MSC validation report based on processing that led to GDR3

prepared by:	Jan Rybizki
approved by:	U. Heiter, M. Fouesneau, O.L. Creevey
reference:	GAIA-C8-TN-MPIA-JR-003-02
issue:	1
revision:	4
date:	2021-09-22
status:	Issued

Abstract

The multiple star classifier (MSC) is a module to infer stellar parameters from BPRP spectra and parallax for the two components of a coeval binary system. In this validation report we analyze the output of processing that led to GDR3 where all sources down to G=18.25 mag have been processed. We use an empirical forward model for the BPRP spectra and have implemented priors depending on position on the sky for the distances and the extinctions. The main results are:

✔ MSC works as expected and returns useful results for applicable binary sources (fluxratio smaller than 5).

✔ Our empirical BPRP forward model works reasonably well and the MSC inferred parameters are validated using two independent validation data sets.

✘ The binary classifications from DSC are problematic, so we need to find an alternative solution to identify valid binary sources offline.

Document History

Issue	Revision	Date	Author	Comment
D	0	2021-06-30	JR	First draft
D	1	2021-07-16	JR	Comments from UH implemented
D	2	2021-08-02	JR	Approved by OC and MFX
D	3	2021-09-16	JR	internal links removed in order to publish with DR3
D	4	2021-09-22	JR	Comments from OC implemented

Contents

1 Summary

MSC overview The multiple star classifier (MSC) is a module to infer stellar parameters from BPRP spectra and parallax for the two components of a coeval binary system with a flux ratio below 5. MSC is now using ExtraTrees, trained on wide binaries for which we artificially summed their BPRP fluxes. The results of the ExtraTree regression are used as initialisation for the double Aeneas MCMC, which uses an BPRP forward model relying on empirically calibrated spectral grid trained on APOGEE sources with known astrophysical parameters.

Goals of processing that led to GDR3

- ✓ MSC works as expected and returns sensible results for applicable binary sources
- ✓ show the achieved accuracy for individual stellar parameters
- ✓ compare to GSP-Phot results and show that MSC generally agrees
- ✓ show the biases that arise for binary sources fit by GSP-Phot
- ✓ validate the MSC uncertainties and found the necessity to inflate them by a factor of 10
- ✓ assess the usability of the MCMC convergence parameters. The mean log posterior has a high predictive power of the reliability of the MSC inferred parameters
- ✓ DSC probabilities are not useful to identify binaries therefore all sources with G smaller 18.25 mag have been processed and binary selection needs to be done offline
- ✓ validate the final inference precision and propose an error inflation of 10
- ✓ propose postprocessing criteria

Main results

1. MSC as an algorithm works as expected. The bias and median absolute deviation for an independent GALAH binary validation set consisting of $\sim$10k sources (and $\sim$1k APOGEE binary sources in parenthesis), when throwing out the 16% worse gof sources are:
   - primary $T_{\rm eff}$ / [K] = -72, 143 (-74, 113)
   - secondary $T_{\rm eff}$ / [K] = -350, 143 (-237, 113)
   - primary $\log g$ / [dex] = 0.20, 0.21 (0.09, 0.14)
   - secondary $\log g$ / [dex] = 0.30, 0.28 (0.07, 0.13)
   - $[\mathrm{Fe/H}]$ / [dex] = 0.19, 0.19 (0.11, 0.13)
   - A_0 / [mag] = 0.01, 0.12 (-0.01, 0,08)
   - distance / [pc]= -95, 58 (-24, 18)
2. the empirical forward model BPRP spectra reduces the bias, but has problems due to sparse sampling at ’high’ extinction values

3. the uncertainty reported by MSC is underestimated by a factor of 10 and has been inflated during postprocessing

Post-processing of the outputs Has been applied according to the following rules:

- exclude sources that miss one of the following: G, BP or RP photometry, parallax, BP or RP spectrum.
- exclude sources if they have NaN values in the parameter estimates
- exclude sources if their fluxratio is outside of 1-5.
- exclude source with a parallax larger than 100 mas.
- white dwarf cut, include sources if:
  G + 5*log10(parallax/100) $>$ 7.8 + 3.2*(G_BP-G_RP)
- uncertainty inflation with a factor of 10 but maintaining the upper and lower limits of the respective parameters
- Rescaling the mcmcDrift[1] with the uncertainty inflation factor
- processing flag == '1' if gof $<$ -1000 or (rescaled) mcmcDrift $>$ 1, else processing flag == '0'

[1] mcmcDrift is a parameter that gives the change of the median parameter value at the beginning and end of the final MCMC chain in terms of the standard deviation of that parameter, averaged over all parameters

2 Acronyms

Acronym	Description
DSC	Discrete Source Classifier
GSP-Phot	General Stellar Parametrizer from Photometry
VDT	Validation Data Table, a small DR3 subset of selected sources
gof	goodness of fit value, `logposterior_msc`, the mean log posterior

3 List of validation test & results

The list below lists all individual tests throughout this document and their individual results. (It can be used to rapidly access some particular tests.)

List of validation tests and results

4 Overview of the module

The multiple star classifier (MSC) is a module to infer stellar parameters from BPRP spectra and parallax for the two components of a coeval binary system with a fluxratio below 5. MSC is using ExtraTrees, trained on wide binaries for which we artificially added their BPRP spectra. The results of the ExtraTree regression are used as initialisation for the double Aeneas MCMC chain which uses a BPRP forward model which was trained on empirical BPRP spectra from an APOGEE sample that had all the necessary astrophysical parameters.

As input per source, MSC needs a BPRP spectrum and the associated uncertainties. It will normalise the spectrum but use the total flux as a seperate piece of information. MSC also needs a parallax and parallax uncertainty. MSC selects sources to process based on DSC's binary probabilities > 0.1. It also processes all sources with G $<$ 18.25 mag. The latter conditions were added in the light of weak DSC classification. This resulted in $\sim$348M sources, of which (as a rough estimate) $\sim$20 % could be applicable MSC sources.

For GDR3 we want to publish $T_{\rm eff}$ and $\log g$ of the primary and the secondary components and a common A_0, $[{\rm Fe/H}]$ and distance per system. Those parameters are sampled in the MCMC and also the 100 thinned samples will be published in GDR3 together with MCMC quality indicators. Latent variables which will also be published are the fluxratio and A_G. Fluxratio is defined as the BPRP flux of the primary divided by the BPRP flux of the secondary as coming from the MSC forward model of the individual components. The primary is defined as the brighter component in BPRP flux. This means only fluxratios greater or equal to 1 are possible and we limit MSC to a maximum fluxratio of 5.

The objective for this run is to provide final data products for the GDR3 release.

4.1 MSC in detail

In the following we explain the steps to achieve the MSC inference as outlined above. We start with an ExtraTree that is trained on empirical binary data. The outcome of which is fed into an MCMC algorithm (double Aeneas, similar to GSPPhot but for 2 stellar components, that share the same $[{\rm Fe/H}]$, A_0 and distance). A general problem for MSC is to get training data because binary sources with stellar parameters given for each component in the literature are less than 3k. In order to increase this sample, we queried Gaia EDR3 for resolved wide binaries and inferred their individual parameters. This is done via isochrone fitting the colour and absolute magnitude of each component exploiting that both should have the same age. We only fit $T_{\rm eff}$, $\log g$ and $[{\rm Fe/H}]$ taking the A_0 from an extinction map prior while the distance is well constrained by the parallax. The isochrones and extinction map was taken from Rybizki et al. (2020). The exact routine can be inspected here: [2]. We combine the BPRP fluxes of both components (just adding them and ignoring the problems of the noise properties) and used this sample (see Figure 6) for empirically training the ExtraTrees, but also for validation and as a prior.

The ExtraTree results for log10($T_{\rm eff\,1}$), log10($T_{\rm eff\,2}$), $\log g_1$, $\log g_2$, A_0, $[{\rm Fe/H}]$ and log10(distance) are used as initialisation for the double Aeneas algorithm. We use the empirical BPRP spectra grid (see Section 4.2) to sample the 7 parameter space (distance is only scaling the total flux) posterior given the data (BPRP fluxes, total flux, parallax). In order to decrease the parameter space we restrict the parameter space to $3.477 <$ log10($T_{\rm eff}$) < 3.903, $2.0 < \log g < 5.2$, -1.0

[2] https://github.com/jan-rybizki/isochrone_fitting_example/blob/master/notebook/5params_inference_DR2_wide_binaries.ipynb

$<$ [Fe/H] < 0.5, $0 \leq A_0 \leq 5$ and $1 <$ log10(distance) < 4. We further employ the following priors:

- The exponentially decreasing space density distance prior (Bailer-Jones et al., 2018) with the lengthscale being sky dependent. It was created from the GeDR3mock Rybizki et al. (2020) catalog in order to reproduce the selection cuts of MSC. Unfortunately the parallax SNR > 5 cut is included, but actually all sources irrespective of parallax SNR cut were processed due to a wrong JAVA implementation (this means there is an inconsistency between the prior sample and the sample to which MSC is applied). The actual ADQL query is as follows:

```
SELECT AVG(1.0/parallax)/3.0 AS Lprior,
MAX(a0) as a0max,
ivo_healpix_index(5, ra, dec) AS hpx
FROM gedr3mock.main
WHERE parallax/parallax_error > 5
AND phot_g_mean_mag < 18
And teff_val > 3000 and teff_val < 8000
and logg > 2 and logg < 5.5
and a0 < 5
and feh > -1 and feh < 0.5
GROUP BY hpx
Order by hpx
```

  And a plot of the retrieved data can be seen in Figure 1.

- a Gaussian prior on [Fe/H] with mean = 0 and standard deviation of 0.2 dex

- extinction prior of exp($-A_0/1$ mag) with the further addition of an $A_{0,\mathrm{max}}$ depending on the maximum A_0 coming from the GeDR3mock query from above. Figure 2 shows the distribution over the sky.

- a BPRP fluxratio prior that peaks towards equal luminosity binaries in the range of 1<fluxratio<5. The distribution depicted in figure 3 is coming from the wide binary sample.

- an Kiel diagram prior for both components coming from the wide binaries primary component as shown in figure 4

For those 7 parameters the MCMC chains are reported in GDR3. But we also report fluxratio and A_G, together with quality indicators: mean mcmc drift (this is the mean drift over all parameters in the MCMC chain measured in terms of the respective parameters standard deviation), mean acceptance rate and mean posterior value.

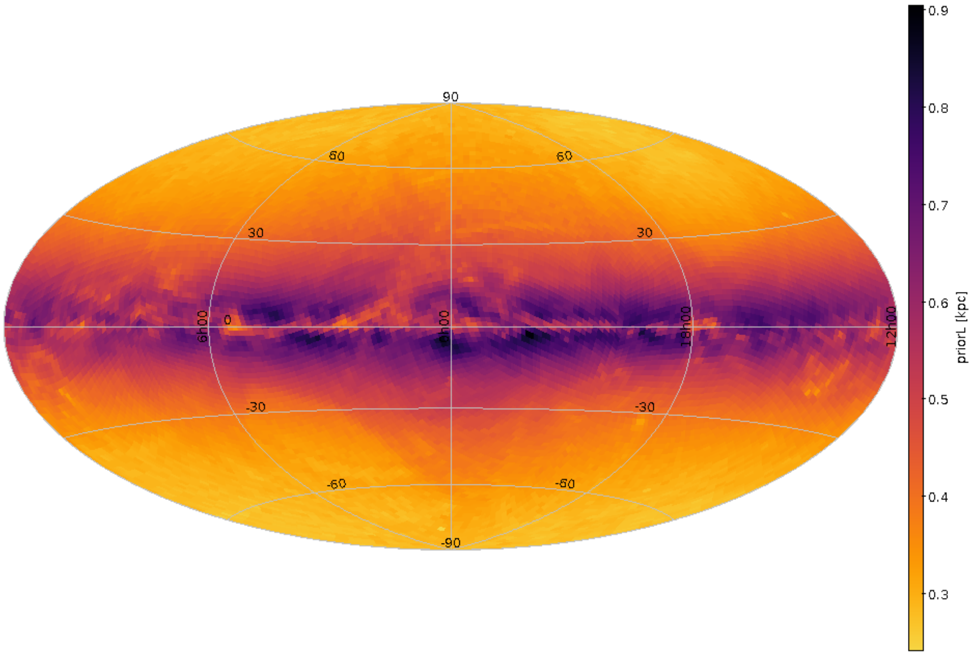

FIGURE 1: MSC distance prior in HEALpix level 5 in Galactic coordinates and aitoff projection.

The likelihood of MSC is Gaussian and combines the parallax, BPRP total flux and individual BPRP fluxes (240 pixels) and associated uncertainties (total BPRP uncertainty inflated by a factor of 2) with the forward model predicted quantities. A total of 242 data point are compared which of course down-weights the contribution from the parallax and the total flux measurement.

DSC is not able to robustly predict suitable binaries for MSC. Therefore we will need to sort out offline which sources are actually unresolved binaries in the applicable fluxratio range. In fact most sources down to 18.25 mag G have been processed.

It was shown that we have biased results due to a mismatch in simulated BPRP spectra model grid (which we used for our forward model) and observed BPRP spectra. Therefore we construct an empirical BPRP spectra model grid which we will explain in the following section.

4.2 Empirical BPRP spectral model grid

We constructed a training catalog of ∼80k APOGEE sources which all had distance and A_0 estimates from StarHorse (Queiroz et al., 2020), $\mathcal{T}_{\rm eff}$, $\log g$ and [M/H] estimates from the ASPCAP pipeline (Jönsson et al., 2020) and BPRP spectra in GaiaDR3. We cleaned those of known binary stars and cut them to the required parameter ranges. We uncertainty sample the parameters

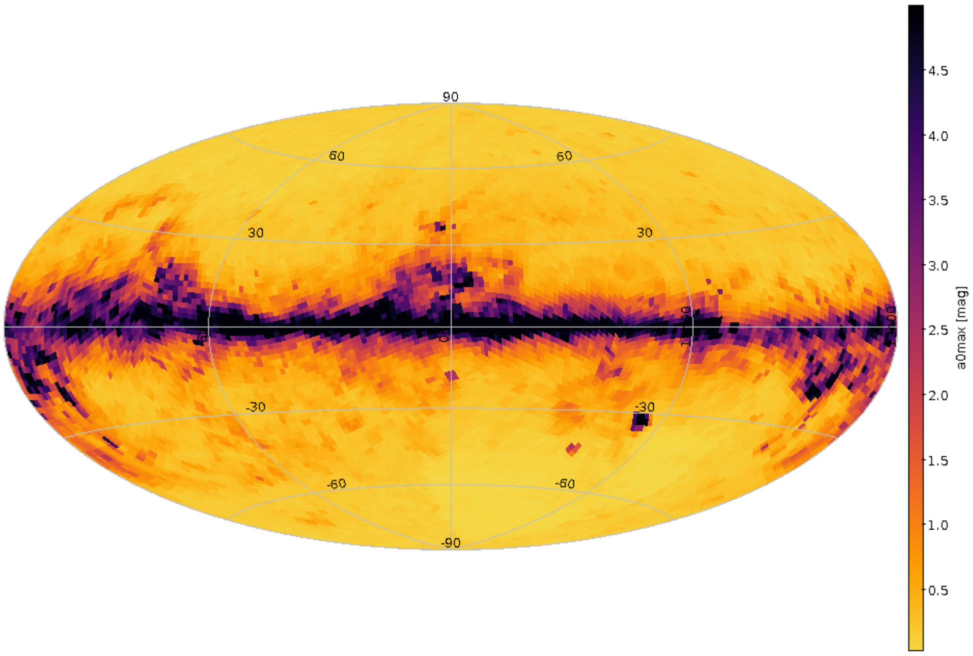

FIGURE 2: MSC maximum A_0prior in HEALpix level 5 in Galactic coordinates and aitoff projection.

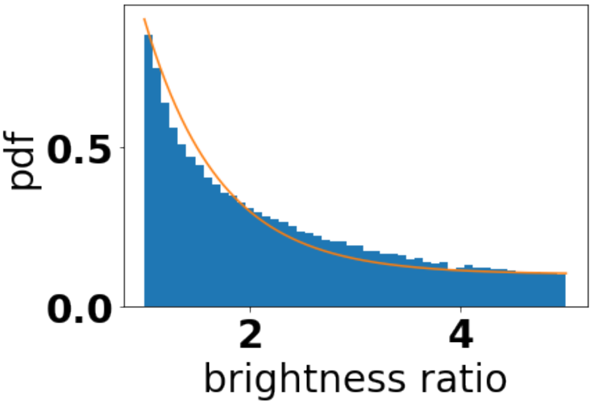

FIGURE 3: MSC fluxratio prior (solid orange line) compared with the binned probability density function (pdf) of fluxratio from the wide binary sample used as the empirical sample of reference (blue histogram).

and BPRP spectra of those sources 5 times (assuming normal distributions, cutting at physical limits, i.e negative distances, extinction and fluxes) and add all together inflating the sample by a factor of 6. We shift all fluxes to a distance of 10 pc in order to have a training set for absolute fluxes. Now we train an ExtraTree to predict BPRP spectra for specific parameter combinations of $\mathcal{T}_{\rm eff}$, $\log g$, [M/H] and A_0. We set aside a validation sample of the APOGEE sources and

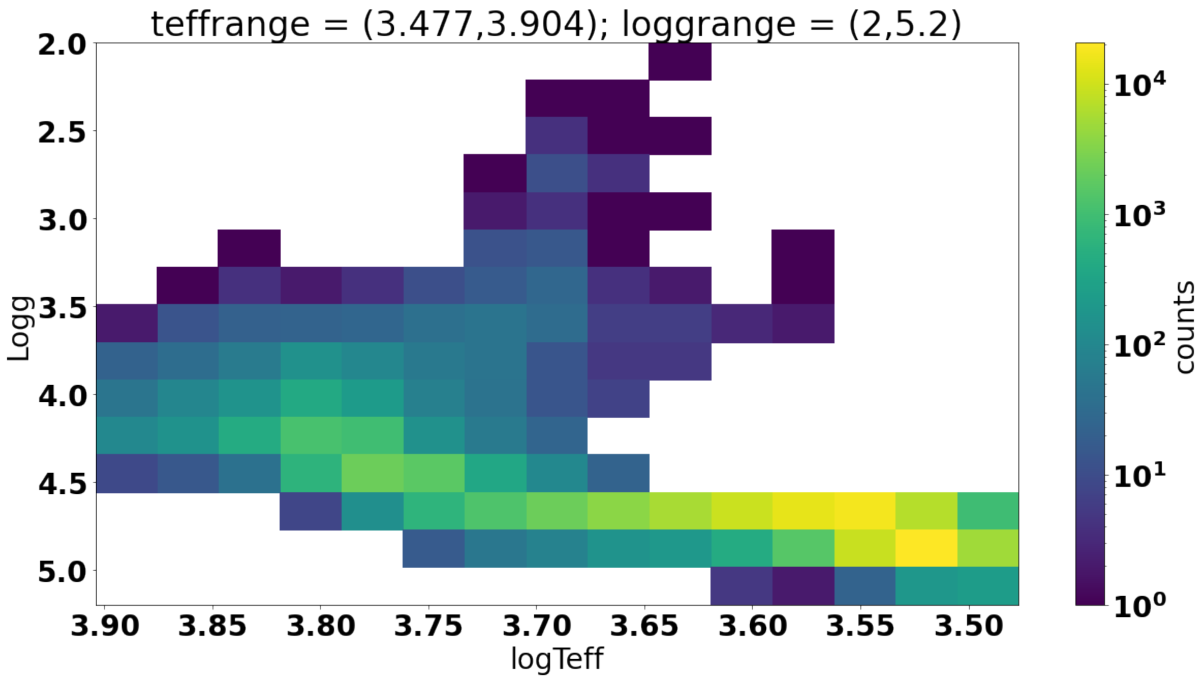

FIGURE 4: Kiel diagram for the primary component in the wide binary sample. 2d binned distribution is used as a prior for each component of the double aeneas inference. Giants are rare in the wide binary sample but we also expect MSC to not find many binaries including giants since the fluxratio limit of 5 is easily exceeded for those.

compare BPRP predictions of our empirical forward model with the PHEONIX forward model interpolation scheme used previously in MSC. The likelihood from the empirical model predictions compared to the real BPRP spectra outperforms the PHOENIX forward model and a general bias in the latter is found with too little flux in the bluest 50 pixel of the spectrum. See a representative example comparison in Figure 5.

With this ExtraTree we predict BPRP spectra for all BPRP grid points that are needed for the forward model interpolation scheme of MSC.

We used this opportunity of the forward model grid update to also change the spacing and ranges of the parameters: The old PHOENIX grid had 255,416 gridpoints (with a few holes) distributed as:

- $\log g$ range: -0.5,5.5 dex in 0.5 dex steps
- T_{eff} range: 3,000 - 10,000K in steps of 100 to 200K
- [M/H] range: -2.5 - 0.5 dex in steps of 0.5 dex
- A_0 range: 0 - 10 mag in steps of 0.1 to 0.25 dex

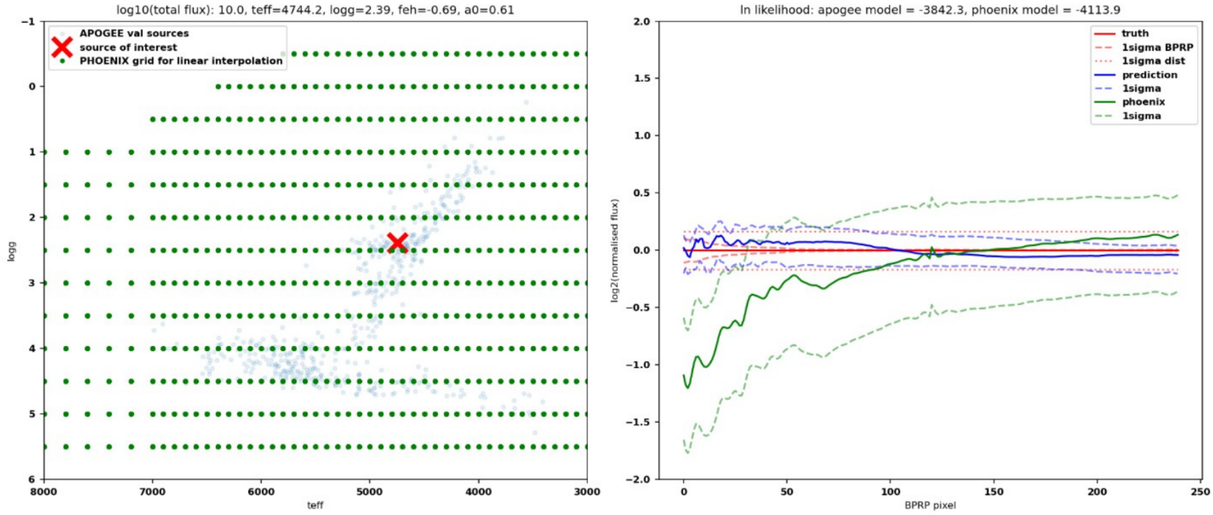

FIGURE 5: Empirical BPRP forward model ExtraTree validation showing the labels/astrophysical parameters in the left panel and the predicted feature / BPRP spectra in the right panel. The left panel shows the Kiel diagramm plane with the PHOENIX model grid in green and the APOGEE validation sources in blue (the training sample is about a factor of 100 times bigger) with the exact parameters of the source given in the title. The right panel shows the predicted BPRP pixel flux normalised to the true BPRP spectrum with the distance uncertainty and BPRP flux uncertainty shown as red dotted and red dashed line. From sampling the uncertainties in $\mathcal{T}_{\rm eff}$, $\log g$, [M/H] and A_0 we can derive 1 sigma uncertainties of the predictions of the empirical model (blue lines) and the PHOENIX forward model (green lines) as used in MSC.

The new BPRP empirical (APOGEE trained) grid has 384,054 gridpoints distributed over smaller ranges but usually with a finer grid:

- $\log g$ range: 2.0 - 5.2 dex in steps of 0.1dex
- $\mathcal{T}_{\rm eff}$ range: 3,000 - 8,000K in steps of 100 to 200K
- [M/H] grid points are: -1. -0.7 -0.5 -0.3 -0.2 -0.1 0. 0.1 0.2 0.3 0.5 dex
- A_0 grid points are: 0. 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1. 1.2 1.4 1.6 1.8 2. 2.4 2.8 3.2 3.6 4. 4.4 5. mag.

In figure 6 the training set and the parameter ranges are shown. As we can see the training data gets very sparse towards 'high' extinction (above A_0 = 3 mag). Together with the fact that APOGEE astrophysical parameters also have an uncertainty attached to them the empirical model grid should have a noisier BPRP spectrum distribution, compared to the PHOENIX model that relies on simulations. This hypothesis is supported by tests that show that for short

MCMC chains and suboptimal initialisation of the chain PHOENIX outperforms the empirical forward model especially in the 'weak' parameters ([M/H], $\log g$) and A_0. A way for improvement would be an improved training set covering larger parameter space. An alternative would be to use an empirical model for 0 extinction and apply the extinction via simulations. A further improvement could be reached by using a domain adaptation method as for example used in O'Briain et al. (2021).

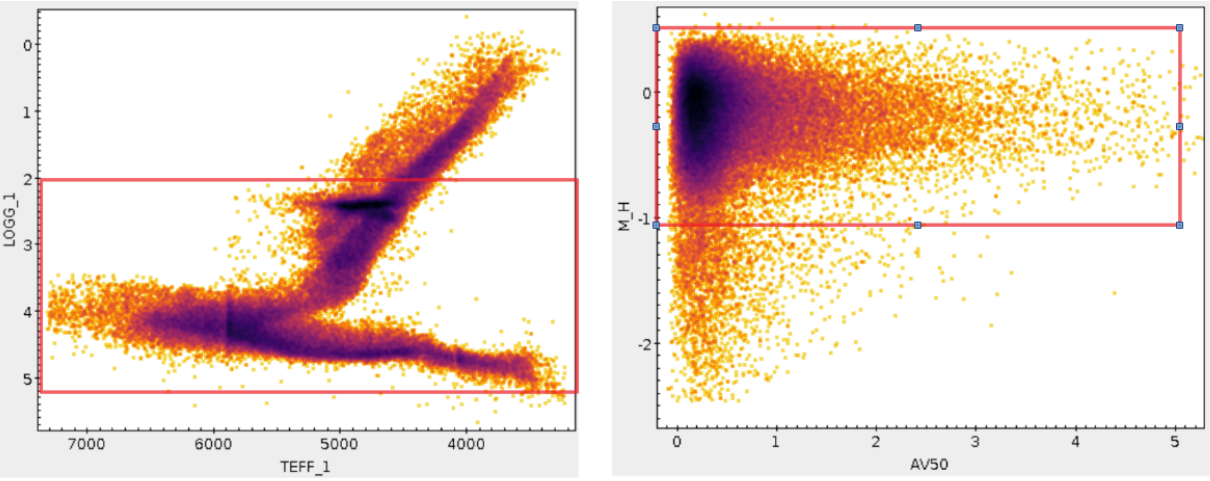

FIGURE 6: Training sample of APOGEE sources used for the empirical BPRP spectra generation. Parameter ranges of MSC are indicated in the red boxes.

5 Validation Tests & Results

5.1 Used data

For processing that led to GDR3 validation we use the following data-sets.

name	entries	description	link
MSC VDT (Validation Data Table)	12 307 146	Output of processing that led to GDR3	internal
Validation binary table	2 684	$T_{\rm eff}$, $\log g$ for both components	El-Badry et al. (20
Known binary sample	15 799	binaries with no stellar parameters	internal
ExtraTree Training	106 083	Sources used for ExtraTree training	El-Badry et al. (20
GSPPhot results	12 306 476	as seen by MSC in processing that led to GDR3	MSC VDT
DSC results	12 306 360	as seen by MSC in processing that led to GDR3	MSC VDT
GALAH binary validation	11 263	GALAH binaries with fluxratio < 5	Traven+'20

The Validation binary catalog is a literature compilation mainly consisting of APOGEE dwarfs (and we will call it 'APOGEE binary' sample) from (El-Badry et al., 2018). We added extinction values using the 3D extinction map from Rybizki et al. (2020). We also calculated luminosity ratios (in G) for each system using PARSEC isochrones with the given astrophysical parameters.

The known binary sample was compiled from the binary training catalog which consists of:

- Southworth (2015) 195 eclipsing binaries (EB)
- (Stassun & Torres, 2016) 158 EB's
- `http://caleb.eastern.edu/` 305 EB's
- El-Badry et al. (2018) 2418 spectroscopic dwarf binaries from APOGEE.

These all do have stellar parameters for individual components which is useful for validation and empirical training. We supplement these using catalogs which do not have stellar parameters for each individual component but simply list known binaries which could be helpful for validation and classification:

- Kepler Mission. VII. Eclipsing binaries in DR3 (Kirk et al., 2016) 2876 EB's
- ASAS, NSVS, and LINEAR detached eclipsing binaries (Lee, 2015) 2138 EB's
- SDSS WD main-sequence binaries (Rebassa-Mansergas et al., 2010) 1602 WDMS binaries
- 9th Catalogue of Spectroscopic Binary Orbits (Pourbaix et al., 2004) 3713 spectroscopic binaries
- Close binary systems from SDSS DR4 (Silvestri et al., 2006) 746 detached mainly WDMS binaries
- M dwarf-white dwarf binary systems (Silvestri et al., 2005) 203 WDMD binaries.
- Auxiliary binary sources used in validation: 1766 sources.

After a crossmatch with Gaia DR3 we are left with a known binary sample of 15 799 sources quoted in section 5.1. It does include the APOGEE binary sample, but not the following GALAH binary sample.

The GALAH binary validation catalogue is taken from the publication `https://ui.adsabs.harvard.edu/abs/2020A&A...638A.145T` (citing does somehow not work for the ADS BibTex entry). We only use sources with a GALAH fluxratio (ratio_1) of smaller than 5 and crop the sample to common parameter ranges. The A_0 value is determined from GALAH reported E(B-V) values. For the conversion we use an ExtraTree trained on simulated BP/RP model grid parameters.

5.2 DSC interface

VT:1 DSC interface ✔

Objectives: see how the binary probabilities can be used for MSC processing

Dataset: known binary sample, VDT

Result: Binary classification, purity high, completeness very low $\sim 5\%$

We only investigate the main CombinedProb from the DSC classifier which uses the combined results of Allosmod and Specmode and is therefore using astrometry, photometry and BPRP spectra. From Figure 7 we see that the binaries have a slightly higher probability to be classified as such than normal stars. In numbers: For all known binaries the CombinedProb is indicating a single star ($p<0.1$) for 13 097 sources[3], only classifying 372 sources as probable binaries. That's only 3 %. On the other hand those are probably quite pure because for all 12M validation sources only a fraction of 0.4 % are classified as binaries (most of those could actually be binaries, we do not know). When only looking at valid MSC sources with a fluxratio below 5 from the validation binary table, then DSC assigns a CombinedProb of greater 0.1 in 5 % of cases.

So at least a few binaries can be identified like this. But for the bulk we need an alternative selection process. Therefore we decided to simply process all sources with $G < 18.25$ mag. We will try to sort out in an offline analysis which sources are suitable for MSC and which are not. We might find out about this together with other NSS modules in GAIA when making the DR3 science verification paper on binaries.

5.3 MSC resulting parameter ranges

VT:2 ExtraTree training parameter ranges compared to predicted values ✔

Objectives: See if the predictions are within the training set

Dataset: ExtraTree Training, VDT

Result: All parameters within within marginalized training ranges

This test makes sure that the values reported by MSC are within the training sample parameter ranges. This is the case. Because we are showing the median values of the extra tree the VDT distribution is narrower and usually more centered than the training parameter distribution. As already mentioned the wide binaries sample is very local (within 1kpc) and also low extinction as can be seen in Figure 8.

VT:3 double Aeneas parameter ranges compared to ExtraTree initialisations ✔

[3] The lower overall number compared to the number stated above could be due to part of the sources not being part of the VDT.

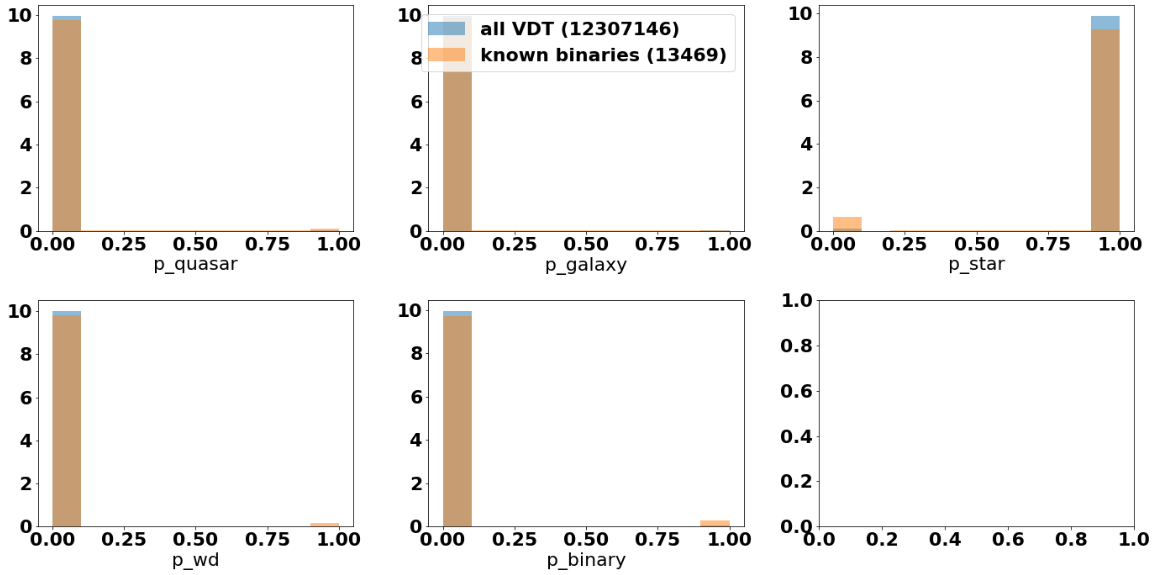

FIGURE 7: DSC probabilities for the whole VDT in blue and for the training set binaries in orange. In the legend the total number of each sample is given. We need to keep in mind that most of the binaries here are probably not applicable for MSC (e.g. because they are outside of the fluxratio range).

Objectives: See how the predictions spread after initialisation

Dataset: VDT

Result: All parameters distribute well

Here we compare the ExtraTree predicted distribution (which have been shown in figure 8), with the double aeneas result. Beware that the ExtraTree result is used as initialisation for the double Aeneas MCMC chain. As we can see from Figure 9 the double Aeneas algorithm populates a broader parameter range, as expected. The distance distribution of the ExtraTree is much smaller, due to the local training sample, and the MCMC has a larger distance range. Likewise for $\log g$, feh and A_0. In $\log g$ we see that the grid points of the forward model (BPRP simulations) are overdense (i.e. logg = 4.5, 4.6, 4.7, 4.8). We also see those peaks in the metallicity estimate, there are unphysical peaks at each grid point value (-0.5, -0.3, -0.2, 0.0, 0.1, 0.3, 0.5, with double peaks at -0.1 and +0.2). (Remember that there are no grid points +-0.4.).

5.4 Uncertainty distribution

VT:4 Look at the uncertainty distributions ✔

Objectives: Check if uncertainties make sense

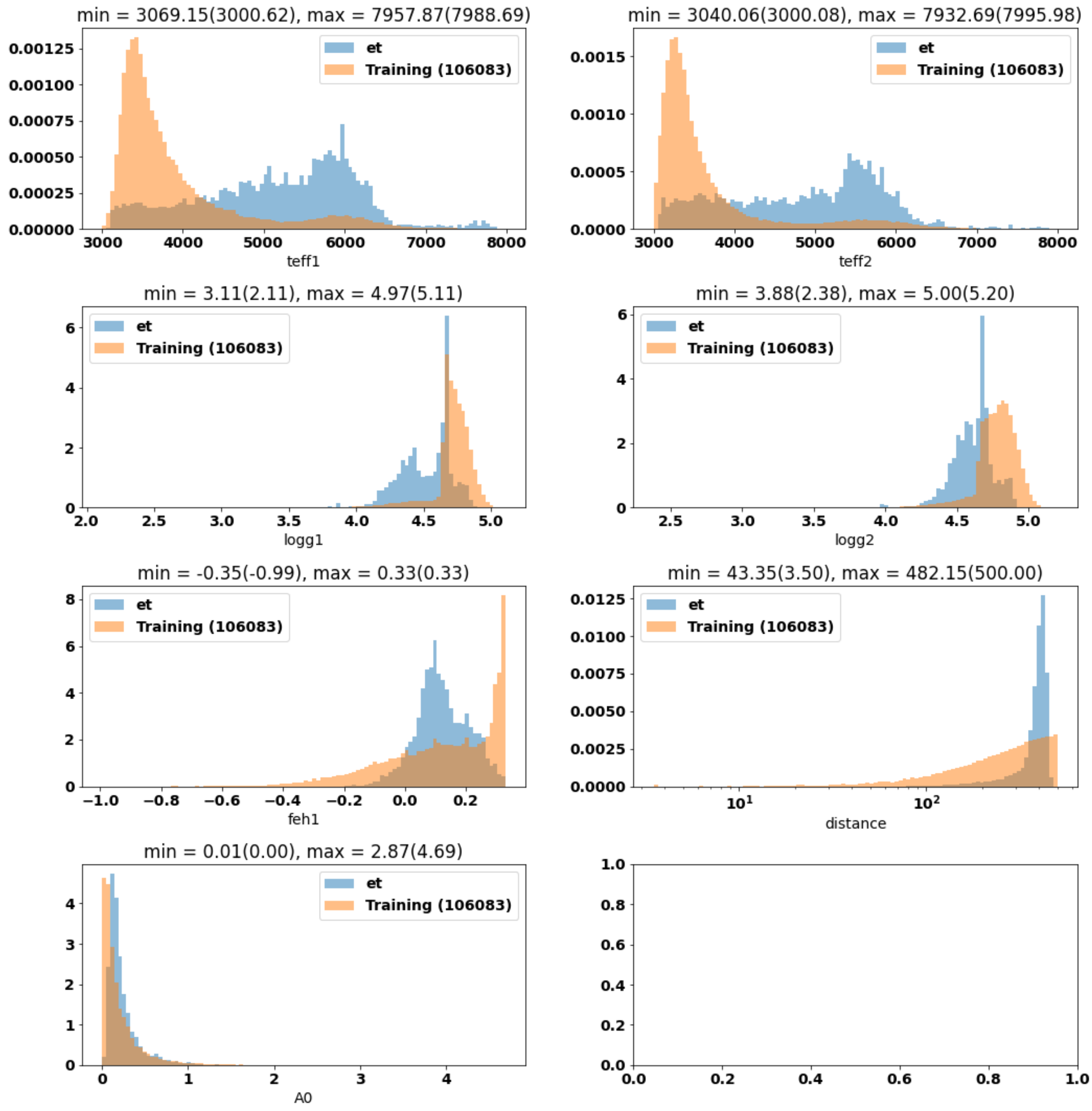

FIGURE 8: Binned distributions of stellar parameters for the whole VDT ExtraTree (et) results in blue compared to the training set orange used to train the ExtraTree model. The xlabel shows the parameter and the number behind 'training' in the label shows how many sources had a non-nan value for this parameter. The min max values are given as subplot titles with ExtraTree first and training set values in brackets. The range of training set values should always be wider and should include the VDT range.

Dataset: VDT

Result: ExtraTree uncertainties are inconsistent but not required for GDR3. Double Aeneas behaves as expected.

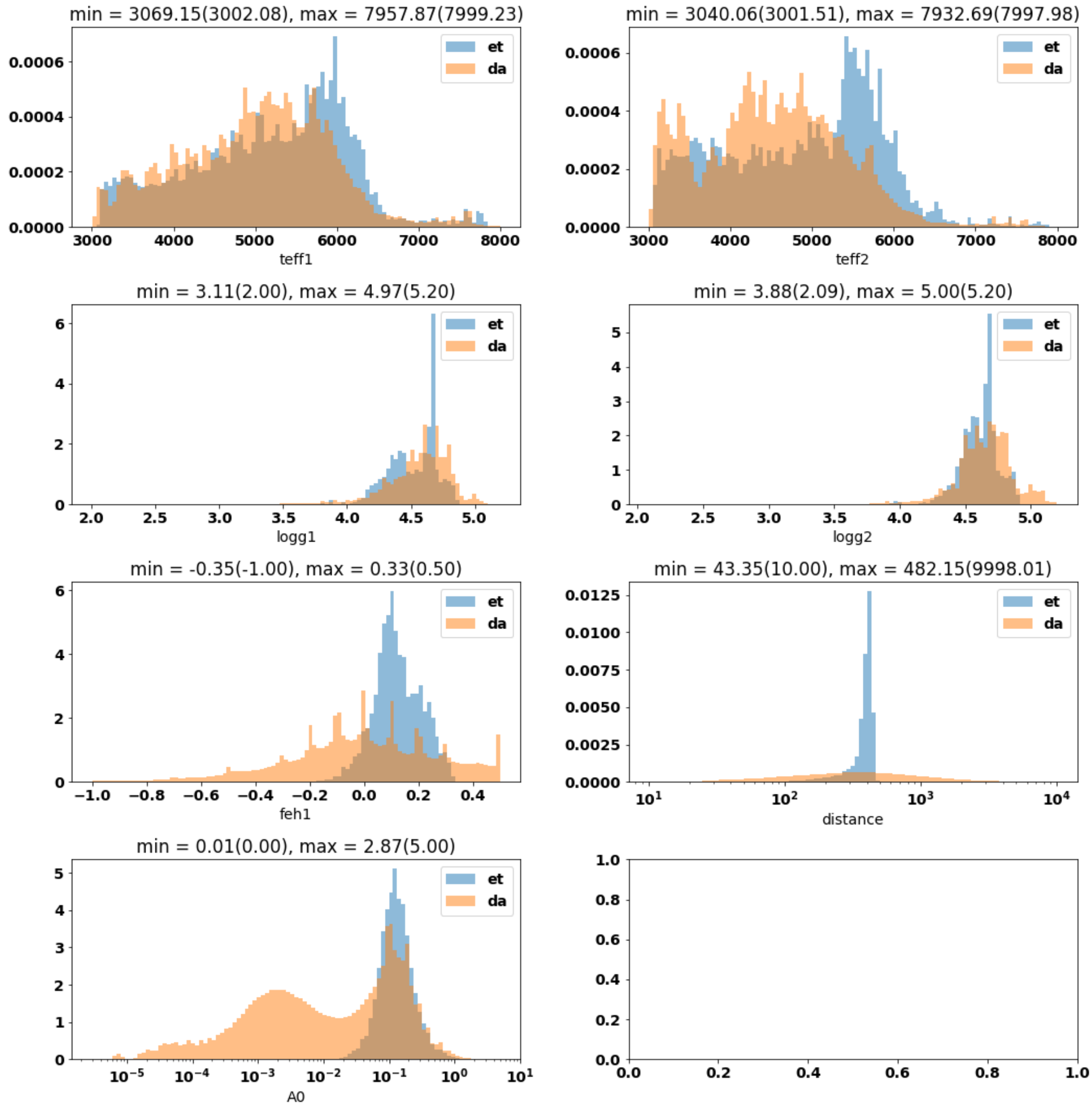

FIGURE 9: Binned distributions of stellar parameters for the whole VDT ExtraTree (et) results in blue compared to the double Aeneas (da) results. The min max values are given as subplot titles with et first and da values in brackets.

This test just looks at the confidence intervals. With a quick look one should be able to see problems in the parameter inference. In figure 10 we see that for most parameters the errors look good. The error distributions of double Aeneas are better behaved, i.e. more symmetrically distributed around zero, than for the ExtraTrees which are not always zero-centered. Only for distance we expect an asymmetric distribution in double Aeneas (due to larger errors above the median value), which is the case.

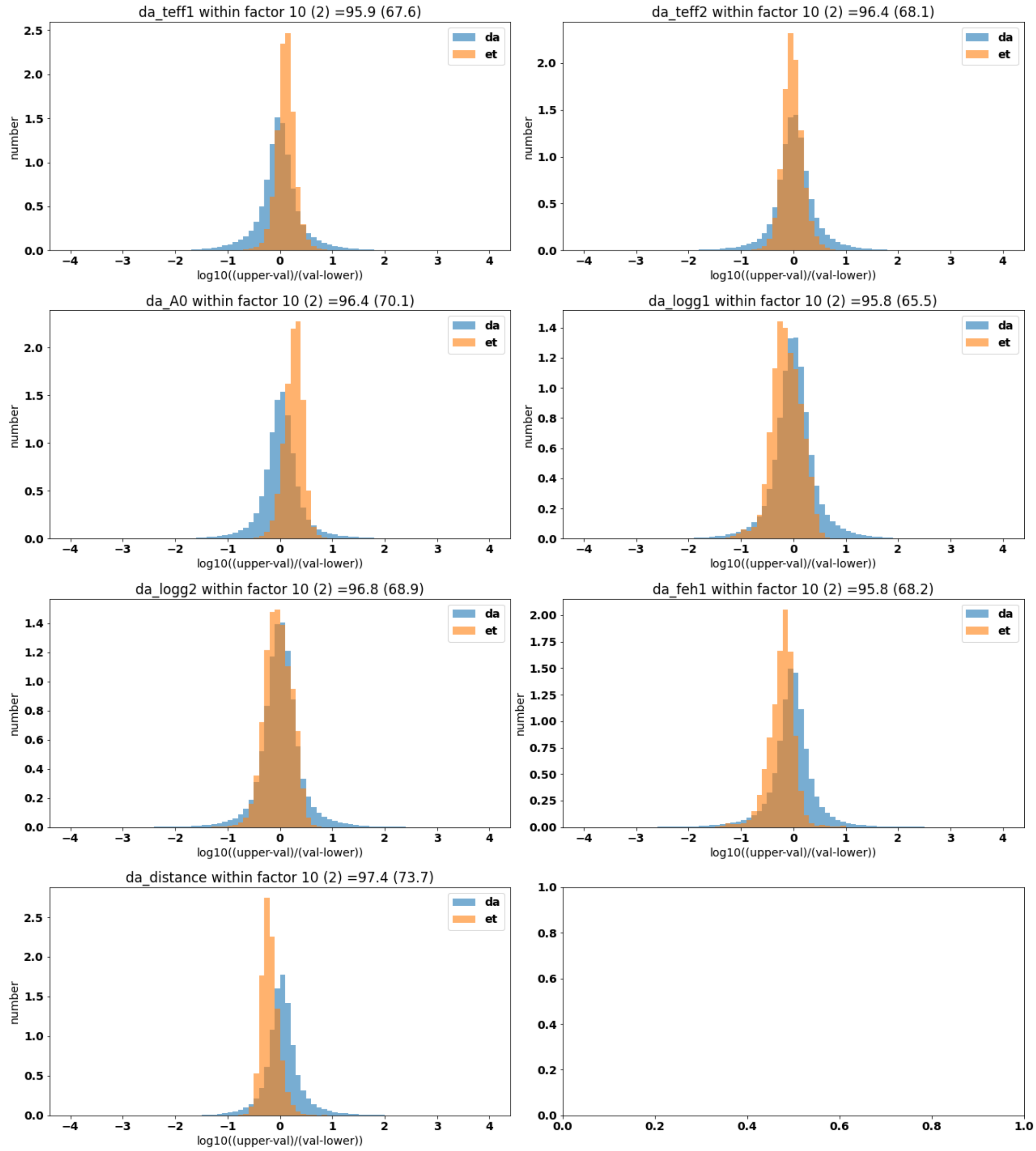

FIGURE 10: Histograms of the log10 ratios of upper-to-lower confidence intervals for double Aeneas results in blue and ExtraTree results in orange. Numbers in the subtitle show the percentage of sources where the upper-to-lower ratio of confidence intervals is within a factor of 10 (2) for double Aeneas.

5.5 Parameter value validation (APOGEE)

IVT:5 Check results with independent validation data ✔

Objectives: Look at bias and median absolute deviation of the results

Dataset: validation binary table as defined in Sec. 5.1, VDT

Result: Results on independent validation sample, where we can extract applicable MSC sources are good enough for publication

Here we compare how MSC inferred parameters perform on a completely independent set of validation sources. Bear in mind that the validation set mainly contains APOGEE binary dwarfs classified using high resolution spectroscopy and eclipsing binaries for which BPRP spectra might be corrupted depending on the observational epochs and whether they coincide with eclipses or not. The parameter ranges spanned are cropped to the common validity range and only sources with a luminosity range of lower than 5 are retained. This results in 1 370 validation sources.

The comparison is shown in figure 11. $\mathcal{T}_{\mathrm{eff}\,1}$ has a good correlation with the validation. The bias is -152K and the median absolute error (mae) is 135K. For $\mathcal{T}_{\mathrm{eff}\,2}$ the bias is -314K and the mae increases to 236K. For Teff1 and Teff2, the values most deviating from the 1-to-1 line have low MSC inferred values and low gof values. $\log g_1$ and $\log g_2$ have a bias of 0.09 and 0.08, respectively. The respective mae is 0.14 and 0.13. The metallicity has a bias of 0.11 dex and an mae of 0.14 dex. For logg and metallicity the MSC values seem to be biased towards solar values. There seem to be low distance estimates for a few sources with high inverse parallax. These do have a low mean log posterior value. A few of those actually have low astrometric fidelities, i.e. might be corrupted astrometric solutions. It could also be that our ExtraTree inferred distances result in low initial values and the MCMC does not find a way out of this. The extinctions look OK. But high A_0 values are usually not inferred correctly and the mean log posterior for higher extinctions is usually quite low, even if it is near the extinction prior value. This value was deduced from a 3D extinction map at the corresponding inverse parallax distance and should only be used as an order of magnitude check.

5.6 Parameter value validation (GALAH)

IVT:6 Check results with independent validation data ✔

Objectives: Look at bias and median absolute deviation of the results

Dataset: GALAH binary validation table as defined in Sec. 5.1, VDT

Result: Results on independent validation sample, where we can extract applicable MSC sources are good enough for publication

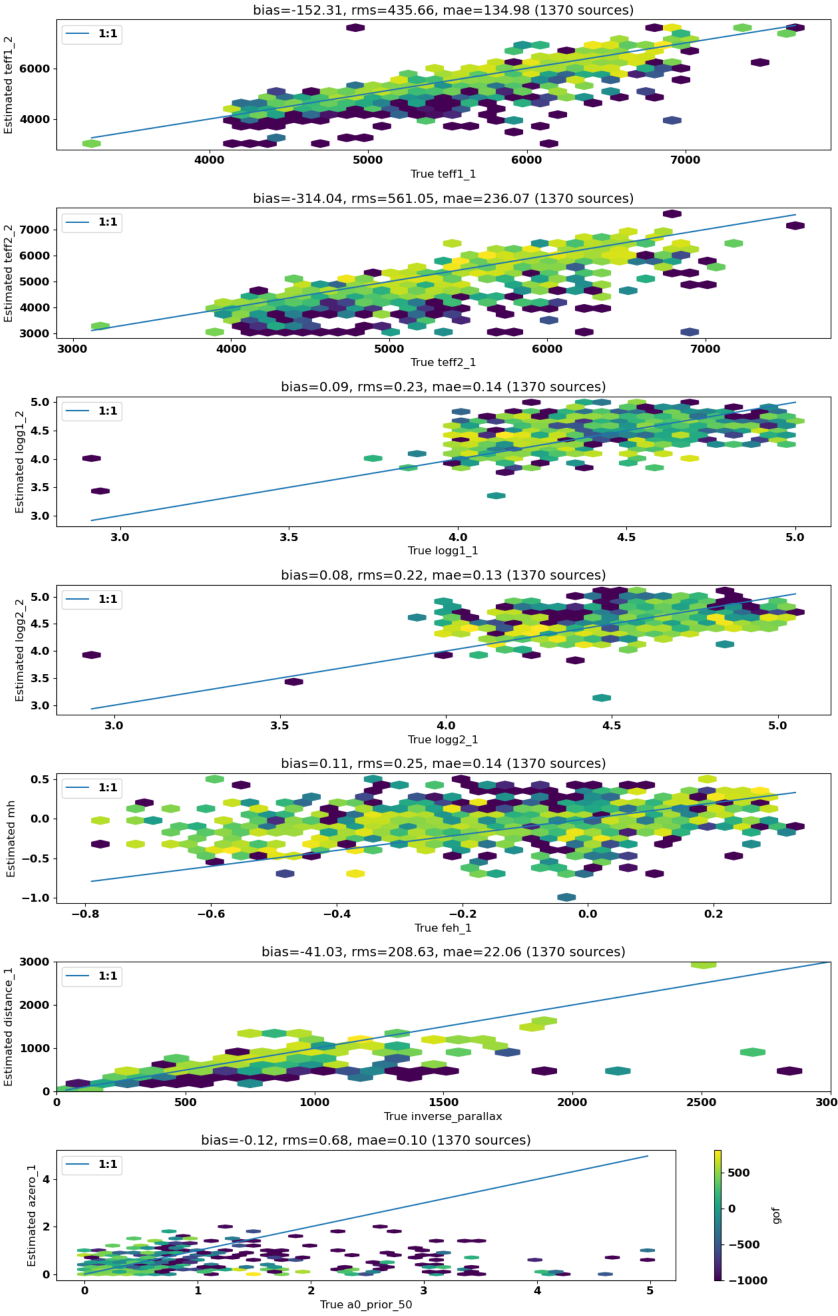

FIGURE 11: MSC inferred values on the y-axis vs. APOGEE literature values on the x-axis for sources with common parameter range (including fluxratio smaller 5). Shown are the 7 MCMC sampled parameters and the respective 1:1 line. Color-coded is the average mean log posterior per hexbin, 'gof', with the colorscale given in the bottom panel. The colorscale is truncated at -1000 but sources with lower values are still included in the plot.

For this test we use Galah binaries with a Galah based fluxratio of smaller than 5 and also sort the primary and secondary component according to the flux as in MSC. We crop the parameter range to a common minimum. This results in 11 263 validation sources.

The comparison is shown in figure 12. $\mathcal{T}_{\mathrm{eff}\,1}$ has a good correlation with the validation. The bias is -139K and the median absolute error (mae) is 165K. For $\mathcal{T}_{\mathrm{eff}\,2}$ the bias is -418K and the mae increases to 393K. $\log g_1$ and $\log g_2$ have a bias of 0.24 and 0.35, respectively. The respective mae is 0.23 and 0.31. The metallicity has a bias of 0.21 dex and an mae of 0.20 dex. The same as for the APOGEE sample can be seen here. In addition, the most deviating logg and metallicity values have also low gof values. There seem to be low distance estimates for a few sources with high inverse parallax. These do have a low mean log posterior value. A few of those actually have low astrometric fidelities, i.e. might be corrupted astrometric solutions. It could also be that our ExtraTree inferred distances result in low initial values and the MCMC does not find a way out of this. These wrongly inferred distances might correspond to the badly fitted areas in other parameters as well (too high $\log g$, though this would exactly be the wrong trend as lower $\log g$ values would increase the luminosity of the object, which might be the reason for the bad mean log posterior).

The extinctions look OK with a bias of -0.01 and a mae of 0.12 but these values are dominated by nearby and low-extinction sources. For high literature A_0 values the mean log posterior is usually quite low.

5.7 Mean log posterior cuts

VT:7 check if the mean log posterior of MSC can be used to find reliable results ✔

Objectives: look at mean log posterior distribution

Dataset: GALAH and APOGEE binary validation table as defined in Sec. 5.1, VDT

Result: mean log posterior is a good measure of fit quality

First we look at the distribution of mean log posteriors, which we use interchangeably with gof, for the APOGEE and the GALAH validation sample.

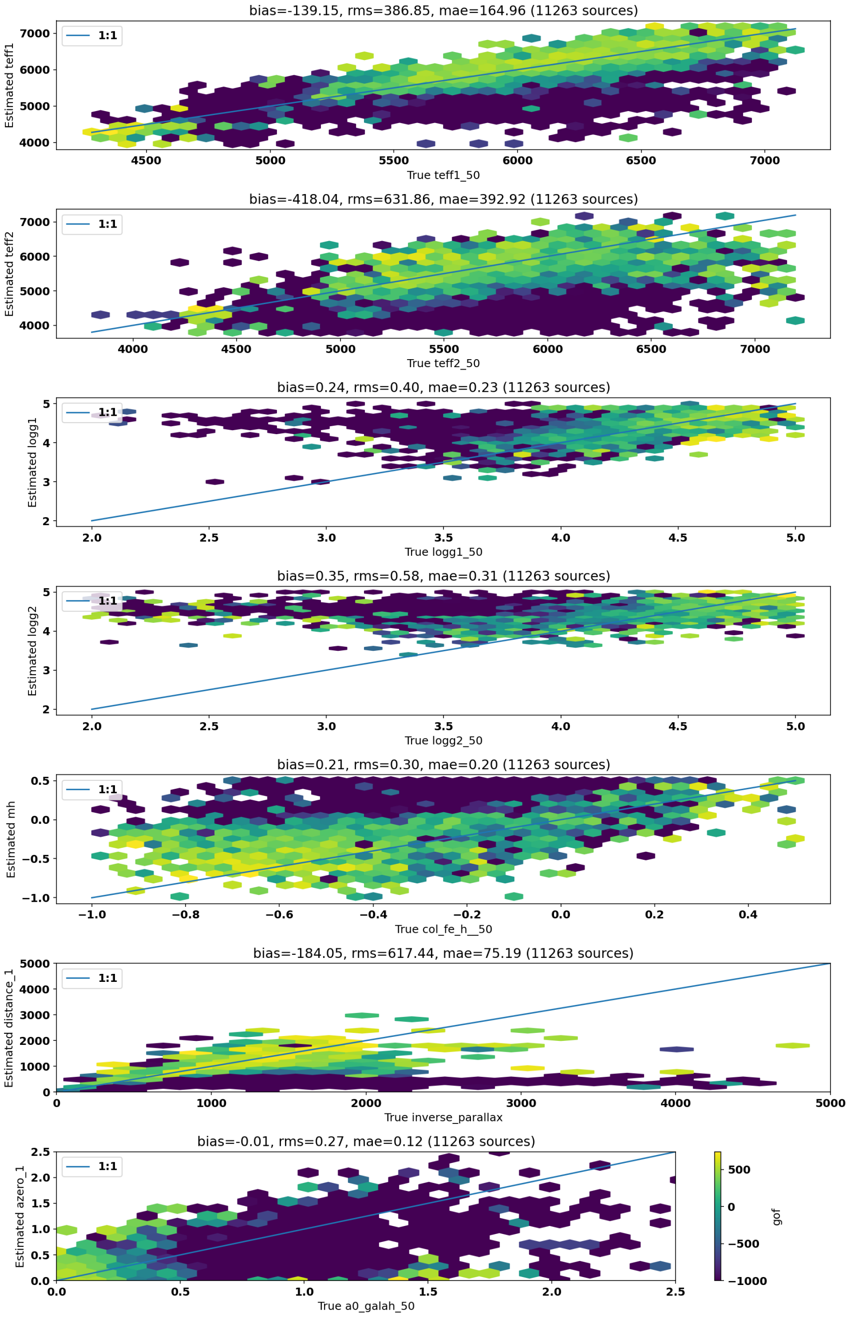

FIGURE 12: MSC inferred values on the y-axis vs. GALAH literature values on the x-axis for sources with common parameter range (including fluxratio smaller 5). Shown are the 7 MCMC sampled parameters and the respective 1:1 line. Color-coded is the average mean log posterior per hexbin, 'gof', with the colorscale given in the bottom panel. The colorscale is truncated at -1000 but sources with lower values are still included in the plot.

cut-off percentile	mean log posterior		sourcecount	
val data	GALAH	APOGEE	GALAH	APOGEE
0 (min)	-201 147	-36 211	11 263	1 370
5 (-2σ)	-4 437	-2 322	10 699	1 301
16 (-1σ)	-596	-226	9 461	1 149
50 (median)	488	511	5 637	686
84 (+1σ)	689	684	1 814	221
95 (+2σ)	752	758	567	69
100 (max)	887	885	1	1

As shown in the above table the tail of very negative (very poor fits) gof values goes quite negative with worst 5% of the sources being below -4437 and -2322, for GALAH and APOGEE, respectively. But around half of both validation source samples have better gof values than ~ 500. Converging to a similar best gof value of ~ 885.

The next table shows how the bias of the validation samples evolve when successively sorting out the poor gof sources.

	bias					
gof cut off percentile	0	5	16	50	84	95
teff1 APOGEE	-152	-124	-74	-17	2	7
teff1 GALAH	-139	-118	-72	-6	21	10
teff2 APOGEE	-314	-282	-237	-143	-76	-33
teff2 GALAH	-418	-392	-350	-245	-144	-60
logg1 APOGEE	0.09	0.08	0.09	0.09	0.07	0.06
logg1 GALAH	0.24	0.22	0.20	0.17	0.13	0.12
logg2 APOGEE	0.08	0.07	0.07	0.07	0.06	0.05
logg2 GALAH	0.35	0.33	0.30	0.24	0.17	0.15
[M/H] APOGEE	0.11	0.11	0.11	0.10	0.09	0.08
[M/H] GALAH	0.21	0.20	0.19	0.19	0.19	0.18
distance APOGEE	-41	-33	-24	-5	-2	-2
distance GALAH	-184	-148	-95	-49	-16	-9
A_0 bias APOGEE	-0.12	-0.08	-0.01	0.04	0.04	0.03
A_0 bias GALAH	-0.01	0.00	0.01	0.02	0.01	0.01

As we see, the bias generally reduces significantly when increasing the gof threshold, almost vanishing for a few parameters (like teff1, teff2, distance and A_0) when only looking at the 5% best fit sources. A general systematic deviation with respect to both validation samples seem to be that MSC overpredicts both $\log g$ and the [M/H] values. Physically that makes sense as a higher metallicity increases the luminosity which can be counterbalanced by a higher surface gravity which reduces the luminosity. It might be that our metallicity prior, a Gauss with

N(0,0.2), induces this offset. Still the results look very encouraging with the best 50% of both samples already having reasonably low biases. For the differences between both validation samples we need to recall that both use independent methods in inferring their parameters and on independent wavelength ranges (GALAH in the optical and APOGEE in the infrared). Additionally the samples probe different parameter regimes with the GALAH sample covering a larger volume.

The following table shows the root mean squared error (rms). The same decrease of rms with increased gof can be seen, with GALAH heaving slightly worse values throughout, except for A_0 but APOGEE values on this parameter are only values from a 3D extinction map using the inverse parallax as distance.

rms

gof cut off percentile	0	5	16	50	84	95
teff1 APOGEE	436	366	279	187	168	126
teff1 GALAH	387	348	273	192	144	135
teff2 APOGEE	561	506	435	309	217	192
teff2 GALAH	632	592	536	417	310	258
logg1 APOGEE	0.23	0.22	0.22	0.20	0.17	0.13
logg1 GALAH	0.40	0.35	0.33	0.29	0.25	0.24
logg2 APOGEE	0.22	0.21	0.21	0.19	0.17	0.14
logg2 GALAH	0.58	0.54	0.50	0.45	0.38	0.36
[M/H] APOGEE	0.25	0.25	0.23	0.19	0.15	0.12
[M/H] GALAH	0.30	0.29	0.27	0.24	0.22	0.21
distance APOGEE	209	184	168	52	22	14
distance GALAH	617	553	277	152	47	25
A_0 APOGEE	0.68	0.61	0.42	0.17	0.15	0.21
A_0 GALAH	0.27	0.24	0.21	0.19	0.15	0.13

The last table shows the median absolute error (mae) which again improves with gof cut-off, with slight advantages for APOGEE validation sources.

mae

gof cut off percentile	0	5	16	50	84	95
teff1 APOGEE	135	130	113	84	70	59
teff1 GALAH	165	158	143	116	92	82
teff2 APOGEE	236	222	197	84	70	59
teff2 GALAH	393	378	350	256	179	156
logg1 APOGEE	0.14	0.14	0.14	0.12	0.10	0.11
logg1 GALAH	0.23	0.22	0.21	0.18	0.15	0.14
logg2 APOGEE	0.13	0.13	0.13	0.12	0.09	0.09
logg2 GALAH	0.31	0.30	0.28	0.23	0.18	0.15
[M/H] APOGEE	0.14	0.14	0.13	0.11	0.09	0.09
[M/H] GALAH	0.20	0.20	0.19	0.18	0.18	0.17
distance APOGEE	22	21	18	12	5	5
distance GALAH	75	69	58	34	13	8
A_0 bias APOGEE	0.10	0.10	0.08	0.07	0.05	0.05
A_0 bias GALAH	0.12	0.11	0.11	0.10	0.08	0.07

In essence the mean log posterior is a very valuable indicator of reliability of the MSC inferred parameters. We flag results that have a worse than -1000 gof with '1' in the postprocessing to indicate non-reliability. A cut in gof may exclude high extinction results, as well as low $\log g$ systems disproportionally.

5.8 Mean log posterior over observables and inferred parameters

VT:8 check how the mean log posterior of MSC is affected by position in observable space and where it gets projected in inferred parameters space ✔

Objectives: look at mean log posterior distribution with different parameters

Dataset: GALAH and APOGEE binary validation table as defined in Sec. 5.1, VDT

Result: mean log posterior depends on CAMD (Colour Absolute Magnitude Diagram) and sky position

In Figure 13 we see that MSC produces good fits for sources near the binary sequence. For sources on the giant branch the mean log posterior deteriorates significantly with only few exceptions.

In Figure 14 we see the gof value with sky position. A pattern emerges that suggests, that high-extincted areas have worse gofs than out-of-plane sky areas.

In Figure 15 we see that almost all high extinction and high metallicity sources have a bad gof. The working hypothesis is that due to the sparse sampling of the high extinction region the

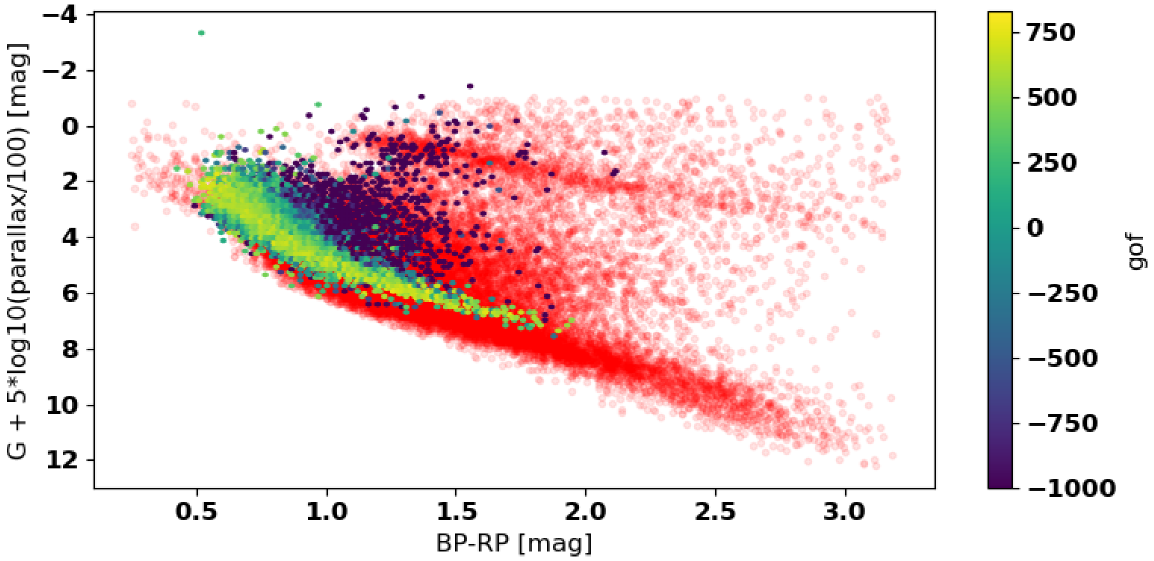

FIGURE 13: CAMD of GALAH sources color coded by MSC gof value. In red a random subset of GaiaEDR3 sources is shown to illustrate the usual CAMD positions

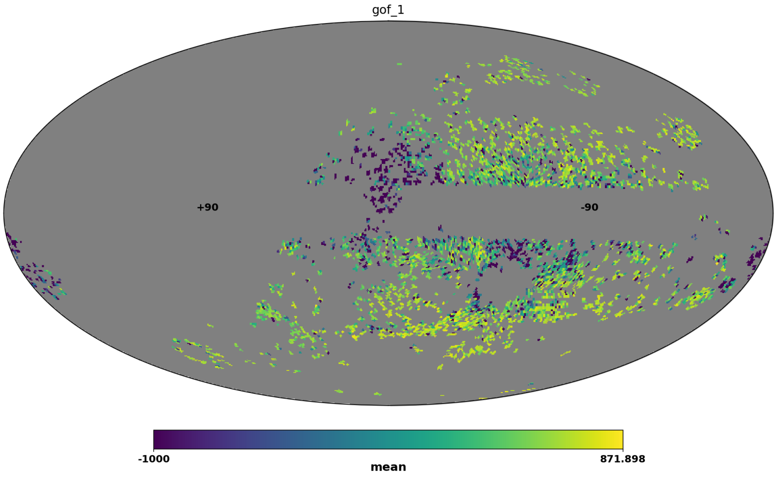

FIGURE 14: Sky of GALAH sources color coded by MSC gof value, projected in mollweide and galactic coordinates with longitude increasing to the left.

empirical BPRP model does produce bad fits in that area. For the high metallicity target the reason for bad gof values might be that the mcmc converges towards high metallicity values in order to reach the elevated CAMD position that would normally be fit by a giant star. But giants seem to be strongly disfavoured by MSC inference.

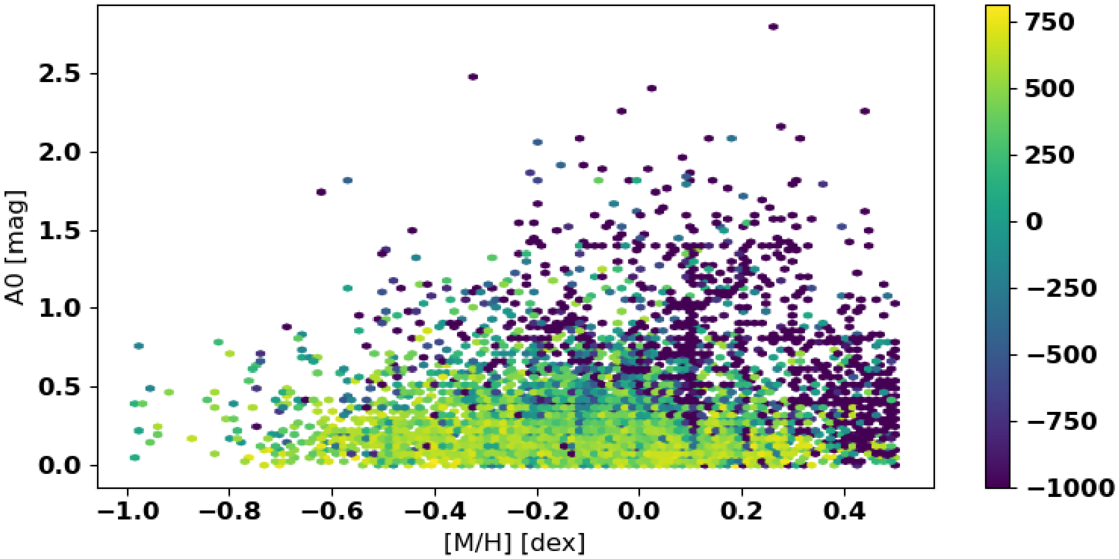

FIGURE 15: MSC inferred extinction and metallicity values color-coded by gof.

In Figure 16 we see that most poorly fit sources end up with a teff1 of around 5 000 K and slightly elevated logg1 value. As already discussed poor fits are expected for high-extinction sources and for giants.

For APOGEE sources the picture is the same. It might be interesting to see if the gof is different for genuine binary sources in comparison to truly single sources.

5.9 crossvalidation of validation samples

VT:9 Check if APOGEE and GALAH estimates agree with each other ✘

Objectives: Assess reliability of the validation data

Dataset: GALAH binary validation table as defined in Sec. 5.1, APOGEE binary validation table

Result: The parameters do not agree particularly well with each other.

Both catalogues only have 26 sources in common. While GALAH report 16,50 and 84 percentiles of their inferred values APOGEE does not report an error estimate. We plot the comparison in Figure 17. As can be seen from the reported statistics, the biases and rms values are really high. The reported uncertainties of GALAH are not compatible with the 1:1 line. Especially for the metallicity one would have hoped that the derived quantities from both studies are

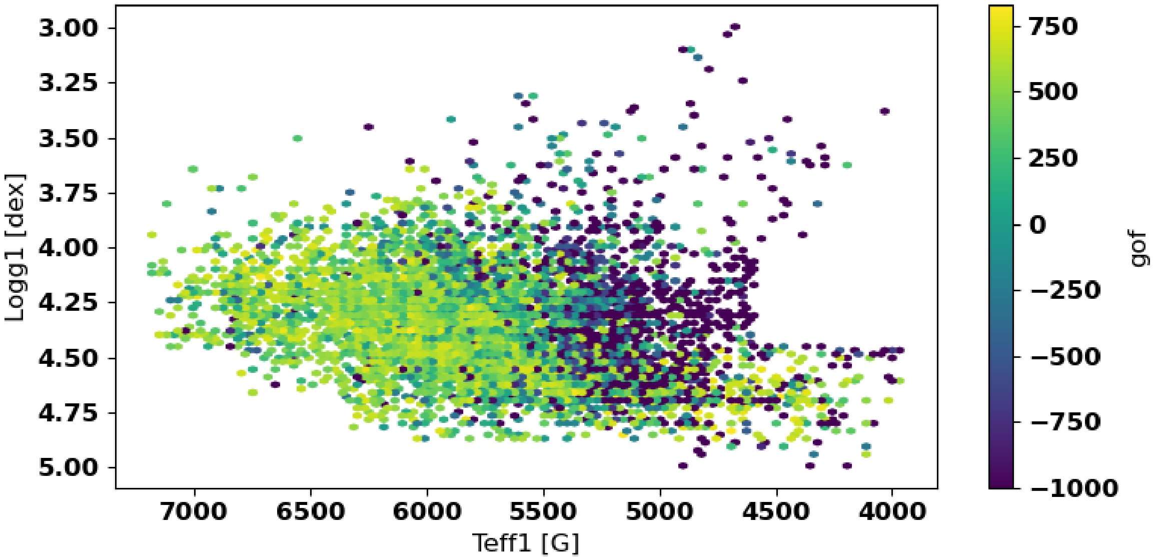

FIGURE 16: MSC inferred Kiel diagramm color-coded by gof.

similar. But this is not the case indicating systematic differences in the modelling that have a strong impact on the derived parameters. Since these are the only two catalogs that have astrophysical parameter values for both components of binary systems it is hard to assess which one is right.

5.10 Fluxratio

VT:10 Check MSC estimated fluxratios ✔

Objectives: See if the fluxratios make sense

Dataset: GALAH binary validation table as defined in Sec. 5.1, VDT

Result: The fluxratio is too uncertain to be compared directly, but internally it makes sense and also in comparison with the CAMD position

In Figure 18 we show the GALAH sources in the CAMD color-coded by the inferred fluxratio value. We can see a trend with CAMD position that the equal-mass binary sequence tends to have low fluxratios as well, increasing for sources on the single star main-sequence.

The same applies for APOGEE sources as seen in Figure 19

On the other hand, when directly comparing GALAH fluxratios (which are not tied to the G magnitude fluxratio as it is for MSC) we do not see a good correlation. The reason ist most

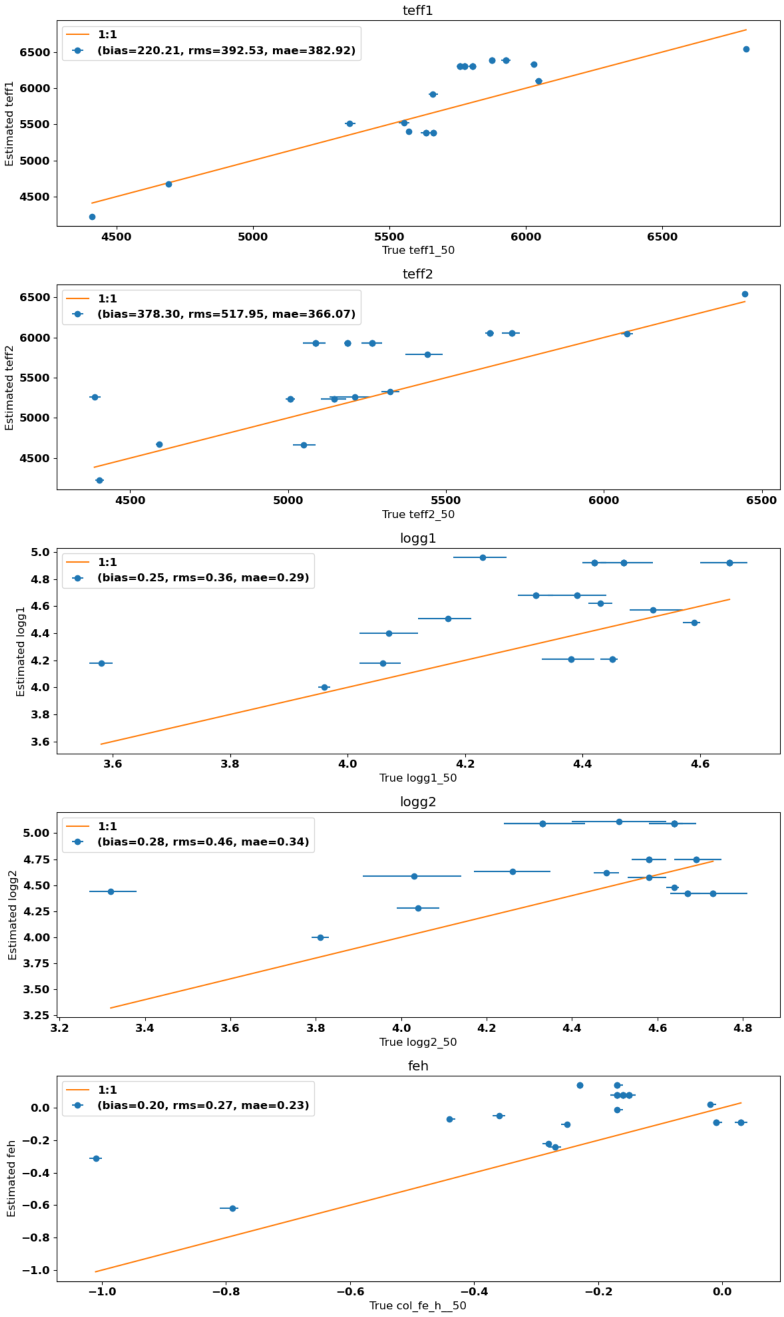

FIGURE 17: APOGEE inferred values on the y-axis vs. GALAH inferred values on the x-axis for 26 sources both samples have in common. Shown are both $\mathcal{T}_{\rm eff}$ both $\log g$ and the $[{\rm Fe/H}]$ comparisons and the respective 1:1 line. Only GALAH reported uncertainties which are shown as horizontal errorbars.

probably that the fluxratio is a very sensitive parameter to the stellar parameters of the individual components which are only weakly constrained, especially for the secondary component. We

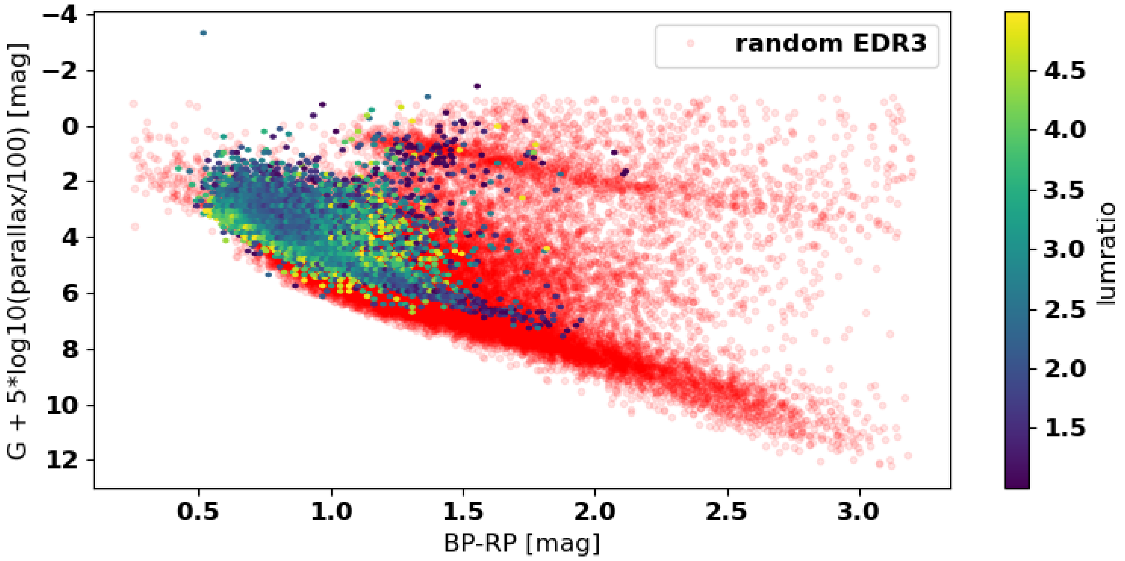

FIGURE 18: CAMD of GALAH binaries color coded by fluxratio value. In red a random subset of GaiaEDR3 with high parallax SNR is shown to guide the eye on stellar population loci.

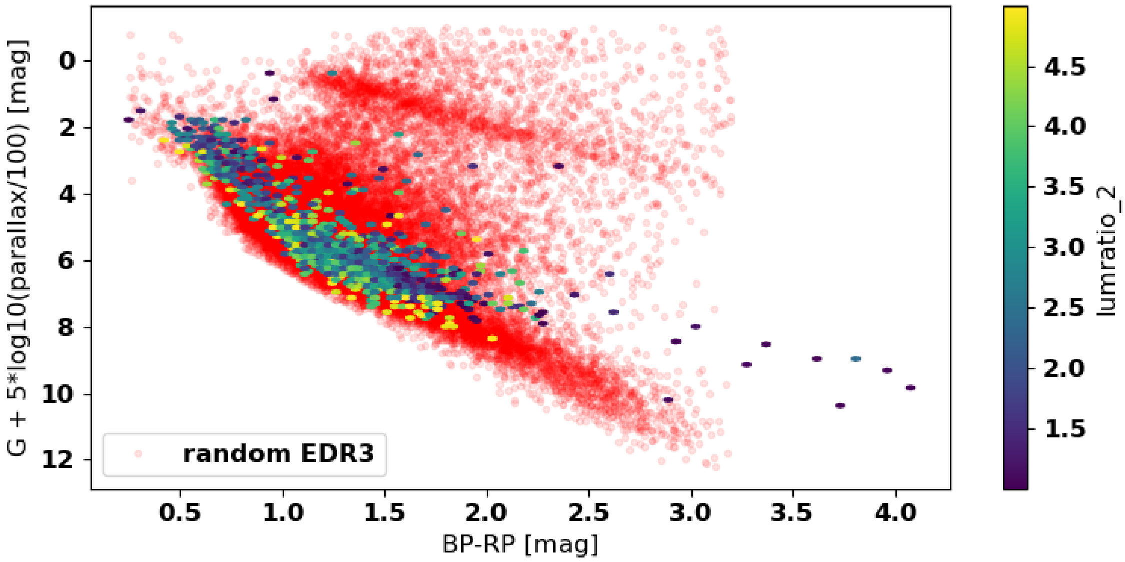

FIGURE 19: CAMD of APOGEE binaries color coded by fluxratio value. In red a random subset of GaiaEDR3 with high parallax SNR is shown to guide the eye on stellar population loci.

also checked that when using the MSC inferred stellar parameters and using these together with parsec isochrones to derive flux values that this is close to the MSC inferred fluxratio values.

5.11 Error inflation

VT:11 Check MSC estimated uncertainty values ✔

Objectives: define an error inflation for postprocessing

Dataset: GALAH binary validation table as defined in Sec. 5.1, VDT

Result: The uncertainty estimated by MSC is roughly a factor of 10 too low and is corrected for in the postprocessing

In order to assess the reported uncertainty of MSC, we compare to GALAH values. MSC reports upper and lower confidence intervals using the 16th and the 84th percentile of the parameter distribution in the MCMC chain. The reported values are much too low which is probably due to local minima in our posterior landscape induced by the empirical forward model BPRP spectral grid. When inflating the uncertainties we use fixed error inflation factors for all parameters for simplicity. We change the upper and lower values according to the following calculations: new_upper = ((upper - median) * error_inflation) + median. If new_upper then exceeds the upper limit of the parameter range it will be set to that upper limit instead. Likewise for the lower confidence interval.

In the left panels of Figure 20 the histogram of bias corrected differences between GALAH and MSC estimates are shown with their respective rms values in the legend. On the right hand side the same is shown normalised by the MSC inflated uncertainty added in quadrature to the GALAH uncertainty and compared to a Gaussian. We show this for an inflation factor of 10 and for the 50% MSC sources with the best gof. In the title of the right hand panels the 16,50 and 84 percentiles of the added uncertainties are shown. Our objective was to get the median relatively close to the rms value on the left. We see that the bias corrected differences histogram in the left panels of Figure 20 compare well with the Gaussian distribution except for logg2 which is slightly skewed and for A_0 which is more peaked than the Gaussian. For the right hand-side the uncertainty normalised version of the bias corrected differences are not fit well by a Gaussian. They are usually more peaked and have longer tails of strongly underestimated errors, even though we are applying an error inflation of 10 throughout this section. When cutting at higher gof values the right hand panels look closer to a Gaussian but still retain a few sources in the long tails.

In the following tables we are giving the rms values of the bias corrected differences and the median of the in quadrature added uncertainties (in which MSC has already been error inflated) for different cut-off values of the gof.

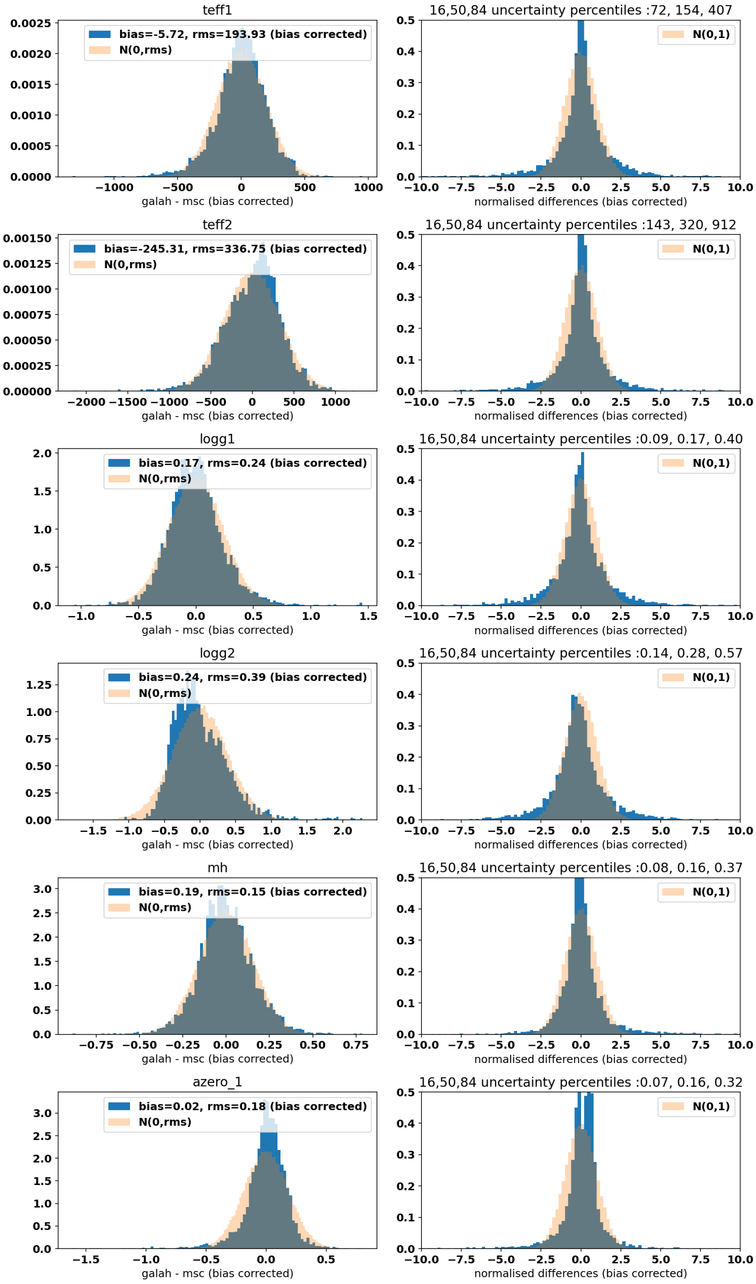

FIGURE 20: Left panels are the histograms of the biased corrected differences between galah and msc results, i.e. the mean is 0. For comparison a Gaussian with the same standard deviation as the rms values is shown. The title of the left panel show for which parameter that row is plotted. The right panels show the same bias corrected differences but normalised with the errors of GALAH and MSC added in quadrature. For comparison a Gaussian with mean 0 and standard deviation of 1 is shown in orange. In the title the 1sigma and median percentiles of the symmetrised MSC and galah added uncertainties are given. We were aiming to get the median close to the rms value in the left panel. For teff1 for examples rms of the bias corrected differences is 194K and the median uncertainty value is 154K. The MSC uncertainty has been inflated by a factor of 10. We are only showing the sources with the 50% best gof values.

inflated error for different gof cut-offs

gof cut-off percentile	0	5	16	50	84	95
teff1 rms	361	328	264	194	142	135
teff1 uncertainty	147	148	149	154	137	129
teff2 rms	474	443	405	337	274	250
teff2 uncertainty	314	317	318	320	277	244
logg1 rms	0.32	0.28	0.26	0.24	0.21	0.21
logg1 uncertainty	0.16	0.16	0.17	0.17	0.16	0.14
logg2 rms	0.46	0.43	0.41	0.39	0.33	0.32
logg2 uncertainty	0.27	0.28	0.28	0.28	0.25	0.22
[M/H] rms	0.22	0.20	0.19	0.15	0.12	0.11
[M/H] uncertainty	0.16	0.16	0.16	0.16	0.13	0.11
A_0 rms	0.27	0.24	0.21	0.18	0.15	0.13
A_0 uncertainty	0.16	0.16	0.16	0.16	0.14	0.12

We see that for samples only including the best 50% as in Figure 20 the numbers agree reasonably well, except for the $\log g$ values which are slightly underestimated by the uncertainties. When including all sources then the MSC uncertainties do not capture the true uncertainties even when inflating the error by a factor of 10. From Figure 17 we know that GALAH uncertainties are probably also underestimated.

For the postprocessing we inflate the error by a factor of 10 which should capture the true uncertainties reasonably well for the well-fit part of the MSC sample. We did a similar test for the APOGEE sample and found the inflation of 10 to fit that sample uncertainties quite good as well.

5.12 GSP-Phot comparison

VT:12 Compare to GSP-Phot results ✔

Objectives: Look if general trends in parameters agree

Dataset: validation binary table, VDT MSC, VDT GSP-Phot

Result: general agreement in trends

Here we want to assess how MSC and GSP-Phot inferred parameters compare to each other. On the one hand we want them to correlate somewhat. On the other hand we want to see that for known binary systems MSC results differ from GSP-Phot parameters and that especially stellar parameters which are shared among the two components, namely metallicity, extinction and distance, have different and unbiased results in MSC.

First we only look at the bulk structure for all VDT sources in common between MSC and GSP-

Phot (irrespective whether they are binary or single star systems, on the order of 12M sources) while only comparing to the primary component's stellar parameters. As seen in Figure 21 the parameters generally correlate. For $\log g$ and [M/H] the forward model grids are strongly visible in both MSC and GSP-Phot, weaker grid features are also visible in $\mathcal{T}_{\rm eff}$. In extinction we see that GSP-Phot assigns 0 to a large number of sources, which have a broader range of extinctions in MSC. Generally MSC assigns less giant gravities than GSP-Phot, cf. Fig. 24. For the metallicities the grid features are quite strong, but there seems to be a class of sources for which feh correlates for both algorithms, albeit not on the 1:1 line.

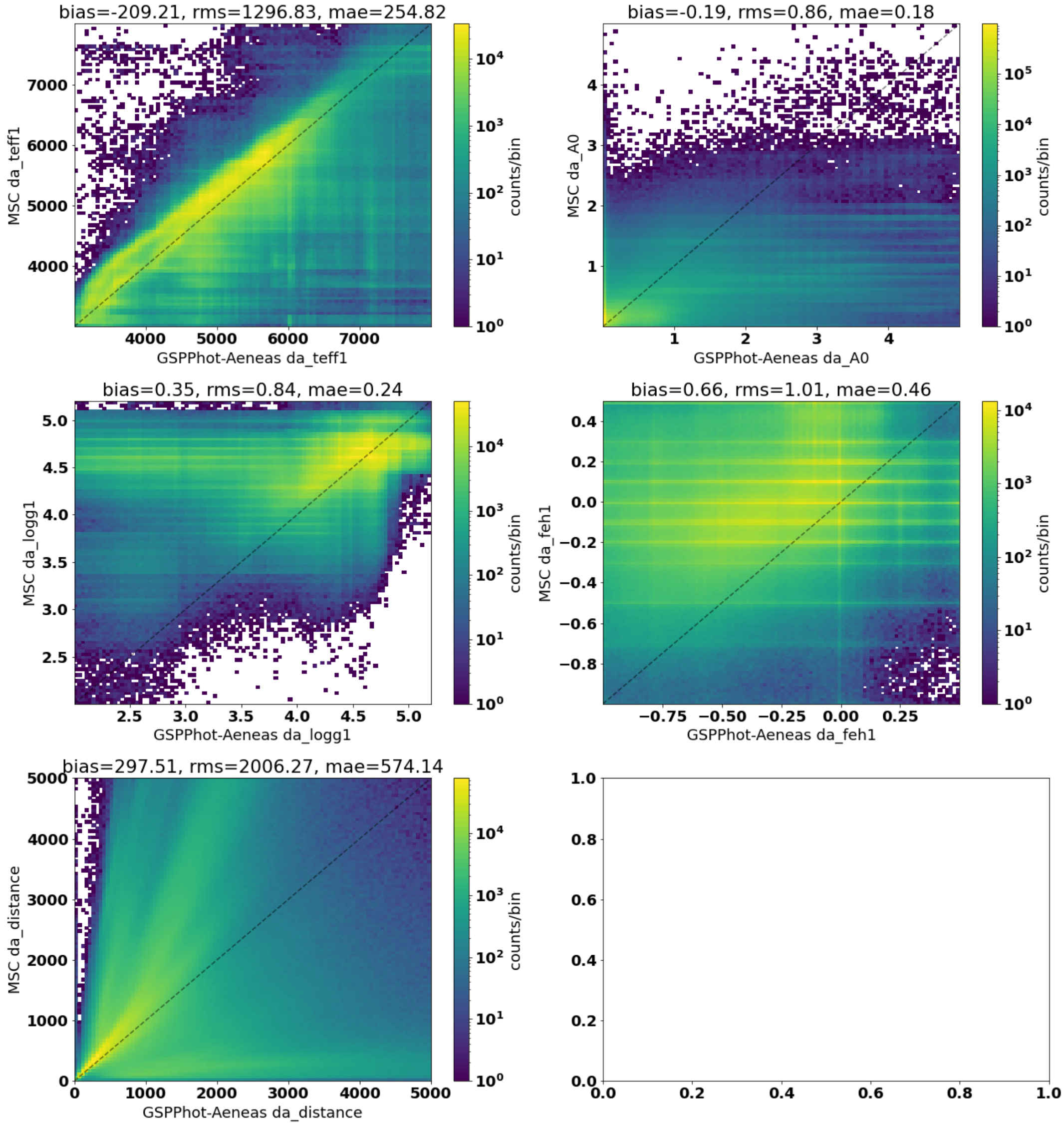

FIGURE 21: 2d histograms of all Validation sources in common of MSC and GSP-Phot with logarithmic color scheme. The bias, rms and mae are written in the title of each subpanel. The grey dashed line marks the 1:1 relation. This plot is not intended for quantitative analysis but rather for qualitative agreement.

VT:13 Compare GSP-Phot results for validation sample and assess bias with fluxratio ✔

Objectives: How much better does MSC perform on binaries and what are the GSP-Phot biases

Dataset: GALAH and APOGEE sample, MSC and GSP-Phot results

Result: MSC slightly better on binaries, the biases of GSP-Phot seem to result from the one-component assumption it relies on

Here we compare for the best 84% gof sources of MSC results the performance of MSC and GSP-Phot with our two validation sets. First the comparison with GALAH is shown in Figure 22. The left hand side shows MSC results and the right hand side the respective GSP-Phot results for the different parameters. GSP-Phot is actually performing better than MSC in teff1 and A_0. For logg1 and distances, GSP-Phot struggles to find the true values because the sources are nominally too bright for a single component fit. Therefore the distances are over and the surface gravities under-estimated. For metallicity both modules have a significant bias, each in another direction but they both do correlate with literature metallicity.

For the same comparison with the APOGEE sample, shown in Figure 23, the picture changes slightly. Now MSC does a better job on all parameters. Do not trust A_0 values from APOGEE samples though, because these are just values from the 3D extinction map. Again the biases of GSP-Phot go into the right direction with $\log g$ being too low and distances too high.

In order to get an overview how that changes with luminosity ratio we bin the sample in luminosity ratios as derived by MSC in the following tables.

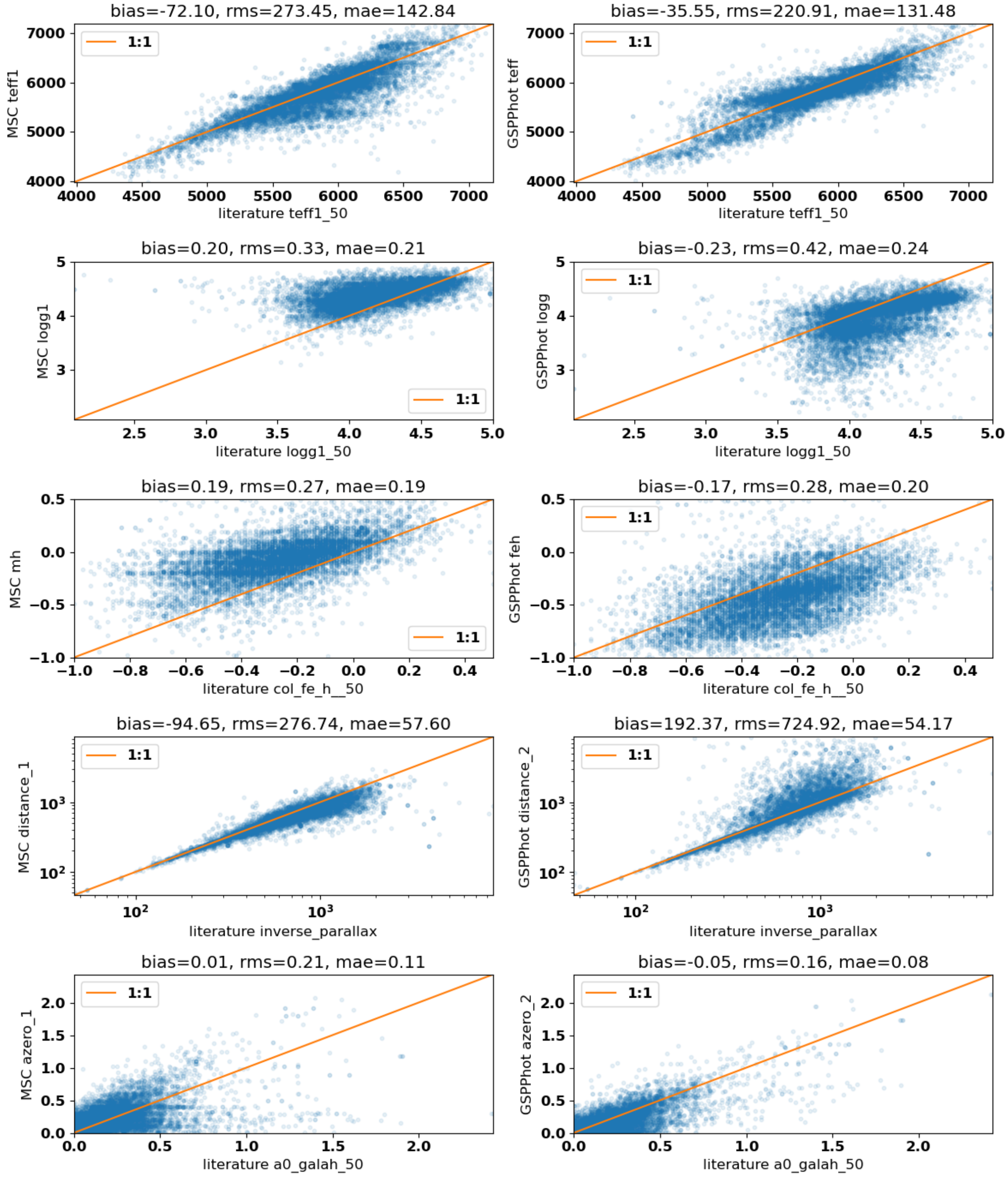

FIGURE 22: Comparison of MSC (left) and GSP-Phot (right) parameters with galah binary parameters for the 84% best MSC gof sources. Statistics of the comparison are given in the titles and the 1:1 relation is shown in orange.

First we look at the biases in comparison to GALAH:

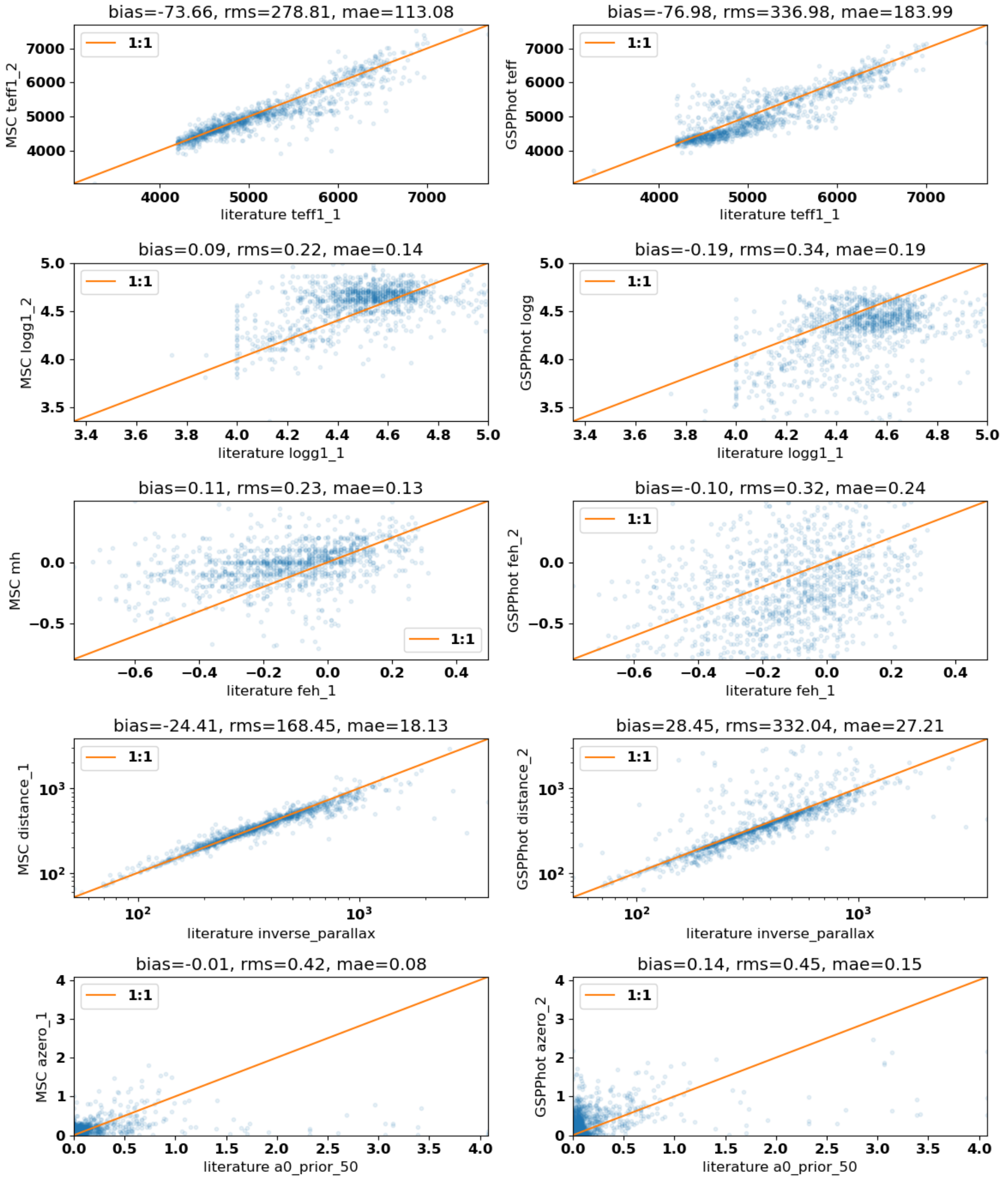

FIGURE 23: Comparison of MSC (left) and GSP-Phot (right) parameters with apogee binary parameters for the 84% best MSC gof sources. Statistics of the comparison are given in the titles and the 1:1 relation is shown in orange.

bias for GALAH, 84% best MSC gof values

sourcecount	3069	2688	2672	1032
fluxratio bins	5-3	3-2	2-1.3	1.3-1
teff1 MSC	-44	-40	-89	-196
teff1 GSP-Phot	-5	-56	-50	-66
logg1 MSC	0.21	0.20	0.21	0.19
logg1 GSP-Phot	-0.17	-0.24	-0.23	-0.25
[M/H] MSC	0.21	0.18	0.18	0.19
[M/H] GSP-Phot	-0.17	-0.19	-0.16	-0.11
distance MSC	-87	-89	-105	-101
distance GSP-Phot	143	215	119	134

Biases for MSC stay basically the same for all luminosity ratio bins except for teff1 where it increases quite strongly for equal brightness binaries. For GSP-Phot the biases are quite low for the 5-3 ratio bin and usually reach a plateau for lower luminosity ratios. Metallicity keeps increasing though for increasing fluxratio in order to increase the luminosity. Also the GSP-Phot distances do not follow the trend, but they could be dominated by extreme outliers

rms for GALAH, 84% best MSC gof values

sourcecount	3069	2688	2672	1032
fluxratio bins	5-3	3-2	2-1.3	1.3-1
teff1 MSC	254	238	281	376
teff1 GSP-Phot	235	225	200	218
logg1 MSC	0.33	0.32	0.34	0.34
logg1 GSP-Phot	0.38	0.43	0.40	0.42
[M/H] MSC	0.29	0.26	0.26	0.27
[M/H] GSP-Phot	0.28	0.28	0.27	0.30
distance MSC	269	279	254	341
distance GSP-Phot	535	729	442	638

For the rms values MSC basically stays constant over the complete luminosity range, though having a high distance uncertainty for the equal brightness binaries and also a very high teff1 uncertainty. GSP-Phot rms value stays relatively constant over the luminosity ratio.

For the APOGEE sample the picture changes slightly:

bias for APOGEE, 84% best MSC gof values

sourcecount	443	266	260	180
fluxratio bins	5-3	3-2	2-1.3	1.3-1
teff1 MSC	-56	-61	-90	-112
teff1 GSP-Phot	-42	-123	-123	-30
logg1 MSC	0.11	0.07	0.07	0.08
logg1 GSP-Phot	-0.10	-0.15	-0.23	-0.14
[M/H] MSC	0.14	0.12	0.08	0.07
[M/H] GSP-Phot	-0.16	-0.14	-0.10	0.12
distance MSC	-44	-14	-19	1
distance GSP-Phot	6	0	27	20

For the effective temperature, the negative bias increases for MSC with decreasing fluxratio. For GSP-Phot on the other hand the temperature bias is lowest for high (secondary does not play a role) and low fluxratios (both components do have the same temperature). For surface gravity the MSC bias is constant and the negative GSP-Phot bias decreases again (the bias gets worse) for the intermediate fluxratio bins . For metallicity the positive MSC bias decreases with fluxratio. For GSP-Phot the negative bias gets less negative with decreasing fluxratio and entirely swapping signs for equal brightness binaries. For the distances, the negative MSC bias shrinks with fluxratio, while for GSP-Phot the bias is higher for lower fluxratios.

rms for APOGEE, 84% best MSC gof values

sourcecount	443	266	260	180
fluxratio bins	5-3	3-2	2-1.3	1.3-1
teff1 MSC	276	288	290	256
teff1 GSP-Phot	406	314	289	231
logg1 MSC	0.25	0.21	0.20	0.18
logg1 GSP-Phot	0.25	0.26	0.35	0.21
[M/H] MSC	0.26	0.24	0.19	0.19
[M/H] GSP-Phot	0.34	0.29	0.29	0.31
distance MSC	252	81	83	72
distance GSP-Phot	383	144	194	223

For MSC the rms values usually decrease for lower fluxratios. For GSP-Phot this is only the case for teff1. For for logg the 2-1.3 fluxratio bin has the highest value and the equal-brightness bin the lowest. Metallicity is more or less constant and distance has low rms values for the intermediate fluxratio bins.

Overall some of the biases are physically understandable, but are very dependent on the validation sample. We also need to mention that MSC inferred fluxratio does not show a strong correlation with e.g the luminosity ratio from the GALAH sample. This is due to the strong effect of the luminosity ratio to slight parameter changes in teff and logg of the individual components. We checked that fluxratio is internally consistent with the given stellar parameters, though.

5.13 Physical properties of individual components

VT:14 Teff logg plane ✔

Objectives: Look if the Kiel diagram is well populated

Dataset: VDT, APOGEE Validation and training set, VDT GSP-Phot

Result: Kiel diagram source distribution looks physical, while MSC and GSP-Phot employ different stellar models.

Here we want to see the consistency of $\mathcal{T}_{\rm eff}$ and $\log g$ inferred values. In Figure 24 we show the Kiel diagram for the primary in the left panel, for the secondary in the middle panel and for

GSP-Phot in the right panel. The color distributions are the VDT results for MSC or GSP-Phot. In red we show the ExtraTree training set and in black our validation sample. MSC distribution has an imprint of the Kiel diagram prior which forces the MCMC to stay within the borders (implementation in JAVA is a 2d binned histogram, hard-coded). We also see grid features in the distribution. But the bulk of sources seem to populate the relevant parameter space. GSP-Phot also shows prior features but different to MSC (their priors differ). Both modules suppress predictions below the main-sequence.

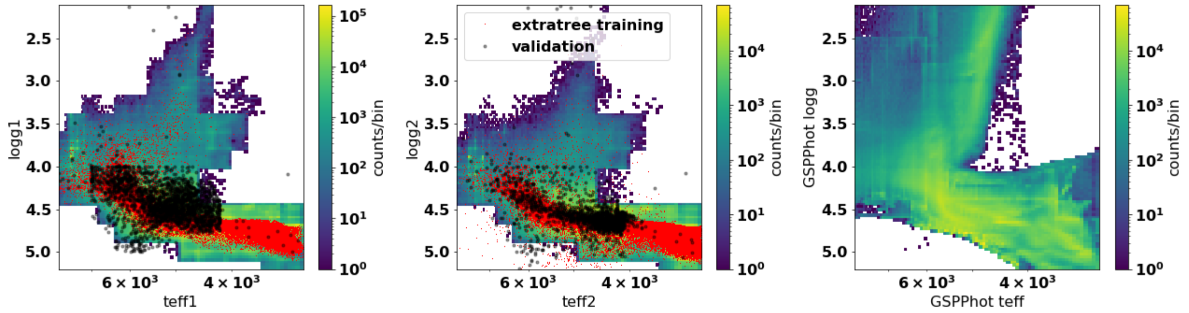

FIGURE 24: 2D histograms (in log10 density) of the teff, logg plane for the primary, secondary and GSPPhot results from left to right. In red the data used for the ExtraTree training is shown and black illustrates the validation data.

VT:15 Teff A0 plane ✔

Objectives: Check physical integrity of results

Dataset: VDT, APOGEE Validation and training set, VDT GSP-Phot

Result: Distribution looks good

Similarly we would like to investigate the teff-extinction degeneracy. The plots in Figure 25 are not particularly helpful in that respect. It is interesting to see the different distributions MSC and GSP-Phot are producing.

6 Conclusions

- MSC runs. Double Aeneas inference works well and is using normalised BPRP spectra, total BPRP flux and parallax information.
- MSC results on applicable binaries are good especially for sources with high gof.
- a strategy is still needed to separate applicable binary sources (fluxratio below 5) from the rest. For DR3 we simply processed all sources with G $<$ 18.25 mag.

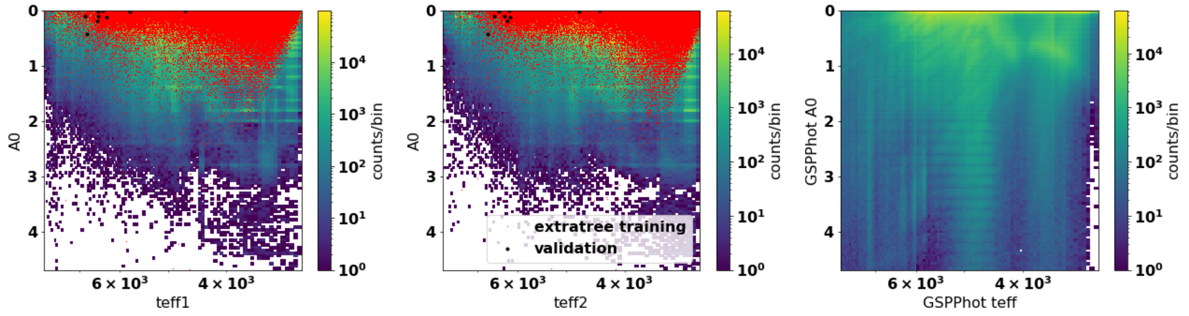

FIGURE 25: 2D histograms (in log10 density) of the $T_{\rm eff}$, A_0 plane for the primary, secondary and GSP-Phot results from left to right. In red the data used for the ExtraTree training is shown and black illustrates the validation data, which has very few entries for A_0.

References

Bailer-Jones, C.A.L., Rybizki, J., Fouesneau, M., Mantelet, G., Andrae, R., 2018, AJ, 156, 58, ADS Link

El-Badry, K., Ting, Y.S., Rix, H.W., et al., 2018, MNRAS, 476, 528, ADS Link

Jönsson, H., Holtzman, J.A., Allende Prieto, C., et al., 2020, AJ, 160, 120, ADS Link

Kirk, B., Conroy, K., Prša, A., et al., 2016, AJ, 151, 68, ADS Link

Lee, C.H., 2015, MNRAS, 453, 3474, ADS Link

O'Briain, T., Ting, Y.S., Fabbro, S., et al., 2021, ApJ, 906, 130, ADS Link

Pourbaix, D., Tokovinin, A.A., Batten, A.H., et al., 2004, A&A, 424, 727, ADS Link

Queiroz, A.B.A., Anders, F., Chiappini, C., et al., 2020, A&A, 638, A76

Rebassa-Mansergas, A., Gänsicke, B.T., Schreiber, M.R., Koester, D., Rodríguez-Gil, P., 2010, MNRAS, 402, 620, ADS Link

Rybizki, J., Demleitner, M., Bailer-Jones, C., et al., 2020, PASP, 132, 074501, ADS Link

Silvestri, N.M., Hawley, S.L., Oswalt, T.D., 2005, AJ, 129, 2428, ADS Link

Silvestri, N.M., Hawley, S.L., West, A.A., et al., 2006, AJ, 131, 1674, ADS Link

Southworth, J., 2015, In: Rucinski, S.M., Torres, G., Zejda, M. (eds.) Living Together: Planets, Host Stars and Binaries, vol. 496 of Astronomical Society of the Pacific Conference Series, 164, ADS Link

Stassun, K.G., Torres, G., 2016, AJ, 152, 180, ADS Link