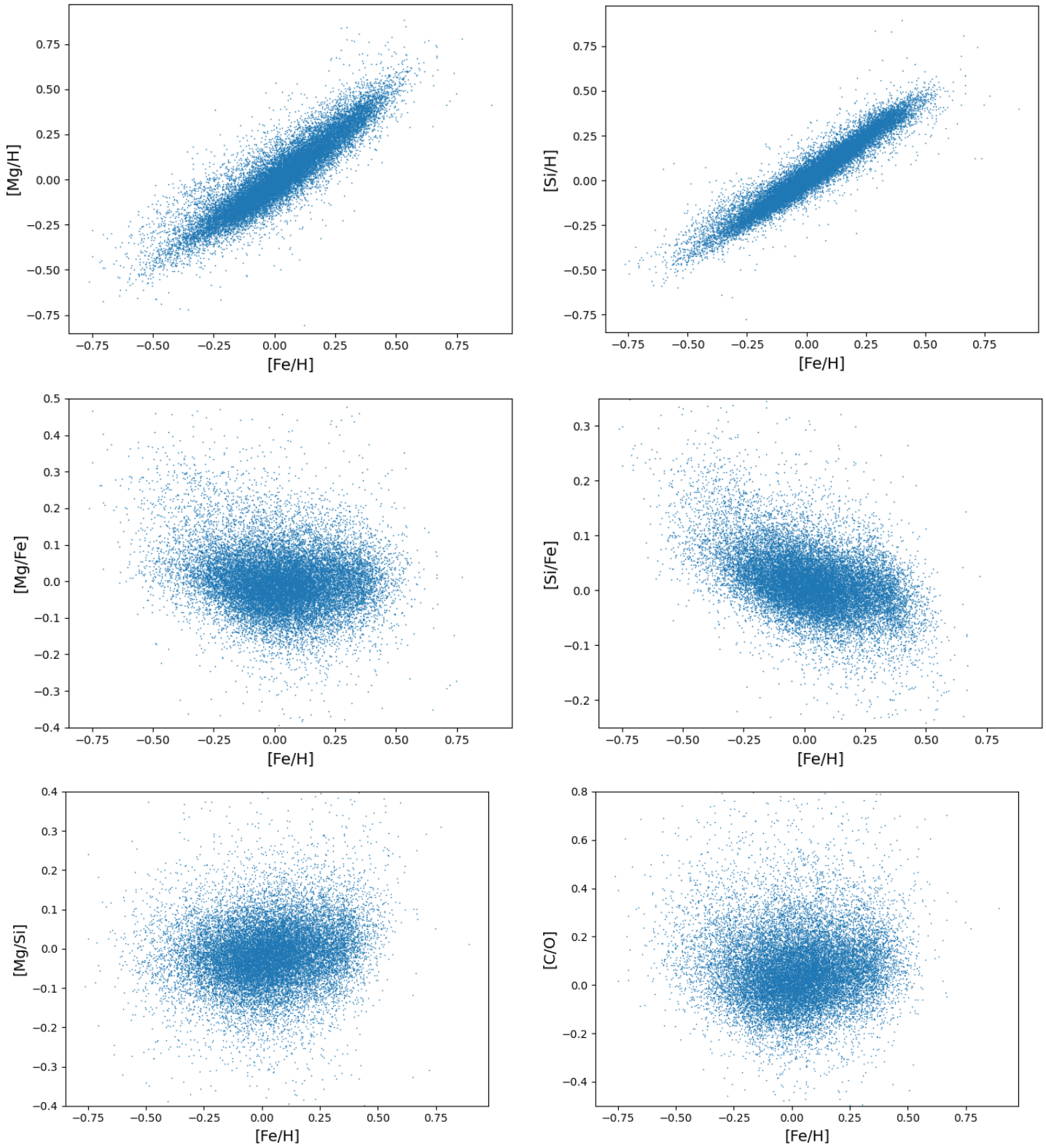
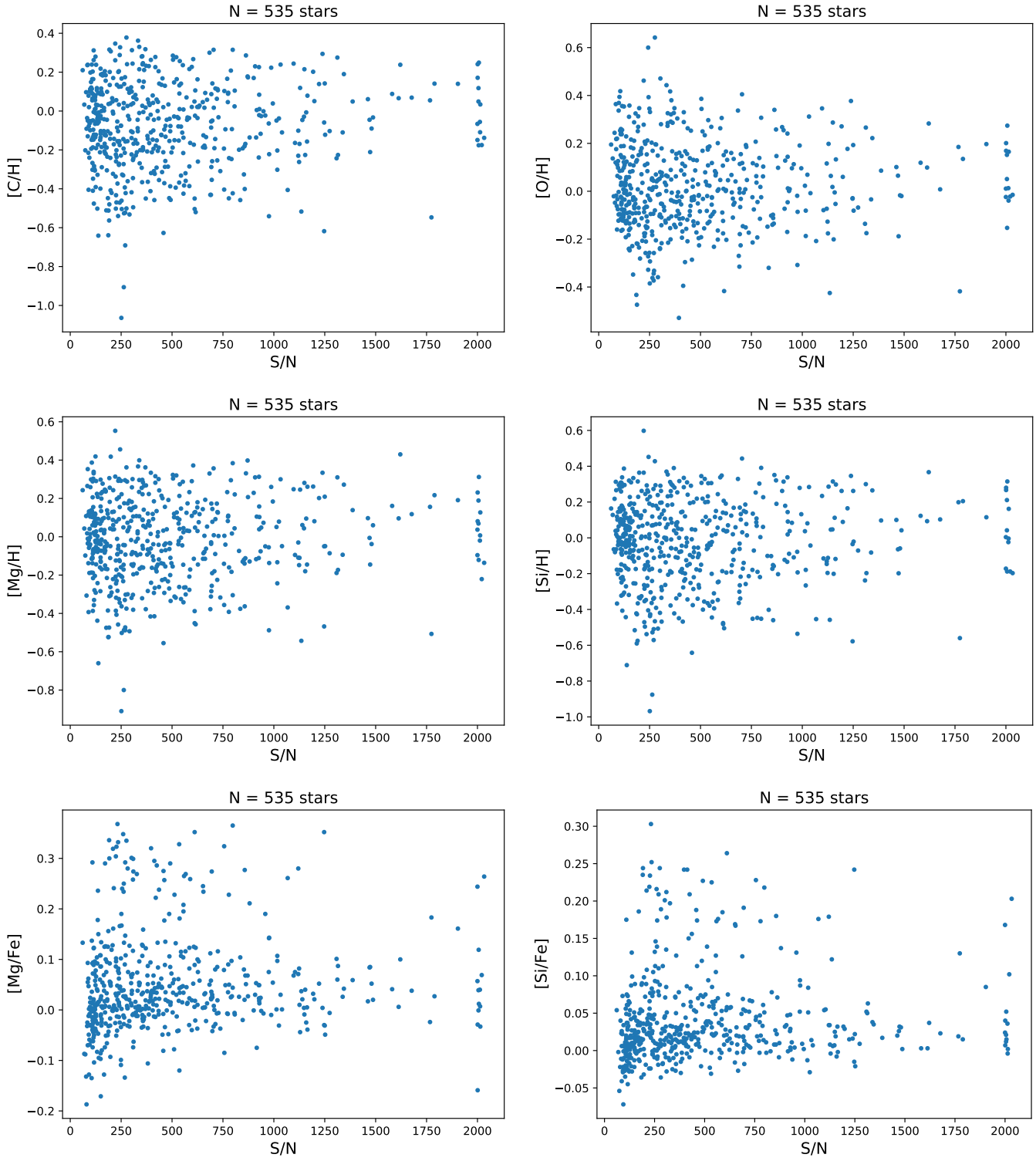


FIGURE B.6: Plots of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O] depending on metallicity ([Fe/H]).

B.3 HARPS sample

B.3.1 Dependency of chemical abundances with S/N



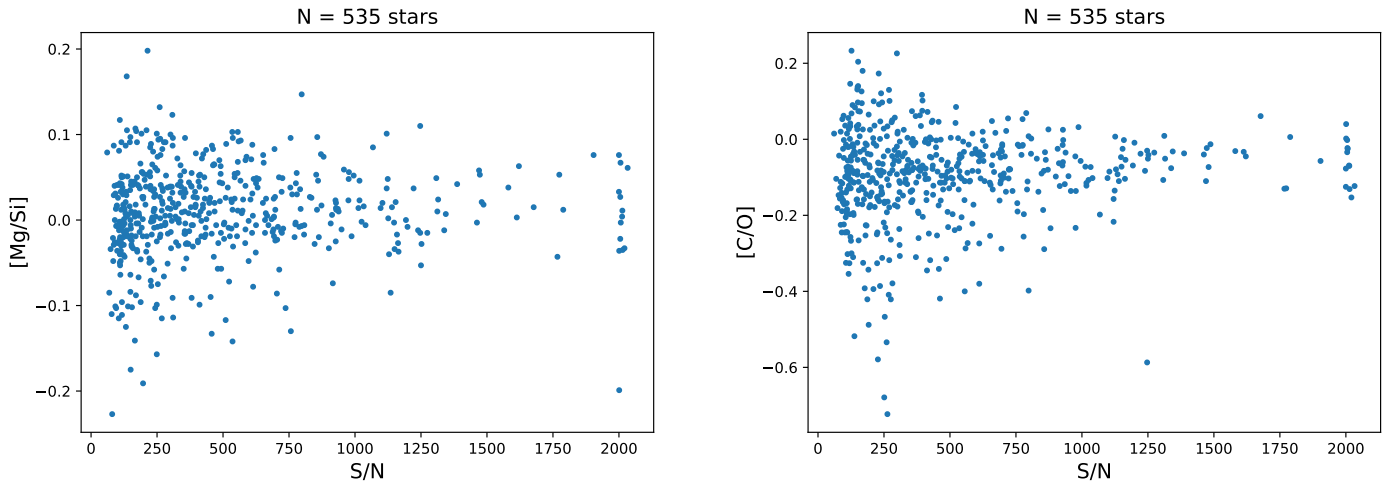
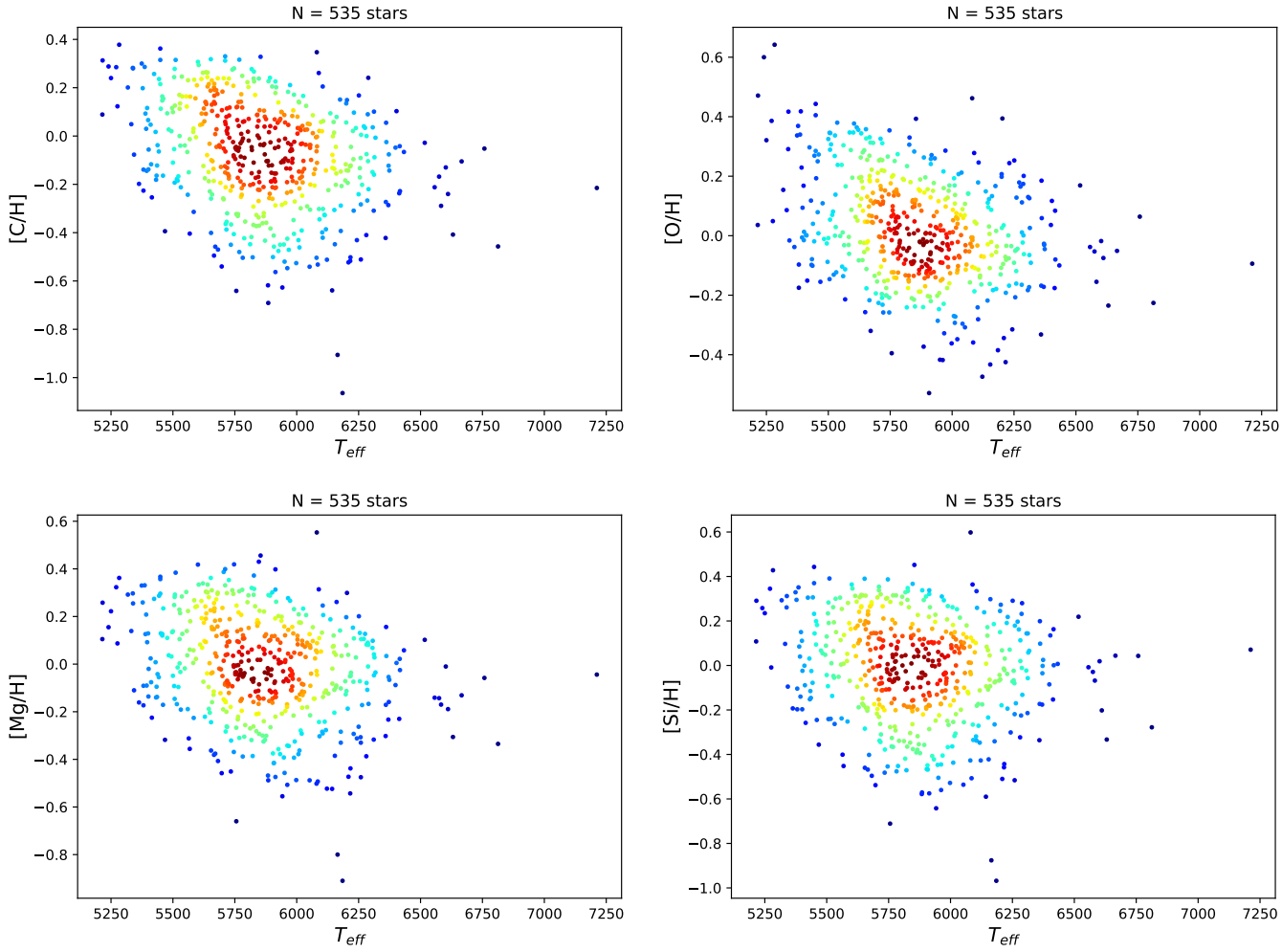


FIGURE B.7: Plots of chemical abundances depending on S/N, for [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O].

B.3.2 Temperature dependency



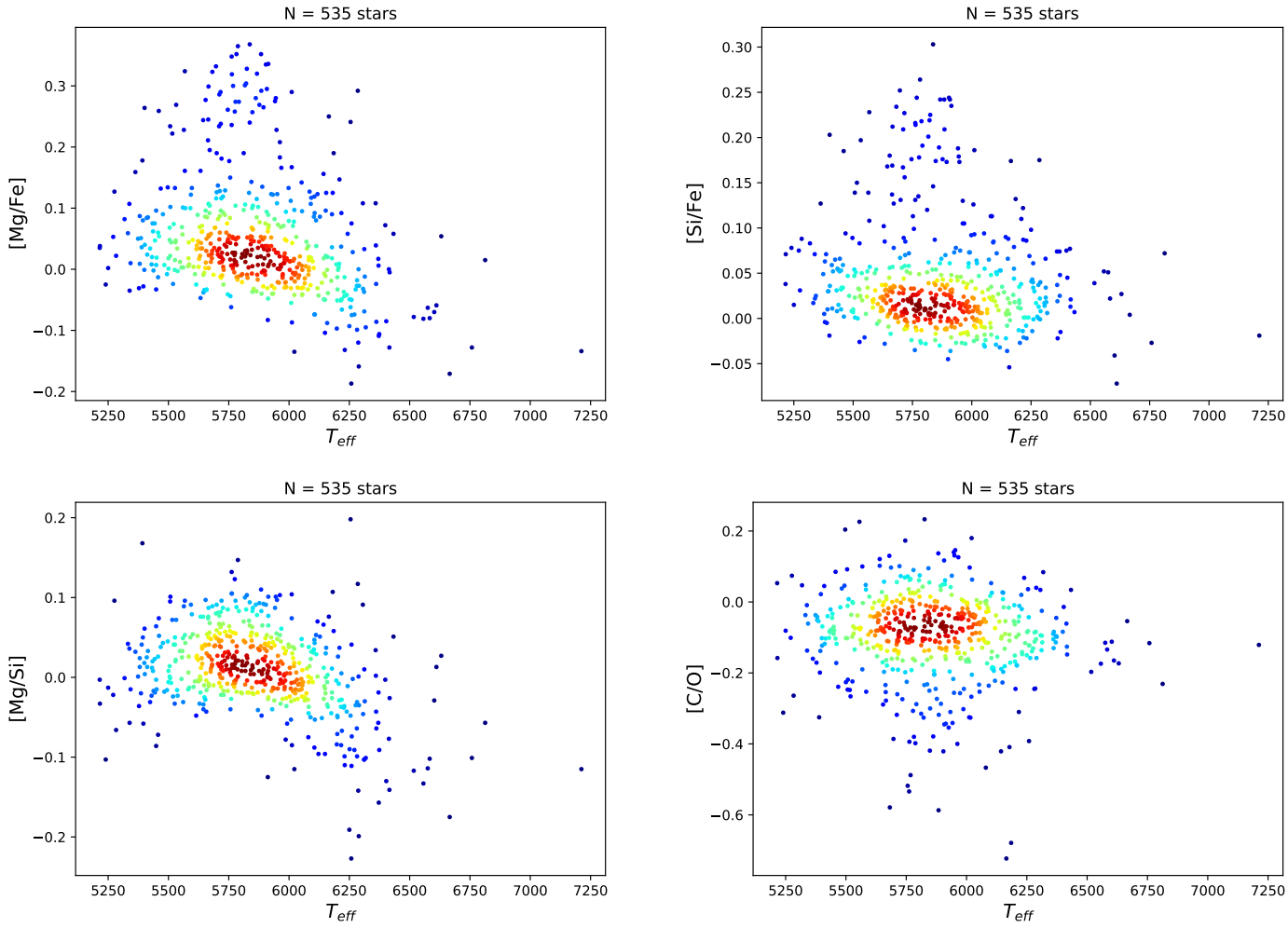
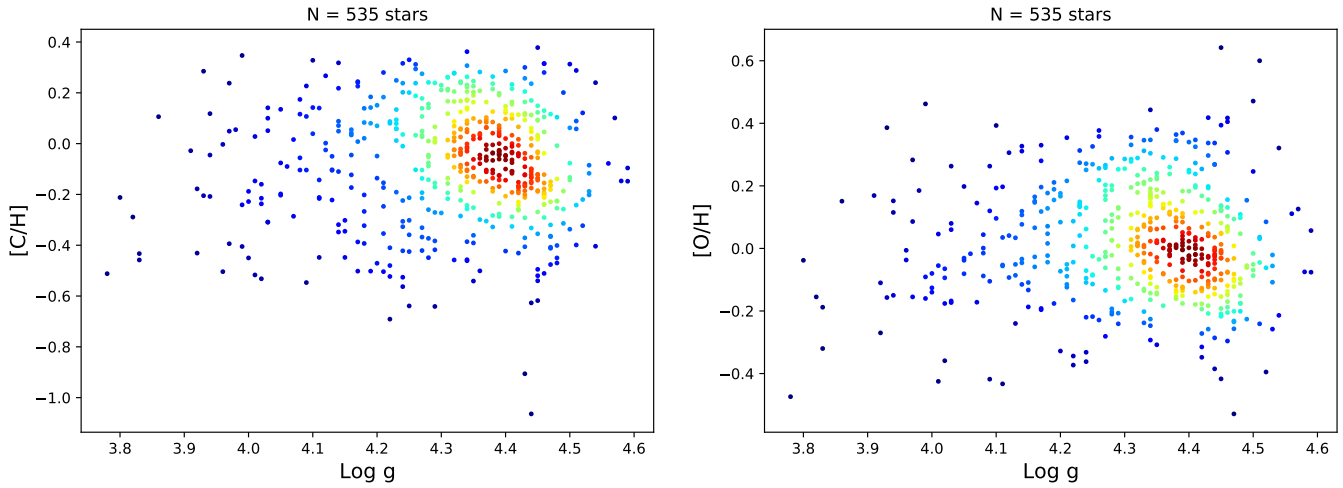


FIGURE B.8: Plots of chemical abundances depending on T_{eff}, for [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O].

B.3.3 Surface gravity dependency



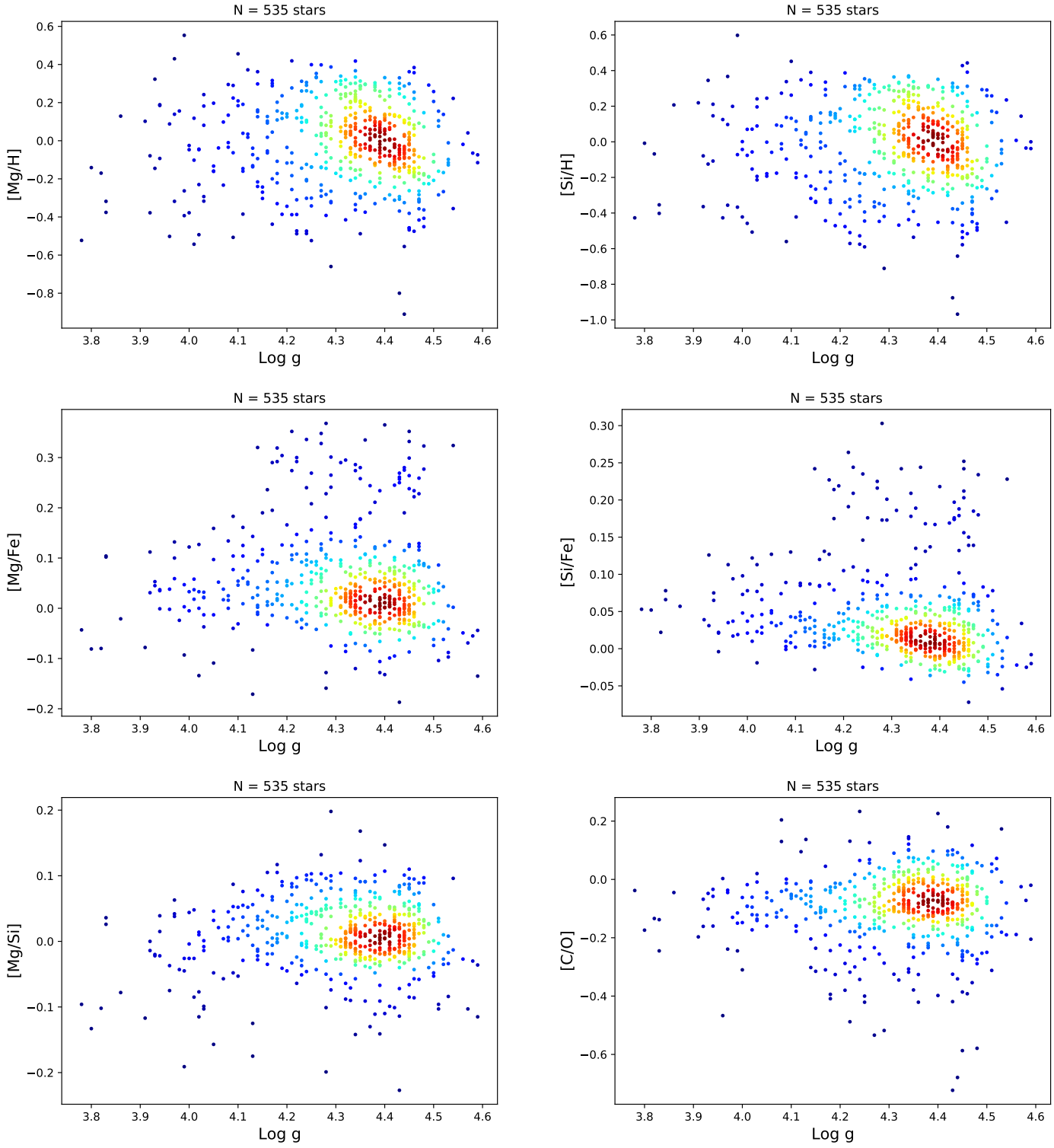
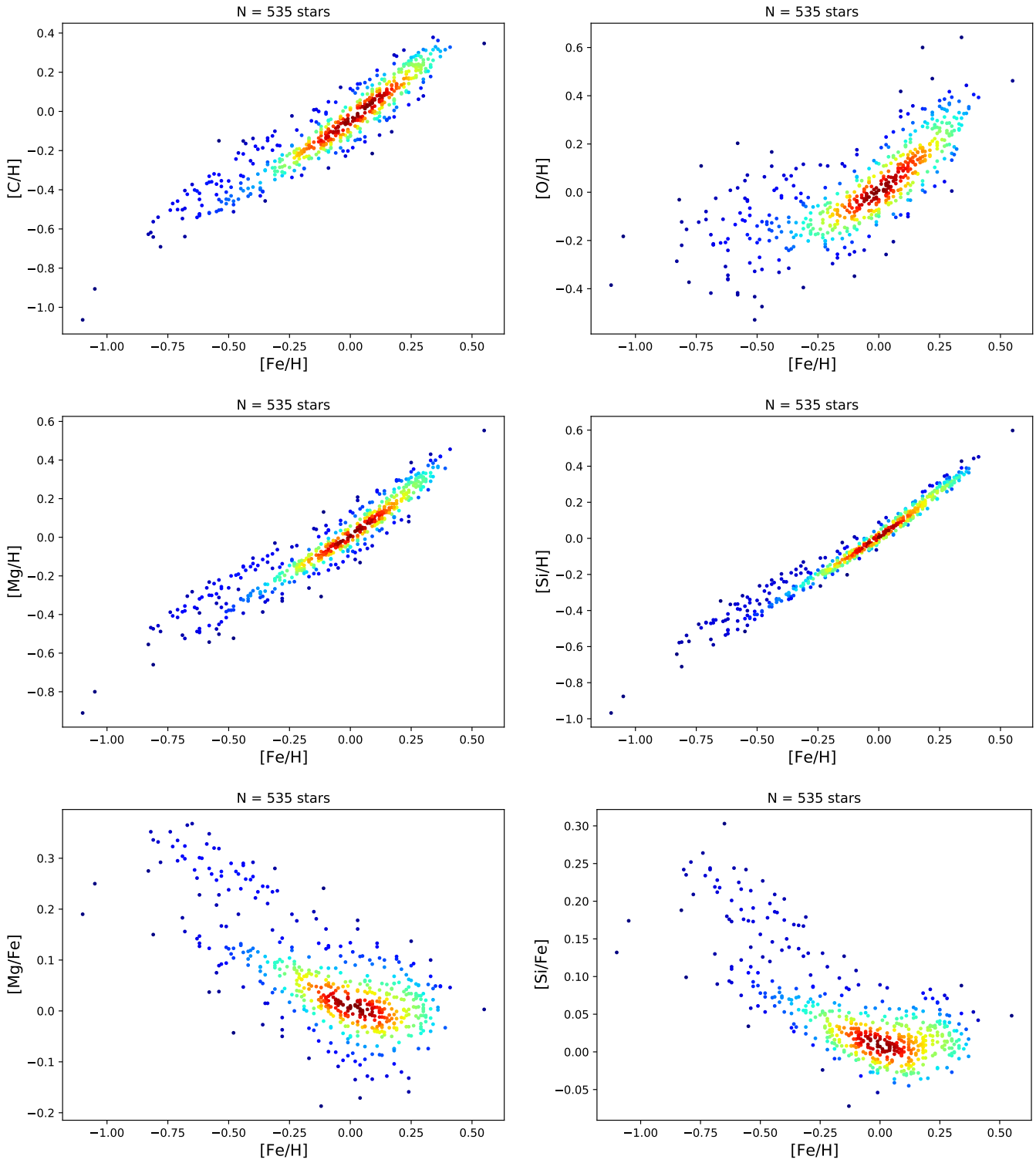


FIGURE B.9: Plots of chemical abundances depending on $\log g$, for [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O].

B.3.4 Metallicity dependency



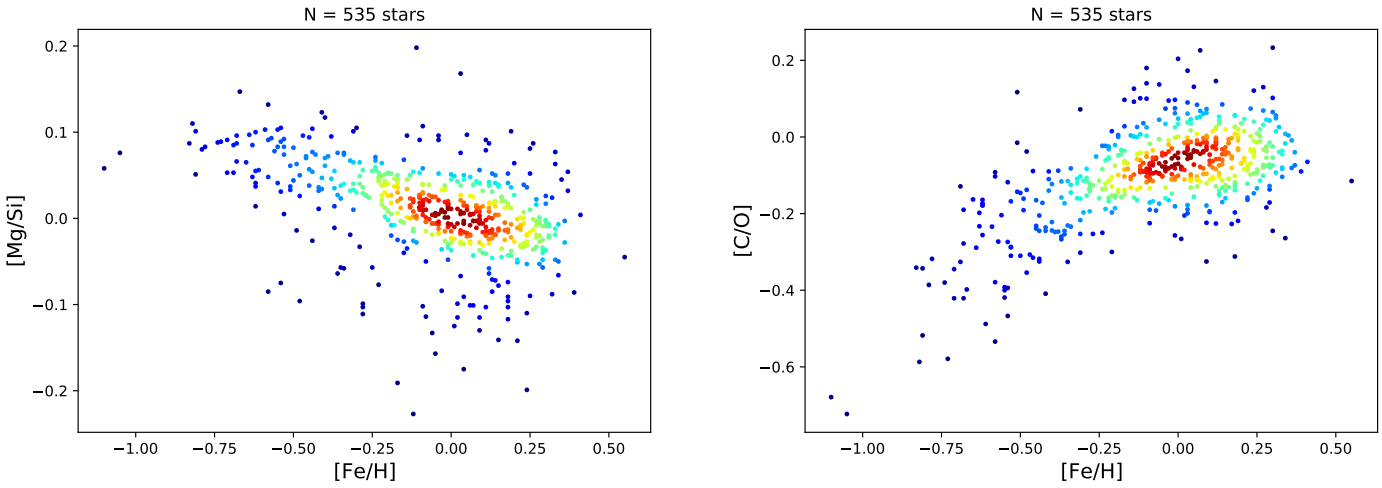


FIGURE B.10: Plots of chemical abundances depending on metallicity, for [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O].

B.4 A brief check on the Galactic chemical evolution

Similar to what was done for APOGEE, in this section we briefly check if the Galactic chemical evolution is behaving as expected through the [C/Fe] and [O/Fe] ratios for the total GALAH sample, the Solar Analogue stars (SAS) sample, and the HARPS sample.

Be it for the GALAH, SAS or HARPS sample, it is possible to realise that for higher values of metallicity, there is a decrease on the [C/Fe] and [O/Fe] abundance ratios. In the same way, for lower metallicities, stars show higher [C/Fe] and [O/Fe] abundance ratios. This behaviour is consistent with the Galactic chemical evolution mentioned before. Stars from the thick disk are older, with lower metallicities and higher [α/Fe] abundances, as they formed after SN II exploded but before smaller stars had time to reach the SN Ia stage. The same cannot be said about the stars from the thin disk, however, as they are younger and formed after SN Ia exploded and enriched the interstellar medium with iron.

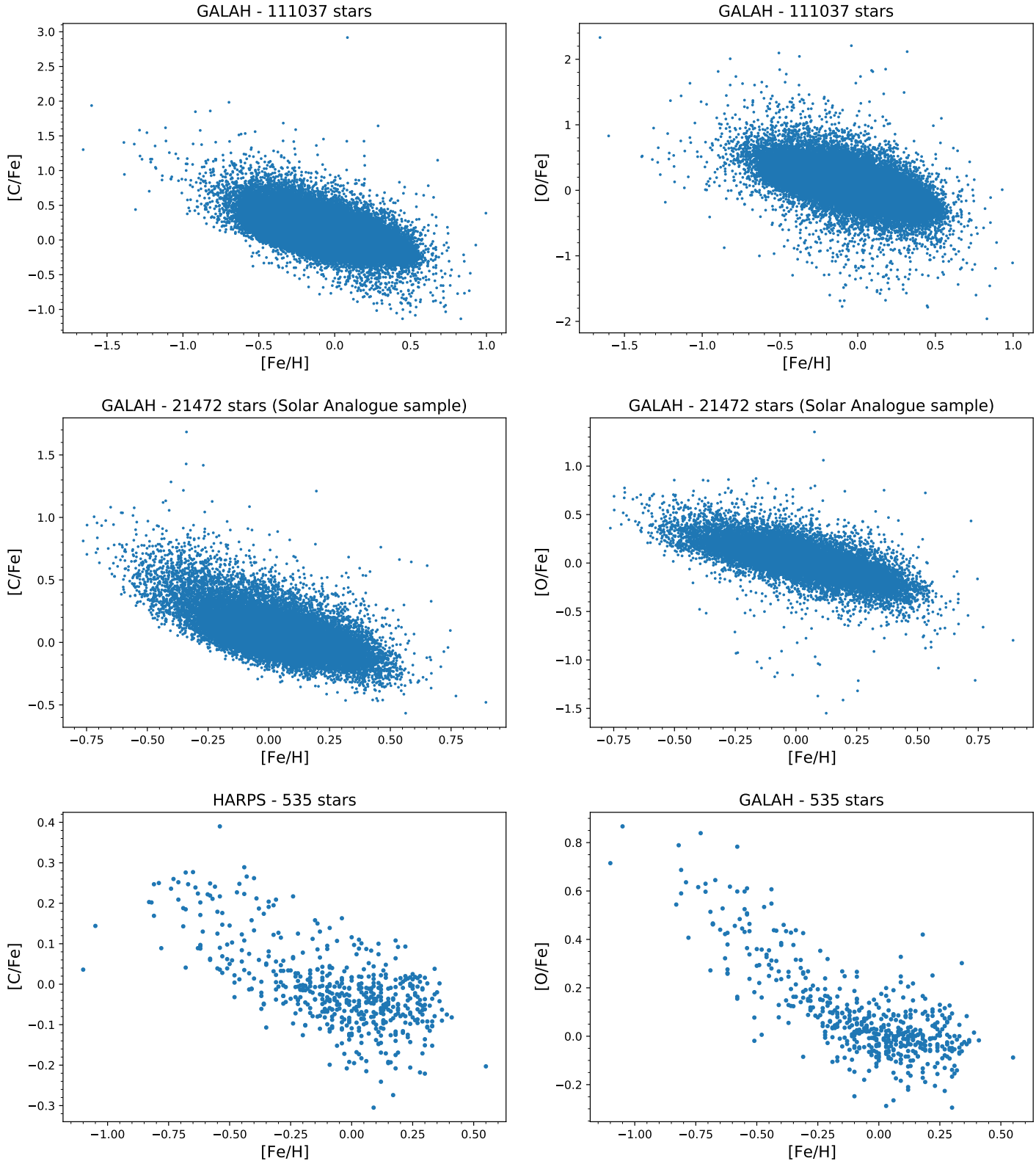


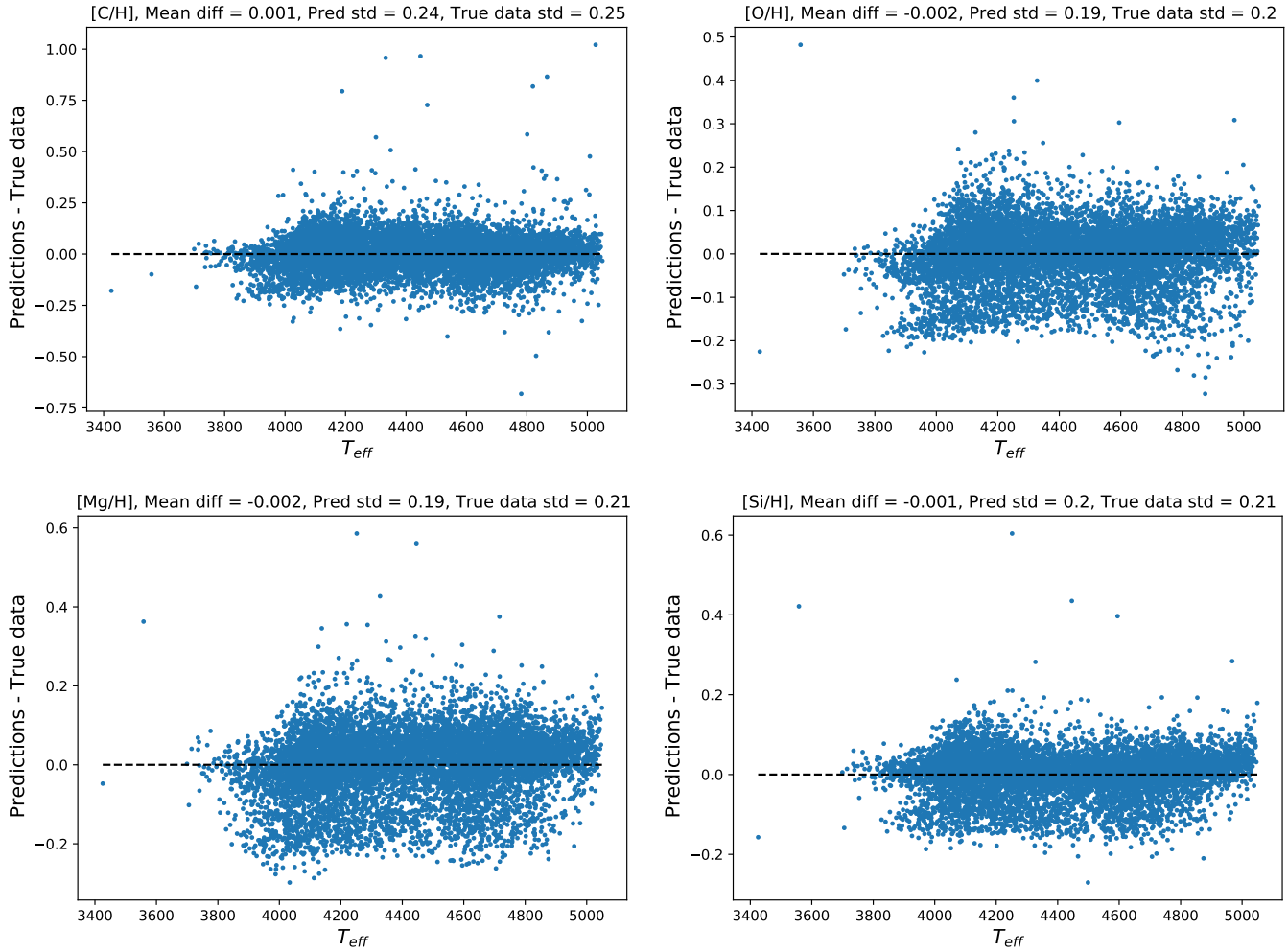
FIGURE B.11: Distribution of [C/Fe] (left) and [O/Fe] (right) abundance ratios with metallicity, for total GALAH sample (top), for the Solar Analogue stars (middle) and for the HARPS sample (bottom).

Appendix C

Random Forest results

C.1 APOGEE

C.1.1 Temperature



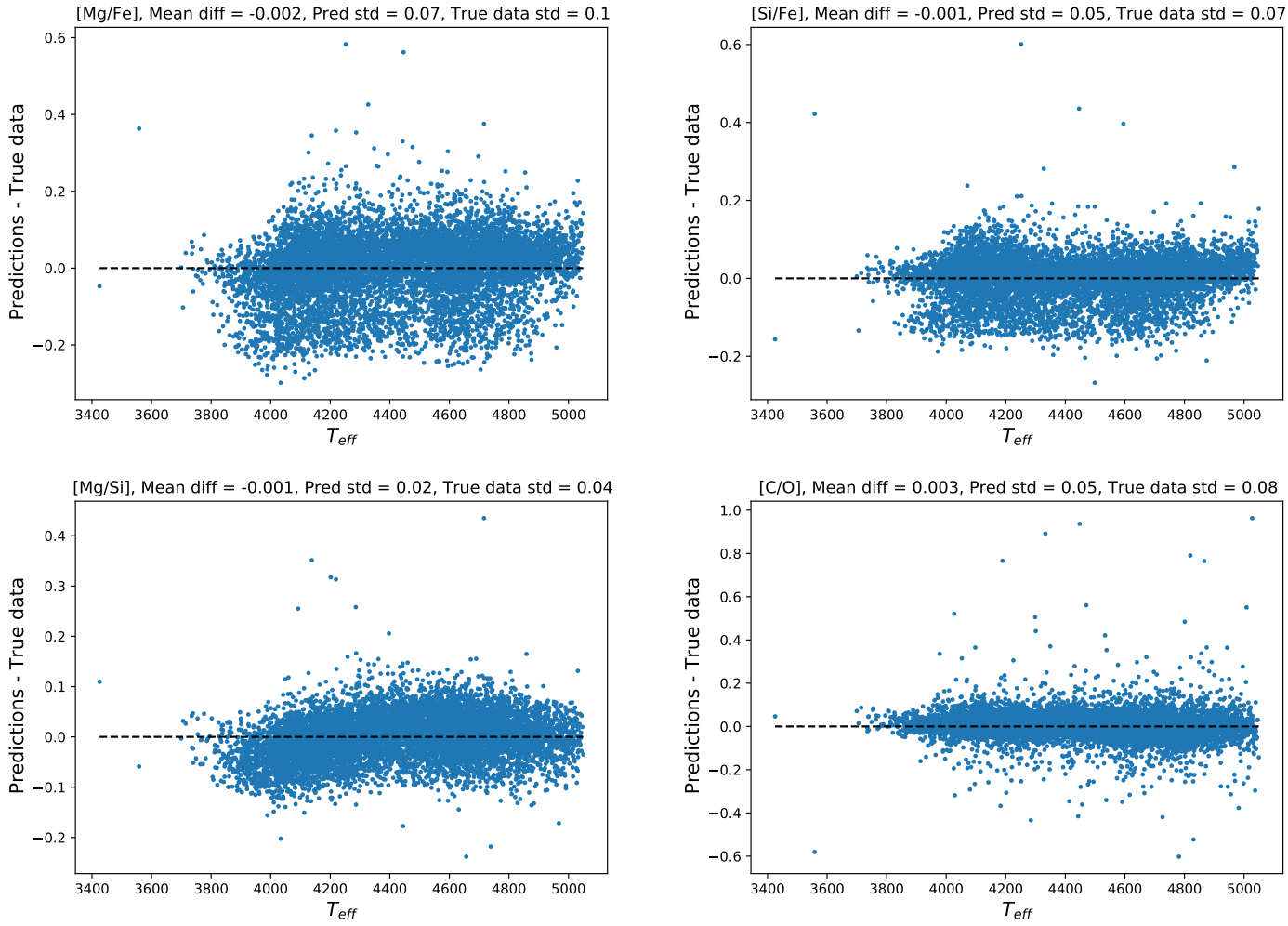
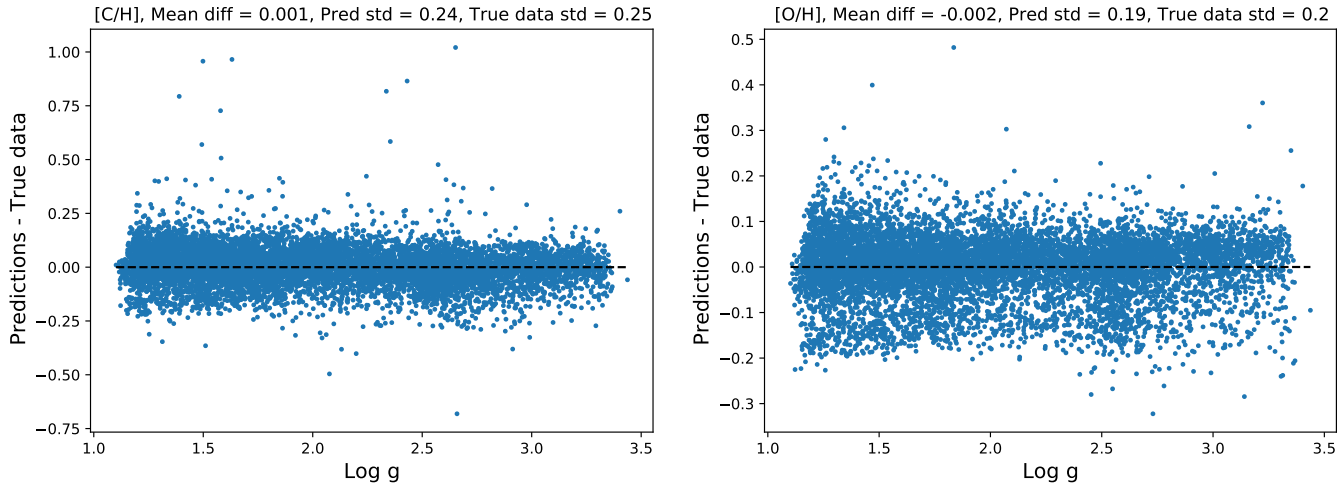


FIGURE C.1: Difference between the predicted values and the true values, depending on temperature, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data.

C.1.2 Log g



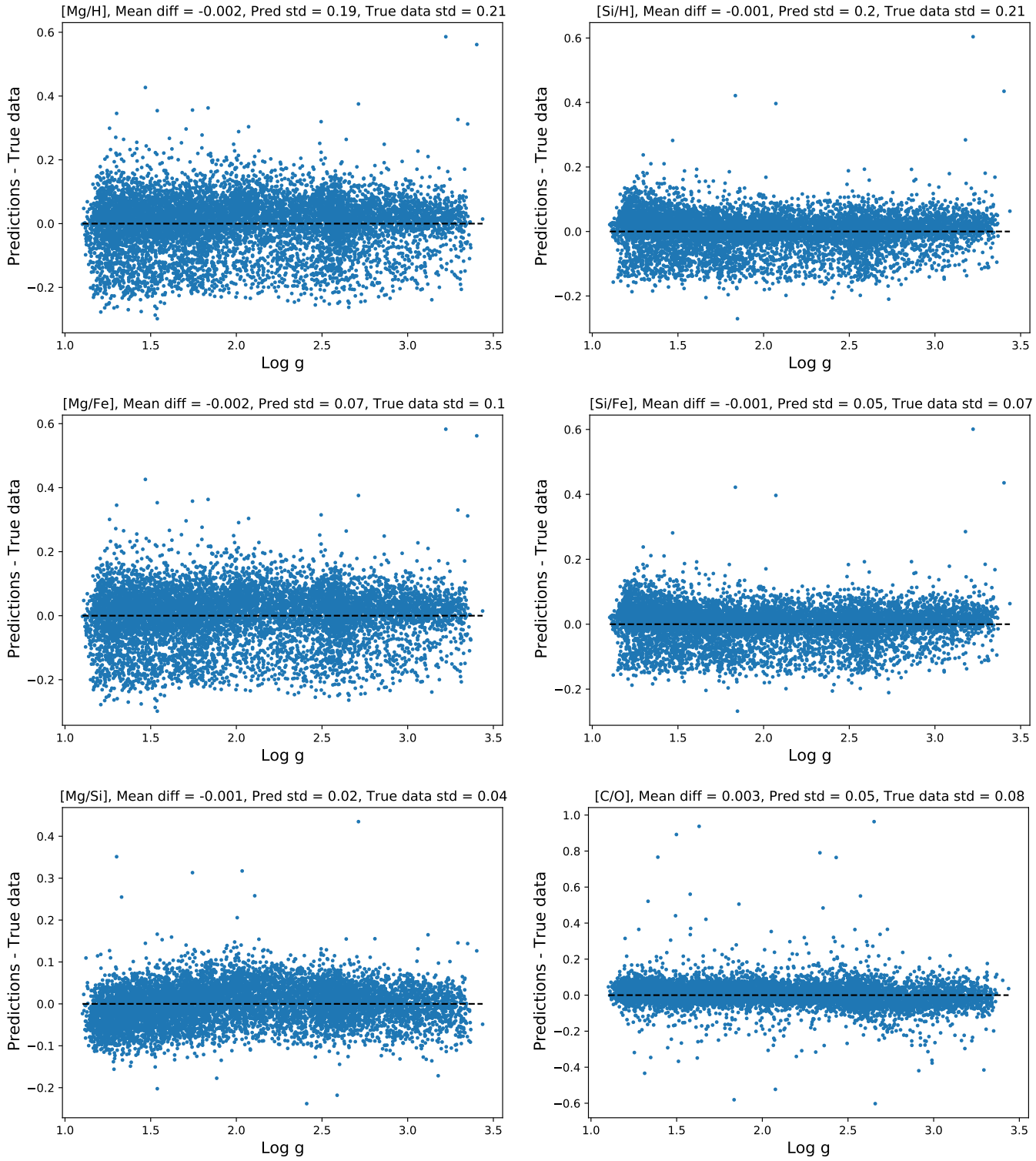
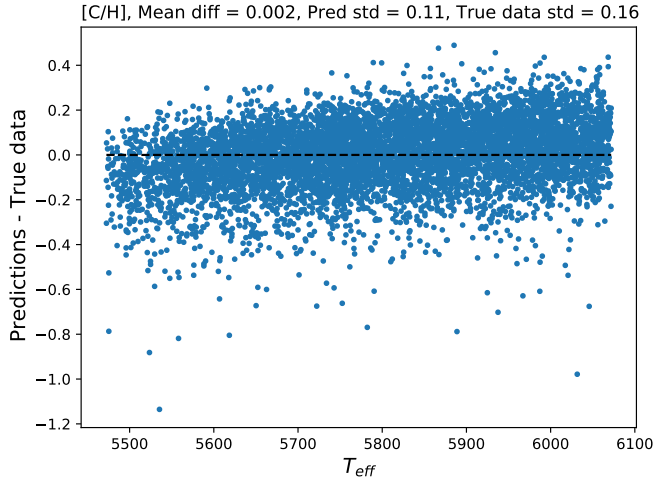
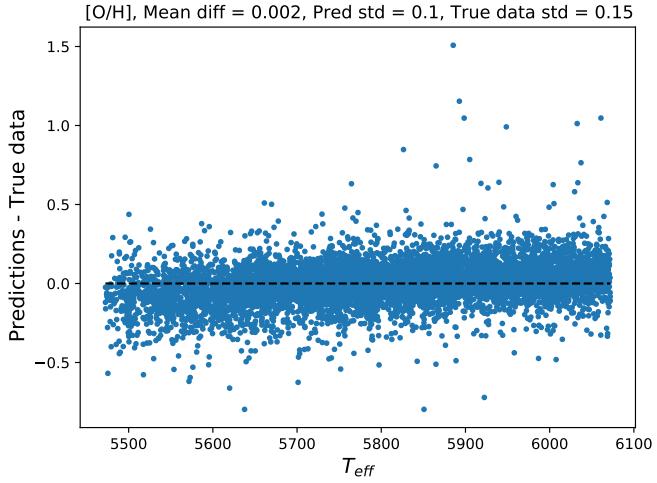


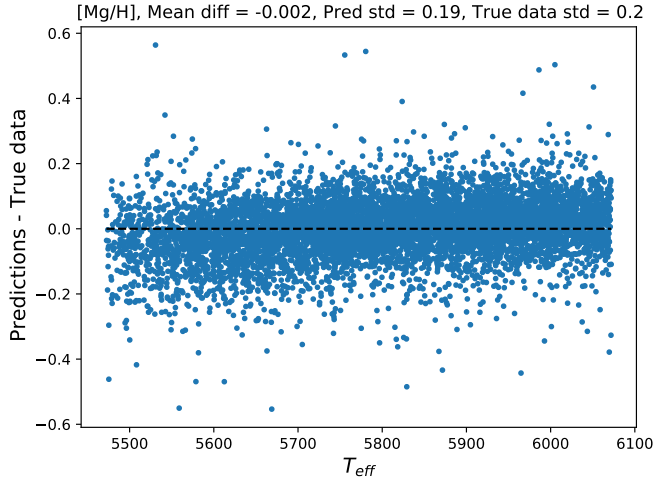
FIGURE C.2: Difference between the predicted values and the true values, depending on $\log g$, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data.

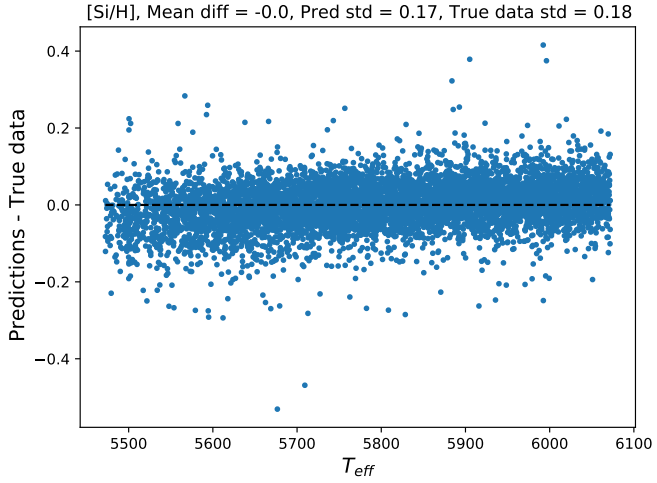
C.2 GALAH

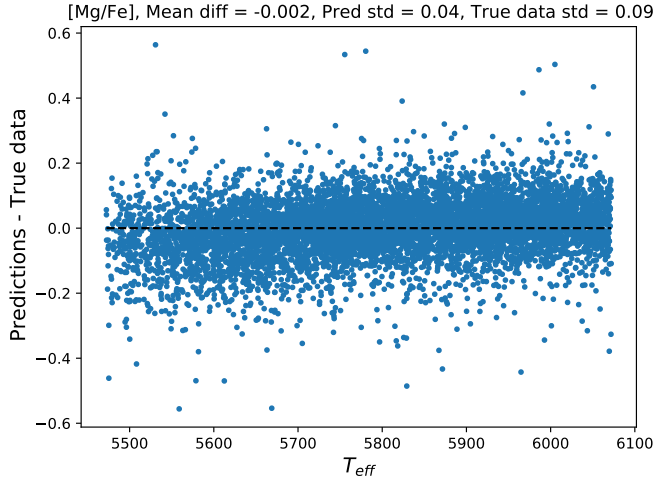
C.2.1 Temperature

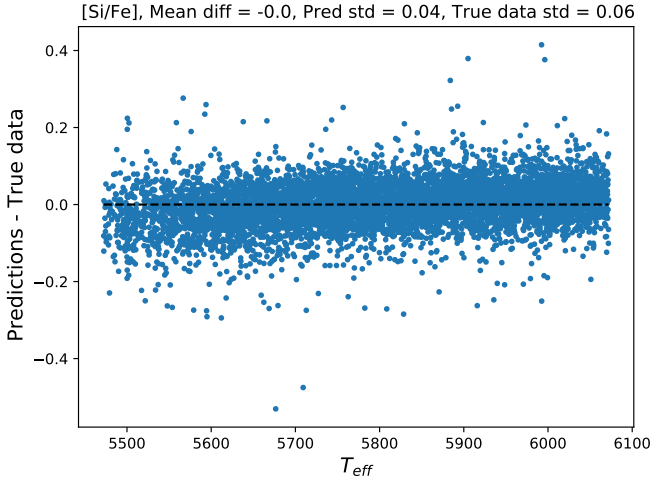












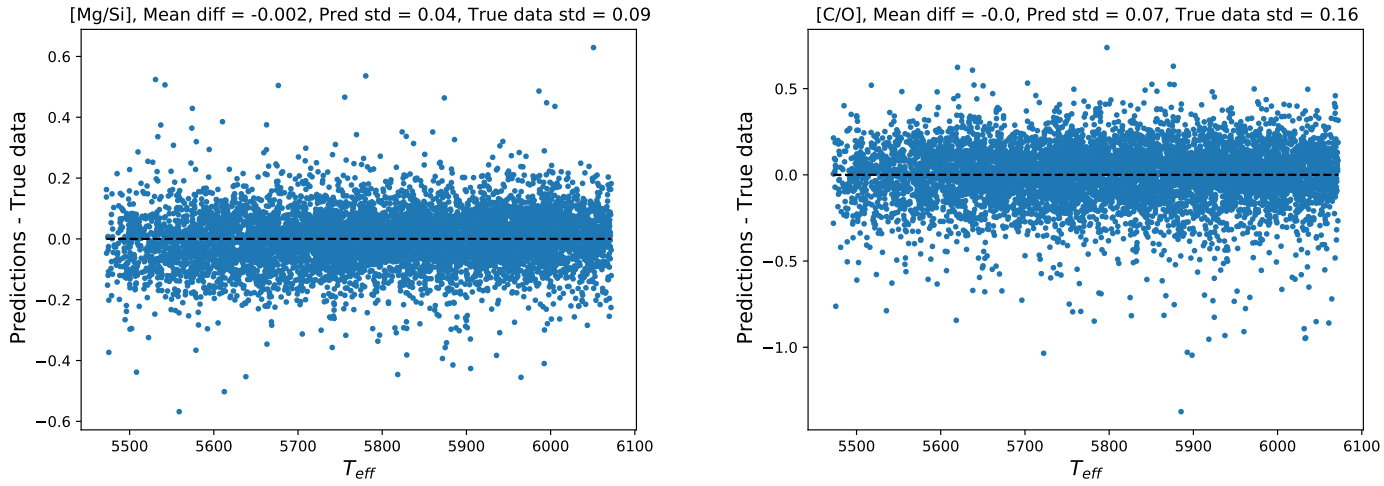
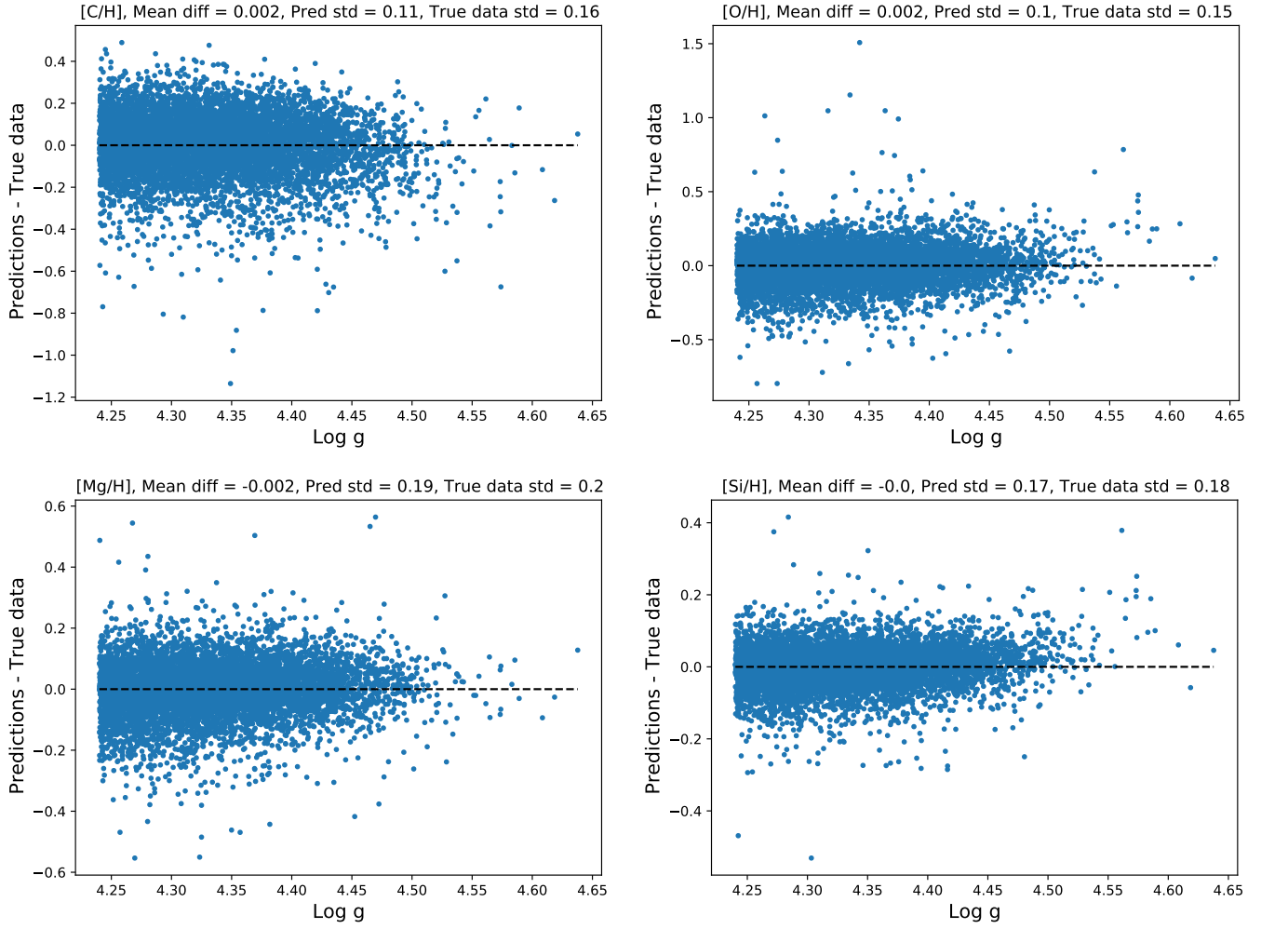


FIGURE C.3: Difference between the predicted values and the true values, depending on temperature, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data.

C.2.2 Log g



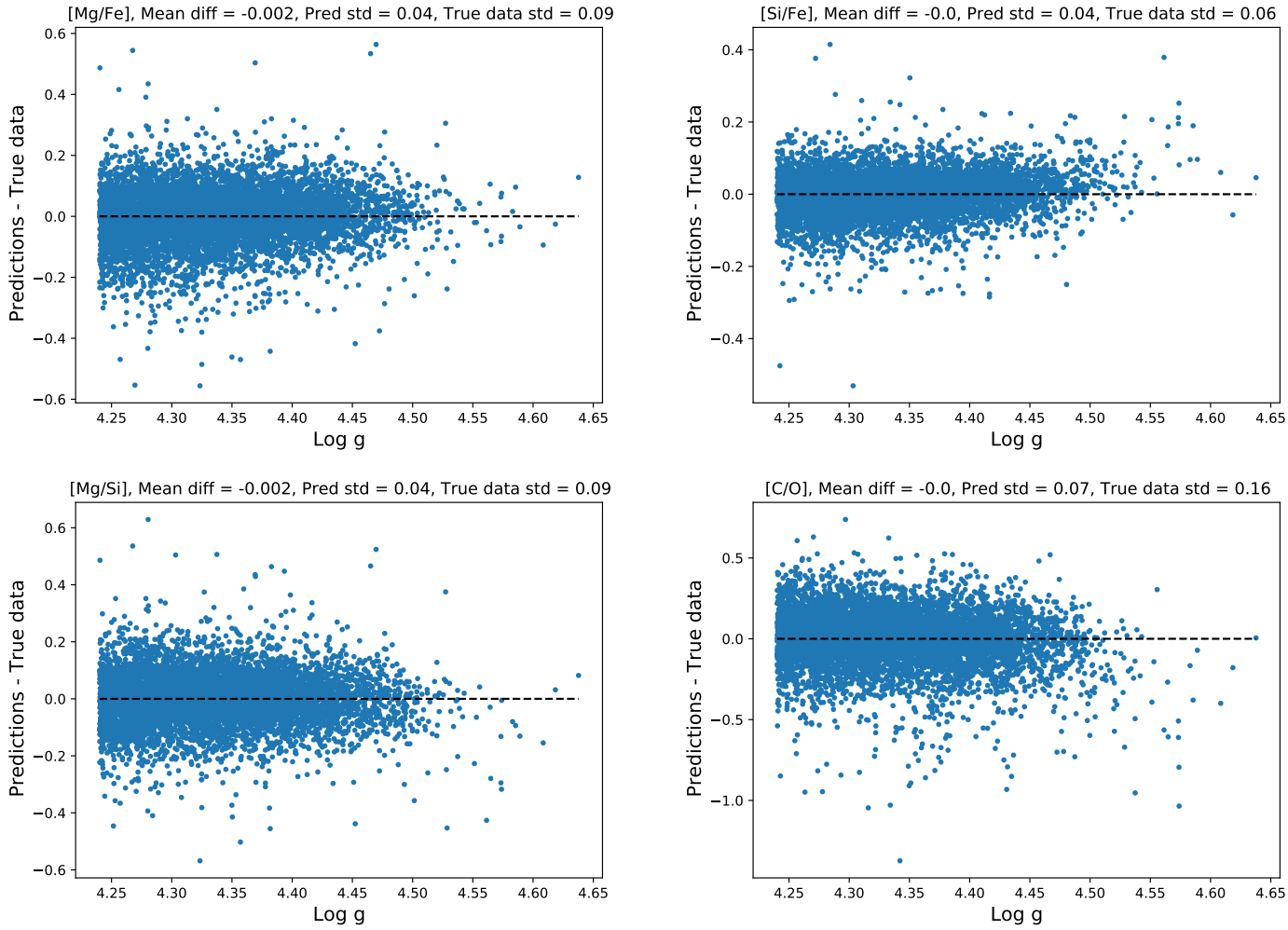


FIGURE C.4: Difference between the predicted values and the true values, depending on $\log g$, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], and abundance ratios of [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data.

C.3 HARPS

C.3.1 Temperature

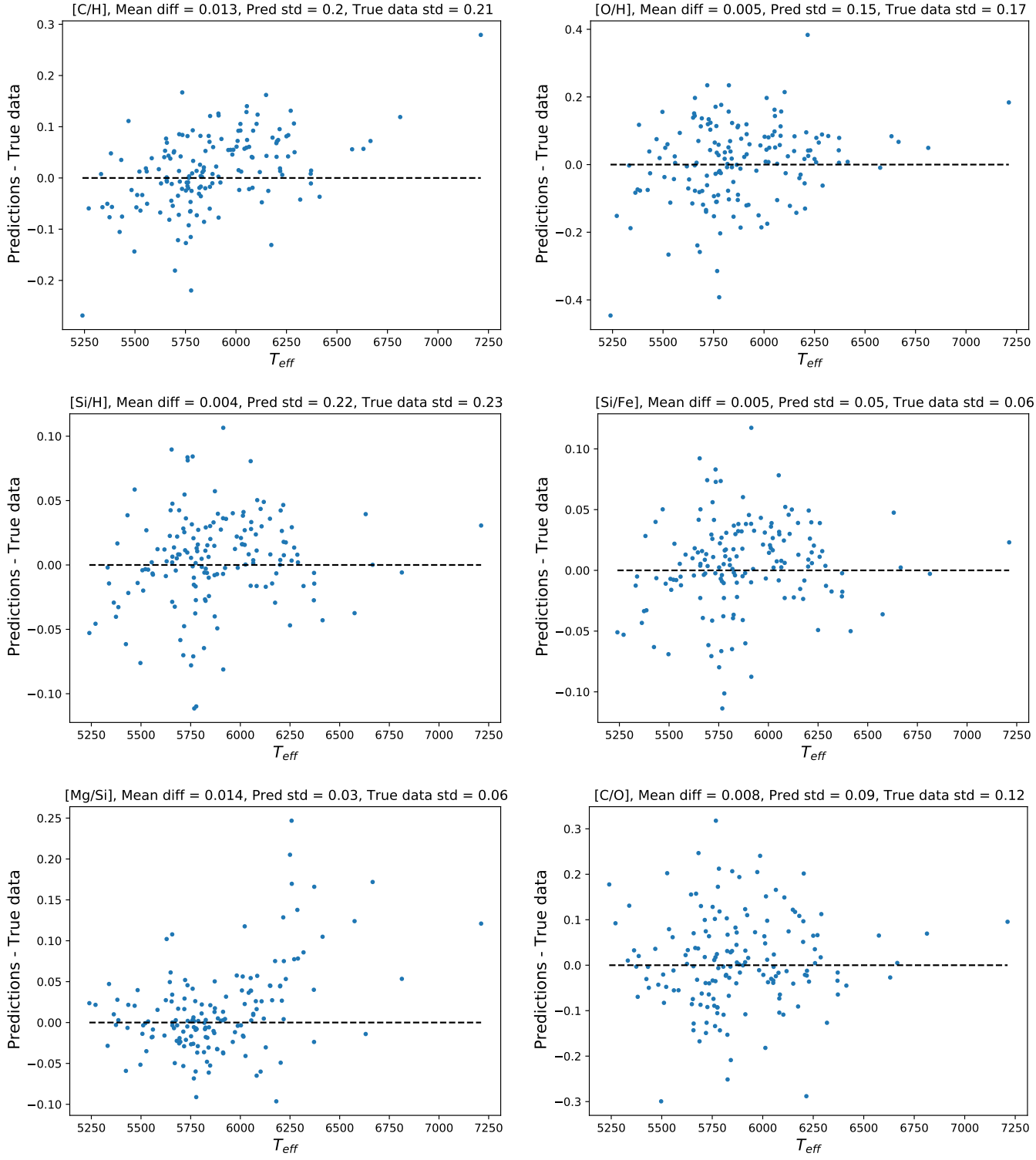


FIGURE C.5: Difference between the predicted values and the true values, depending on temperature, for the chemical abundances of [C/H], [O/H], [Si/H], and abundance ratios of [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data.

C.3.2 Log g

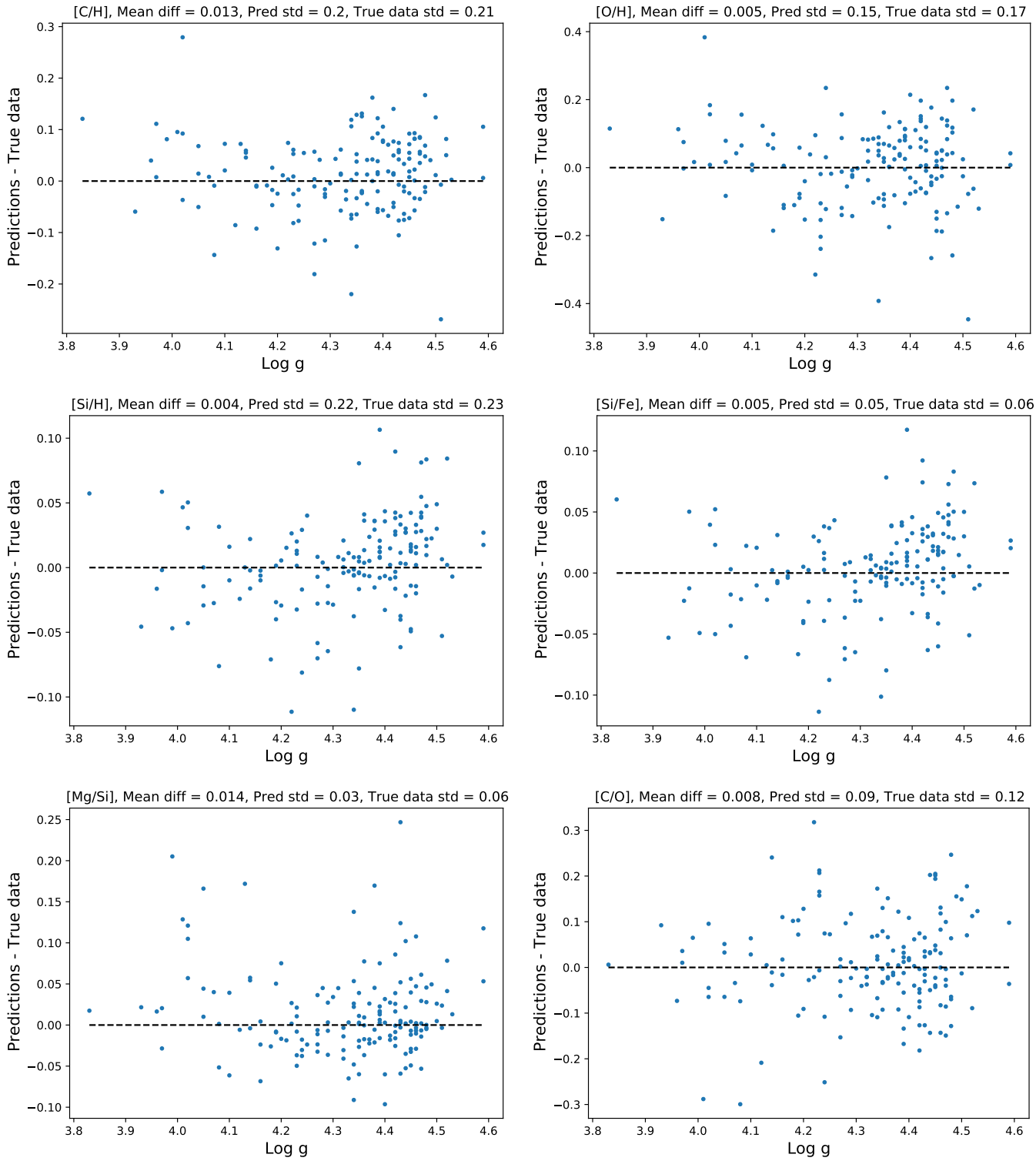


FIGURE C.6: Difference between the predicted values and the true values, depending on $\log g$, for the chemical abundances of [C/H], [O/H], [Si/H], and abundance ratios of [Si/Fe], [Mg/Si] and [C/O]. The black line marks values of predictions equal to the true data.

Appendix D

Uncertainties of predicted abundances

Similar to the real abundance values, the predicted abundances also have uncertainties, which must reflect not only the abundance uncertainties, but also the systematic error from the RF model itself. The precision of the predicted abundances must take into account the uncertainties of the input parameters ([Fe/H] and total galactic velocity) needed to predict the abundances.

D.1 Precision of abundance values

Through bootstrapping, we start by estimating the precision of the input parameters. Assuming a normal distribution where the mean is the real value of the parameter and the sigma is its error, we randomly draw 5000 values, for both [Fe/H] and total galactic velocity. We then apply these 5000 pairs of input parameters on the RF model, in order to predict the chemical abundances.

Subsequently, for each star and each abundance, we do the standard deviation of the 5000 results, which will gives us the uncertainty of the abundance value. We do not, however, do the mean of the results to estimate the abundance value. For lower abundance values, there are more stars with higher abundances than with lower. The similar is true for very high abundances, where there are more stars with lower abundances than higher. Hence, when generating the 5000 values for these extreme cases and then applying them in the model, this would lead to a bias in the mean and would not reflect the real chemical abundances. Therefore, only the precision is estimated and already takes into account the input parameters' uncertainties.

D.2 Determining the systematic error

The model itself has an uncertainty associated to it, since the model could be given input parameters with no error and the results would not be all the same. To determine this systematic error, we start by building 100 random forest models similar to the already existing one, with the only

difference being that these models do not have the `random_state` parameter fixed. Without this parameter fixed, there is no replicability of the results and every time we apply the data to the model, the predicted abundances will be different each time. Applying the data to these models, we get a set of 100 predicted values for each abundance in each star. For each set of 100 values, we do the standard deviation. This will be the systematic error, which is the error inherent to the RF model itself.

D.3 Final error of predicted abundances

By accounting for the systematic error (σ_{syst}) and for the uncertainty of the [Fe/H] and total galactic velocity (σ_{MC}), we can estimate the final error of the predicted abundances through a propagation uncertainty:

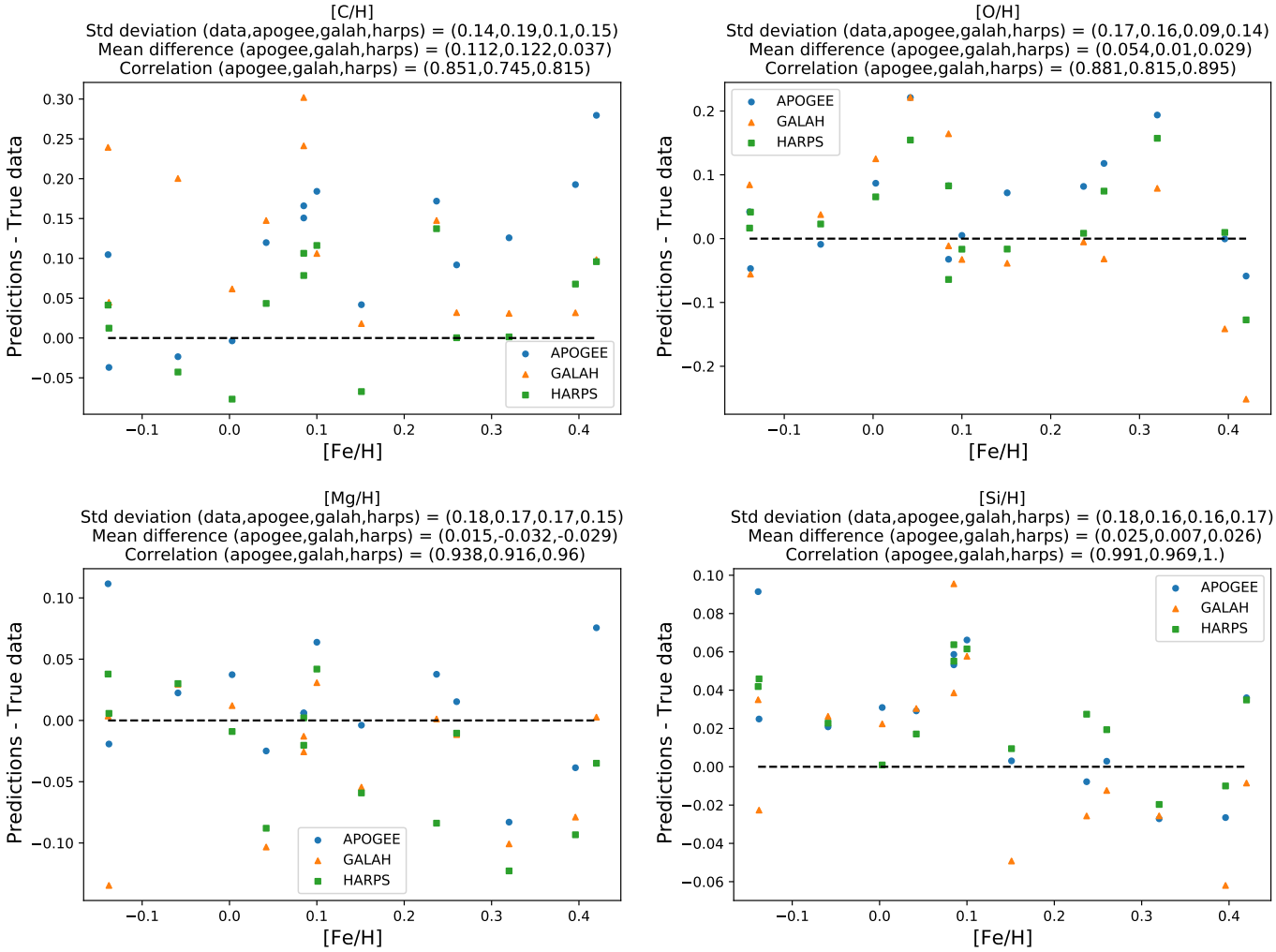
$$\sigma_{final} = \sqrt{\sigma_{syst}^2 + \sigma_{MC}^2} \tag{D.1}$$

This is done for both APOGEE and HARPS data, to estimate the errors of the abundances predicted with RFA and RFH, respectively. The abundance values come not from the mean of the results, but from applying the data sets to the respective RF model, with the `random_state` parameter fixed.

Appendix E

Testing the models with external data

E.1 Testing the models with data from Adibekyan et al. (2021)



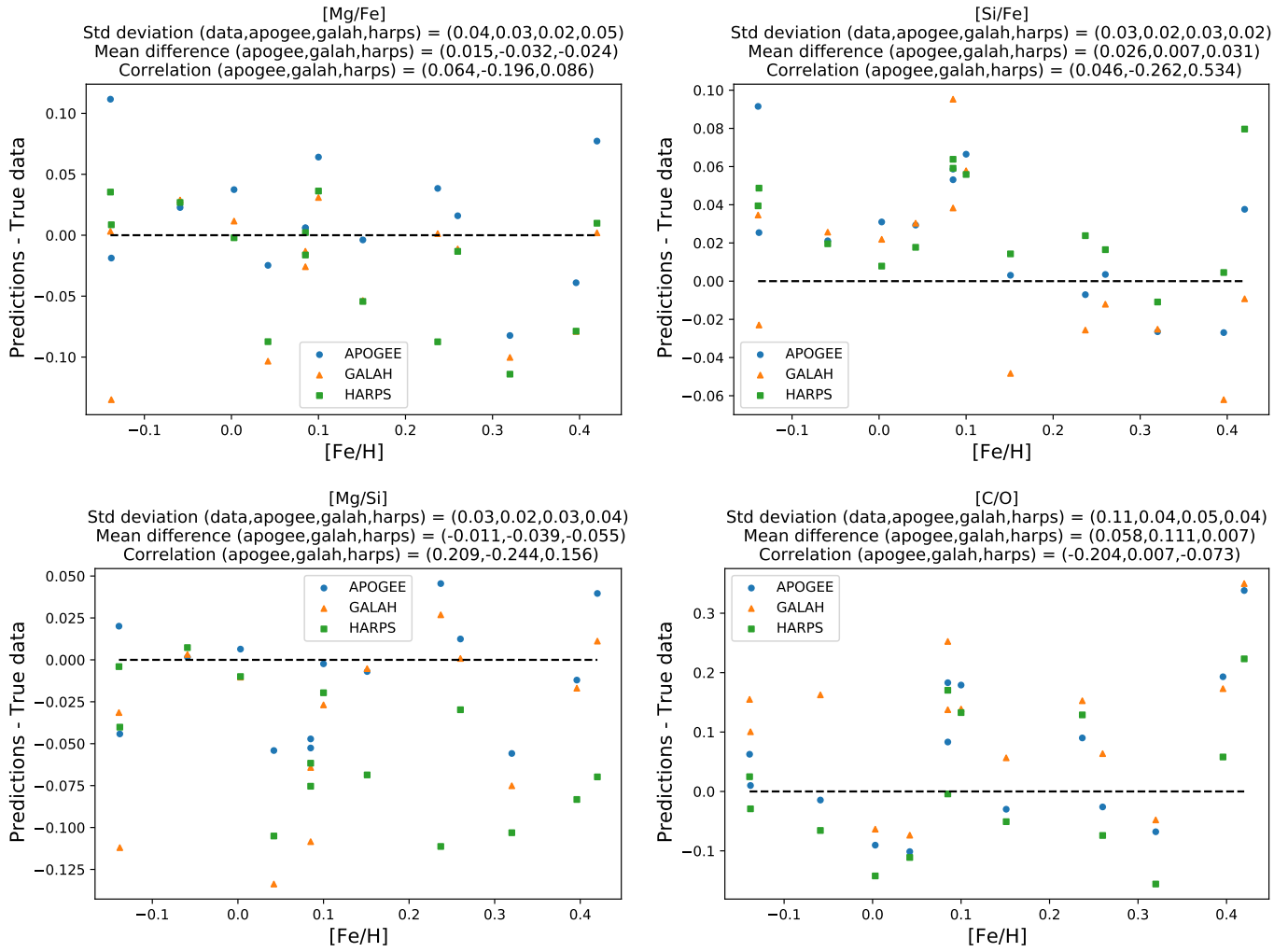


FIGURE E.1: Predictions of Adibekyan et al.(2021)'s data, using the RFA, RFG and RFH models. Y axis is the difference between the predicted values and the true values, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The title includes the standard deviation of true data and predictions, mean difference of *Predictions* − *True data* values, and correlation between prediction and true data, for all samples. The black line marks values of predictions equal to the true data.

TABLE E.1: Mean difference values for Adibekyan et al. (2021)'s data.

	Mean Difference							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.0147	0.0254	0.1118	0.054	0.0151	0.0257	-0.0107	0.0578
GALAH	0.078	0.1311	0.1514	0.056	-0.055	-0.002	-0.053	0.0954
HARPS	0.0835	0.1486	0.134	0.156	-0.0347	0.0304	-0.0651	-0.022

TABLE E.2: Standard deviation values for Adibekyan et al. (2021)'s data.

	Standard deviation							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
Data	0.18164	0.17916	0.13917	0.17162	0.0447	0.02997	0.03081	0.10851
APOGEE	0.16547	0.15704	0.18862	0.15556	0.02968	0.02274	0.01905	0.0351
GALAH	0.13275	0.11697	0.07815	0.05657	0.06185	0.01885	0.06964	0.07405
HARPS	0.13721	0.15812	0.12701	0.14969	0.04316	0.01529	0.04294	0.03341

TABLE E.3: R^2 score for the predictions of Adibekyan et al. (2021)'s data.

	R^2 score for the predictions							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.9237	0.945	-0.0407	0.6724	-0.2699	-0.9793	-0.3108	-0.5838
GALAH	0.4854	0.285	-0.9045	0.1507	-3.9573	-0.9625	-7.4625	-1.5338
HARPS	0.6384	0.26	-0.1733	-0.4064	-1.1869	-1.2515	-7.008	-0.1852

TABLE E.4: Correlation between predictions and true data, for Adibekyan et al. (2021)'s data.

	Correlation between predictions and true data							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.938	0.991	0.851	0.881	0.064	0.046	0.209	-0.204
GALAH	0.862	0.981	0.491	0.425	0.013	-0.378	0.242	-0.337
HARPS	0.887	0.981	0.882	0.805	0.026	-0.004	-0.319	0.242

E.2 Testing the models with data from Maldonado et al. (2020)

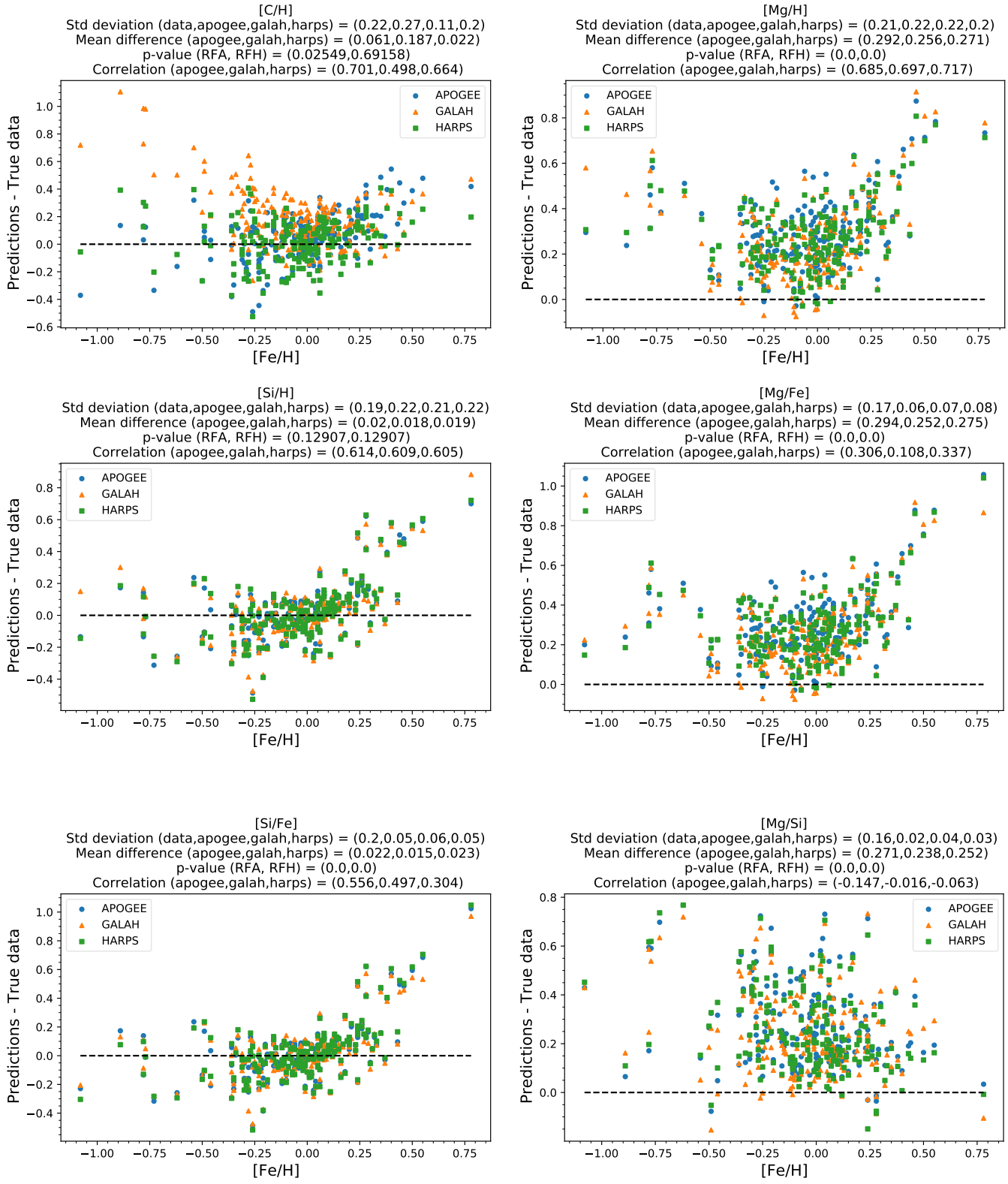


FIGURE E.2: Predictions of Maldonado et al. (2020)'s data, using the RFA, RFG and RFH models. Y axis is the difference between the predicted values and the true values, for the chemical abundances of [C/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe] and [Mg/Si]. The title includes the standard deviation of true data and predictions, mean difference of *Predictions* − *True data* values, and correlation between prediction and true data, for all samples. The black line marks values of predictions equal to the true data.

TABLE E.5: Mean difference values for Maldonado et al. (2020)'s data.

	Mean Difference							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.2905	0.0182	0.0591	—	0.2935	0.0212	0.2723	—
GALAH	0.2558	0.018	0.1867	—	0.2523	0.0145	0.2378	—
HARPS	0.2708	0.019	0.0219	—	0.2745	0.0227	0.2519	—

TABLE E.6: Standard deviation values for Maldonado et al. (2020)'s data.

	Standard deviation							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
Data	0.21199	0.1882	0.22431	—	0.1719	0.20184	0.16239	—
APOGEE	0.2191	0.22052	0.26323	—	0.05353	0.0464	0.01696	—
GALAH	0.21825	0.20946	0.11075	—	0.06894	0.05769	0.04405	—
HARPS	0.20287	0.22049	0.20219	—	0.07532	0.04819	0.03423	—

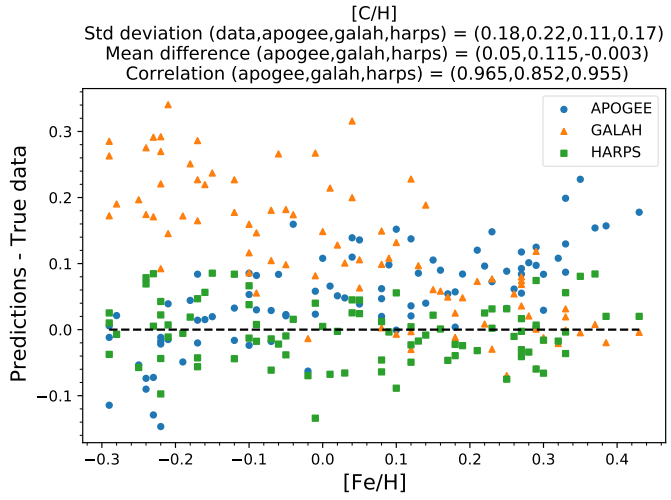
TABLE E.7: R^2 score for the predictions of Maldonado et al. (2020)'s data.

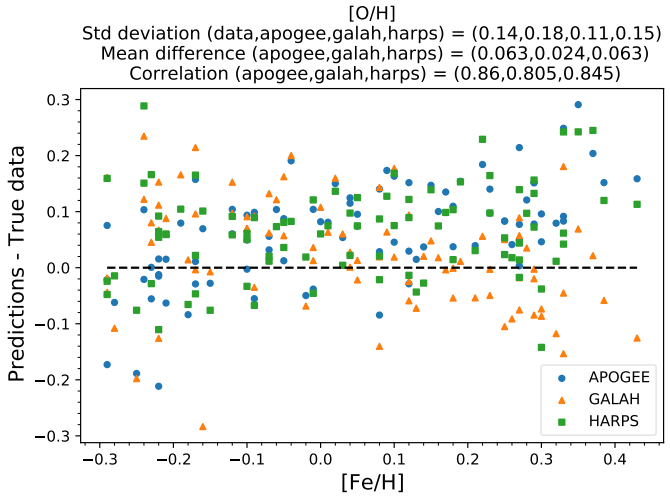
	R^2 score for the predictions							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	-1.3746	0.1631	0.3217	—	-2.7809	0.1695	-2.8177	—
GALAH	-1.1072	0.0975	-0.5203	—	-2.117	0.2022	-2.2149	—
HARPS	-1.1234	0.0866	0.4687	—	-2.4086	0.0923	-2.4811	—

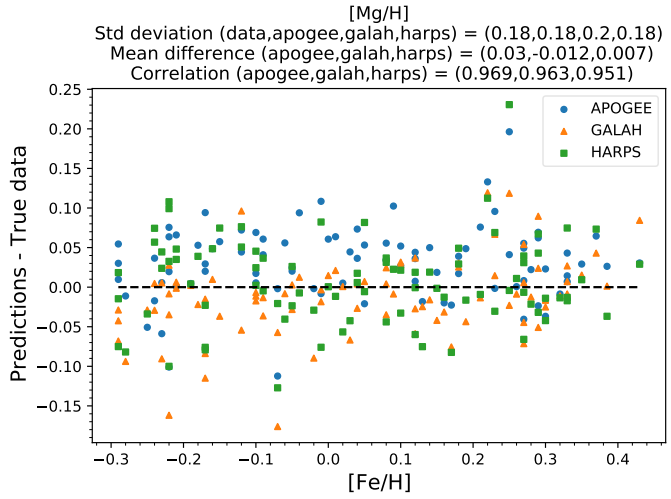
TABLE E.8: Correlation between predictions and true data, for Maldonado et al. (2020)s's data.

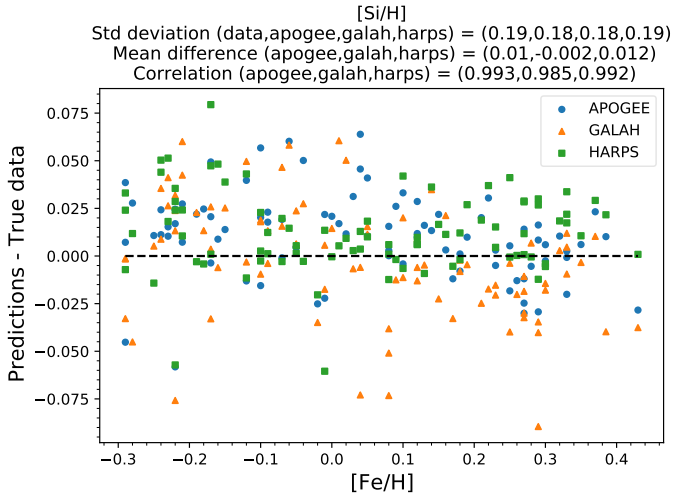
	Correlation between predictions and true data							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.709	0.608	0.687	—	0.383	0.601	0.056	—
GALAH	0.697	0.609	0.498	—	0.108	0.497	-0.016	—
HARPS	0.717	0.605	0.664	—	0.337	0.304	-0.063	—

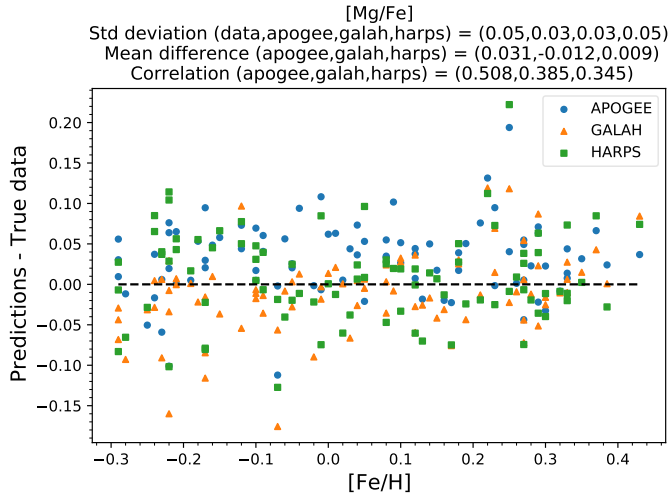
E.3 Testing the models with data from González Hernández et al. (2010)

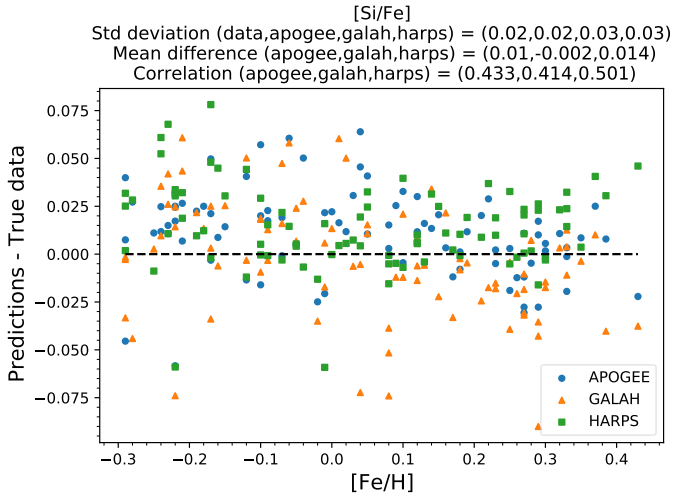












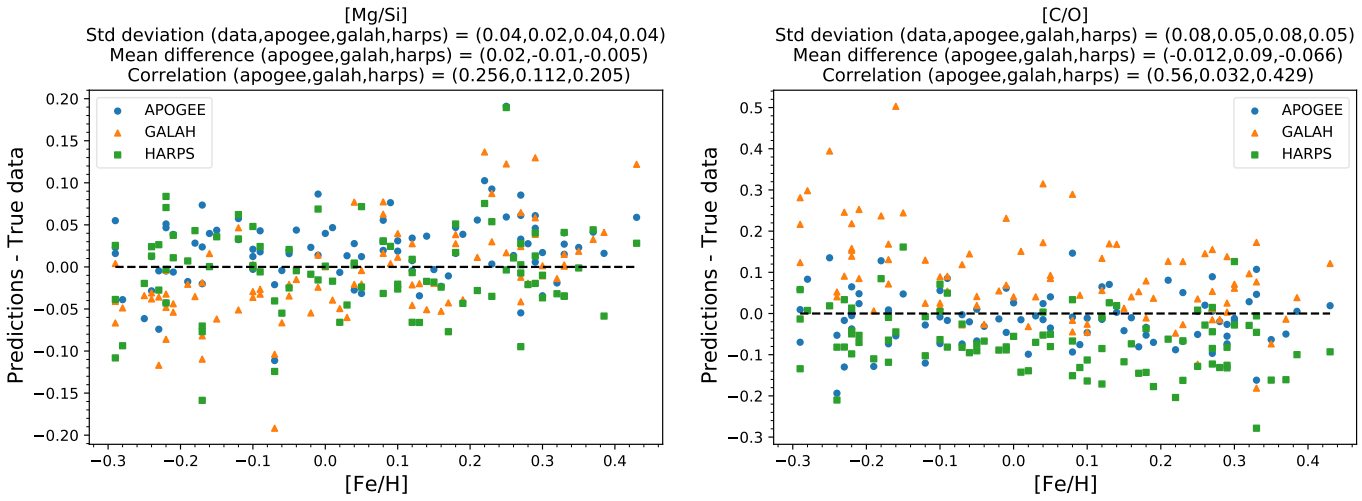


FIGURE E.3: Predictions of González Hernández et al. (2010)'s data, using the RFA, RFG and RFH models. Y axis is the difference between the predicted values and the true values, for the chemical abundances of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The title includes the standard deviation of true data and predictions, mean difference of *Predictions* − *True data* values, and correlation between prediction and true data, for all samples. The black line marks values of predictions equal to the true data.

TABLE E.9: Mean difference values for González Hernández et al. (2010)'s data.

	Mean difference							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.0309	0.0112	0.0497	0.0625	0.0309	0.0113	0.0197	-0.0128
GALAH	-0.012	-0.0024	0.1145	0.0244	-0.0121	-0.0025	-0.0096	0.0902
HARPS	0.007	0.0124	-0.0032	0.0632	0.0089	0.0143	-0.0054	-0.0664

TABLE E.10: Standard deviation values for González Hernández et al. (2010)'s data

	Standard deviation							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
Data	0.18158	0.19003	0.17737	0.14218	0.05399	0.02386	0.04226	0.07759
APOGEE	0.18169	0.18332	0.22294	0.17477	0.03169	0.02322	0.01547	0.05245
GALAH	0.19656	0.17889	0.10693	0.11314	0.03213	0.03212	0.03801	0.07718
HARPS	0.18149	0.1887	0.1741	0.15093	0.0471	0.02689	0.03534	0.04898

TABLE E.11: R^2 score for the predictions of González Hernández et al. (2010)'s data

	R^2 score for the predictions							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.8984	0.9838	0.761	0.4209	-0.1442	-0.0161	-0.3285	0.2566
GALAH	0.9212	0.9743	0.2452	0.5092	0.1101	-0.6438	-0.6858	-2.3666
HARPS	0.9069	0.9834	0.9321	0.4655	-0.0933	-0.2029	-0.4214	-0.6452

TABLE E.12: Correlation between predictions and true data, for González Hernández et al. (2010)'s data

	Correlation between predictions and true data							
	[Mg/H]	[Si/H]	[C/H]	[O/H]	[Mg/Fe]	[Si/Fe]	[Mg/Si]	[C/O]
APOGEE	0.964	0.992	0.961	0.852	0.391	0.499	0.032	0.567
GALAH	0.963	0.985	0.852	0.805	0.385	0.414	0.112	0.032
HARPS	0.951	0.992	0.955	0.845	0.345	0.501	0.205	0.429

Appendix F

Additional tests to the ML models

F.1 SWEET-Cat database

Contrary of the four tests performed in Section 2.5, we now apply the three models to the SWEET-Cat[1] data, which does not have known chemical abundances associated.

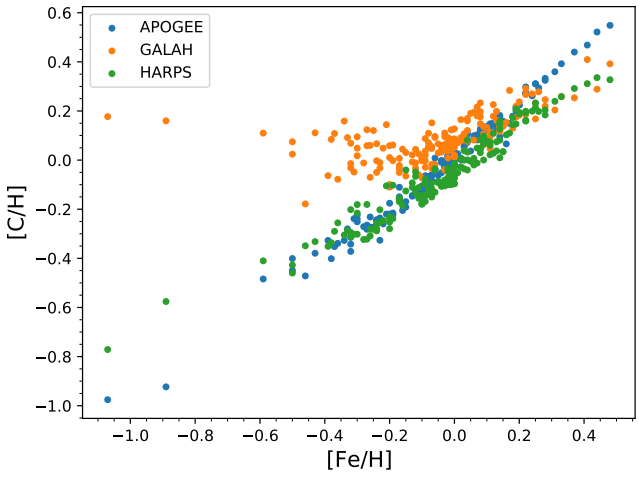
The SWEET-Cat is a catalog of stellar parameters for stars with planets listed in the Extrasolar Planets Encyclopaedia and described in Santos et al. (2013), with improvements made by Sousa et al. (2021). It compiles sets of atmospheric parameters (including T_{eff}, $\log g$, and [Fe/H]) derived using the same uniform methodology whenever possible (e.g. Santos et al. 2004, Sousa et al. 2008) or, if not possible, atmospheric parameters already published in the literature.

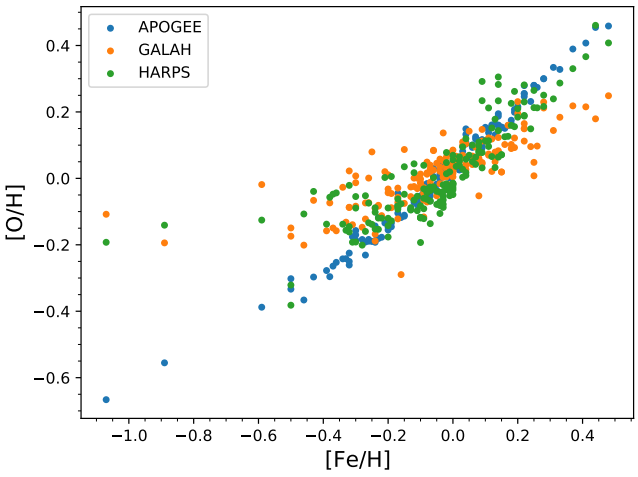
In this case, we focus mainly on cooler stars. From the SWEET-Cat database, we extract all the stars with $T_{eff} < 4000$ K and match them with data from the The Extrasolar Planets Encyclopaedia[2] to get their planets' data, leaving a sample of 211 stars. From SWEET-Cat we get the metallicity, coordinates, proper motions and parallax. Using GAIA information, we get the radial velocity to compute the galactic total velocity. Next, we use the random forest models RFA, RFG and RFH (applied to APOGEE, GALAH and HARPS, respectively) to predict the abundance ratios. All SWEET-Cat data subset is taken from literature.

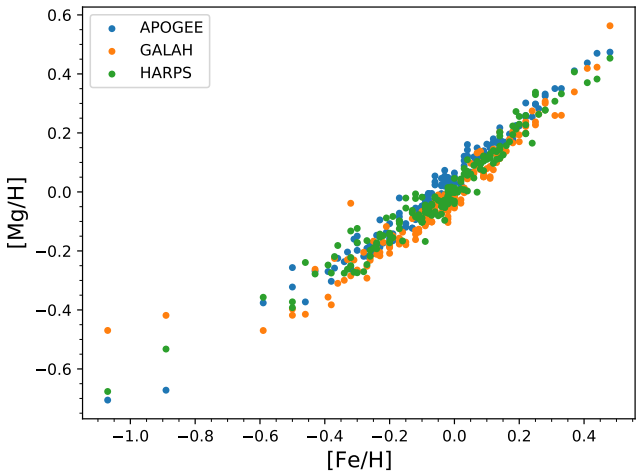
As we can see from Figure F.1, for the [Mg/H] and [Si/H] abundance, the three models are in agreement and have predictions very close to each other. However, the same is not true for the [C/H], [O/H] and [C/O] abundances, where it is possible to notice the GALAH model predictions have a different behaviour than that of the APOGEE and HARPS models.

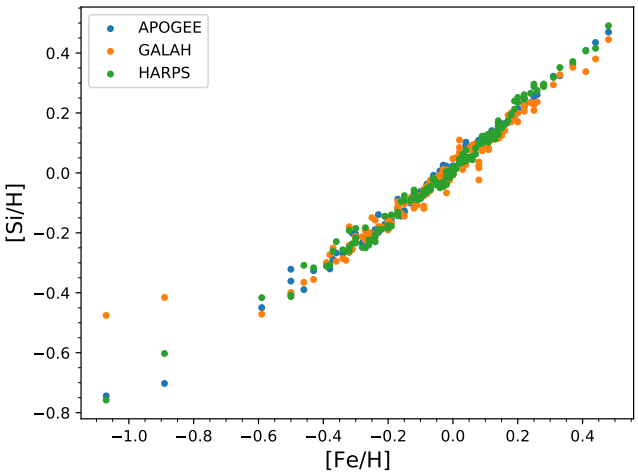
For the [Mg/Fe], [Si/Fe] and [Mg/Si] abundance ratios, there is not a very clear distinction, although one could argue that for [Mg/Fe] and [Mg/Si], the RFG model tends to lower predicted values when comparing with the RFA and RFH abundance predictions.

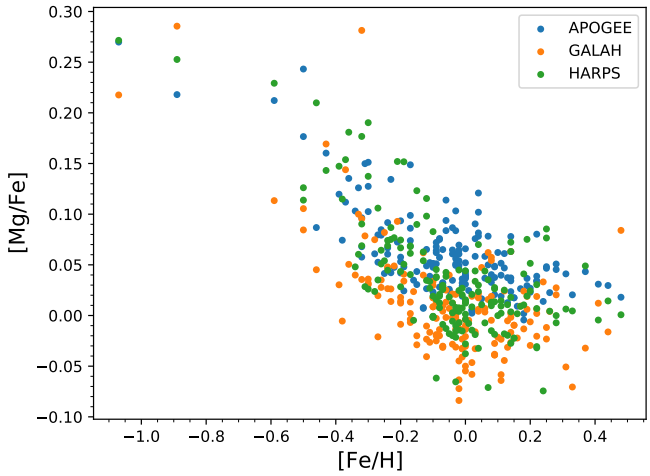
[1] http://sweetcat.iastro.pt/
[2] exoplanet.eu

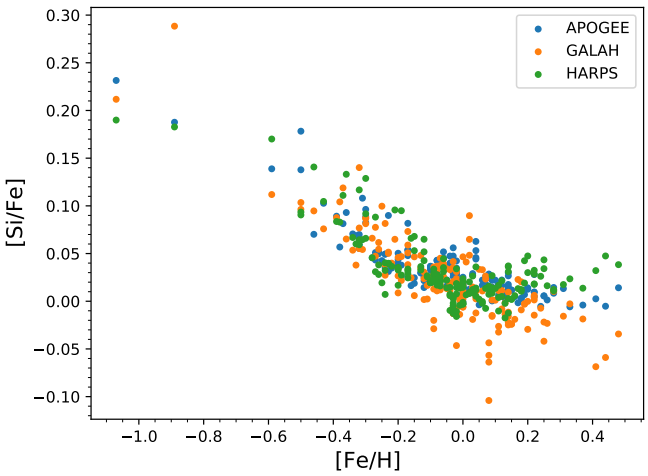

APOGEE
GALAH
HARPS
[C/H]
[Fe/H]


APOGEE
GALAH
HARPS
[O/H]
[Fe/H]


APOGEE
GALAH
HARPS
[Mg/H]
[Fe/H]


APOGEE
GALAH
HARPS
[Si/H]
[Fe/H]


APOGEE
GALAH
HARPS
[Mg/Fe]
[Fe/H]


APOGEE
GALAH
HARPS
[Si/Fe]
[Fe/H]

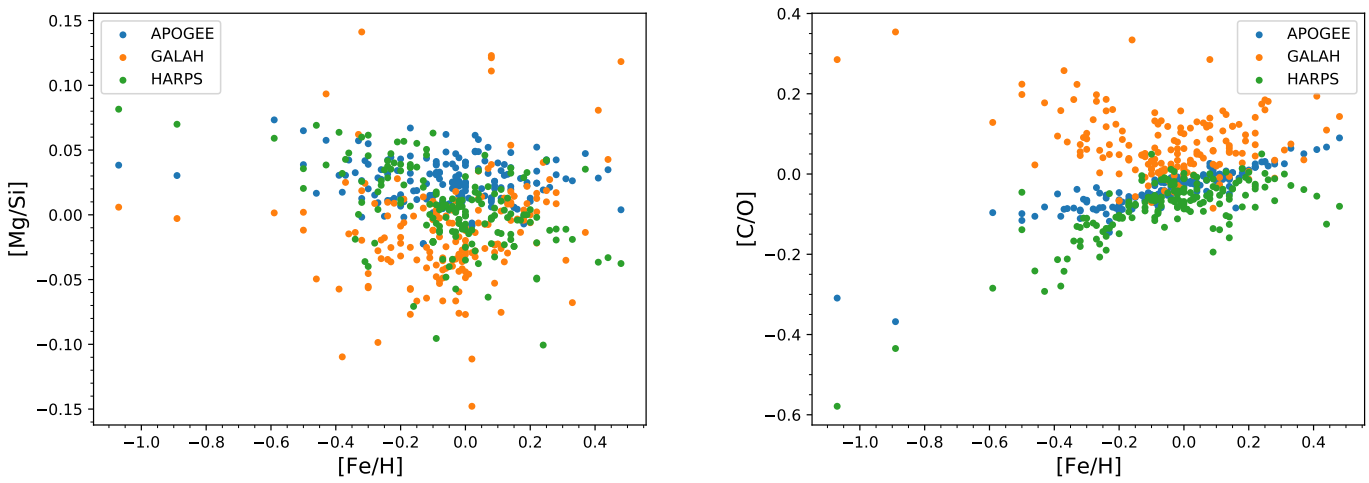
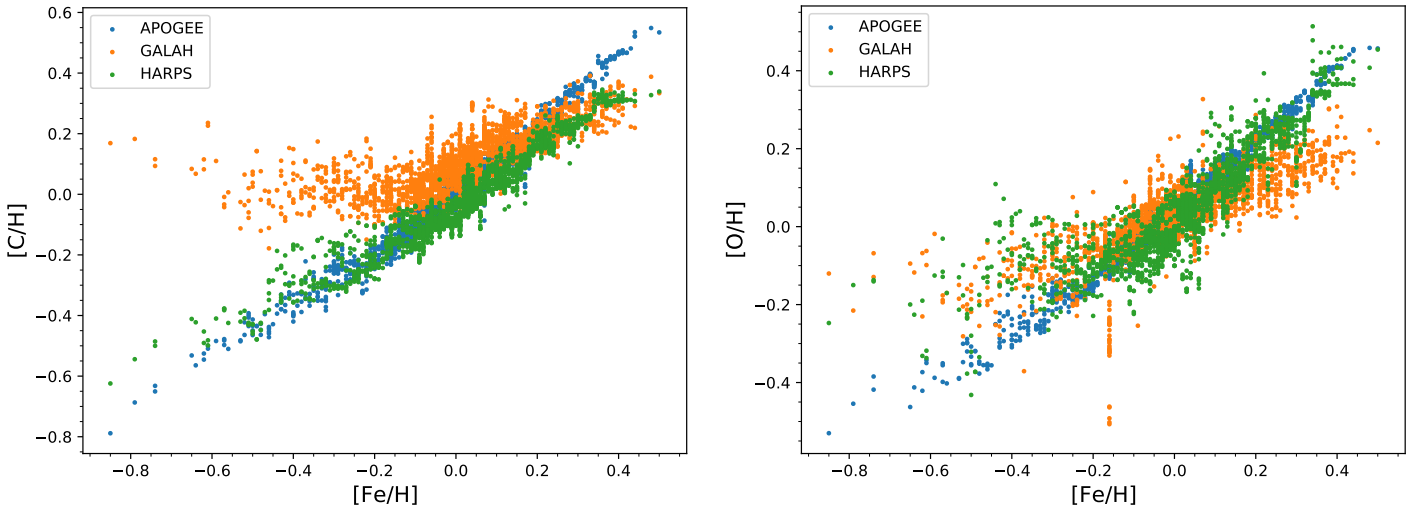


FIGURE F.1: Abundance predictions of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O], for 211 stars with $T_{eff} < 4000$ K from SWEET-Cat dataset, using the APOGEE, GALAH and HARPS random forest model.

To make sure this was not a behaviour particular to just cooler stars, we do the same procedure and apply the three models to all the SWEET-Cat stars. As we can see from Figure F.2, it is possible to make the same conclusions.



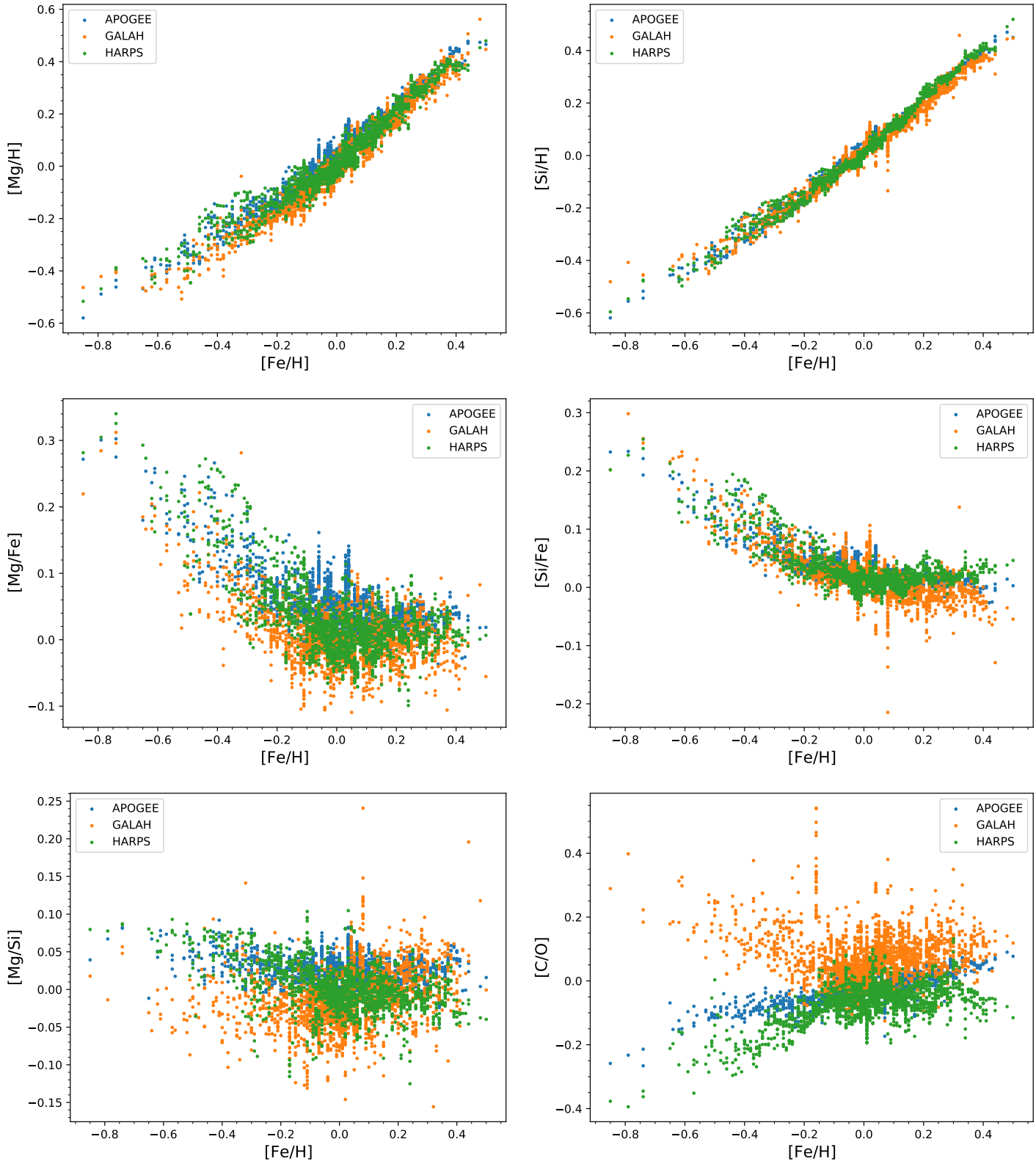


FIGURE F.2: Abundance predictions of [C/H], [O/H], [Mg/H], [Si/H], [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O], for all stars from SWEET-Cat dataset, using the APOGEE, GALAH and HARPS random forest model.

F.2 Predicting the [Mg/Si] abundance ratio

As we have seen in Sections 2.4 and 2.5, usually the models can predict the individual abundances with some degree of success. However, the same cannot be said of abundance ratios such as [Mg/Fe], [Si/Fe], [Mg/Si] and [C/O]. The [Mg/Si] for example is an important mineralogical ratio as it has an effect in the structure and composition of rocky planets (explained in more detailed in Chapter 3). Since we cannot accurately predict directly this ratio, we might be able to compute through the Mg and Si abundances, since these abundances have the best prediction results:

$$\begin{aligned}
[Mg/H] - [Si/H] &= \log\left(\frac{Mg}{H}\right)_{star} - \log\left(\frac{Mg}{H}\right)_{Sun} - \log\left(\frac{Si}{H}\right)_{star} + \log\left(\frac{Si}{H}\right)_{Sun} \\
&= \log\left(\frac{Mg}{Si}\right)_{star} - \log\left(\frac{Mg}{Si}\right)_{Sun} \\
&= [Mg/Si]
\end{aligned} \tag{F.1}$$

In that sense, we plot the [Mg/Si] determined through the Mg and Si abundance predictions, depending on the predicted [Mg/Si] using the RFA, RFG and RFH models. We apply this in Nissen (2015)'s and Adibekyan et al. (2021)'s data. The result are present in Figure F.3 and F.4, respectively. As we can see, regardless of the sample, the difference between both values of [Mg/Si] are at most in the order of 10^{-9}, meaning that it is not strictly necessary to build the model to predict the [Mg/Si] abundance ratio, because we can compute it with the [Mg/H] and [Si/H] abundances.

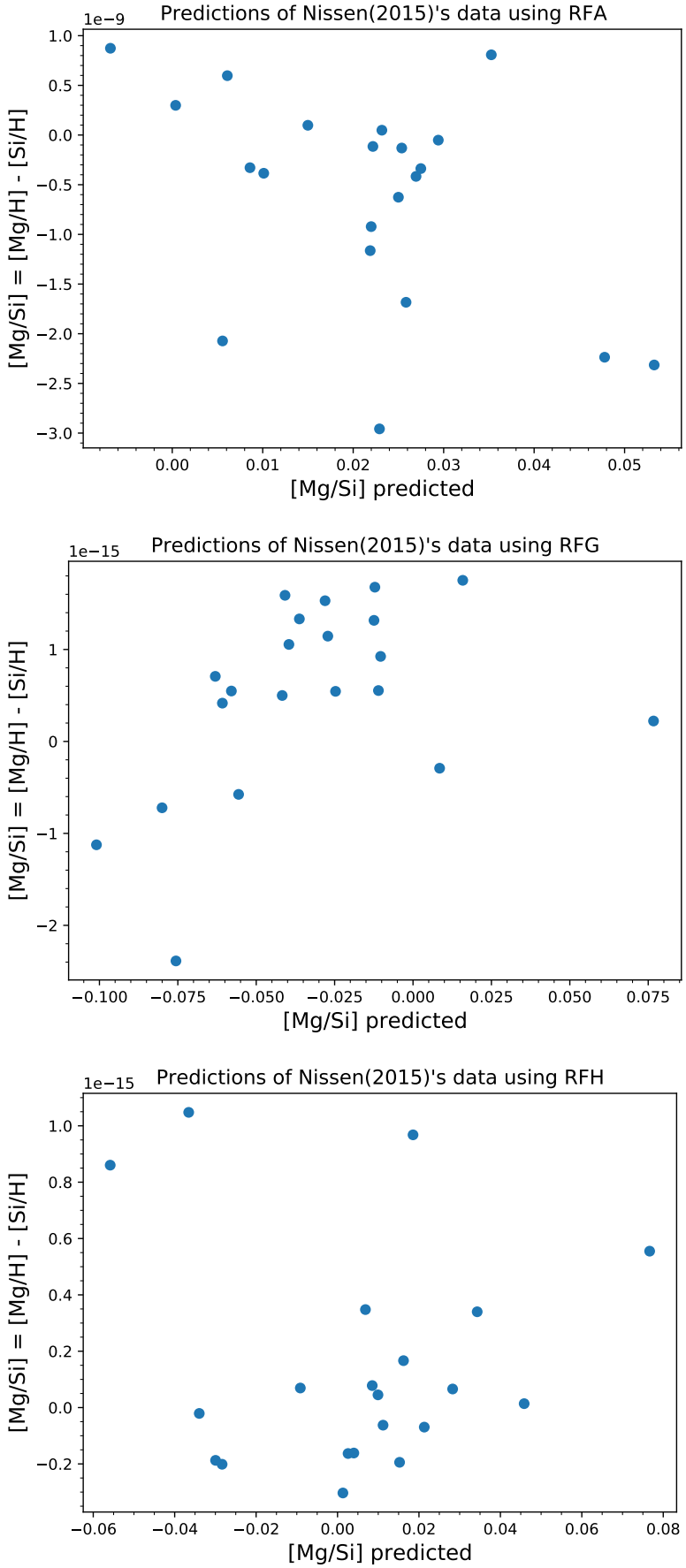


FIGURE F.3: [Mg/Si] determined from the difference of Mg and Si abundances depending on the predicted values of [Mg/Si] from Nissen (2015)'s data, using RFA (top), RFG (middle) and RFH (bottom) models.

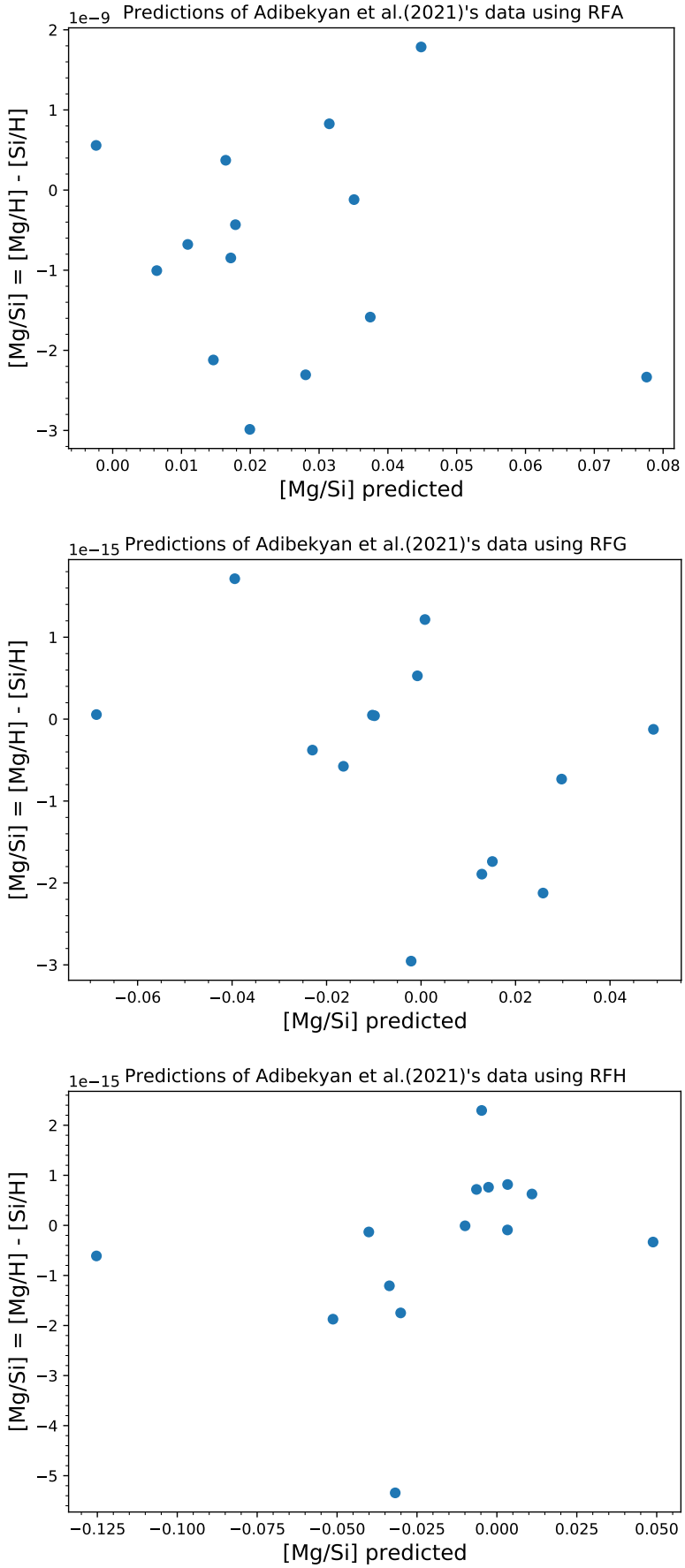


FIGURE F.4: Same as Figure F.3, but for Adibekyan et al. (2021)'s data.

F.3 Predicting the C and O abundances

Like [Mg/Si], the [C/O] ratio cannot also be precisely predicted by the RF models. However, contrary to [Mg/Si], we cannot determine this ratio through the C and O abundances, as these are not so well predicted by the models like the Mg and Si abundances are. Therefore, to check our abundances, we compare our result obtained after applying the RFA, RFG and RFH models to Adibekyan et al. (2021)'s data, with predictions done by the author, using a different model. We do this for the [C/H] and [O/H] abundances. Figure F.5 shows the results for C abundances (left) and O abundances (right).

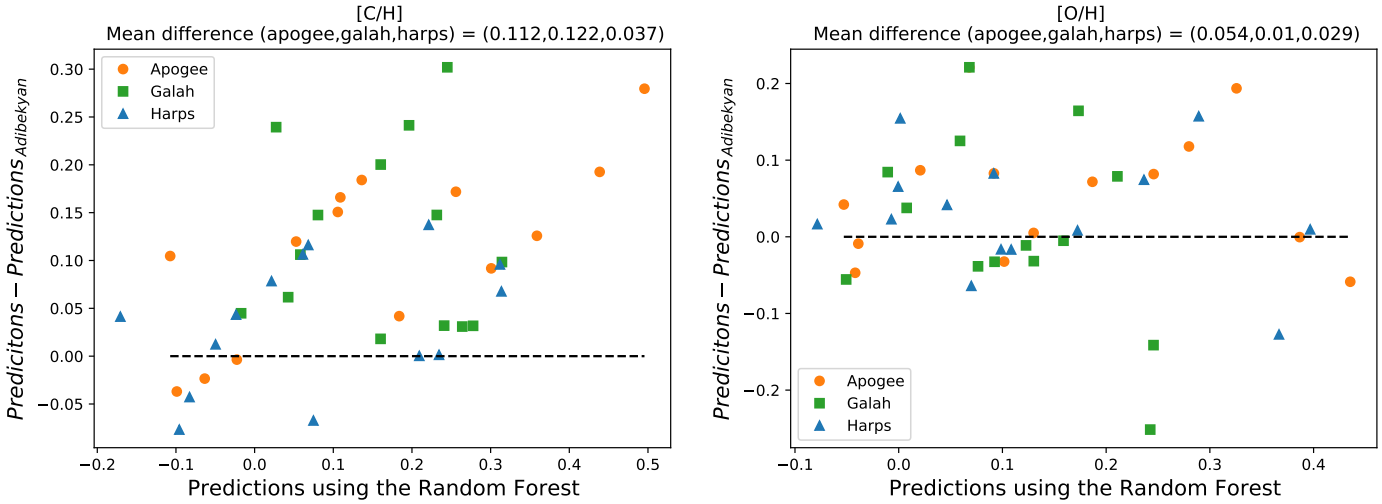


FIGURE F.5: Comparing the predictions of [C/H] and [O/H] abundances using the RFA, RFG and RFH models and the predictions of Adibekyan et al. (2021). Y axis is the difference between our predicted values and those of the author. X axis represents our values obtained using the RF models. Title includes information about the mean value of $Predictions - Predictions_{Adibekyan}$.

By analysing the plots, we see that on the [C/H] predictions, GALAH results seems to get farther away from the author's predictions, for higher abundances to higher values. It is also the model with higher mean difference value, which matches with the tendency to produce results farther away from the author's predictions, for higher abundance values. Similarly, results from the RFA model also show this trend.

On the other hand, predictions for the O abundance from the RFA, RFG and RFH models seem more similar to those of the author's results. The mean difference values are also very similar between themselves, which means the four models have a better agreement on the predictions.

So, while for the O abundance, the RFA, RFG and RFH models' results are compatible with those of the author's independent models, the same can't be said for C abundances. This leads to a need of extra care when trying to predict [C/H].

F.4 Compare HARPS and Nissen (2015) data

When testing the different models using Nissen (2015)'s sample, although the R^2 score for the HARPS model in Table 2.9 pointed to bad predictions, the plots in Figure 2.14 showed the predictions were around

the null line (when the predictions equals the true data) and the standard deviation was never higher than 0.1, which should indicate better results. To explore this, we select the common stars between the HARPS sample and Nissen (2015)'s data and plot the individual abundances depending on metallicity, as it can bee seen Figure F.6. The mean difference between Nissen's and HARPS's data, for the different abundances, can be see on Table F.1.

We note that Nissen (2015)'s sample has a higher metallicity than HARPS, for the same stars. Yet, when looking at the abundance values, it is possible to understand that the difference in values might be because of a different calibration on Nissen (2015)'s work, leading to determining a different value for the same chemical abundances of the same stars.

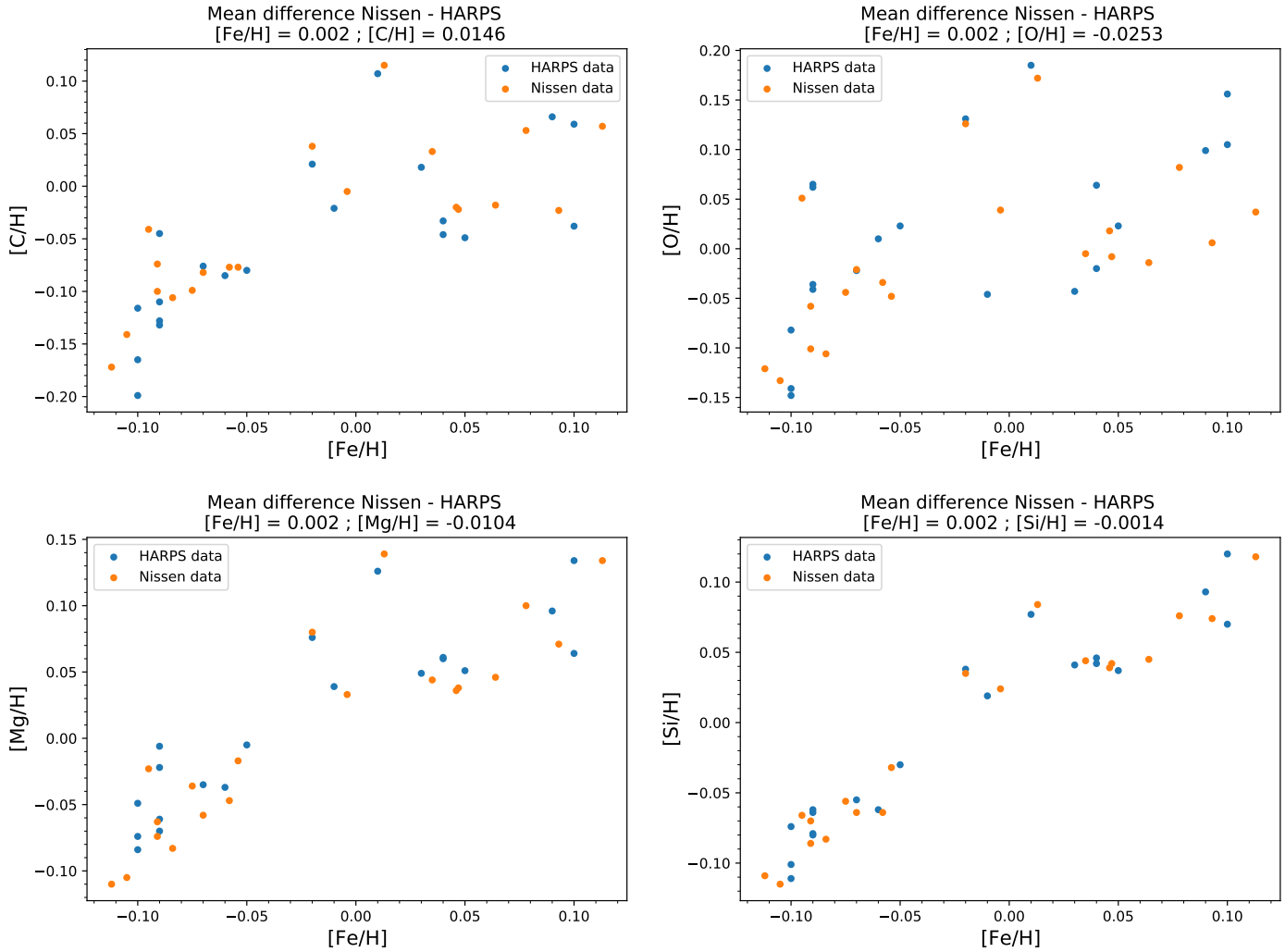


FIGURE F.6: Comparing metallicity and individual abundances of common stars between Nissen (2015)'s data and the HARPS sample.

TABLE F.1: Mean values for the difference between Nissen (2015) and HARPS's abundances.

	[Fe/H]	[C/H]	[O/H]	[Mg/H]	[Si/H]
Nissen - HARPS	0.002	0.01455	-0.0253	-0.0104	-0.00145

F.5 Predicting Mg and Si abundances with HARPS 1111 stars

When selecting the best sample from the HARPS data set in Section 2.2.3, we started with 1111 stars but considering only starts with Fe, Mg, Si, C and O led us to a sample of 535 stars (of which 8 did not have Gaia information). We have seen from Sections 2.4 and 2.5 that the models usually give better predictions to [Mg/H] and [Si/H]. We want to test how would the predictions be for these two abundances if we considered the whole HARPS sample. Moreover, the constrains on the smaller sample excluded the cooler stars. To make sure a smaller range in temperatures does not affect the results, we built a new RF model and train it in the original 1111 stars from HARPS which has roughly twice the size. The inputs remain the same (metallicity and total galactic velocity) but now we aim to predict only the [Mg/H] and [Si/H] abundances. We apply this model in the HARPS sample (Figure F.7) and with Adibekyan et al. (2021)'s data (Figure F.8).

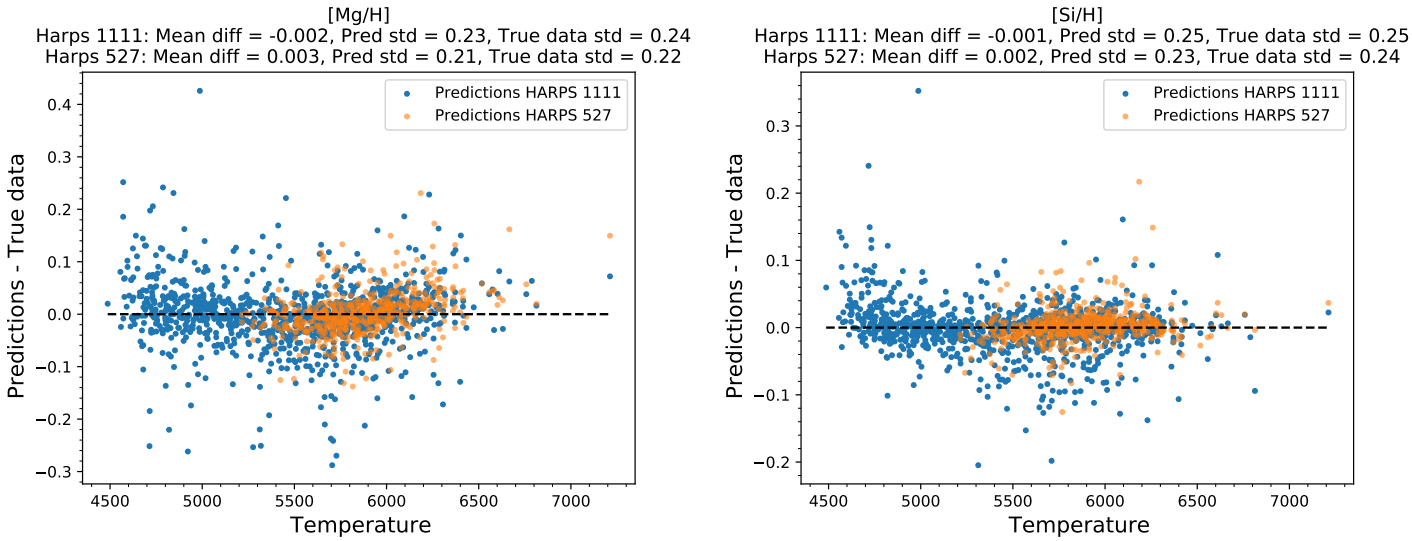


FIGURE F.7: Comparison of difference between predictions and true data, depending on effective temperature, for both HARPS RF models. Both models were applied to the total HARPS sample to predict [Mg/H] abundances (left) and [Si/H] abundances (right). Title includes information on the mean difference between predicted and true values and standard deviation of predictions and true data.

Although the results from applying the HARPS 1111 model have a smaller mean difference value than the results from the RFH model (HARPS with 527 stars), they have a higher standard deviation value, meaning that the predictions have a higher dispersion. Yet both values of mean difference and standard deviation are similar enough that we can conclude a higher range of temperatures does not affect the predictions.

When looking at the results from Adibekyan et al. (2021)'s data in Figure F.8, we see this is equally true. The mean difference for RFH results is higher but its standard deviation is smaller, meaning less dispersion, for both abundance predictions. However, the values for both models were close enough that we can conclude the effective temperature does not affect the predictions.

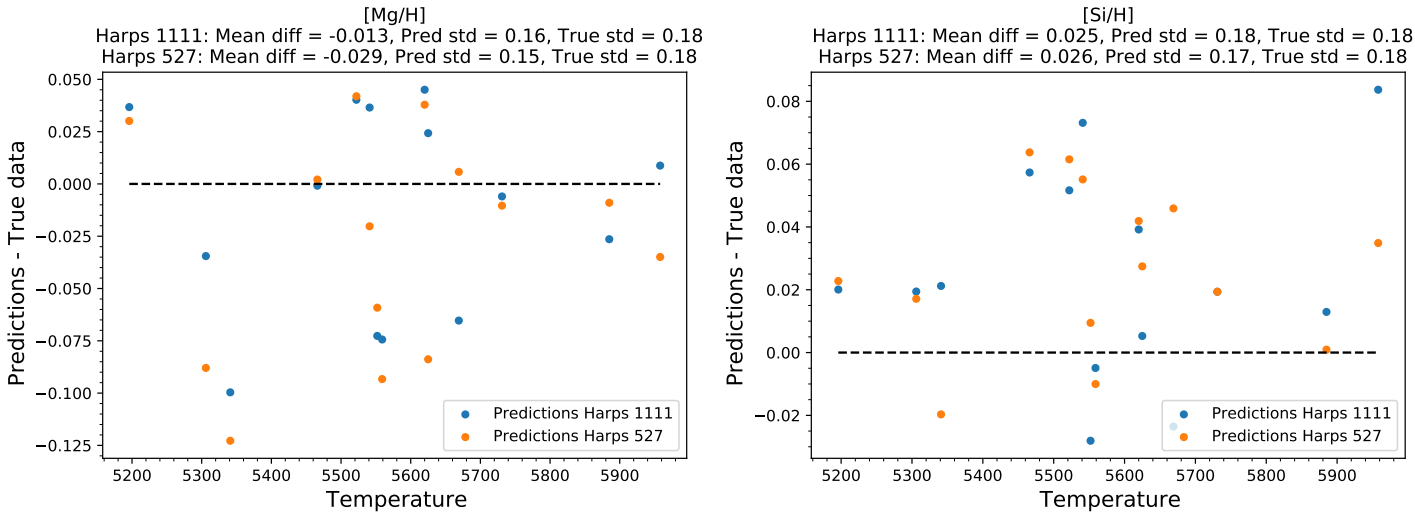


FIGURE F.8: Comparison of difference between predictions and true data, depending on effective temperature, for both HARPS RF models. Both models were applied to Adibekyan et al. (2021)'s data to predict [Mg/H] abundances (left) and [Si/H] abundances (right). Title includes information on the mean difference between predicted and true values and standard deviation of predictions and true data.

Additionally, to make sure halving the sample does not affect the range of values for the [Mg/H] and [Si/H], we plot the distributions of both individual abundances of the two HARPS sample. By observing Figure F.9, it is possible to verify that, although the counts are unsurprisingly smaller, the range of values remains about the same.

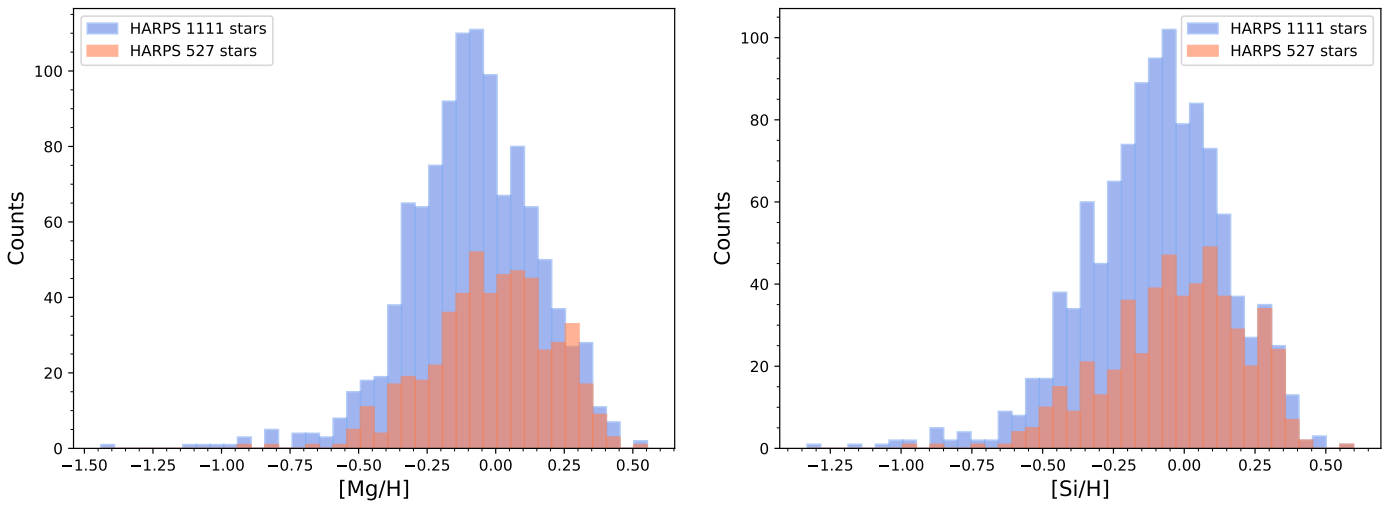


FIGURE F.9: Distributions of [Mg/H] and [Si/H] abundances, for the HARPS samples with 527 stars (red) and 1111 stars (blue).

F.6 Predicting only [Mg/H] and [Si/H]

As mentioned in the previous section, the Mg and Si are the best predicted abundances. In order to see if the presence of other outputs in the models affects the results or if each output is independent of the others, we create new models to test this hypothesis.

For the three samples of APOGEE, GALAH and HARPS, we build a new model for each. Again, we have as inputs the metallicity and total galactic velocity of the stars but just [Mg/H] and [Si/H] as outputs, as that is what we want to predict. After training the models, we apply them to their corresponding sample (APOGEE, GALAH or HARPS) to predict the values of Mg and Si abundances and compare the results with those for the original models. Figure F.10 shows results for the Mg and Si abundances, obtained with the old and new APOGEE, GALAH and HARPS models.

For both abundances and for all models, the standard deviation value for the predictions is always smaller than the one for the original data, meaning that the predictions have less dispersion. The mean value of Predictions – True data is also very close to 0, pointing to reasonable predictions. Furthermore, the standard deviation and mean difference values for the old random forest and the new random forest models are very similar to each other, regardless of which abundance are we predicting, with a maximum difference of 0.0019 for the mean difference and 0.0032 for the standard deviation. This implies that the difference between models is minimum and the results are roughly the same.

Therefore, we conclude that the predictions of [Mg/H] and [Si/H] can be assumed as independent of the other outputs.

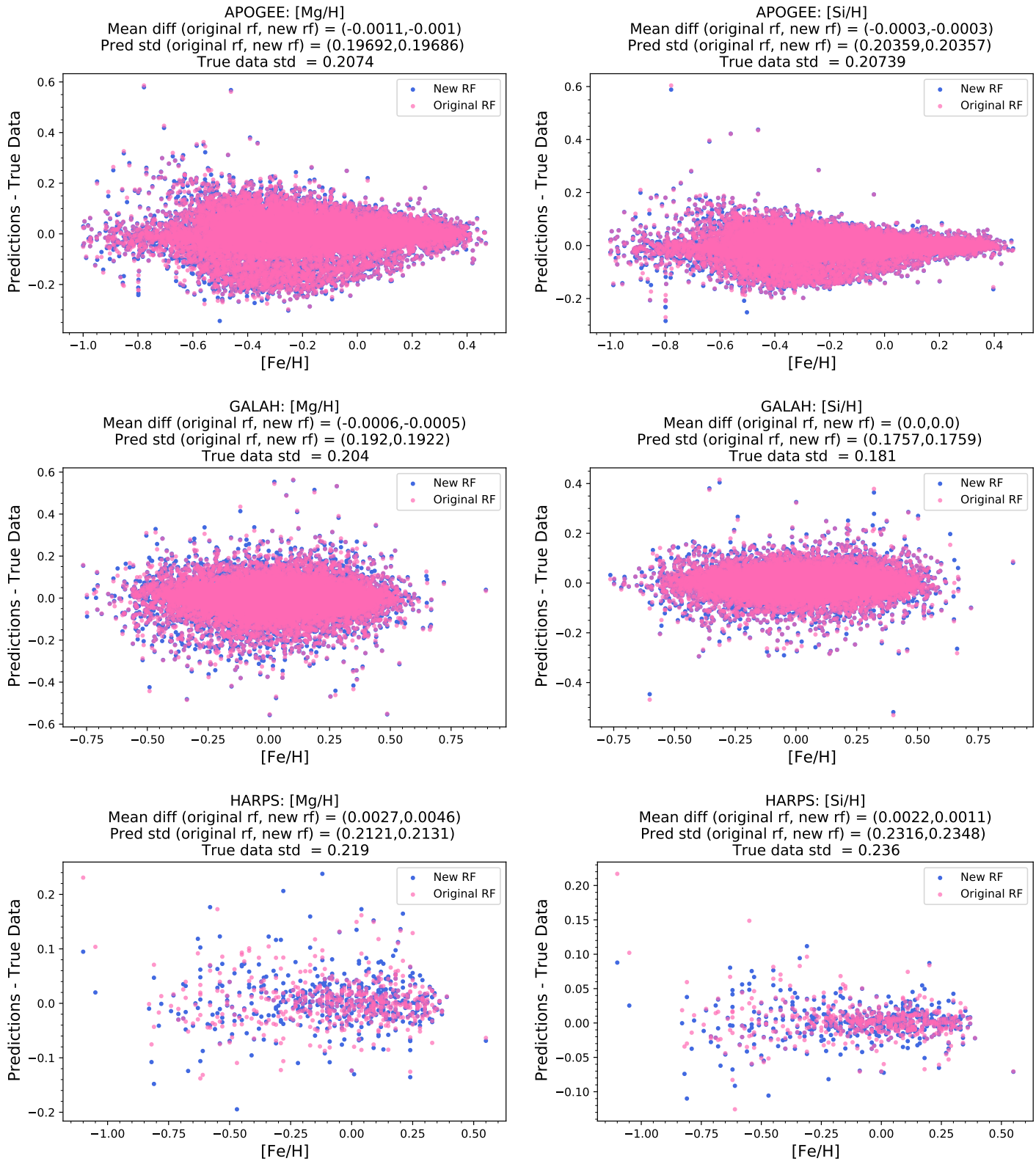


FIGURE F.10: Comparing predictions from RF with individual abundances and abundance ratios as outputs and from RF with only [Mg/H] and [Si/H] as outputs, for the APOGEE (top), GALAH (middle) and HARPS (bottom) samples. Left column are the results for the Mg abundances and right column are for the Si abundances. Title includes information on the mean difference and standard deviation of both predictions and true data.

Bibliography

A. Wolszczan and D. A. Frail, "A planetary system around the millisecond pulsar PSR1257 + 12," , vol. 355, no. 6356, pp. 145–147, Jan. 1992. [Cited on page 1.]

M. Mayor and D. Queloz, "A Jupiter-mass companion to a solar-type star," , vol. 378, no. 6555, pp. 355–359, Nov. 1995. [Cited on page 1.]

N. C. Santos, S. C. C. Barros, O. D. S. Demangeon, and J. P. Faria, *Detection and Characterization Methods of Exoplanets*, 2020, p. 189. [Cited on page 1.]

J. Sahlmann, "Astrometric exoplanet surveys in practice," *IAU Focus Meeting*, vol. 29A, pp. 217–218, Oct. 2016. [Cited on page 2.]

O. Struve, "Proposal for a project of high-precision stellar radial velocity work," *The Observatory*, vol. 72, pp. 199–200, Oct. 1952. [Cited on page 3.]

J. Rosen and L. Gothard, *Encyclopedia of Physical Science*, ser. Facts on File Science Library. Facts On File, 2010. [Online]. Available: https://books.google.pt/books?id=avyQ64LIJa0C [Cited on page 3.]

A. Cumming, G. W. Marcy, and R. P. Butler, "The Lick Planet Search: Detectability and Mass Thresholds," , vol. 526, no. 2, pp. 890–915, Dec. 1999. [Cited on page 3.]

R. W. Hilditch, *An Introduction to Close Binary Stars*, 2001. [Cited on page 3.]

J. T. Wright and A. W. Howard, "Efficient Fitting of Multiplanet Keplerian Models to Radial Velocity and Astrometry Data," , vol. 182, no. 1, pp. 205–215, May 2009. [Cited on page 3.]

J. J. Lissauer and I. de Pater, *Fundamental Planetary Science*, 2013. [Cited on pages 3 and 5.]

D. Queloz, G. W. Henry, J. P. Sivan, S. L. Baliunas, J. L. Beuzit, R. A. Donahue, M. Mayor, D. Naef, C. Perrier, and S. Udry, "No planet for HD 166435," , vol. 379, pp. 279–287, Nov. 2001. [Cited on page 4.]

M. Mayor, F. Pepe, D. Queloz, F. Bouchy, G. Rupprecht, G. Lo Curto, G. Avila, W. Benz, J. L. Bertaux, X. Bonfils, T. Dall, H. Dekker, B. Delabre, W. Eckert, M. Fleury, A. Gilliotte, D. Gojak, J. C. Guzman, D. Kohler, J. L. Lizon, A. Longinotti, C. Lovis, D. Megevand, L. Pasquini, J. Reyes, J. P. Sivan, D. Sosnowska, R. Soto, S. Udry, A. van Kesteren, L. Weber, and U. Weilenmann, "Setting New Standards with HARPS," *The Messenger*, vol. 114, pp. 20–24, Dec. 2003. [Cited on pages 4, 18, 19, and 77.]

R. Cosentino, C. Lovis, F. Pepe, A. Collier Cameron, D. W. Latham, E. Molinari, S. Udry, N. Bezawada, M. Black, A. Born, N. Buchschacher, D. Charbonneau, P. Figueira, M. Fleury, A. Galli, A. Gallie, X. Gao, A. Ghedina, C. Gonzalez, M. Gonzalez, J. Guerra, D. Henry, K. Horne, I. Hughes, D. Kelly, M. Lodi, D. Lunney, C. Maire, M. Mayor, G. Micela, M. P. Ordway, J. Peacock, D. Phillips, G. Piotto, D. Pollacco, D. Queloz, K. Rice, C. Riverol, L. Riverol, J. San Juan, D. Sasselov, D. Segransan, A. Sozzetti, D. Sosnowska, B. Stobie, A. Szentgyorgyi, A. Vick, and L. Weber, "Harps-N: the new planet hunter at TNG," in *Ground-based and Airborne Instrumentation for Astronomy IV*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, I. S. McLean, S. K. Ramsay, and H. Takami, Eds., vol. 8446, Sep. 2012, p. 84461V. [Cited on pages 4 and 16.]

F. Pepe, A. C. Cameron, D. W. Latham, E. Molinari, S. Udry, A. S. Bonomo, L. A. Buchhave, D. Charbonneau, R. Cosentino, C. D. Dressing, X. Dumusque, P. Figueira, A. F. M. Fiorenzano, S. Gettel, A. Harutyunyan, R. D. Haywood, K. Horne, M. Lopez-Morales, C. Lovis, L. Malavolta, M. Mayor, G. Micela, F. Motalebi, V. Nascimbeni, D. Phillips, G. Piotto, D. Pollacco, D. Queloz, K. Rice, D. Sasselov, D. Ségransan, A. Sozzetti, A. Szentgyorgyi, and C. A. Watson, "An Earth-sized planet with an Earth-like density," , vol. 503, no. 7476, pp. 377–380, Nov. 2013. [Cited on page 4.]

F. Pepe, S. Cristiani, R. Rebolo, N. C. Santos, H. Dekker, D. Mégevand, F. M. Zerbi, A. Cabral, P. Molaro, P. Di Marcantonio, M. Abreu, M. Affolter, M. Aliverti, C. Allende Prieto, M. Amate, G. Avila, V. Baldini, P. Bristow, C. Broeg, R. Cirami, J. Coelho, P. Conconi, I. Coretti, G. Cupani, V. D'Odorico, V. De Caprio, B. Delabre, R. Dorn, P. Figueira, A. Fragoso, S. Galeotta, L. Genolet, R. Gomes, J. I. González Hernández, I. Hughes,

O. Iwert, F. Kerber, M. Landoni, J. L. Lizon, C. Lovis, C. Maire, M. Mannetta, C. Martins, M. A. Monteiro, A. Oliveira, E. Poretti, J. L. Rasilla, M. Riva, S. Santana Tschudi, P. Santos, D. Sosnowska, S. Sousa, P. Spanò, F. Tenegi, G. Toso, E. Vanzella, M. Viel, and M. R. Zapatero Osorio, "ESPRESSO — An Echelle SPectrograph for Rocky Exoplanets Search and Stable Spectroscopic Observations," *The Messenger*, vol. 153, pp. 6–16, Sep. 2013. [Cited on pages 4 and 16.]

F. Pepe, S. Cristiani, R. Rebolo, N. C. Santos, H. Dekker, A. Cabral, P. Di Marcantonio, P. Figueira, G. Lo Curto, C. Lovis, M. Mayor, D. Mégevand, P. Molaro, M. Riva, M. R. Zapatero Osorio, M. Amate, A. Manescau, L. Pasquini, F. M. Zerbi, V. Adibekyan, M. Abreu, M. Affolter, Y. Alibert, M. Aliverti, R. Allart, C. Allende Prieto, D. Álvarez, D. Alves, G. Avila, V. Baldini, T. Bandy, S. C. C. Barros, W. Benz, A. Bianco, F. Borsa, V. Bourrier, F. Bouchy, C. Broeg, G. Calderone, R. Cirami, J. Coelho, P. Conconi, I. Coretti, C. Cumani, G. Cupani, V. D'Odorico, M. Damasso, S. Deiries, B. Delabre, O. D. S. Demangeon, X. Dumusque, D. Ehrenreich, J. P. Faria, A. Fragoso, L. Genolet, M. Genoni, R. Génova Santos, J. I. González Hernández, I. Hughes, O. Iwert, F. Kerber, J. Knudstrup, M. Landoni, B. Lavie, J. Lillo-Box, J. L. Lizon, C. Maire, C. J. A. P. Martins, A. Mehner, G. Micela, A. Modigliani, M. A. Monteiro, M. J. P. F. G. Monteiro, M. Moschetti, M. T. Murphy, N. Nunes, L. Oggioni, A. Oliveira, M. Oshagh, E. Pallé, G. Pariani, E. Poretti, J. L. Rasilla, J. Rebordão, E. M. Redaelli, S. Santana Tschudi, P. Santin, P. Santos, D. Ségransan, T. M. Schmidt, A. Segovia, D. Sosnowska, A. Sozzetti, S. G. Sousa, P. Spanò, A. Suárez Mascareño, H. Tabernero, F. Tenegi, S. Udry, and A. Zanutta, "ESPRESSO at VLT. On-sky performance and first results," , vol. 645, p. A96, Jan. 2021. [Cited on pages 4 and 16.]

G. W. Henry, G. Marcy, R. P. Butler, and S. S. Vogt, "HD 209458," , vol. 7307, p. 1, Nov. 1999. [Cited on page 4.]

D. Charbonneau, T. M. Brown, D. W. Latham, and M. Mayor, "Detection of Planetary Transits Across a Sun-like Star," , vol. 529, no. 1, pp. L45–L48, Jan. 2000. [Cited on page 4.]

A. Baglin, M. Auvergne, L. Boisnard, T. Lam-Trong, P. Barge, C. Catala, M. Deleuil, E. Michel, and W. Weiss, "CoRoT: a high precision photometer for stellar ecolution and exoplanet finding," in *36th COSPAR Scientific Assembly*, vol. 36, Jan. 2006, p. 3749. [Cited on page 5.]

A. Baglin, M. Auvergne, P. Barge, E. Michel, C. Catala, M. Deleuil, and W. Weiss, "The CoRoT mission and its scientific objectives," in *Fifty Years of Romanian Astrophysics*, ser. American Institute of Physics Conference Series, C. Dumitrache, N. A. Popescu, M. D. Suran, and V. Mioc, Eds., vol. 895, Mar. 2007, pp. 201–209. [Cited on page 5.]

A. Léger, O. Grasset, B. Fegley, F. Codron, A. Albarede, P. Barge, R. Barnes, P. Cance, S. Carpy, F. Catalano, C. Cavarroc, O. Demangeon, S. Ferraz-Mello, P. Gabor, J.-M. Grießmeier, J. Leibacher, G. Libourel, A.-S. Maurin, S. Raymond, D. Rouan, B. Samuel, L. Schaefer, J. Schneider, P. Schuller, F. Selsis, and C. Sotin, "The extreme physical properties of the corot-7b super-earth," *Icarus*, vol. 213, no. 1, pp. 1–11, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019103511000534 [Cited on page 5.]

S. C. C. Barros, J. M. Almenara, M. Deleuil, R. F. Diaz, S. Csizmadia, J. Cabrera, S. Chaintreuil, A. Collier Cameron, A. Hatzes, R. Haywood, A. F. Lanza, S. Aigrain, R. Alonso, P. Bordé, F. Bouchy, H. J. Deeg, A. Erikson, M. Fridlund, S. Grziwa, D. Gandolfi, T. Guillot, E. Guenther, A. Leger, C. Moutou, M. Ollivier, T. Pasternacki, M. Pätzold, H. Rauer, D. Rouan, A. Santerne, J. Schneider, and G. Wuchterl, "Revisiting the transits of CoRoT-7b at a lower activity level," , vol. 569, p. A74, Sep. 2014. [Cited on page 5.]

W. J. Borucki, D. Koch, G. Basri, N. Batalha, T. Brown, D. Caldwell, J. Caldwell, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore, E. W. Dunham, A. K. Dupree, T. N. Gautier, J. C. Geary, R. Gilliland, A. Gould, S. B. Howell, J. M. Jenkins, Y. Kondo, D. W. Latham, G. W. Marcy, S. Meibom, H. Kjeldsen, J. J. Lissauer, D. G. Monet, D. Morrison, D. Sasselov, J. Tarter, A. Boss, D. Brownlee, T. Owen, D. Buzasi, D. Charbonneau, L. Doyle, J. Fortney, E. B. Ford, M. J. Holman, S. Seager, J. H. Steffen, W. F. Welsh, J. Rowe, H. Anderson, L. Buchhave, D. Ciardi, L. Walkowicz, W. Sherry, E. Horch, H. Isaacson, M. E. Everett, D. Fischer, G. Torres, J. A. Johnson, M. Endl, P. MacQueen, S. T. Bryson, J. Dotson, M. Haas, J. Kolodziejczak, J. Van Cleve, H. Chandrasekaran, J. D. Twicken, E. V. Quintana, B. D. Clarke, C. Allen, J. Li, H. Wu, P. Tenenbaum, E. Verner, F. Bruhweiler, J. Barnes, and A. Prsa, "Kepler Planet-Detection Mission: Introduction and First Results," *Science*, vol. 327, no. 5968, p. 977, Feb. 2010. [Cited on pages 5 and 16.]

A. Vanderburg, P. Rowden, S. Bryson, J. Coughlin, N. Batalha, K. A. Collins, D. W. Latham, S. E. Mullally, K. D. Colón, C. Henze, C. X. Huang, and S. N. Quinn, "A habitable-zone earth-sized planet rescued from false positive status,"

The Astrophysical Journal, vol. 893, no. 1, p. L27, apr 2020. [Online]. Available: https://doi.org/10.3847/2041-8213/ab84e5 [Cited on pages 5 and 16.]

G. R. Ricker, D. W. Latham, R. K. Vanderspek, K. A. Ennico, G. Bakos, T. M. Brown, A. J. Burgasser, D. Charbonneau, L. D. Deming, J. P. Doty, E. W. Dunham, J. L. Elliot, M. J. Holman, S. Ida, J. M. Jenkins, J. G. Jernigan, N. Kawai, G. P. Laughlin, J. J. Lissauer, F. Martel, D. D. Sasselov, R. H. Schingler, S. Seager, G. Torres, S. Udry, J. S. Villasenor, J. N. Winn, and S. P. Worden, "The Transiting Exoplanet Survey Satellite (TESS)," in *American Astronomical Society Meeting Abstracts #213*, ser. American Astronomical Society Meeting Abstracts, vol. 213, Jan. 2009, p. 403.01. [Cited on pages 5 and 16.]

H. Rauer, C. Catala, C. Aerts, T. Appourchaux, W. Benz, A. Brandeker, J. Christensen-Dalsgaard, M. Deleuil, L. Gizon, M. J. Goupil, M. Güdel, E. Janot-Pacheco, M. Mas-Hesse, I. Pagano, G. Piotto, D. Pollacco, Ċ. Santos, A. Smith, J. C. Suárez, R. Szabó, S. Udry, V. Adibekyan, Y. Alibert, J. M. Almenara, P. Amaro-Seoane, M. A.-v. Eiff, M. Asplund, E. Antonello, S. Barnes, F. Baudin, K. Belkacem, M. Bergemann, G. Bihain, A. C. Birch, X. Bonfils, I. Boisse, A. S. Bonomo, F. Borsa, I. M. Brandão, E. Brocato, S. Brun, M. Burleigh, R. Burston, J. Cabrera, S. Cassisi, W. Chaplin, S. Charpinet, C. Chiappini, R. P. Church, S. Csizmadia, M. Cunha, M. Damasso, M. B. Davies, H. J. Deeg, R. F. Díaz, S. Dreizler, C. Dreyer, P. Eggenberger, D. Ehrenreich, P. Eigmüller, A. Erikson, R. Farmer, S. Feltzing, F. de Oliveira Fialho, P. Figueira, T. Forveille, M. Fridlund, R. A. García, P. Giommi, G. Giuffrida, M. Godolt, J. Gomes da Silva, T. Granzer, J. L. Grenfell, A. Grotsch-Noels, E. Günther, C. A. Haswell, A. P. Hatzes, G. Hébrard, S. Hekker, R. Helled, K. Heng, J. M. Jenkins, A. Johansen, M. L. Khodachenko, K. G. Kislyakova, W. Kley, U. Kolb, N. Krivova, F. Kupka, H. Lammer, A. F. Lanza, Y. Lebreton, D. Magrin, P. Marcos-Arenal, P. M. Marrese, J. P. Marques, J. Martins, S. Mathis, S. Mathur, S. Messina, A. Miglio, J. Montalban, M. Montalto, M. J. P. F. G. Monteiro, H. Moradi, E. Moravveji, C. Mordasini, T. Morel, A. Mortier, V. Nascimbeni, R. P. Nelson, M. B. Nielsen, L. Noack, A. J. Norton, A. Ofir, M. Oshagh, R. M. Ouazzani, P. Pápics, V. C. Parro, P. Petit, B. Plez, E. Poretti, A. Quirrenbach, R. Ragazzoni, G. Raimondo, M. Rainer, D. R. Reese, R. Redmer, S. Reffert, B. Rojas-Ayala, I. W. Roxburgh, S. Salmon, A. Santerne, J. Schneider, J. Schou, S. Schuh, H. Schunker, A. Silva-Valio, R. Silvotti, I. Skillen, I. Snellen, F. Sohl, S. G. Sousa, A. Sozzetti, D. Stello, K. G. Strassmeier, M. Švanda, G. M. Szabó, A. Tkachenko, D. Valencia, V. Van Grootel, S. D. Vauclair, P. Ventura, F. W. Wagner, N. A. Walton, J. Weingrill, S. C. Werner, P. J. Wheatley,

and K. Zwintz, "The PLATO 2.0 mission," *Experimental Astronomy*, vol. 38, no. 1-2, pp. 249–330, Nov. 2014. [Cited on pages 5 and 79.]

H. Rauer, C. Aerts, J. Cabrera, and PLATO Team, "The PLATO Mission," *Astronomische Nachrichten*, vol. 337, no. 8-9, p. 961, Sep. 2016. [Cited on pages 5 and 79.]

C. Catala and Plato Team, "PLATO: Planetary Transits and Oscillations of Stars," in *The CoRoT Mission Pre-Launch Status - Stellar Seismology and Planet Finding*, ser. ESA Special Publication, M. Fridlund, A. Baglin, J. Lochard, and L. Conroy, Eds., vol. 1306, Nov. 2006, p. 497. [Cited on page 5.]

I. Roxburgh, C. Catala, and PLATO Consortium, "The PLATO mission concept," *Communications in Asteroseismology*, vol. 150, p. 357, Jun. 2007. [Cited on page 5.]

G. Christie, "Detecting Exoplanets by Gravitational Microlensing using a Small Telescope," *Society for Astronomical Sciences Annual Symposium*, vol. 25, p. 97, May 2006. [Cited on page 6.]

P. Kalas, "Direct imaging of massive extrasolar planets," in *The Astrophysics of Planetary Systems: Formation, Structure, and Dynamical Evolution*, A. Sozzetti, M. G. Lattanzi, and A. P. Boss, Eds., vol. 276, Nov. 2011, pp. 279–286. [Cited on page 6.]

L. Pueyo, *Direct Imaging as a Detection Technique for Exoplanets*, 2018, p. 10. [Cited on page 6.]

W. A. Traub and B. R. Oppenheimer, *Direct Imaging of Exoplanets*, 2010, pp. 111–156. [Cited on page 6.]

B. Gauza, V. J. S. Béjar, A. Pérez-Garrido, M. R. Zapatero Osorio, N. Lodieu, R. Rebolo, E. Pallé, and G. Nowak, "Discovery of a Young Planetary Mass Companion to the Nearby M Dwarf VHS J125601.92-125723.9," , vol. 804, no. 2, p. 96, May 2015. [Cited on page 6.]

A. Vigan, M. Bonnefoy, C. Ginski, H. Beust, R. Galicher, M. Janson, J. Baudino, E. Buenzli, J. Hagelberg, V. D'Orazi, S. Desidera, A. Maire, R. Gratton, J. Sauvage, G. Chauvin, C. Thalmann, L. Malo, G. Salter, A. Zurlo, J. Antichi, A. Baruffolo, P. Baudoz, P. Blanchard, A. Boccaletti, J. Beuzit, M. Carle, R. Claudi, A. Costille, A. Delboulbe, K. Dohlen, C. Dominik, M. Feldt, T. Fusco, L. Gluck, J. Girard, E. Giro, C. Gry, T. Henning, N. Hubin, E. Hugot, M. Jaquet, M. Kasper, A. Lagrange, M. Langlois, D. Le

Mignant, M. Llored, F. Madec, P. Martinez, D. Mawet, D. Mesa, J. Milli, D. Mouillet, T. Moulin, C. Moutou, A. Origne, A. Pavlov, D. Perret, C. Petit, J. Pragt, P. Puget, P. Rabou, S. Rochat, R. Roelfsema, B. Salasnich, H. Schmid, A. Sevin, R. Siebenmorgen, A. Smette, E. Stadler, M. Suarez, M. Turatto, S. Udry, F. Vakili, Z. Wahhaj, L. Weber, and F. Wildi, "First light of the vlt planet finder sphere i. detection and characterization of the substellar companion gj 758 b," *Astronomy and Astrophysics*, vol. 587, pp. 1–20, Mar. 2016, copyright 2016 ESO. First published in Astronomy and Astrophysics, 587, A55, 2016, published by EDP Sciences. The original publication is available at https://doi.org/10.1051/0004-6361/201526465. [Cited on page 6.]

M. Bonnefoy, A. Zurlo, J. L. Baudino, P. Lucas, D. Mesa, A. L. Maire, A. Vigan, R. Galicher, D. Homeier, F. Marocco, R. Gratton, G. Chauvin, F. Allard, S. Desidera, M. Kasper, C. Moutou, A. M. Lagrange, J. Antichi, A. Baruffolo, J. Baudrand, J. L. Beuzit, A. Boccaletti, F. Cantalloube, M. Carbillet, J. Charton, R. U. Claudi, A. Costille, K. Dohlen, C. Dominik, D. Fantinel, P. Feautrier, M. Feldt, T. Fusco, P. Gigan, J. H. Girard, L. Gluck, C. Gry, T. Henning, M. Janson, M. Langlois, F. Madec, Y. Magnard, D. Maurel, D. Mawet, M. R. Meyer, J. Milli, O. Moeller-Nilsson, D. Mouillet, A. Pavlov, D. Perret, P. Pujet, S. P. Quanz, S. Rochat, G. Rousset, A. Roux, B. Salasnich, G. Salter, J. F. Sauvage, H. M. Schmid, A. Sevin, C. Soenke, E. Stadler, M. Turatto, S. Udry, F. Vakili, Z. Wahhaj, and F. Wildi, "First light of the VLT planet finder SPHERE. IV. Physical and chemical properties of the planets around HR8799," , vol. 587, p. A58, Mar. 2016. [Cited on page 6.]

A. Zurlo, A. Vigan, R. Galicher, A. L. Maire, D. Mesa, R. Gratton, G. Chauvin, M. Kasper, C. Moutou, M. Bonnefoy, S. Desidera, L. Abe, D. Apai, A. Baruffolo, P. Baudoz, J. Baudrand, J. L. Beuzit, P. Blancard, A. Boccaletti, F. Cantalloube, M. Carle, E. Cascone, J. Charton, R. U. Claudi, A. Costille, V. de Caprio, K. Dohlen, C. Dominik, D. Fantinel, P. Feautrier, M. Feldt, T. Fusco, P. Gigan, J. H. Girard, D. Gisler, L. Gluck, C. Gry, T. Henning, E. Hugot, M. Janson, M. Jaquet, A. M. Lagrange, M. Langlois, M. Llored, F. Madec, Y. Magnard, P. Martinez, D. Maurel, D. Mawet, M. R. Meyer, J. Milli, O. Moeller-Nilsson, D. Mouillet, A. Origné, A. Pavlov, C. Petit, P. Puget, S. P. Quanz, P. Rabou, J. Ramos, G. Rousset, A. Roux, B. Salasnich, G. Salter, J. F. Sauvage, H. M. Schmid, C. Soenke, E. Stadler, M. Suarez, M. Turatto, S. Udry, F. Vakili, Z. Wahhaj, F. Wildi,

and J. Antichi, "First light of the VLT planet finder SPHERE. III. New spectrophotometry and astrometry of the HR 8799 exoplanetary system," , vol. 587, p. A57, Mar. 2016. [Cited on page 6.]

Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, A. G. A. Brown, A. Vallenari, C. Babusiaux, C. A. L. Bailer-Jones, U. Bastian, M. Biermann, D. W. Evans, L. Eyer, F. Jansen, C. Jordi, S. A. Klioner, U. Lammers, L. Lindegren, X. Luri, F. Mignard, D. J. Milligan, C. Panem, V. Poinsignon, D. Pourbaix, S. Randich, G. Sarri, P. Sartoretti, H. I. Siddiqui, C. Soubiran, V. Valette, F. van Leeuwen, N. A. Walton, C. Aerts, F. Arenou, M. Cropper, R. Drimmel, E. Høg, D. Katz, M. G. Lattanzi, W. O'Mullane, E. K. Grebel, A. D. Holland, C. Huc, X. Passot, L. Bramante, C. Cacciari, J. Castañeda, L. Chaoul, N. Cheek, F. De Angeli, C. Fabricius, R. Guerra, J. Hernández, A. Jean-Antoine-Piccolo, E. Masana, R. Messineo, N. Mowlavi, K. Nienartowicz, D. Ordóñez-Blanco, P. Panuzzo, J. Portell, P. J. Richards, M. Riello, G. M. Seabroke, P. Tanga, F. Thévenin, J. Torra, S. G. Els, G. Gracia-Abril, G. Comoretto, M. Garcia-Reinaldos, T. Lock, E. Mercier, M. Altmann, R. Andrae, T. L. Astraatmadja, I. Bellas-Velidis, K. Benson, J. Berthier, R. Blomme, G. Busso, B. Carry, A. Cellino, G. Clementini, S. Cowell, O. Creevey, J. Cuypers, M. Davidson, J. De Ridder, A. de Torres, L. Delchambre, A. Dell'Oro, C. Ducourant, Y. Frémat, M. García-Torres, E. Gosset, J. L. Halbwachs, N. C. Hambly, D. L. Harrison, M. Hauser, D. Hestroffer, S. T. Hodgkin, H. E. Huckle, A. Hutton, G. Jasniewicz, S. Jordan, M. Kontizas, A. J. Korn, A. C. Lanzafame, M. Manteiga, A. Moitinho, K. Muinonen, J. Osinde, E. Pancino, T. Pauwels, J. M. Petit, A. Recio-Blanco, A. C. Robin, L. M. Sarro, C. Siopis, M. Smith, K. W. Smith, A. Sozzetti, W. Thuillot, W. van Reeven, Y. Viala, U. Abbas, A. Abreu Aramburu, S. Accart, J. J. Aguado, P. M. Allan, W. Allasia, G. Altavilla, M. A. Álvarez, J. Alves, R. I. Anderson, A. H. Andrei, E. Anglada Varela, E. Antiche, T. Antoja, S. Antón, B. Arcay, A. Atzei, L. Ayache, N. Bach, S. G. Baker, L. Balaguer-Núñez, C. Barache, C. Barata, A. Barbier, F. Barblan, M. Baroni, D. Barrado y Navascués, M. Barros, M. A. Barstow, U. Becciani, M. Bellazzini, G. Bellei, A. Bello García, V. Belokurov, P. Bendjoya, A. Berihuete, L. Bianchi, O. Bienaymé, F. Billebaud, N. Blagorodnova, S. Blanco-Cuaresma, T. Boch, A. Bombrun, R. Borrachero, S. Bouquillon, G. Bourda, H. Bouy, A. Bragaglia, M. A. Breddels, N. Brouillet, T. Brüsemeister, B. Bucciarelli, F. Budnik, P. Burgess, R. Burgon, A. Burlacu, D. Busonero, R. Buzzi, E. Caffau, J. Cambras, H. Campbell, R. Cancelliere, T. Cantat-Gaudin, T. Carlucci, J. M. Carrasco, M. Castellani, P. Charlot, J. Charnas, P. Charvet, F. Chassat, A. Chiavassa,

M. Clotet, G. Cocozza, R. S. Collins, P. Collins, G. Costigan, F. Crifo, N. J. G. Cross, M. Crosta, C. Crowley, C. Dafonte, Y. Damerdji, A. Dapergolas, P. David, M. David, P. De Cat, F. de Felice, P. de Laverny, F. De Luise, R. De March, D. de Martino, R. de Souza, J. Debosscher, E. del Pozo, M. Delbo, A. Delgado, H. E. Delgado, F. di Marco, P. Di Matteo, S. Diakite, E. Distefano, C. Dolding, S. Dos Anjos, P. Drazinos, J. Durán, Y. Dzigan, E. Ecale, B. Edvardsson, H. Enke, M. Erdmann, D. Escolar, M. Espina, N. W. Evans, G. Eynard Bontemps, C. Fabre, M. Fabrizio, S. Faigler, A. J. Falcão, M. Farràs Casas, F. Faye, L. Federici, G. Fedorets, J. Fernández-Hernández, P. Fernique, A. Fienga, F. Figueras, F. Filippi, K. Findeisen, A. Fonti, M. Fouesneau, E. Fraile, M. Fraser, J. Fuchs, R. Furnell, M. Gai, S. Galleti, L. Galluccio, D. Garabato, F. García-Sedano, P. Garé, A. Garofalo, N. Garralda, P. Gavras, J. Gerssen, R. Geyer, G. Gilmore, S. Girona, G. Giuffrida, M. Gomes, A. González-Marcos, J. González-Núñez, J. J. González-Vidal, M. Granvik, A. Guerrier, P. Guillout, J. Guiraud, A. Gúrpide, R. Gutiérrez-Sánchez, L. P. Guy, R. Haigron, D. Hatzidimitriou, M. Haywood, U. Heiter, A. Helmi, D. Hobbs, W. Hofmann, B. Holl, G. Holland, J. A. S. Hunt, A. Hypki, V. Icardi, M. Irwin, G. Jevardat de Fombelle, P. Jofré, P. G. Jonker, A. Jorissen, F. Julbe, A. Karampelas, A. Kochoska, R. Kohley, K. Kolenberg, E. Kontizas, S. E. Koposov, G. Kordopatis, P. Koubsky, A. Kowalczyk, A. Krone-Martins, M. Kudryashova, I. Kull, R. K. Bachchan, F. Lacoste-Seris, A. F. Lanza, J. B. Lavigne, C. Le Poncin-Lafitte, Y. Lebreton, T. Lebzelter, S. Leccia, N. Leclerc, I. Lecoeur-Taibi, V. Lemaitre, H. Lenhardt, F. Leroux, S. Liao, E. Licata, H. E. P. Lindstrøm, T. A. Lister, E. Livanou, A. Lobel, W. Löffler, M. López, A. Lopez-Lozano, D. Lorenz, T. Loureiro, I. MacDonald, T. Magalhães Fernandes, S. Managau, R. G. Mann, G. Mantelet, O. Marchal, J. M. Marchant, M. Marconi, J. Marie, S. Marinoni, P. M. Marrese, G. Marschalkó, D. J. Marshall, J. M. Martín-Fleitas, M. Martino, N. Mary, G. Matijevič, T. Mazeh, P. J. McMillan, S. Messina, A. Mestre, D. Michalik, N. R. Millar, B. M. H. Miranda, D. Molina, R. Molinaro, M. Molinaro, L. Molnár, M. Moniez, P. Montegriffo, D. Monteiro, R. Mor, A. Mora, R. Morbidelli, T. Morel, S. Morgenthaler, T. Morley, D. Morris, A. F. Mulone, T. Muraveva, I. Musella, J. Narbonne, G. Nelemans, L. Nicastro, L. Noval, C. Ordénovic, J. Ordieres-Meré, P. Osborne, C. Pagani, I. Pagano, F. Pailler, H. Palacin, L. Palaversa, P. Parsons, T. Paulsen, M. Pecoraro, R. Pedrosa, H. Pentikäinen, J. Pereira, B. Pichon, A. M. Piersimoni, F. X. Pineau, E. Plachy,

G. Plum, E. Poujoulet, A. Prša, L. Pulone, S. Ragaini, S. Rago, N. Rambaux, M. Ramos-Lerate, P. Ranalli, G. Rauw, A. Read, S. Regibo, F. Renk, C. Reylé, R. A. Ribeiro, L. Rimoldini, V. Ripepi, A. Riva, G. Rixon, M. Roelens, M. Romero-Gómez, N. Rowell, F. Royer, A. Rudolph, L. Ruiz-Dern, G. Sadowski, T. Sagristà Sellés, J. Sahlmann, J. Salgado, E. Salguero, M. Sarasso, H. Savietto, A. Schnorhk, M. Schultheis, E. Sciacca, M. Segol, J. C. Segovia, D. Segransan, E. Serpell, I. C. Shih, R. Smareglia, R. L. Smart, C. Smith, E. Solano, F. Solitro, R. Sordo, S. Soria Nieto, J. Souchay, A. Spagna, F. Spoto, U. Stampa, I. A. Steele, H. Steidelmüller, C. A. Stephenson, H. Stoev, F. F. Suess, M. Süveges, J. Surdej, L. Szabados, E. Szegedi-Elek, D. Tapiador, F. Taris, G. Tauran, M. B. Taylor, R. Teixeira, D. Terrett, B. Tingley, S. C. Trager, C. Turon, A. Ulla, E. Utrilla, G. Valentini, A. van Elteren, E. Van Hemelryck, M. van Leeuwen, M. Varadi, A. Vecchiato, J. Veljanoski, T. Via, D. Vicente, S. Vogt, H. Voss, V. Votruba, S. Voutsinas, G. Walmsley, M. Weiler, K. Weingrill, D. Werner, T. Wevers, G. Whitehead, Ł. Wyrzykowski, A. Yoldas, M. Žerjal, S. Zucker, C. Zurbach, T. Zwitter, A. Alecu, M. Allen, C. Allende Prieto, A. Amorim, G. Anglada-Escudé, V. Arsenijevic, S. Azaz, P. Balm, M. Beck, H. H. Bernstein, L. Bigot, A. Bijaoui, C. Blasco, M. Bonfigli, G. Bono, S. Boudreault, A. Bressan, S. Brown, P. M. Brunet, P. Bunclark, R. Buonanno, A. G. Butkevich, C. Carret, C. Carrion, L. Chemin, F. Chéreau, L. Corcione, E. Darmigny, K. S. de Boer, P. de Teodoro, P. T. de Zeeuw, C. Delle Luche, C. D. Domingues, P. Dubath, F. Fodor, B. Frézouls, A. Fries, D. Fustes, D. Fyfe, E. Gallardo, J. Gallegos, D. Gardiol, M. Gebran, A. Gomboc, A. Gómez, E. Grux, A. Gueguen, A. Heyrovsky, J. Hoar, G. Iannicola, Y. Isasi Parache, A. M. Janotto, E. Joliet, A. Jonckheere, R. Keil, D. W. Kim, P. Klagyivik, J. Klar, J. Knude, O. Kochukhov, I. Kolka, J. Kos, A. Kutka, V. Lainey, D. LeBouquin, C. Liu, D. Loreggia, V. V. Makarov, M. G. Marseille, C. Martayan, O. Martinez-Rubi, B. Massart, F. Meynadier, S. Mignot, U. Munari, A. T. Nguyen, T. Nordlander, P. Ocvirk, K. S. O'Flaherty, A. Olias Sanz, P. Ortiz, J. Osorio, D. Oszkiewicz, A. Ouzounis, M. Palmer, P. Park, E. Pasquato, C. Peltzer, J. Peralta, F. Péturaud, T. Pieniluoma, E. Pigozzi, J. Poels, G. Prat, T. Prod'homme, F. Raison, J. M. Rebordao, D. Risquez, B. Rocca-Volmerange, S. Rosen, M. I. Ruiz-Fuertes, F. Russo, S. Sembay, I. Serraller Vizcaino, A. Short, A. Siebert, H. Silva, D. Sinachopoulos, E. Slezak, M. Soffel, D. Sosnowska, V. Straižys, M. ter Linden, D. Terrell, S. Theil, C. Tiede, L. Troisi, P. Tsalmantza, D. Tur, M. Vaccari, F. Vachier, P. Valles, W. Van Hamme, L. Veltz, J. Virtanen, J. M. Wallut, R. Wichmann, M. I. Wilkinson, H. Ziaeepour, and S. Zschocke, "The Gaia mission," , vol. 595, p. A1, Nov. 2016.

[Cited on page 6.]

M. Perryman, J. Hartman, G. Á. Bakos, and L. Lindegren, "Astrometric Exoplanet Detection with Gaia," , vol. 797, no. 1, p. 14, Dec. 2014. [Cited on page 6.]

A. Sozzetti, P. Giacobbe, M. G. Lattanzi, G. Micela, and G. Tinetti, "The Gaia Astrometric Survey of Nearby M Dwarfs: A Treasure Trove for Exoplanet Astrophysics," in *AAS/Division for Extreme Solar Systems Abstracts*, ser. AAS/Division for Extreme Solar Systems Abstracts, vol. 2, Sep. 2011, p. 41.01. [Cited on page 6.]

A. Thiabaud, U. Marboeuf, Y. Alibert, I. Leya, and K. Mezger, "Elemental ratios in stars vs planets," , vol. 580, p. A30, Aug. 2015. [Cited on page 7.]

D. G. Russell, "On the Need for a Classification System for Consistent Characterization of the Composition of Planetary Bodies," *arXiv e-prints*, p. arXiv:2108.03343, Aug. 2021. [Cited on page 7.]

C. D. Dressing, D. Charbonneau, X. Dumusque, S. Gettel, F. Pepe, A. Collier Cameron, D. W. Latham, E. Molinari, S. Udry, L. Affer, A. S. Bonomo, L. A. Buchhave, R. Cosentino, P. Figueira, A. F. M. Fiorenzano, A. Harutyunyan, R. D. Haywood, J. A. Johnson, M. Lopez-Morales, C. Lovis, L. Malavolta, M. Mayor, G. Micela, F. Motalebi, V. Nascimbeni, D. F. Phillips, G. Piotto, D. Pollacco, D. Queloz, K. Rice, D. Sasselov, D. Ségransan, A. Sozzetti, A. Szentgyorgyi, and C. Watson, "The Mass of Kepler-93b and The Composition of Terrestrial Planets," , vol. 800, no. 2, p. 135, Feb. 2015. [Cited on pages xvii, 7, and 8.]

D. Rubie, S. Jacobson, A. Morbidelli, D. O'Brien, E. Young, J. de Vries, F. Nimmo, H. Palme, and D. Frost, "Accretion and differentiation of the terrestrial planets with implications for the compositions of early-formed solar system bodies and accretion of water," *Icarus*, vol. 248, pp. 89–108, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019103514005545 [Cited on page 7.]

A. Aguichine, O. Mousis, B. Devouard, and T. Ronnet, "Rocklines as Cradles for Refractory Solids in the Protosolar Nebula," , vol. 901, no. 2, p. 97, Oct. 2020. [Cited on page 7.]

C. Dorn, A. Khan, K. Heng, J. A. D. Connolly, Y. Alibert, W. Benz, and P. Tackley, "Can we constrain the interior structure of rocky exoplanets from mass and radius measurements?" , vol. 577, p. A83, May 2015. [Cited on pages 7, 14, 15, 16, 53, and 74.]

C. Dorn, N. R. Hinkel, and J. Venturini, "Bayesian analysis of interiors of HD 219134b, Kepler-10b, Kepler-93b, CoRoT-7b, 55 Cnc e, and HD 97658b using stellar abundance proxies," , vol. 597, p. A38, Jan. 2017. [Cited on pages 7, 15, and 55.]

Oxford-Reference, "Williamina Fleming." [Online]. Available: https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095823407 [Cited on page 8.]

E. C. Pickering, "The Draper Catalogue of stellar spectra photographed with the 8-inch Bache telescope as a part of the Henry Draper memorial," *Annals of Harvard College Observatory*, vol. 27, pp. 1–388, Jan. 1890. [Cited on page 8.]

D. Sobel, "The glass universe: How the ladies of the harvard observatory took the measure of the stars. new york: Viking, 2016. 324 pp. hardcover (isbn 9780670016952)." *Polymath*, vol. 7, pp. 35–38, 06 2017. [Cited on page 9.]

I. Ridpath, "A dictionary of astronomy, 3rd edition," 2018. [Online]. Available: https://www.oxfordreference.com/view/10.1093/acref/9780199609055.001.0001/acref-9780199609055 [Cited on page 9.]

M. J. Pecaut and E. E. Mamajek, "Intrinsic Colors, Temperatures, and Bolometric Corrections of Pre-main-sequence Stars," , vol. 208, no. 1, p. 9, Sep. 2013. [Cited on page 9.]

G. Ledrew, "The Real Starry Sky," , vol. 95, p. 32, Feb. 2001. [Cited on page 9.]

T. J. Henry, W.-C. Jao, J. P. Subasavage, T. D. Beaulieu, P. A. Ianna, E. Costa, and R. A. Méndez, "The Solar Neighborhood. XVII. Parallax Results from the CTIOPI 0.9 m Program: 20 New Members of the RECONS 10 Parsec Sample," , vol. 132, no. 6, pp. 2360–2371, Dec. 2006. [Cited on page 9.]

J. G. Winters, T. J. Henry, W.-C. Jao, J. P. Subasavage, J. P. Chatelain, K. Slatten, A. R. Riedel, M. L. Silverstein, and M. J. Payne, "The Solar Neighborhood. XLV. The Stellar Multiplicity Rate of M Dwarfs Within 25 pc," , vol. 157, no. 6, p. 216, Jun. 2019. [Cited on page 9.]

C. H. Payne, "Stellar Atmospheres; a Contribution to the Observational Study of High Temperature in the Reversing Layers of Stars." Ph.D. dissertation, RADCLIFFE COLLEGE., Jan. 1925. [Cited on page 9.]

P. Moore, *The Guinness book of astronomy*, 1992. [Cited on page 9.]

A. Önehag, U. Heiter, B. Gustafsson, N. Piskunov, B. Plez, and A. Reiners, "M-dwarf metallicities. A high-resolution spectroscopic study in the near infrared," , vol. 542, p. A33, Jun. 2012. [Cited on pages 9 and 10.]

N. Afram and S. V. Berdyugina, "Complexity of magnetic fields on red dwarfs," , vol. 629, p. A83, Sep. 2019. [Cited on page 9.]

O. Kochukhov, "Magnetic fields of M dwarfs," , vol. 29, no. 1, p. 1, Dec. 2021. [Cited on pages 9 and 13.]

A. S. Rajpurohit, F. Allard, S. Rajpurohit, R. Sharma, G. D. C. Teixeira, O. Mousis, and K. Rajpurohit, "Exploring the stellar properties of M dwarfs with high-resolution spectroscopy from the optical to the near-infrared," , vol. 620, p. A180, Dec. 2018. [Cited on pages xvii, 9, and 10.]

R. J. Rutten, *Radiative Transfer in Stellar Atmospheres*, 2003. [Cited on page 10.]

X. Bonfils, X. Delfosse, S. Udry, N. C. Santos, T. Forveille, and D. Ségransan, "Metallicity of M dwarfs. I. A photometric calibration and the impact on the mass-luminosity relation at the bottom of the main sequence," , vol. 442, no. 2, pp. 635–642, Nov. 2005. [Cited on page 10.]

B. Rojas-Ayala, K. R. Covey, P. S. Muirhead, and J. P. Lloyd, "Metal-rich M-Dwarf Planet Hosts: Metallicities with K-band Spectra," , vol. 720, no. 1, pp. L113–L118, Sep. 2010. [Cited on page 10.]

J. L. Bean, G. F. Benedict, and M. Endl, "Metallicities of M Dwarf Planet Hosts from Spectral Synthesis," , vol. 653, no. 1, pp. L65–L68, Dec. 2006. [Cited on page 10.]

J. L. Bean, C. Sneden, P. H. Hauschildt, C. M. Johns-Krull, and G. F. Benedict, "Accurate M Dwarf Metallicities from Spectral Synthesis: A Critical Test of Model Atmospheres," , vol. 652, no. 2, pp. 1604–1616, Dec. 2006. [Cited on page 10.]

V. M. Woolf and G. Wallerstein, "Calibrating M Dwarf Metallicities Using Molecular Indices," , vol. 118, no. 840, pp. 218–226, Feb. 2006. [Cited on page 10.]

D. Souto, K. Cunha, D. A. García-Hernández, O. Zamora, C. A. Prieto, V. V. Smith, S. Mahadevan, C. Blake, J. A. Johnson, H. Jönsson, M. Pinsonneault, J. Holtzman,

S. R. Majewski, M. Shetrone, J. Teske, D. Nidever, R. Schiavon, J. Sobeck, A. E. G. Pérez, Y. G. M. Chew, and K. Stassun, "Chemical abundances of m-dwarfs from the apogee survey. i. the exoplanet hosting stars kepler-138 and kepler-186," *The Astrophysical Journal*, vol. 835, no. 2, p. 239, jan 2017. [Online]. Available: https://doi.org/10.3847/1538-4357/835/2/239 [Cited on page 10.]

J. Maldonado, G. Micela, M. Baratella, V. D'Orazi, L. Affer, K. Biazzo, A. F. Lanza, A. Maggio, J. I. González Hernández, M. Perger, M. Pinamonti, G. Scandariato, A. Sozzetti, D. Locci, C. Di Maio, A. Bignamini, R. Claudi, E. Molinari, R. Rebolo, I. Ribas, B. Toledo-Padrón, E. Covino, S. Desidera, E. Herrero, J. C. Morales, A. Suárez-Mascareño, I. Pagano, A. Petralia, G. Piotto, and E. Poretti, "HADES RV programme with HARPS-N at TNG. XII. The abundance signature of M dwarf stars with planets," , vol. 644, p. A68, Dec. 2020. [Cited on pages 11, 51, and 52.]

H. T. Ishikawa, W. Aoki, T. Kotani, M. Kuzuhara, M. Omiya, A. Reiners, and M. Zechmeister, "Elemental abundances of M dwarfs based on high-resolution near-infrared spectra: Verification by binary systems," , vol. 72, no. 6, p. 102, Dec. 2020. [Cited on page 11.]

S. Lindgren, U. Heiter, and A. Seifahrt, "Metallicity determination of M dwarfs. High-resolution infrared spectroscopy," , vol. 586, p. A100, Feb. 2016. [Cited on page 11.]

F. C. Adams and G. Laughlin, "A dying universe: the long-term fate and evolutionof astrophysical objects," *Reviews of Modern Physics*, vol. 69, no. 2, pp. 337–372, Apr. 1997. [Cited on page 12.]

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J.-P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J.-F., Carron, J., Challinor, A., Chiang, H. C., Chluba, J., Colombo, L. P. L., Combet, C., Contreras, D., Crill, B. P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J.-M., Di Valentino, E., Diego, J. M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Fantaye, Y., Farhang, M., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Fraisse, A. A., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga,

K., Génova-Santos, R. T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Handley, W., Hansen, F. K., Herranz, D., Hildebrandt, S. R., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T. S., Knox, L., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lagache, G., Lamarre, J.-M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Le Jeune, M., Lemos, P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lilley, M., Lindholm, V., López-Caniego, M., Lubin, P. M., Ma, Y.-Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M.-A., Molinari, D., Montier, L., Morgante, G., Moss, A., Natoli, P., Nørgaard-Nielsen, H. U., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Peiris, H. V., Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J.-L., Rachen, J. P., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E. P. S., Sirignano, C., Sirri, G., Spencer, L. D., Sunyaev, R., Suur-Uski, A.-S., Tauber, J. A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valenziano, L., Valiviita, J., Van Tent, B., Vibert, L., Vielva, P., Villa, F., Vittorio, N., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Zacchei, A., and Zonca, A., "Planck 2018 results - vi. cosmological parameters," *A&A*, vol. 641, p. A6, 2020. [Online]. Available: https://doi.org/10.1051/0004-6361/201833910 [Cited on page 12.]

A. L. Shields, S. Ballard, and J. A. Johnson, "The habitability of planets orbiting M-dwarf stars," , vol. 663, p. 1, Dec. 2016. [Cited on pages 12 and 13.]

M. Lingam and A. Loeb, "Reduced Diversity of Life around Proxima Centauri and TRAPPIST-1," , vol. 846, no. 2, p. L21, Sep. 2017. [Cited on pages 12 and 13.]

J. C. Tarter, P. R. Backus, R. L. Mancinelli, J. M. Aurnou, D. E. Backman, G. S. Basri, A. P. Boss, A. Clarke, D. Deming, L. R. Doyle, E. D. Feigelson, F. Freund, D. H. Grinspoon, R. M. Haberle, I. Hauck, Steven A., M. J. Heath, T. J. Henry, J. L. Hollingsworth, M. M. Joshi, S. Kilston, M. C. Liu, E. Meikle, I. N. Reid, L. J. Rothschild, J. Scalo, A. Segura, C. M. Tang, J. M. Tiedje, M. C. Turnbull, L. M. Walkowicz, A. L. Weber, and R. E. Young,

"A Reappraisal of The Habitability of Planets around M Dwarf Stars," *Astrobiology*, vol. 7, no. 1, pp. 30–65, Mar. 2007. [Cited on page 12.]

S. Ballard, "Constraining M Dwarf Planetary System Lifetimes with the TESS Yield," in *American Astronomical Society Meeting Abstracts #233*, ser. American Astronomical Society Meeting Abstracts, vol. 233, Jan. 2019, p. 326.01. [Cited on page 12.]

S.-S. Huang, "Occurrence of Life in the Universe," *American Scientist*, vol. 47, no. 3, pp. 397–402, Sep. 1959. [Cited on page 13.]

M. H. Hart, "The evolution of the atmosphere of the earth," , vol. 33, no. 1, pp. 23–39, Jan. 1978. [Cited on page 13.]

R. K. Kopparapu, "A Revised Estimate of the Occurrence Rate of Terrestrial Planets in the Habitable Zones around Kepler M-dwarfs," , vol. 767, no. 1, p. L8, Apr. 2013. [Cited on page 13.]

R. K. Kopparapu, R. M. Ramirez, J. SchottelKotte, J. F. Kasting, S. Domagal-Goldman, and V. Eymet, "Habitable Zones around Main-sequence Stars: Dependence on Planetary Mass," , vol. 787, no. 2, p. L29, Jun. 2014. [Cited on page 13.]

R. K. Kopparapu, *The Habitable Zone: The Climatic Limits of Habitability*, 2018, p. 58. [Cited on page 13.]

L. Affer, G. Micela, M. Damasso, M. Perger, I. Ribas, A. Suárez Mascareño, J. I. González Hernández, R. Rebolo, E. Poretti, J. Maldonado, G. Leto, I. Pagano, G. Scandariato, R. Zanmar Sanchez, A. Sozzetti, A. S. Bonomo, L. Malavolta, J. C. Morales, A. Rosich, A. Bignamini, R. Gratton, S. Velasco, D. Cenadelli, R. Claudi, R. Cosentino, S. Desidera, P. Giacobbe, E. Herrero, M. Lafarga, A. F. Lanza, E. Molinari, and G. Piotto, "HADES RV program with HARPS-N at the TNG GJ 3998: An early M-dwarf hosting a system of super-Earths," , vol. 593, p. A117, Oct. 2016. [Cited on page 13.]

M. Gillon, E. Jehin, S. M. Lederer, L. Delrez, J. de Wit, A. Burdanov, V. Van Grootel, A. J. Burgasser, A. H. M. J. Triaud, C. Opitom, B.-O. Demory, D. K. Sahu, D. Bardalez Gagliuffi, P. Magain, and D. Queloz, "Temperate Earth-sized planets transiting a nearby ultracool dwarf star," , vol. 533, no. 7602, pp. 221–224, May 2016. [Cited on page 13.]

M. Gillon, A. H. M. J. Triaud, B.-O. Demory, E. Jehin, E. Agol, K. M. Deck, S. M. Lederer, J. de Wit, A. Burdanov, J. G. Ingalls, E. Bolmont, J. Leconte, S. N. Raymond, F. Selsis,

M. Turbet, K. Barkaoui, A. Burgasser, M. R. Burleigh, S. J. Carey, A. Chaushev, C. M. Copperwheat, L. Delrez, C. S. Fernandes, D. L. Holdsworth, E. J. Kotze, V. Van Grootel, Y. Almleaky, Z. Benkhaldoun, P. Magain, and D. Queloz, "Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1," , vol. 542, no. 7642, pp. 456–460, Feb. 2017. [Cited on page 13.]

E. V. Quintana, T. Barclay, S. N. Raymond, J. F. Rowe, E. Bolmont, D. A. Caldwell, S. B. Howell, S. R. Kane, D. Huber, J. R. Crepp, J. J. Lissauer, D. R. Ciardi, J. L. Coughlin, M. E. Everett, C. E. Henze, E. Horch, H. Isaacson, E. B. Ford, F. C. Adams, M. Still, R. C. Hunter, B. Quarles, and F. Selsis, "An Earth-Sized Planet in the Habitable Zone of a Cool Star," *Science*, vol. 344, no. 6181, pp. 277–280, Apr. 2014. [Cited on page 13.]

G. Anglada-Escudé, P. J. Amado, J. Barnes, Z. M. Berdiñas, R. P. Butler, G. A. L. Coleman, I. de La Cueva, S. Dreizler, M. Endl, B. Giesers, S. V. Jeffers, J. S. Jenkins, H. R. A. Jones, M. Kiraga, M. Kürster, M. J. López-González, C. J. Marvin, N. Morales, J. Morin, R. P. Nelson, J. L. Ortiz, A. Ofir, S.-J. Paardekooper, A. Reiners, E. Rodríguez, C. Rodríguez-López, L. F. Sarmiento, J. P. Strachan, Y. Tsapras, M. Tuomi, and M. Zechmeister, "A terrestrial planet candidate in a temperate orbit around Proxima Centauri," , vol. 536, no. 7617, pp. 437–440, Aug. 2016. [Cited on page 13.]

R. Luque, E. Pallé, D. Kossakowski, S. Dreizler, J. Kemmer, N. Espinoza, J. Burt, G. Anglada-Escudé, V. J. S. Béjar, J. A. Caballero, K. A. Collins, K. I. Collins, M. Cortés-Contreras, E. Díez-Alonso, F. Feng, A. Hatzes, C. Hellier, T. Henning, S. V. Jeffers, L. Kaltenegger, M. Kürster, J. Madden, K. Molaverdikhani, D. Montes, N. Narita, G. Nowak, A. Ofir, M. Oshagh, H. Parviainen, A. Quirrenbach, S. Reffert, A. Reiners, C. Rodríguez-López, M. Schlecker, S. Stock, T. Trifonov, J. N. Winn, M. R. Zapatero Osorio, M. Zechmeister, P. J. Amado, D. R. Anderson, N. E. Batalha, F. F. Bauer, P. Bluhm, C. J. Burke, R. P. Butler, D. A. Caldwell, G. Chen, J. D. Crane, D. Dragomir, C. D. Dressing, S. Dynes, J. M. Jenkins, A. Kaminski, H. Klahr, T. Kotani, M. Lafarga, D. W. Latham, P. Lewin, S. McDermott, P. Montañés-Rodríguez, J. C. Morales, F. Murgas, E. Nagel, S. Pedraz, I. Ribas, G. R. Ricker, P. Rowden, S. Seager, S. A. Shectman, M. Tamura, J. Teske, J. D. Twicken, R. Vanderspeck, S. X. Wang, and B. Wohler, "Planetary system around the nearby M dwarf GJ 357 including a transiting, hot, Earth-sized planet optimal for atmospheric characterization," , vol. 628, p. A39, Aug. 2019. [Cited on page 13.]

M. Damasso, "A low-mass planet candidate orbiting Proxima Centauri at a distance of 1.5 au," in *AAS/Division for Extreme Solar Systems Abstracts*, ser. AAS/Division for Extreme Solar Systems Abstracts, vol. 51, Aug. 2019, p. 102.03. [Cited on pages 13 and 14.]

A. Reiners, "Observations of Cool-Star Magnetic Fields," *Living Reviews in Solar Physics*, vol. 9, no. 1, p. 1, Feb. 2012. [Cited on page 13.]

G. A. Feiden and B. Chaboyer, "Magnetic Inhibition of Convection and the Fundamental Properties of Low-mass Stars. I. Stars with a Radiative Core," , vol. 779, no. 2, p. 183, Dec. 2013. [Cited on page 13.]

D. J. Mullan and J. MacDonald, "Are Magnetically Active Low-Mass M Dwarfs Completely Convective?" , vol. 559, no. 1, pp. 353–371, Sep. 2001. [Cited on page 13.]

G. Chabrier, J. Gallardo, and I. Baraffe, "Evolution of low-mass star and brown dwarf eclipsing binaries," , vol. 472, no. 2, pp. L17–L20, Sep. 2007. [Cited on page 13.]

G. A. Feiden and B. Chaboyer, "Magnetic Inhibition of Convection and the Fundamental Properties of Low-mass Stars. II. Fully Convective Main-sequence Stars," , vol. 789, no. 1, p. 53, Jul. 2014. [Cited on page 13.]

J. MacDonald and D. J. Mullan, "Surface Magnetic Field Strengths: New Tests of Magnetoconvective Models of M Dwarfs," , vol. 787, no. 1, p. 70, May 2014. [Cited on page 13.]

S. J. Schmidt, J. L. Prieto, K. Z. Stanek, B. J. Shappee, N. Morrell, D. C. Bardalez Gagliuffi, C. S. Kochanek, J. Jencson, T. W. S. Holoien, U. Basu, J. F. Beacom, D. M. Szczygieł, G. Pojmanski, J. Brimacombe, M. Dubberley, M. Elphick, S. Foale, E. Hawkins, D. Mullins, W. Rosing, R. Ross, and Z. Walker, "Characterizing a Dramatic ΔV ~-9 Flare on an Ultracool Dwarf Found by the ASAS-SN Survey," , vol. 781, no. 2, p. L24, Feb. 2014. [Cited on page 13.]

Y. A. Yamashiki, H. Maehara, V. Airapetian, Y. Notsu, T. Sato, S. Notsu, R. Kuroki, K. Murashima, H. Sato, K. Namekata, T. Sasaki, T. B. Scott, H. Bando, S. Nashimoto, F. Takagi, C. Ling, D. Nogami, and K. Shibata, "Impact of Stellar Superflares on Planetary Habitability," , vol. 881, no. 2, p. 114, Aug. 2019. [Cited on page 13.]

Y. A. Yamashiki, H. Maehara, V. Airapetian, Y. Notsu, T. Sato, S. Notsu, R. Shimizu, N. Kimura, T. Sasaki, and K. Shibata, "Evaluation system of Stellar Superflares impacts on Planetary Habitability," in *AGU Fall Meeting Abstracts*, vol. 2020, Dec. 2020, pp. P007–0004. [Cited on page 13.]

S.-P. Moschou, J. J. Drake, O. Cohen, J. D. Alvarado-Gómez, C. Garraffo, and F. Fraschetti, "The Stellar CME-Flare Relation: What Do Historic Observations Reveal?" , vol. 877, no. 2, p. 105, Jun. 2019. [Cited on page 13.]

H. Lammer, H. I. M. Lichtenegger, Y. N. Kulikov, J.-M. Grießmeier, N. Terada, N. V. Erkaev, H. K. Biernat, M. L. Khodachenko, I. Ribas, T. Penz, and F. Selsis, "Coronal Mass Ejection (CME) Activity of Low Mass M Stars as An Important Factor for The Habitability of Terrestrial Exoplanets. II. CME-Induced Ion Pick Up of Earth-like Exoplanets in Close-In Habitable Zones," *Astrobiology*, vol. 7, no. 1, pp. 185–207, Feb. 2007. [Cited on page 13.]

M. L. Khodachenko, I. Ribas, H. Lammer, J.-M. Grießmeier, M. Leitner, F. Selsis, C. Eiroa, A. Hanslmeier, H. K. Biernat, C. J. Farrugia, and H. O. Rucker, "Coronal Mass Ejection (CME) Activity of Low Mass M Stars as An Important Factor for The Habitability of Terrestrial Exoplanets. I. CME Impact on Expected Magnetospheres of Earth-Like Exoplanets in Close-In Habitable Zones," *Astrobiology*, vol. 7, no. 1, pp. 167–184, Feb. 2007. [Cited on page 13.]

K. G. Kislyakova, L. Fossati, C. P. Johnstone, L. Noack, T. Lüftinger, V. V. Zaitsev, and H. Lammer, "Effective Induction Heating around Strongly Magnetized Stars," , vol. 858, no. 2, p. 105, May 2018. [Cited on page 14.]

K. G. Kislyakova, L. Noack, C. P. Johnstone, V. V. Zaitsev, L. Fossati, H. Lammer, M. L. Khodachenko, P. Odert, and M. Güdel, "Magma oceans and enhanced volcanism on TRAPPIST-1 planets due to induction heating," *Nature Astronomy*, vol. 1, pp. 878–885, Oct. 2017. [Cited on page 14.]

T. Penz, G. Micela, and H. Lammer, "Influence of the evolving stellar X-ray luminosity distribution on exoplanetary mass loss," , vol. 477, no. 1, pp. 309–314, Jan. 2008. [Cited on page 14.]

S. L. Grimm, B.-O. Demory, M. Gillon, C. Dorn, E. Agol, A. Burdanov, L. Delrez, M. Sestovic, A. H. M. J. Triaud, M. Turbet, É. Bolmont, A. Caldas, J. de Wit, E. Jehin, J. Leconte,

S. N. Raymond, V. Van Grootel, A. J. Burgasser, S. Carey, D. Fabrycky, K. Heng, D. M. Hernandez, J. G. Ingalls, S. Lederer, F. Selsis, and D. Queloz, "The nature of the TRAPPIST-1 exoplanets," , vol. 613, p. A68, May 2018. [Cited on page 14.]

V. Adibekyan, "Formation and Evolution of Exoplanets in Different Environments," in *Non-Stable Universe: Energetic Resources, Activity Phenomena, and Evolutionary Processes*, ser. Astronomical Society of the Pacific Conference Series, A. M. Mickaelian, H. A. Harutyunian, and E. H. Nikoghosyan, Eds., vol. 511, Jul. 2017, p. 70. [Cited on page 14.]

V. Z. Adibekyan, S. G. Sousa, N. C. Santos, E. Delgado Mena, J. I. González Hernández, G. Israelian, M. Mayor, and G. Khachatryan, "Chemical abundances of 1111 FGK stars from the HARPS GTO planet search program. Galactic stellar populations and planets," , vol. 545, p. A32, Sep. 2012. [Cited on pages 14, 19, 31, 32, 39, 42, and 55.]

N. C. Santos, V. Adibekyan, C. Dorn, C. Mordasini, L. Noack, S. C. C. Barros, E. Delgado-Mena, O. Demangeon, J. P. Faria, G. Israelian, and S. G. Sousa, "Constraining planet structure and composition from stellar chemistry: trends in different stellar populations," , vol. 608, p. A94, Dec. 2017. [Cited on pages 14, 39, 53, 54, 55, 71, 73, 77, and 78.]

N. Madhusudhan, "C/O Ratio as a Dimension for Characterizing Exoplanetary Atmospheres," , vol. 758, no. 1, p. 36, Oct. 2012. [Cited on page 14.]

J. C. Bond, D. P. O'Brien, and D. S. Lauretta, "The Compositional Diversity of Extrasolar Terrestrial Planets. I. In Situ Simulations," , vol. 715, no. 2, pp. 1050–1070, Jun. 2010. [Cited on pages 14, 15, and 54.]

A. Thiabaud, U. Marboeuf, Y. Alibert, N. Cabral, I. Leya, and K. Mezger, "From stellar nebula to planets: The refractory components," , vol. 562, p. A27, Feb. 2014. [Cited on pages 14 and 15.]

L. A. Rogers and S. Seager, "A Framework for Quantifying the Degeneracies of Exoplanet Interior Compositions," , vol. 712, no. 2, pp. 974–991, Apr. 2010. [Cited on page 14.]

C. T. Unterborn, E. E. Dismukes, and W. R. Panero, "Scaling the Earth: A Sensitivity Analysis of Terrestrial Exoplanetary Interior Models," , vol. 819, no. 1, p. 32, Mar. 2016. [Cited on pages 15 and 16.]

S. Elser, M. R. Meyer, and B. Moore, "On the origin of elemental abundances in the terrestrial planets," , vol. 221, no. 2, pp. 859–874, Nov. 2012. [Cited on page 15.]

A. Thiabaud, U. Marboeuf, Y. Alibert, I. Leya, and K. Mezger, "Elemental ratios in stars vs planets," , vol. 580, p. A30, Aug. 2015. [Cited on page 15.]

K. Lodders, "Solar System Abundances and Condensation Temperatures of the Elements," , vol. 591, no. 2, pp. 1220–1247, Jul. 2003. [Cited on pages 15, 53, 69, 71, 73, and 78.]

T. V. Johnson, O. Mousis, J. I. Lunine, and N. Madhusudhan, "Planetesimal Compositions in Exoplanet Systems," , vol. 757, no. 2, p. 192, Oct. 2012. [Cited on page 15.]

B. Bitsch, A. Crida, A. Morbidelli, W. Kley, and I. Dobbs-Dixon, "Stellar irradiated discs and implications on migration of embedded planets. I. Equilibrium discs," , vol. 549, p. A124, Jan. 2013. [Cited on page 15.]

B. Bitsch, A. Johansen, M. Lambrechts, and A. Morbidelli, "The structure of protoplanetary discs around evolving young stars," , vol. 575, p. A28, Mar. 2015. [Cited on page 15.]

B. Bitsch, M. Lambrechts, and A. Johansen, "The growth of planets by pebble accretion in evolving protoplanetary discs," , vol. 582, p. A112, Oct. 2015. [Cited on page 15.]

B. Bitsch and S. Savvidou, "Influence of grain size and composition on the contraction rates of planetary envelopes and on planetary migration," , vol. 647, p. A96, Mar. 2021. [Cited on page 15.]

E. Delgado Mena, G. Israelian, J. I. González Hernández, J. C. Bond, N. C. Santos, S. Udry, and M. Mayor, "Chemical Clues on the Formation of Planetary Systems: C/O Versus Mg/Si for HARPS GTO Sample," , vol. 725, no. 2, pp. 2349–2358, Dec. 2010. [Cited on pages 15 and 54.]

J. C. Bond, D. S. Lauretta, and D. P. O'Brien, "Making the Earth: Combining dynamics and chemistry in the Solar System," , vol. 205, no. 2, pp. 321–337, Feb. 2010. [Cited on page 15.]

N. C. Santos, V. Adibekyan, C. Mordasini, W. Benz, E. Delgado-Mena, C. Dorn, L. Buchhave, P. Figueira, A. Mortier, F. Pepe, A. Santerne, S. G. Sousa, and S. Udry, "Constraining planet structure from stellar chemistry: the cases of CoRoT-7, Kepler-10, and Kepler-93," , vol. 580, p. L13, Aug. 2015. [Cited on pages 16, 53, 54, 55, and 77.]

V. Adibekyan, C. Dorn, S. G. Sousa, N. C. Santos, B. Bitsch, G. Israelian, C. Mordasini, S. C. C. Barros, E. D. Mena, O. D. S. Demangeon, J. P. Faria, P. Figueira, A. A. Hakobyan, M. Oshagh, B. M. T. B. Soares, M. Kunitomo, Y. Takeda, E. Jofré, R. Petrucci, and E. Martioli, "A compositional link between rocky exoplanets and their host stars," *Science*, vol. 374, no. 6565, pp. 330–332, 2021. [Cited on pages 16, 50, 68, 73, and 74.]

V. Adibekyan, S. G. Sousa, and N. C. Santos, "Characterization of Exoplanet-Host Stars," in *Asteroseismology and Exoplanets: Listening to the Stars and Searching for New Worlds*, T. L. Campante, N. C. Santos, and M. J. P. F. G. Monteiro, Eds., vol. 49, Jan. 2018, p. 225. [Cited on page 16.]

N. C. Santos, G. Israelian, and M. Mayor, "Chemical analysis of 8 recently discovered extra-solar planet host stars," , vol. 363, pp. 228–238, Nov. 2000. [Cited on page 16.]

——, "Spectroscopic [Fe/H] for 98 extra-solar planet-host stars. Exploring the probability of planet formation," , vol. 415, pp. 1153–1166, Mar. 2004. [Cited on pages 16 and 119.]

V. Adibekyan, "Heavy Metal Rules. I. Exoplanet Incidence and Metallicity," *Geosciences*, vol. 9, no. 3, p. 105, Feb. 2019. [Cited on pages 16 and 25.]

D. Naef, M. Mayor, F. Pepe, D. Queloz, N. C. Santos, S. Udry, and M. Burnet, "The CORALIE survey for southern extrasolar planets. V. 3 new extrasolar planets," , vol. 375, pp. 205–218, Aug. 2001. [Cited on page 16.]

E. L. Rickman, D. Ségransan, M. Marmier, S. Udry, F. Bouchy, C. Lovis, M. Mayor, F. Pepe, D. Queloz, N. C. Santos, R. Allart, V. Bonvin, P. Bratschi, F. Cersullo, B. Chazelas, A. Choplin, U. Conod, A. Deline, J. B. Delisle, L. A. Dos Santos, P. Figueira, H. A. C. Giles, M. Girard, B. Lavie, D. Martin, F. Motalebi, L. D. Nielsen, H. Osborn, G. Ottoni, M. Raimbault, J. Rey, T. Roger, J. V. Seidel, M. Stalport, A. Suárez Mascareño, A. Triaud, O. Turner, L. Weber, and A. Wyttenbach, "The CORALIE survey for southern extrasolar planets. XVIII. Three new massive planets and two low-mass brown dwarfs at greater than 5 AU separation," , vol. 625, p. A71, May 2019. [Cited on page 16.]

M. Meyer, "Frequency of Massive Wide-orbit Planets vs. Stellar Mass: SPHERE SHINES on the ESO VLT," in *AAS/Division for Extreme Solar Systems Abstracts*, ser. AAS/Division for Extreme Solar Systems Abstracts, vol. 51, Aug. 2019, p. 101.01. [Cited on page 16.]

S. Desidera, G. Chauvin, M. Bonavita, S. Messina, H. LeCoroller, T. Schmidt, R. Gratton, C. Lazzoni, M. Meyer, J. Schlieder, A. Cheetham, J. Hagelberg, M. Bonnefoy, M. Feldt, A. M. Lagrange, M. Langlois, A. Vigan, T. G. Tan, F. J. Hambsch, M. Millward, J. Alcalá, S. Benatti, W. Brandner, J. Carson, E. Covino, P. Delorme, V. D'Orazi, M. Janson, E. Rigliaco, J. L. Beuzit, B. Biller, A. Boccaletti, C. Dominik, F. Cantalloube, C. Fontanive, R. Galicher, T. Henning, E. Lagadec, R. Ligi, A. L. Maire, F. Menard, D. Mesa, A. Müller, M. Samland, H. M. Schmid, E. Sissa, M. Turatto, S. Udry, A. Zurlo, R. Asensio-Torres, T. Kopytova, E. Rickman, L. Abe, J. Antichi, A. Baruffolo, P. Baudoz, J. Baudrand, P. Blanchard, A. Bazzon, T. Buey, M. Carbillet, M. Carle, J. Charton, E. Cascone, R. Claudi, A. Costille, A. Deboulbé, V. De Caprio, K. Dohlen, D. Fantinel, P. Feautrier, T. Fusco, P. Gigan, E. Giro, D. Gisler, L. Gluck, N. Hubin, E. Hugot, M. Jaquet, M. Kasper, F. Madec, Y. Magnard, P. Martinez, D. Maurel, D. Le Mignant, O. Möller-Nilsson, M. Llored, T. Moulin, A. Origné, A. Pavlov, D. Perret, C. Petit, J. Pragt, P. Puget, P. Rabou, J. Ramos, F. Rigal, S. Rochat, R. Roelfsema, G. Rousset, A. Roux, B. Salasnich, J. F. Sauvage, A. Sevin, C. Soenke, E. Stadler, M. Suarez, L. Weber, and F. Wildi, "The SPHERE infrared survey for exoplanets (SHINE). I. Sample definition and target characterization," , vol. 651, p. A70, Jul. 2021. [Cited on page 16.]

F. Pepe, M. Mayor, B. Delabre, D. Kohler, D. Lacroix, D. Queloz, S. Udry, W. Benz, J.-L. Bertaux, and J.-P. Sivan, "HARPS: a new high-resolution spectrograph for the search of extrasolar planets," in *Optical and IR Telescope Instrumentation and Detectors*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, M. Iye and A. F. Moorwood, Eds., vol. 4008, Aug. 2000, pp. 582–592. [Cited on pages 16 and 19.]

G. Rupprecht, F. Pepe, M. Mayor, D. Queloz, F. Bouchy, G. Avila, W. Benz, J.-L. Bertaux, X. Bonfils, T. Dall, B. Delabre, H. Dekker, W. Eckert, M. Fleury, A. Gilliotte, D. Gojak, J. C. Guzman, D. Kohler, J.-L. Lizon, G. Lo Curto, A. Longinotti, C. Lovis, D. Megevand, L. Pasquini, J. Reyes, J.-P. Sivan, D. Sosnowska, R. Soto, S. Udry, A. Van Kesteren, L. Weber, and U. Weilenmann, "The exoplanet hunter HARPS: performance and first results," in *Ground-based Instrumentation for Astronomy*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, A. F. M. Moorwood and M. Iye, Eds., vol. 5492, Sep. 2004, pp. 148–159. [Cited on pages 16 and 19.]

S. B. Howell, C. Sobeck, M. Haas, M. Still, T. Barclay, F. Mullally, J. Troeltzsch, S. Aigrain, S. T. Bryson, D. Caldwell, W. J. Chaplin, W. D. Cochran, D. Huber, G. W. Marcy,

A. Miglio, J. R. Najita, M. Smith, J. D. Twicken, and J. J. Fortney, "The K2 Mission: Characterization and Early Results," , vol. 126, no. 938, p. 398, Apr. 2014. [Cited on page 16.]

S. R. Majewski, R. P. Schiavon, P. M. Frinchaboy, C. Allende Prieto, R. Barkhouser, D. Bizyaev, B. Blank, S. Brunner, A. Burton, R. Carrera, S. D. Chojnowski, K. Cunha, C. Epstein, G. Fitzgerald, A. E. García Pérez, F. R. Hearty, C. Henderson, J. A. Holtzman, J. A. Johnson, C. R. Lam, J. E. Lawler, P. Maseman, S. Mészáros, M. Nelson, D. C. Nguyen, D. L. Nidever, M. Pinsonneault, M. Shetrone, S. Smee, V. V. Smith, T. Stolberg, M. F. Skrutskie, E. Walker, J. C. Wilson, G. Zasowski, F. Anders, S. Basu, S. Beland, M. R. Blanton, J. Bovy, J. R. Brownstein, J. Carlberg, W. Chaplin, C. Chiappini, D. J. Eisenstein, Y. Elsworth, D. Feuillet, S. W. Fleming, J. Galbraith-Frew, R. A. García, D. A. García-Hernández, B. A. Gillespie, L. Girardi, J. E. Gunn, S. Hasselquist, M. R. Hayden, S. Hekker, I. Ivans, K. Kinemuchi, M. Klaene, S. Mahadevan, S. Mathur, B. Mosser, D. Muna, J. A. Munn, R. C. Nichol, R. W. O'Connell, J. K. Parejko, A. C. Robin, H. Rocha-Pinto, M. Schultheis, A. M. Serenelli, N. Shane, V. Silva Aguirre, J. S. Sobeck, B. Thompson, N. W. Troup, D. H. Weinberg, and O. Zamora, "The Apache Point Observatory Galactic Evolution Experiment (APOGEE)," , vol. 154, no. 3, p. 94, Sep. 2017. [Cited on pages 16 and 17.]

S. Randich, G. Gilmore, and Gaia-ESO Consortium, "The Gaia-ESO Large Public Spectroscopic Survey," *The Messenger*, vol. 154, pp. 47–49, Dec. 2013. [Cited on pages 16 and 77.]

M. Steinmetz, T. Zwitter, A. Siebert, F. G. Watson, K. C. Freeman, U. Munari, R. Campbell, M. Williams, G. M. Seabroke, R. F. G. Wyse, Q. A. Parker, O. Bienaymé, S. Roeser, B. K. Gibson, G. Gilmore, E. K. Grebel, A. Helmi, J. F. Navarro, D. Burton, C. J. P. Cass, J. A. Dawe, K. Fiegert, M. Hartley, K. S. Russell, W. Saunders, H. Enke, J. Bailin, J. Binney, J. Bland-Hawthorn, C. Boeche, W. Dehnen, D. J. Eisenstein, N. W. Evans, M. Fiorucci, J. P. Fulbright, O. Gerhard, U. Jauregi, A. Kelz, L. Mijović, I. Minchev, G. Parmentier, J. Peñarrubia, A. C. Quillen, M. A. Read, G. Ruchti, R. D. Scholz, A. Siviero, M. C. Smith, R. Sordo, L. Veltz, S. Vidrih, R. von Berlepsch, B. J. Boyle, and E. Schilbach, "The Radial Velocity Experiment (RAVE): First Data Release," , vol. 132, no. 4, pp. 1645–1668, Oct. 2006. [Cited on page 16.]

R. Deshpande, C. H. Blake, C. F. Bender, S. Mahadevan, R. C. Terrien, J. K. Carlberg, G. Zasowski, J. Crepp, A. S. Rajpurohit, C. Reylé, D. L. Nidever, D. P. Schneider, C. Allende Prieto, D. Bizyaev, G. Ebelke, S. W. Fleming, P. M. Frinchaboy, J. Ge, F. Hearty, J. Hernández, E. Malanushenko, V. Malanushenko, S. R. Majewski, R. Marchwinski, D. Muna, D. Oravetz, K. Pan, R. P. Schiavon, M. Shetrone, A. Simmons, K. G. Stassun, J. C. Wilson, and J. P. Wisniewski, "The SDSS-III APOGEE Radial Velocity Survey of M Dwarfs. I. Description of the Survey and Science Goals," , vol. 146, no. 6, p. 156, Dec. 2013. [Cited on pages 17 and 77.]

J. C. Wilson, F. R. Hearty, M. F. Skrutskie, S. R. Majewski, J. A. Holtzman, D. Eisenstein, J. Gunn, B. Blank, C. Henderson, S. Smee, M. Nelson, D. Nidever, J. Arns, R. Barkhouser, J. Barr, S. Beland, M. A. Bershady, M. R. Blanton, S. Brunner, A. Burton, L. Carey, M. Carr, J. P. Colque, J. Crane, G. J. Damke, J. Davidson, J. W., J. Dean, F. Di Mille, K. W. Don, G. Ebelke, M. Evans, G. Fitzgerald, B. Gillespie, M. Hall, A. Harding, P. Harding, R. Hammond, D. Hancock, C. Harrison, S. Hope, T. Horne, J. Karakla, C. Lam, F. Leger, N. MacDonald, P. Maseman, J. Matsunari, S. Melton, T. Mitcheltree, T. O'Brien, R. W. O'Connell, A. Patten, W. Richardson, G. Rieke, M. Rieke, A. Roman-Lopes, R. P. Schiavon, J. S. Sobeck, T. Stolberg, R. Stoll, M. Tembe, J. D. Trujillo, A. Uomoto, M. Vernieri, E. Walker, D. H. Weinberg, E. Young, B. Anthony-Brumfield, D. Bizyaev, B. Breslauer, N. De Lee, J. Downey, S. Halverson, J. Huehnerhoff, M. Klaene, E. Leon, D. Long, S. Mahadevan, E. Malanushenko, D. C. Nguyen, R. Owen, J. R. Sánchez-Gallego, C. Sayres, N. Shane, S. A. Shectman, M. Shetrone, D. Skinner, F. Stauffer, and B. Zhao, "The Apache Point Observatory Galactic Evolution Experiment (APOGEE) Spectrographs," , vol. 131, no. 999, p. 055001, May 2019. [Cited on page 17.]

M. Shetrone, D. Bizyaev, J. E. Lawler, C. Allende Prieto, J. A. Johnson, V. V. Smith, K. Cunha, J. Holtzman, A. E. García Pérez, S. Mészáros, J. Sobeck, O. Zamora, D. A. García-Hernández, D. Souto, D. Chojnowski, L. Koesterke, S. Majewski, and G. Zasowski, "The SDSS-III APOGEE Spectral Line List for H-band Spectroscopy," , vol. 221, no. 2, p. 24, Dec. 2015. [Cited on page 17.]

J. A. Holtzman, M. Shetrone, J. A. Johnson, C. A. Prieto, F. Anders, B. Andrews, T. C. Beers, D. Bizyaev, M. R. Blanton, J. Bovy, R. Carrera, S. D. Chojnowski, K. Cunha, D. J. Eisenstein, D. Feuillet, P. M. Frinchaboy, J. Galbraith-Frew, A. E. G. Pérez, D. A. García-Hernández, S. Hasselquist, M. R. Hayden, F. R. Hearty, I. Ivans, S. R. Majewski,

S. Martell, S. Meszaros, D. Muna, D. Nidever, D. C. Nguyen, R. W. O'Connell, K. Pan, M. Pinsonneault, A. C. Robin, R. P. Schiavon, N. Shane, J. Sobeck, V. V. Smith, N. Troup, D. H. Weinberg, J. C. Wilson, W. M. Wood-Vasey, O. Zamora, and G. Zasowski, "ABUNDANCES, STELLAR PARAMETERS, AND SPECTRA FROM THE SDSS-III/APOGEE SURVEY," *The Astronomical Journal*, vol. 150, no. 5, p. 148, oct 2015. [Online]. Available: https://doi.org/10.1088/0004-6256/150/5/148 [Cited on page 17.]

A. E. García Pérez, C. Allende Prieto, J. A. Holtzman, M. Shetrone, S. Mészáros, D. Bizyaev, R. Carrera, K. Cunha, D. A. García-Hernández, J. A. Johnson, S. R. Majewski, D. L. Nidever, R. P. Schiavon, N. Shane, V. V. Smith, J. Sobeck, N. Troup, O. Zamora, D. H. Weinberg, J. Bovy, D. J. Eisenstein, D. Feuillet, P. M. Frinchaboy, M. R. Hayden, F. R. Hearty, D. C. Nguyen, R. W. O'Connell, M. H. Pinsonneault, J. C. Wilson, and G. Zasowski, "ASPCAP: The APOGEE Stellar Parameter and Chemical Abundances Pipeline," , vol. 151, no. 6, p. 144, Jun. 2016. [Cited on page 17.]

S. R. Majewski, APOGEE Team, and APOGEE-2 Team, "The Apache Point Observatory Galactic Evolution Experiment (APOGEE) and its successor, APOGEE-2," *Astronomische Nachrichten*, vol. 337, no. 8-9, p. 863, Sep. 2016. [Cited on page 17.]

H. Jönsson, J. A. Holtzman, C. Allende Prieto, K. Cunha, D. A. García-Hernández, S. Hasselquist, T. Masseron, Y. Osorio, M. Shetrone, V. Smith, G. S. Stringfellow, D. Bizyaev, B. Edvardsson, S. R. Majewski, S. Mészáros, D. Souto, O. Zamora, R. L. Beaton, J. Bovy, J. Donor, M. H. Pinsonneault, V. J. Poovelil, and J. Sobeck, "APOGEE Data and Spectral Analysis from SDSS Data Release 16: Seven Years of Observations Including First Results from APOGEE-South," , vol. 160, no. 3, p. 120, Sep. 2020. [Cited on pages xix, 17, 32, 33, 83, and 84.]

R. Ahumada, C. A. Prieto, A. Almeida, F. Anders, S. F. Anderson, B. H. Andrews, B. Anguiano, R. Arcodia, E. Armengaud, M. Aubert, S. Avila, V. Avila-Reese, C. Badenes, C. Balland, K. Barger, J. K. Barrera-Ballesteros, S. Basu, J. Bautista, R. L. Beaton, T. C. Beers, B. I. T. Benavides, C. F. Bender, M. Bernardi, M. Bershady, F. Beutler, C. M. Bidin, J. Bird, D. Bizyaev, G. A. Blanc, M. R. Blanton, M. Boquien, J. Borissova, J. Bovy, W. N. Brandt, J. Brinkmann, J. R. Brownstein, K. Bundy, M. Bureau, A. Burgasser, E. Burtin, M. Cano-Díaz, R. Capasso, M. Cappellari, R. Carrera, S. Chabanier, W. Chaplin, M. Chapman, B. Cherinka, C. Chiappini, P. Doohyun Choi, S. D. Chojnowski,

H. Chung, N. Clerc, D. Coffey, J. M. Comerford, J. Comparat, L. da Costa, M.-C. Cousinou, K. Covey, J. D. Crane, K. Cunha, G. d. S. Ilha, Y. S. Dai, S. B. Damsted, J. Darling, J. Davidson, James W., R. Davies, K. Dawson, N. De, A. de la Macorra, N. De Lee, A. B. d. A. Queiroz, A. Deconto Machado, S. de la Torre, F. Dell'Agli, H. du Mas des Bourboux, A. M. Diamond-Stanic, S. Dillon, J. Donor, N. Drory, C. Duckworth, T. Dwelly, G. Ebelke, S. Eftekharzadeh, A. Davis Eigenbrot, Y. P. Elsworth, M. Eracleous, G. Erfanianfar, S. Escoffier, X. Fan, E. Farr, J. G. Fernández-Trincado, D. Feuillet, A. Finoguenov, P. Fofie, A. Fraser-McKelvie, P. M. Frinchaboy, S. Fromenteau, H. Fu, L. Galbany, R. A. Garcia, D. A. García-Hernández, L. A. G. Oehmichen, J. Ge, M. A. G. Maia, D. Geisler, J. Gelfand, J. Goddy, V. Gonzalez-Perez, K. Grabowski, P. Green, C. J. Grier, H. Guo, J. Guy, P. Harding, S. Hasselquist, A. J. Hawken, C. R. Hayes, F. Hearty, S. Hekker, D. W. Hogg, J. A. Holtzman, D. Horta, J. Hou, B.-C. Hsieh, D. Huber, J. A. S. Hunt, J. I. Chitham, J. Imig, M. Jaber, C. E. J. Angel, J. A. Johnson, A. M. Jones, H. Jönsson, E. Jullo, Y. Kim, K. Kinemuchi, I. Kirkpatrick, Charles C., G. W. Kite, M. Klaene, J.-P. Kneib, J. A. Kollmeier, H. Kong, M. Kounkel, D. Krishnarao, I. Lacerna, T.-W. Lan, R. R. Lane, D. R. Law, J.-M. Le Goff, H. W. Leung, H. Lewis, C. Li, J. Lian, L. Lin, D. Long, P. Longa-Peña, B. Lundgren, B. W. Lyke, J. Ted Mackereth, C. L. MacLeod, S. R. Majewski, A. Manchado, C. Maraston, P. Martini, T. Masseron, K. L. Masters, S. Mathur, R. M. McDermid, A. Merloni, M. Merrifield, S. Mészáros, A. Miglio, D. Minniti, R. Minsley, T. Miyaji, F. G. Mohammad, B. Mosser, E.-M. Mueller, D. Muna, A. Muñoz-Gutiérrez, A. D. Myers, S. Nadathur, P. Nair, K. Nandra, J. C. do Nascimento, R. J. Nevin, J. A. Newman, D. L. Nidever, C. Nitschelm, P. Noterdaeme, J. E. O'Connell, M. D. Olmstead, D. Oravetz, A. Oravetz, Y. Osorio, Z. J. Pace, N. Padilla, N. Palanque-Delabrouille, P. A. Palicio, H.-A. Pan, K. Pan, J. Parker, R. Paviot, S. Peirani, K. P. Ramŕez, S. Penny, W. J. Percival, I. Perez-Fournon, I. Pérez-Ràfols, P. Petitjean, M. M. Pieri, M. Pinsonneault, V. J. Poovelil, J. T. Povick, A. Prakash, A. M. Price-Whelan, M. J. Raddick, A. Raichoor, A. Ray, S. B. Rembold, M. Rezaie, R. A. Riffel, R. Riffel, H.-W. Rix, A. C. Robin, A. Roman-Lopes, C. Román-Zúñiga, B. Rose, A. J. Ross, G. Rossi, K. Rowlands, K. H. R. Rubin, M. Salvato, A. G. Sánchez, L. Sánchez-Menguiano, J. R. Sánchez-Gallego, C. Sayres, A. Schaefer, R. P. Schiavon, J. S. Schimoia, E. Schlafly, D. Schlegel, D. P. Schneider, M. Schultheis, A. Schwope, H.-J. Seo, A. Serenelli, A. Shafieloo, S. J. Shamsi, Z. Shao, S. Shen, M. Shetrone, R. Shirley, V. S. Aguirre, J. D. Simon, M. F. Skrutskie, A. Slosar, R. Smethurst, J. Sobeck, B. C. Sodi, D. Souto, D. V. Stark, K. G. Stassun,

M. Steinmetz, D. Stello, J. Stermer, T. Storchi-Bergmann, A. Streblyanska, G. S. Stringfellow, A. Stutz, G. Suárez, J. Sun, M. Taghizadeh-Popp, M. S. Talbot, J. Tayar, A. R. Thakar, R. Theriault, D. Thomas, Z. C. Thomas, J. Tinker, R. Tojeiro, H. H. Toledo, C. A. Tremonti, N. W. Troup, S. Tuttle, E. Unda-Sanzana, M. Valentini, J. Vargas-González, M. Vargas-Magaña, J. A. Vázquez-Mata, M. Vivek, D. Wake, Y. Wang, B. A. Weaver, A.-M. Weijmans, V. Wild, J. C. Wilson, R. F. Wilson, N. Wolthuis, W. M. Wood-Vasey, R. Yan, M. Yang, C. Yèche, O. Zamora, P. Zarrouk, G. Zasowski, K. Zhang, C. Zhao, G. Zhao, Z. Zheng, Z. Zheng, G. Zhu, and H. Zou, "The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra," , vol. 249, no. 1, p. 3, Jul. 2020. [Cited on page 17.]

G. M. De Silva, K. C. Freeman, J. Bland-Hawthorn, S. Martell, E. W. de Boer, M. Asplund, S. Keller, S. Sharma, D. B. Zucker, T. Zwitter, B. Anguiano, C. Bacigalupo, D. Bayliss, M. A. Beavis, M. Bergemann, S. Campbell, R. Cannon, D. Carollo, L. Casagrande, A. R. Casey, G. Da Costa, V. D'Orazi, A. Dotter, L. Duong, A. Heger, M. J. Ireland, P. R. Kafle, J. Kos, J. Lattanzio, G. F. Lewis, J. Lin, K. Lind, U. Munari, D. M. Nataf, S. O'Toole, Q. Parker, W. Reid, K. J. Schlesinger, A. Sheinis, J. D. Simpson, D. Stello, Y. S. Ting, G. Traven, F. Watson, R. Wittenmyer, D. Yong, and M. Žerjal, "The GALAH survey: scientific motivation," , vol. 449, no. 3, pp. 2604–2617, May 2015. [Cited on pages 18 and 77.]

A. Sheinis, B. Anguiano, M. Asplund, C. Bacigalupo, S. Barden, M. Birchall, J. Bland-Hawthorn, J. Brzeski, R. Cannon, D. Carollo, S. Case, A. Casey, V. Churilov, C. Warrick, R. Dean, G. De Silva, V. D'Orazi, L. Duong, T. Farrell, K. Fiegert, K. Freeman, F. Gabriella, L. Gers, M. Goodwin, D. Gray, A. Green, R. Heald, J. Heijmans, M. Ireland, D. Jones, P. Kafle, S. Keller, U. Klauser, Y. Kondrat, J. Kos, J. Lawrence, S. Lee, S. Mali, S. Martell, D. Mathews, D. Mayfield, S. Miziarski, R. Muller, N. Pai, R. Patterson, E. Penny, D. Orr, K. Schlesinger, S. Sharma, K. Shortridge, J. Simpson, S. Smedley, G. Smith, D. Stafford, N. Staszak, M. Vuong, L. Waller, E. W. de Boer, P. Xavier, J. Zheng, R. Zhelem, D. Zucker, and T. Zwitter, "First light results from the High Efficiency and Resolution Multi-Element Spectrograph at the Anglo-Australian Telescope," *Journal of Astronomical Telescopes, Instruments, and Systems*, vol. 1, p. 035002, Jul. 2015. [Cited on page 18.]

S. Buder, S. Sharma, J. Kos, A. M. Amarsi, T. Nordlander, K. Lind, S. L. Martell, M. Asplund, J. Bland-Hawthorn, A. R. Casey, G. M. de Silva, V. D'Orazi, K. C. Freeman, M. R. Hayden, G. F. Lewis, J. Lin, K. J. Schlesinger, J. D. Simpson, D. Stello, D. B. Zucker, T. Zwitter, K. L. Beeson, T. Buck, L. Casagrande, J. T. Clark, K. Čotar, G. S. da Costa, R. de Grijs, D. Feuillet, J. Horner, P. R. Kafle, S. Khanna, C. Kobayashi, F. Liu, B. T. Montet, G. Nandakumar, D. M. Nataf, M. K. Ness, L. Spina, T. Tepper-García, Y.-S. Ting, G. Traven, R. Vogrinčič, R. A. Wittenmyer, R. F. G. Wyse, M. Žerjal, M. Žerjal, and Galah Collaboration, "The GALAH+ survey: Third data release," , vol. 506, no. 1, pp. 150–201, Sep. 2021. [Cited on page 18.]

A. Suárez Mascareño, J. I. González Hernández, R. Rebolo, N. Astudillo-Defru, X. Bonfils, F. Bouchy, X. Delfosse, T. Forveille, C. Lovis, M. Mayor, F. Murgas, F. Pepe, N. C. Santos, S. Udry, A. Wünsche, and S. Velasco, "A super-Earth orbiting the nearby M dwarf GJ 536," , vol. 597, p. A108, Jan. 2017. [Cited on page 19.]

E. Delgado Mena, M. Tsantaki, V. Z. Adibekyan, S. G. Sousa, N. C. Santos, J. I. González Hernández, and G. Israelian, "Chemical abundances of 1111 FGK stars from the HARPS GTO planet search program. II. Cu, Zn, Sr, Y, Zr, Ba, Ce, Nd, and Eu," , vol. 606, p. A94, Oct. 2017. [Cited on page 19.]

E. Delgado Mena, M. Tsantaki, V. Zh. Adibekyan, S. G. Sousa, N. C. Santos, J. I. González Hernández, and G. Israelian, "Abundance ratios & ages of stellar populations in HARPS-GTO sample," in *Astrometry and Astrophysics in the Gaia Sky*, A. Recio-Blanco, P. de Laverny, A. G. A. Brown, and T. Prusti, Eds., vol. 330, Apr. 2018, pp. 156–159. [Cited on page 19.]

E. Delgado Mena, A. Moya, V. Adibekyan, M. Tsantaki, J. I. González Hernández, G. Israelian, G. R. Davies, W. J. Chaplin, S. G. Sousa, A. C. S. Ferreira, and N. C. Santos, "Abundance to age ratios in the HARPS-GTO sample with Gaia DR2. Chemical clocks for a range of [Fe/H]," , vol. 624, p. A78, Apr. 2019. [Cited on page 19.]

S. G. Sousa, V. Adibekyan, E. Delgado-Mena, N. C. Santos, B. Rojas-Ayala, B. M. T. B. Soares, H. Legoinha, S. Ulmer-Moll, J. D. Camacho, S. C. C. Barros, O. D. S. Demangeon, S. Hoyer, G. Israelian, A. Mortier, M. Tsantaki, and M. Monteiro, "SWEET-Cat 2.0: The Cat just got SWEETer; Higher quality spectra and precise parallaxes from GAIA eDR3," *arXiv e-prints*, p. arXiv:2109.04781, Sep. 2021. [Cited on pages 19, 68, 78, and 119.]

A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959. [Cited on page 21.]

E. M. Howard, "Machine Learning Algorithms in Astronomy," in *Astronomical Data Analysis Software and Systems XXV*, ser. Astronomical Society of the Pacific Conference Series, N. P. F. Lorente, K. Shortridge, and R. Wayth, Eds., vol. 512, Dec. 2017, p. 245. [Cited on pages 21 and 22.]

D. Baron, "Machine Learning in Astronomy: a practical overview," *arXiv e-prints*, p. arXiv:1904.07248, Apr. 2019. [Cited on pages 21, 22, and 23.]

A. Géron, *Hands-on machine learning with SCIKIT-LEARN, Keras, and TENSORFLOW: Concepts, tools, and techniques to build intelligent systems*, 2nd ed. O'Reilly Media, Inc., 2019. [Cited on pages xvii, 22, 23, and 24.]

D. R. Chong, M. Kim, J. Ahn, and H. Jeong, "Machine learning identification of symmetrized base states of rydberg atoms," Aug 2021. [Cited on page 22.]

S. Auddy and M.-K. Lin, "A Machine Learning Model to Infer Planet Masses from Gaps Observed in Protoplanetary Disks," , vol. 900, no. 1, p. 62, Sep. 2020. [Cited on page 22.]

S. Webb, M. Lochner, D. Muthukrishna, J. Cooke, C. Flynn, A. Mahabal, S. Goode, I. Andreoni, T. Pritchard, and T. M. C. Abbott, "Unsupervised machine learning for transient discovery in deeper, wider, faster light curves," , vol. 498, no. 3, pp. 3077–3094, Nov. 2020. [Cited on page 23.]

A. Mey and M. Loog, "Improvability Through Semi-Supervised Learning: A Survey of Theoretical Results," *arXiv e-prints*, p. arXiv:1908.09574, Aug. 2019. [Cited on page 23.]

X. Yang, Z. Song, I. King, and Z. Xu, "A Survey on Deep Semi-supervised Learning," *arXiv e-prints*, p. arXiv:2103.00550, Feb. 2021. [Cited on page 23.]

P. Dubath, L. Rimoldini, M. Süveges, J. Blomme, M. López, L. M. Sarro, J. De Ridder, J. Cuypers, L. Guy, I. Lecoeur, K. Nienartowicz, A. Jan, M. Beck, N. Mowlavi, P. De Cat, T. Lebzelter, and L. Eyer, "Random forest automated supervised classification of Hipparcos periodic variable stars," , vol. 414, no. 3, pp. 2602–2617, Jul. 2011. [Cited on pages 24 and 25.]

S. Ulmer-Moll, N. C. Santos, P. Figueira, J. Brinchmann, and J. P. Faria, "Beyond the exoplanet mass-radius relation," , vol. 630, p. A135, Oct. 2019. [Cited on pages 24 and 25.]

E. Melton, "A Random Forest Approach to Identifying Young Stellar Object Candidates in the Lupus Star-forming Region," , vol. 159, no. 5, p. 200, May 2020. [Cited on page 25.]

J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard, "On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data," , vol. 733, no. 1, p. 10, May 2011. [Cited on page 25.]

A. McWilliam, "Abundance Ratios and Galactic Chemical Evolution," , vol. 35, pp. 503–556, Jan. 1997. [Cited on page 31.]

S. E. Woosley and T. A. Weaver, "Type II Supernovae," in *Problems of Collapse and Numerical Relativity*, ser. NATO Advanced Study Institute (ASI) Series C, D. Bancel and M. Signore, Eds., vol. 134, Jan. 1984, p. 27. [Cited on page 31.]

F. W. Giacobbe, "How a Type II Supernova Explodes," *Electronic Journal of Theoretical Physics*, vol. 2, no. 6, pp. 30–38, May 2005. [Cited on page 31.]

J. Whelan and J. Iben, Icko, "Binaries and Supernovae of Type I," , vol. 186, pp. 1007–1014, Dec. 1973. [Cited on page 31.]

X. D. Li and E. P. J. van den Heuvel, "Evolution of white dwarf binaries: supersoft X-ray sources and progenitors of type IA supernovae." , vol. 322, pp. L9–L12, Jun. 1997. [Cited on page 31.]

G. Lü, C. Zhu, Z. Wang, and N. Wang, "An alternative symbiotic channel to Type Ia supernovae," , vol. 396, no. 2, pp. 1086–1095, Jun. 2009. [Cited on page 31.]

R. F. Webbink, "Double white dwarfs as progenitors of R Coronae Borealis stars and type I supernovae." , vol. 277, pp. 355–360, Feb. 1984. [Cited on page 31.]

G. Nelemans, L. R. Yungelson, S. F. Portegies Zwart, and F. Verbunt, "Population synthesis for double white dwarfs . I. Close detached systems," , vol. 365, pp. 491–507, Jan. 2001. [Cited on page 31.]

K. Maguire, *Type Ia Supernovae*, 2017, p. 293. [Cited on page 31.]

T. Bensby and S. Feltzing, "The Galactic thin and thick disks in the context of galaxy formation," in *Chemical Abundances in the Universe: Connecting First Stars to Planets*,

K. Cunha, M. Spite, and B. Barbuy, Eds., vol. 265, Mar. 2010, pp. 300–303. [Cited on pages 31, 32, and 42.]

G. Kordopatis, A. Recio-Blanco, P. de Laverny, G. Gilmore, V. Hill, R. F. G. Wyse, A. Helmi, A. Bijaoui, M. Zoccali, and O. Bienaymé, "A spectroscopic survey of thick disc stars outside the solar neighbourhood," , vol. 535, p. A107, Nov. 2011. [Cited on page 31.]

M. Haywood, P. Di Matteo, M. Lehnert, O. Snaith, F. Fragkoudi, and S. Khoperskov, "Phylogeny of the Milky Way's inner disk and bulge populations: Implications for gas accretion, (the lack of) inside-out thick disk formation, and quenching," , vol. 618, p. A78, Oct. 2018. [Cited on page 32.]

J. L. Johnson and H. Li, "The First Planets: The Critical Metallicity for Planet Formation," , vol. 751, no. 2, p. 81, Jun. 2012. [Cited on page 33.]

A. Mortier, N. C. Santos, A. Sozzetti, M. Mayor, D. Latham, X. Bonfils, and S. Udry, "The frequency of giant planets around metal-poor stars," , vol. 543, p. A45, Jul. 2012. [Cited on page 33.]

A. Niedzielski, G. Nowak, M. Adamów, and A. Wolszczan, "Substellar-mass Companions to the K-dwarf BD+14 4559 and the K-giants HD 240210 and BD+20 2457," , vol. 707, no. 1, pp. 768–777, Dec. 2009. [Cited on page 33.]

M. Perryman, *The Exoplanet Handbook*, 2018. [Cited on page 39.]

S. Bertran de Lis, E. Delgado Mena, V. Z. Adibekyan, N. C. Santos, and S. G. Sousa, "Oxygen abundances in G- and F-type stars from HARPS. Comparison of [OI] 6300 Å and OI 6158 Å," , vol. 576, p. A89, Apr. 2015. [Cited on pages 39 and 69.]

E. Delgado Mena, V. Adibekyan, N. C. Santos, M. Tsantaki, J. I. González Hernández, S. G. Sousa, and S. Bertrán de Lis, "Chemical abundances of 1111 FGK stars from the HARPS GTO planet search program IV. Carbon and C/O ratios for Galactic stellar populations and planet hosts," *arXiv e-prints*, p. arXiv:2109.04844, Sep. 2021. [Cited on page 39.]

S. G. Sousa, N. C. Santos, M. Mayor, S. Udry, L. Casagrande, G. Israelian, F. Pepe, D. Queloz, and M. J. P. F. G. Monteiro, "Spectroscopic parameters for 451 stars in the HARPS GTO planet search program. Stellar [Fe/H] and the frequency of exo-Neptunes," , vol. 487, no. 1, pp. 373–381, Aug. 2008. [Cited on pages 40 and 119.]

S. G. Sousa, N. C. Santos, G. Israelian, C. Lovis, M. Mayor, P. B. Silva, and S. Udry, "Spectroscopic characterization of a sample of metal-poor solar-type stars from the HARPS planet search program. Precise spectroscopic parameters and mass estimation," , vol. 526, p. A99, Feb. 2011. [Cited on page 40.]

S. G. Sousa, N. C. Santos, G. Israelian, M. Mayor, and S. Udry, "Spectroscopic stellar parameters for 582 FGK stars in the HARPS volume-limited sample. Revising the metallicity-planet correlation," , vol. 533, p. A141, Sep. 2011. [Cited on page 40.]

D. Bashi and S. Zucker, "Small Planets in the Galactic Context: Host Star Kinematics, Iron, and Alpha-element Enhancement," , vol. 158, no. 2, p. 61, Aug. 2019. [Cited on page 42.]

D. R. H. Johnson and D. R. Soderblom, "Calculating Galactic Space Velocities and Their Uncertainties, with an Application to the Ursa Major Group," , vol. 93, p. 864, Apr. 1987. [Cited on page 42.]

B. Coşkunoğlu, S. Ak, S. Bilir, S. Karaali, E. Yaz, G. Gilmore, G. M. Seabroke, O. Bienaymé, J. Bland-Hawthorn, R. Campbell, K. C. Freeman, B. Gibson, E. K. Grebel, U. Munari, J. F. Navarro, Q. A. Parker, A. Siebert, A. Siviero, M. Steinmetz, F. G. Watson, R. F. G. Wyse, and T. Zwitter, "Local stellar kinematics from RAVE data - I. Local standard of rest," , vol. 412, no. 2, pp. 1237–1245, Apr. 2011. [Cited on page 42.]

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Cited on page 43.]

P. E. Nissen, "High-precision abundances of elements in solar twin stars. Trends with stellar age and elemental condensation temperature," , vol. 579, p. A52, Jul. 2015. [Cited on pages 47 and 50.]

J. I. González Hernández, G. Israelian, N. C. Santos, S. Sousa, E. Delgado-Mena, V. Neves, and S. Udry, "Searching for the Signatures of Terrestrial Planets in Solar Analogs," , vol. 720, no. 2, pp. 1592–1602, Sep. 2010. [Cited on page 50.]

M. Asplund, N. Grevesse, A. J. Sauval, and P. Scott, "The Chemical Composition of the Sun," , vol. 47, no. 1, pp. 481–522, Sep. 2009. [Cited on pages 54 and 69.]

J. J. Fortney, "On the Carbon-to-oxygen Ratio Measurement in nearby Sun-like Stars: Implications for Planet Formation and the Determination of Stellar Abundances," , vol. 747, no. 2, p. L27, Mar. 2012. [Cited on page 54.]

J. M. Brewer and D. A. Fischer, "C/O and Mg/Si Ratios of Stars in the Solar Neighborhood," , vol. 831, no. 1, p. 20, Nov. 2016. [Cited on page 54.]

V. Z. Adibekyan, N. C. Santos, S. G. Sousa, and G. Israelian, "A new α-enhanced supersolar metallicity population," , vol. 535, p. L11, Nov. 2011. [Cited on page 55.]

T. Bensby, S. Feltzing, and M. S. Oey, "Exploring the Milky Way stellar disk. A detailed elemental abundance study of 714 F and G dwarf stars in the solar neighbourhood," , vol. 562, p. A71, Feb. 2014. [Cited on page 56.]

V. Z. Adibekyan, E. Delgado Mena, S. G. Sousa, N. C. Santos, G. Israelian, J. I. González Hernández, M. Mayor, and A. A. Hakobyan, "Exploring the α-enhancement of metal-poor planet-hosting stars. The Kepler and HARPS samples," , vol. 547, p. A36, Nov. 2012. [Cited on page 57.]

V. Van Eylen, C. Agentoft, M. S. Lundkvist, H. Kjeldsen, J. E. Owen, B. J. Fulton, E. Petigura, and I. Snellen, "An asteroseismic view of the radius valley: stripped cores, not born rocky," , vol. 479, no. 4, pp. 4786–4795, Oct. 2018. [Cited on page 67.]

S. M. Mills and T. Mazeh, "The Planetary Mass-Radius Relation and Its Dependence on Orbital Period as Measured by Transit Timing Variations and Radial Velocities," , vol. 839, no. 1, p. L8, Apr. 2017. [Cited on pages 67 and 78.]

N. C. Santos, S. G. Sousa, A. Mortier, V. Neves, V. Adibekyan, M. Tsantaki, E. Delgado Mena, X. Bonfils, G. Israelian, M. Mayor, and S. Udry, "SWEET-Cat: A catalogue of parameters for Stars With ExoplanETs. I. New atmospheric parameters and masses for 48 stars with planets," , vol. 556, p. A150, Aug. 2013. [Cited on pages 68, 78, and 119.]

L. Suárez-Andrés, G. Israelian, J. I. González Hernández, V. Z. Adibekyan, E. Delgado Mena, N. C. Santos, and S. G. Sousa, "CNO behaviour in planet-harbouring stars. II. Carbon abundances in stars with and without planets using the CH band," , vol. 599, p. A96, Mar. 2017. [Cited on page 69.]

A. Aitta, "Venus' internal structure, temperature and core composition," , vol. 218, no. 2, pp. 967–974, Apr. 2012. [Cited on pages 71 and 78.]

W. F. McDonough, "Compositional Model for the Earth's Core," *Treatise on Geochemistry*, vol. 2, p. 568, Dec. 2003. [Cited on pages 71 and 78.]

A. Khan, C. Liebske, A. Rozel, A. Rivoldini, F. Nimmo, J. A. D. Connolly, A. C. Plesa, and D. Giardini, "A Geophysical Perspective on the Bulk Composition of Mars," *Journal of Geophysical Research (Planets)*, vol. 123, no. 2, pp. 575–611, Feb. 2018. [Cited on pages 71 and 78.]

S. A. Hauck, J.-L. Margot, S. C. Solomon, R. J. Phillips, C. L. Johnson, F. G. Lemoine, E. Mazarico, T. J. McCoy, S. Padovan, S. J. Peale, M. E. Perry, D. E. Smith, and M. T. Zuber, "The curious case of Mercury's internal structure," *Journal of Geophysical Research (Planets)*, vol. 118, no. 6, pp. 1204–1220, Jun. 2013. [Cited on pages 71 and 74.]

L. Noack, I. Snellen, and H. Rauer, "Water in Extrasolar Planets and Implications for Habitability," , vol. 212, no. 1-2, pp. 877–898, Oct. 2017. [Cited on pages 73 and 78.]

W. R. Ward, "Protoplanet Migration by Nebula Tides," , vol. 126, no. 2, pp. 261–281, Apr. 1997. [Cited on pages 73 and 78.]

R. Luger, R. Barnes, E. Lopez, J. Fortney, B. Jackson, and V. Meadows, "Habitable Evaporated Cores: Transforming Mini-Neptunes into Super-Earths in the Habitable Zones of M Dwarfs," *Astrobiology*, vol. 15, no. 1, pp. 57–88, Jan. 2015. [Cited on pages 73 and 78.]

W. Benz, W. L. Slattery, and A. Cameron, "Collisional stripping of mercury's mantle," *Icarus*, vol. 74, no. 3, pp. 516–528, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0019103588901182 [Cited on page 74.]

W. Benz, A. Anic, J. Horner, and J. A. Whitby, "The Origin of Mercury," , vol. 132, no. 2-4, pp. 189–202, Oct. 2007. [Cited on page 74.]

E. Asphaug and A. Reufer, "Mercury and other iron-rich planetary bodies as relics of inefficient accretion," *Nature Geoscience*, vol. 7, no. 8, pp. 564–568, Aug. 2014. [Cited on page 74.]

C. Spalding and F. C. Adams, "The solar wind prevents reaccretion of debris after mercury's giant impact," *The Planetary Science Journal*, vol. 1, no. 1, p. 7, mar 2020. [Online]. Available: https://doi.org/10.3847/psj/ab781f [Cited on page 74.]

M. S. Clement and J. E. Chambers, "Dynamical Avenues for Mercury's Origin. II. In Situ Formation in the Inner Terrestrial Disk," , vol. 162, no. 1, p. 3, Jul. 2021. [Cited on page 74.]

L. Puig, G. L. Pilbratt, A. Heske, I. Escudero Sanz, and P. E. Crouzet, "ARIEL: an ESA M4 mission candidate," in *Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, H. A. MacEwen, G. G. Fazio, M. Lystrup, N. Batalha, N. Siegler, and E. C. Tong, Eds., vol. 9904, Jul. 2016, p. 99041W. [Cited on page 79.]

J. A. Holtzman, M. Shetrone, J. A. Johnson, C. Allende Prieto, F. Anders, B. Andrews, T. C. Beers, D. Bizyaev, M. R. Blanton, J. Bovy, R. Carrera, S. D. Chojnowski, K. Cunha, D. J. Eisenstein, D. Feuillet, P. M. Frinchaboy, J. Galbraith-Frew, A. E. García Pérez, D. A. García-Hernández, S. Hasselquist, M. R. Hayden, F. R. Hearty, I. Ivans, S. R. Majewski, S. Martell, S. Meszaros, D. Muna, D. Nidever, D. C. Nguyen, R. W. O'Connell, K. Pan, M. Pinsonneault, A. C. Robin, R. P. Schiavon, N. Shane, J. Sobeck, V. V. Smith, N. Troup, D. H. Weinberg, J. C. Wilson, W. M. Wood-Vasey, O. Zamora, and G. Zasowski, "Abundances, Stellar Parameters, and Spectra from the SDSS-III/APOGEE Survey," , vol. 150, no. 5, p. 148, Nov. 2015. [Cited on page 81.]