# Gaia Data Release 3: Analysis of the Gaia BP/RP spectra using the General Stellar Parameterizer from Photometry

R. Andrae, M. Fouesneau, R. Sordo, C. A. L. Bailer-Jones, T. E. Dharmawardena, J. Rybizki, F. de Angeli, H. E. P. Lindstrøm, D. J. Marshall, R. Drimmel, et al.

*Astronomy & Astrophysics* manuscript no. AA_2022_43462
June 14, 2022




# Gaia Data Release 3: Analysis of the Gaia BP/RP spectra using the General Stellar Parameterizer from Photometry

R. Andrae[1]⋆, M. Fouesneau[1], R. Sordo[2], C.A.L. Bailer-Jones[1], T.E. Dharmawardena[1], J. Rybizki[1], F. De Angeli[3], H.E.P. Lindstrøm[4,5,6], D.J. Marshall[7], R. Drimmel[4], A.J. Korn[8], C. Soubiran[9], N. Brouillet[9], L. Casamiquela[9,10], H.-W. Rix[1], A. Abreu Aramburu[11], M.A. Álvarez[12], J. Bakker[38], I. Bellas-Velidis[13], A. Bijaoui[14], E. Brugaletta[15], A. Burlacu[16], R. Carballo[17], L. Chaoul[18], A. Chiavassa[14], G. Contursi[14], W.J. Cooper[19,4], O.L. Creevey[14], C. Dafonte[12], A. Dapergolas[13], P. de Laverny[14], L. Delchambre[20], C. Demouchy[21], B. Edvardsson[22], Y. Frémat[23], D. Garabato[12], P. García-Lario[24], M. García-Torres[25], A. Gavel[8], A. Gomez[12], I. González-Santamaría[12], D. Hatzidimitriou[26,13], U. Heiter[8], A. Jean-Antoine Piccolo[18], M. Kontizas[26], G. Kordopatis[14], A.C. Lanzafame[15,27], Y. Lebreton[28,29], E.L. Licata[4], E. Livanou[26], A. Lobel[23], A. Lorca[30], A. Magdaleno Romeo[16], M. Manteiga[31], F. Marocco[32], N. Mary[33], C. Nicolas[18], C. Ordenovic[14], F. Pailler[18], P.A. Palicio[14], L. Pallas-Quintela[12], C. Panem[18], B. Pichon[14], E. Poggio[14,4], A. Recio-Blanco[14], F. Riclet[18], C. Robin[33], R. Santoveña[12], L.M. Sarro[34], M.S. Schultheis[14], M. Segol[21], A. Silvelo[12], I. Slezak[14], R.L. Smart[4], M. Süveges[35], F. Thévenin[14], G. Torralba Elipe[12], A. Ulla[36], E. Utrilla[30], A. Vallenari[2], E. van Dillen[21], H. Zhao[14], and J. Zorec[37]

*(Affiliations can be found after the references)*

Received March 03, 2022; accepted May 04, 2022

**ABSTRACT**

*Context.* The astrophysical characterisation of sources is among the major new data products in the third Gaia data release (DR3). In particular, there are stellar parameters for 471 million sources estimated from low-resolution BP/RP spectra.

*Aims.* We present the General Stellar Parameterizer from Photometry (GSP-Phot), which is part of the astrophysical parameters inference system (Apsis). GSP-Phot is designed to produce a homogeneous catalogue of parameters for hundreds of millions of single non-variable stars based on their astrometry, photometry, and low-resolution BP/RP spectra. These parameters are effective temperature, surface gravity, metallicity, absolute $M_G$ magnitude, radius, distance, and extinction for each star.

*Methods.* GSP-Phot uses a Bayesian forward-modelling approach to simultaneously fit the BP/RP spectrum, parallax, and apparent $G$ magnitude. A major design feature of GSP-Phot is the use of the apparent flux levels of BP/RP spectra to derive, in combination with isochrone models, tight observational constraints on radii and distances. We carefully validate the uncertainty estimates by exploiting repeat Gaia observations of the same source.

*Results.* The data release includes GSP-Phot results for 471 million sources with $G < 19$. Typical differences to literature values are 110 K for $T_{\rm eff}$ and 0.2-0.25 for $\log g$, but these depend strongly on data quality. In particular, GSP-Phot results are significantly better for stars with good parallax measurements ($\varpi/\sigma_\varpi > 20$), mostly within 2kpc. Metallicity estimates exhibit substantial biases compared to literature values and are only useful at a qualitative level. However, we provide an empirical calibration of our metallicity estimates that largely removes these biases. Extinctions $A_0$ and $A_{\rm BP}$ show typical differences from reference values of 0.07-0.09 mag. MCMC samples of the parameters are also available for 95% of the sources.

*Conclusions.* GSP-Phot provides a homogeneous catalogue of stellar parameters, distances, and extinctions that can be used for various purposes, such as sample selections (OB stars, red giants, solar analogues etc.). In the context of asteroseismology or ground-based interferometry, where targets are usually bright and have good parallax measurements, GSP-Phot results should be particularly useful for combined analysis or target selection.

**Key words.** stars: fundamental parameters – methods: data analysis; statistical; surveys; catalogs

## 1. Introduction

The ESA Gaia satellite (Gaia Collaboration et al. 2016) observes nearly two billion sources, most of which are stars residing in our Milky Way galaxy. Its main objective is to measure the parallax and proper motions of these stars with unprecedented accuracy. To achieve this goal, a correction dependent on the source colour is mandatory, and for this a low-resolution BP/RP spectrum is collected for each source. The DPAC Coordination Unit 8 with its astrophysical parameter inference system (CU8 Apsis, Bailer-Jones et al. 2013) classifies and determines the astrophysical parameters for these sources from the Gaia data. This allows more efficient exploitation of the exquisite astrometry and photometry offered by Gaia, for example by enabling appropriate selection criteria tailored to particular science cases. Gaia DR3 (Gaia Collaboration, Vallenari et al. 2022) will provide the first major release of results from CU8 (Creevey et al. 2022; Fouesneau et al. 2022b; Delchambre et al. 2022), including a general validation (Babusiaux et al. 2022).

In this paper, we describe the General Stellar Parameterizer from Photometry (GSP-Phot), which is one module in the CU8

⋆ andrae@mpia-hd.mpg.de

Apsis chain described in Bailer-Jones et al. (2013). GSP-Phot is designed to infer stellar parameters, distances, and line-of-sight extinctions from Gaia's low-resolution BP/RP spectra (Carrasco et al. 2021; De Angeli et al. 2022), astrometry (Lindegren et al. 2021b), and photometry (Riello et al. 2021). In Gaia DR3, the Gaia archive provides GSP-Phot results for 471 million sources with apparent magnitude $G \leq 19$. We also draw attention to a second module of the CU8 Apsis chain, the General Stellar Parameterizer from Spectroscopy (GSP-Spec; Recio-Blanco et al. 2022), which is also designed to characterise single stars in Gaia DR3 but using the higher resolution RVS spectra (Seabroke et al. 2022) instead of the low-resolution BP/RP spectra.

An early version of GSP-Phot was described in Liu et al. (2012) and the core methods were laid out in Bailer-Jones (2010) and Bailer-Jones (2011). Section 2 provides an overview of the current version of GSP-Phot adopted for Gaia DR3 and highlights the improvements over the earlier version in Liu et al. (2012). Section 3 then presents some scientific validation results from GSP-Phot when applied to Gaia DR3 data. Further validation results from GSP-Phot are presented in Creevey et al. (2022) and Fouesneau et al. (2022b). We conclude in Sect. 4.

## 2. GSP-Phot in a nutshell

### 2.1. Main principles

The main goal of GSP-Phot is to characterise all single stars in the Gaia catalogue based on their astrometry, photometry and, most importantly, their low-resolution BP/RP spectra. Those data are available for most sources with $G < 19$ in the Gaia catalogue. We emphasise that the BP/RP spectra are time-averaged mean spectra, which means that any intrinsic time variability is lost. GSP-Phot aims to provide a homogeneously derived catalogue of stellar parameters for non-variable single stars for all Gaia sources for which BP/RP spectra are available (which includes sources whose BP/RP spectra are not published in Gaia DR3). Other modules in the Apsis chain treat stars in binary systems or specific subtypes of stars in more specialised ways (see MSC and Extended Stellar Parametrizers in Creevey et al. 2022; Bailer-Jones et al. 2013). We emphasise that GSP-Phot uses only Gaia data: one objective of GSP-Phot is to attach a consistent set of astrophysical labels to the Gaia data and to also show how well stars can be generically characterised from Gaia data alone. Moreover, using non-Gaia data would fold in systematic errors and selection effects from external catalogues, which would make it more difficult to trace issues back to data sets during validation.

GSP-Phot comprises one main algorithm whose results are published in Gaia DR3 and two support algorithms whose results are used internally but are not published. The main algorithm is called Aeneas (referred to as $q$-method in Bailer-Jones 2011); it fits the measured BP/RP spectra, parallax, and apparent $G$ magnitude (see Sect. 2.2), thereby estimating the stellar parameters. For this optimisation process, Aeneas employs a specific type of Markov-chain Monte-Carlo (MCMC) sampling using an ensemble of walkers (Foreman-Mackey et al. 2013). More specifically, the ensemble MCMC optimises only four fit parameters, namely the stellar age, mass, metallicity (see Sect. 2.3), and the line-of-sight monochromatic extinction $A_0$ at 541.4 nm, where $A_0$ is the extinction parameter from the adopted Fitzpatrick extinction law (Fitzpatrick 1999); see also Sect. 11.2.3.1.4 in the online documentation for details. Other parameters such as distance or the extinction $A_G$ in the $G$ band are derived (see Sect. 2.4). As in Bailer-Jones (2011), GSP-Phot invokes astrophysical prior information; for example a Hertzsprung–Russell diagram (see Sect. 2.5). The two support algorithms provide the initial guess for the MCMC: first, the machine-learning algorithm Extremely Randomised Trees (Geurts et al. 2006) estimates stellar parameters directly from the BP/RP spectra; second, a gradient-descent algorithm (Ilium, Bailer-Jones 2010) further improves this initial parameter estimate. This is necessary because the MCMC alone would require too much computation time to find the best parameters without such an initial guess (see Sect. 2.7).

Within its forward-modelling context, GSP-Phot results are tied to the choice of model SEDs used to create synthetic BP/RP spectra. In its current version, GSP-Phot uses four different sets of model SEDs covering different temperature ranges of stars (see Sect. 2.6). In Sect. 3.1, we briefly investigate various different model SEDs and the extent to which their synthetic BP/RP spectra deviate from real observed BP/RP spectra.

### 2.2. Predicting observables

Combining multiple observables of different kinds is helpful to better constrain the model parameters and extract the maximum information out of all available measurements (e.g. Bailer-Jones 2011; Schönrich & Bergemann 2014). Below, we outline the observable data that are available to constrain our model within the Gaia and GSP-Phot context.

First and foremost, we have the low-resolution BP and RP spectra, which are available for most sources observed by Gaia (De Angeli et al. 2022). These are provided by CU5 in the format of coefficients for an adopted basis representation (Carrasco et al. 2021). Montegriffo et al. (2022) estimate that the spectral resolution, $\frac{\lambda}{\Delta\lambda}$, of BP ranges from 20 to 60 and that of RP from 30 to 50, where the higher resolution is achieved for shorted wavelengths for both BP and RP. For use in CU8, these continuous basis functions are then evaluated on a defined grid of physical wavelengths in order to produce actual sampled spectra in the common format of photon flux within a wavelength range (pixel). DPAC/CU5 also provide covariance matrices for the coefficients of BP and RP. As the CU8 wavelength sampling uses more pixels than coefficients that are provided by CU5, a pixel covariance matrix could be computed, but it would not have full rank and therefore could not be inverted to define a $\chi^2$. For Gaia DR3, CU8 only takes the diagonal elements of the pixel covariance matrix into account, but neglects the correlations. This approximation will be dropped in future versions of GSP-Phot.

Second, Gaia provides an apparent $G$ magnitude, which is available for all sources in the Gaia catalogue. The possibility to exploit the apparent $G$ magnitude was already envisaged in Bailer-Jones (2011) and Liu et al. (2012). Nevertheless, no absolute magnitude was available from their chosen fit parameters, and so the information provided by the apparent $G$ magnitude could not be fully exploited. We resolve this limitation by invoking stellar isochrones as discussed in Sect. 2.3.[1]

Finally, Gaia provides a parallax measurement for most of the sources. This can be used to constrain a distance estimate through an astrometric $\chi^2$ contribution to the total likelihood.

Each of these three observables (BP/RP spectra, apparent $G$, parallax) provides a $\chi^2$. These are summed to obtain a total $\chi^2$,

[1] We cannot exploit the integrated $G_{\rm BP}$ and $G_{\rm RP}$ photometry because these do not provide independent measurements from the dispersed BP/RP spectra themselves. We could use the integrated $G_{\rm RVS}$ magnitudes where available for bright sources, but while the RVS passband is provided in Sartoretti et al. (2022), unfortunately this only became available after our Gaia DR3 processing.

that is, the GSP-Phot likelihood function is constrained by all three observables.

### 2.3. Forward model based on isochrones

The key idea in Bailer-Jones (2011) was to take the apparent $G$ magnitude and make use of the flux conservation equation,

$$G = M_G + A_G + 5\log_{10}(d) - 5\,, \tag{1}$$

to allow information from the spectrum to constrain the distance. However, from the atmospheric parameters $T_{\rm eff}$, $\log g$, and [M/H], it is not possible to uniquely assign an absolute magnitude. This is the well-known problem of inverse isochrone matching. Instead, Bailer-Jones (2011) and Liu et al. (2012) chose to adopt a Hertzsprung–Russell diagram as a prior distribution and marginalise over the unknown absolute $M_G$ magnitude.

In this version of GSP-Phot, we solve this problem by starting from fundamental stellar parameters, namely age, initial mass, and metallicity. Stellar isochrones then uniquely provide us with astrophysically self-consistent absolute $M_G$ magnitude, radius, effective temperature, and surface gravity for the given fundamental parameters (age, mass, [M/H]). The atmospheric parameters are then also used to compute a synthetic model spectrum through multilinear interpolation over a given grid of models (see Sect. 2.6). Given the absolute $M_G$ magnitude provided by the isochrone and the extinction parameter $A_0$, we can compute the extinction $A_G$ from the model SED, the extinction curve, and the $G$ passband, and use that to predict the observed apparent $G$ magnitude from Eq. (1). This prediction of the apparent $G$ magnitude, which has an observational error of a few millimagnitudes, provides a very tight constraint on our model parameters and benefits the estimation of the surface gravity in particular. More precisely, $G$ has measurement errors of a few milli-magnitudes; however, the main uncertainty is likely to be in the $G$ passband estimation used to make model predictions of $M_G$ and $A_G$. We therefore introduce an error floor of 0.05mag (see Eq. (4)) to also account for model errors that may stem from imperfect knowledge of the passband.

For the isochrone models, we adopt a grid of PARSEC 1.2S Colibri S37 models (Tang et al. 2014; Chen et al. 2015; Pastorelli et al. 2020, and references therein) with step sizes of 0.01 between 6.6 and 10.13 in logarithmic age (in years) and 0.03 between -4.15 and 0.80 in [M/H]. These very fine step sizes are required to allow for a computationally efficient 3D linear interpolation (see Sect. 2.7) over age, mass, and metallicity to obtain the derived parameters.

### 2.4. Derived parameters

The four (MCMC) fit parameters are logarithmic age, initial mass, metallicity [M/H], and the parameter $A_0$ in the extinction law (Fitzpatrick 1999). Apart from the four fit parameters, there are several more derived parameters, though.

First, from the fit parameters, the isochrones provide us with derived values of effective temperature $T_{\rm eff}$, surface gravity $\log g$, stellar radius $R$, and absolute $M_G$ magnitude. These additional parameters are derived within the astrophysical models underlying the isochrones themselves and are tabulated in the isochrone data.

Second, coupling the extinction $A_0$ and metallicity [M/H] together with $T_{\rm eff}$ and $\log g$ from isochrones, we compute a model BP/RP spectrum from a library of models (see Sect. 2.6 and Creevey et al. 2022). This is done by computationally efficient 4D linear interpolation (see Sect. 2.7). We then use the fact that our model BP/RP spectra come with absolute flux levels that scale with $\sigma_B T_{\rm eff}^4$ (Stefan–Boltzmann law). Hence, when we use such a model to fit an observed BP/RP spectrum, we obtain an analytic $\chi^2$ solution for the amplitude

$$a = \frac{R^2}{d^2}\,, \tag{2}$$

which is needed to bring the model BP/RP spectrum to the flux scale of the observed BP/RP spectrum. Here, $R$ is the stellar radius and $d$ the distance of the star. As the radius is given by the isochrone, we can directly compute the distance $d$ from Eq. (2) for every MCMC sample. This distance then also enters the likelihood by predicting the measured parallax,

$$\chi^2_{\rm parallax} = \left(\frac{\varpi - \frac{1}{d}}{\sigma_\varpi}\right)^2, \tag{3}$$

and observed apparent $G$ magnitude,

$$\chi^2_{\rm G} = \frac{(G - M_G - A_G - 5\log_{10}(d) + 5)^2}{0.05^2 + \left(\frac{2.5\sigma_f}{f\log 10}\right)^2}, \tag{4}$$

where $f$ and $\sigma_f$ are the measured apparent $G$ flux and its uncertainty and 0.05 acts as an error floor of 50 milli-magnitudes that is added in quadrature to the approximate magnitude error $\frac{2.5\sigma_f}{f\log 10}$ (propagated from flux $f$ and flux error $\sigma_f$). If $\chi^2_{\rm spectra}$ denotes the chi-squared from fitting the observed BP/RP spectrum using the amplitude resulting from Eq. (2), the combined log-likelihood is given by

$$\log\mathcal{L} = \text{const} - \frac{1}{2}\left(\chi^2_{\rm spectra} + \chi^2_{\rm parallax} + \chi^2_{\rm G}\right), \tag{5}$$

ignoring irrelevant normalisation constants. As explained in Sect. 3.3, cases where the distance resulting from Eq. (2) deviates too much from the measured parallax or is inconsistent with the apparent $G$ magnitude have been filtered out of Gaia DR3.

Third, we also need the extinction in the $G$ band, $A_G$, for Eq. (1). This is obtained from the SEDs underlying our grid of model BP/RP spectra. These SEDs cover the wavelength range from 300nm to 1100nm. We apply interstellar extinction to the model grid according to Fitzpatrick (1999) assuming constant $R_0 = 3.1$ (see also Sect. 11.2.3.1.4 in the online documentation). These reddened SEDs are then integrated over the Gaia $G$ passband and the resulting magnitude can be compared to the corresponding value without extinction in order to obtain $A_G$. Thus, we can assign a value of $A_G$ to all models in our model grid. However, $A_G$ is not a free fit parameter. Instead, $A_G$ is submitted to the same 4D linear interpolation as the model BP/RP spectra themselves. In addition to $A_G$, we also compute extinction values $A_{\rm BP}$ and $A_{\rm RP}$ in exactly the same way. From those extinctions, we can compute the reddening $E(G_{\rm BP} - G_{\rm RP}) = A_{\rm BP} - A_{\rm RP}$.

### 2.5. Prior distributions

The full posterior probability distribution sampled by GSP-Phot is given by Eq. (A.2) as derived in Appendix A.1. One might expect us to only put priors on the four fit parameters (age, initial mass, metallicity, and extinction). However, sometimes it is

astrophysically more intuitive to impose priors on derived parameters; for example the Hertzsprung-Russell diagram on temperature and absolute magnitude. There are several prior factors in Eq. (A.2), which we now explain.

First, the prior for $A_G$ is a delta distribution that fixes the value to the extinction obtained from integrating the SED. While this may not behave like a commonly seen prior distribution, it remains a prior distribution from a mathematical point of view. Likewise, the prior for radius is a delta distribution fixing $R$ to the value provided by the isochrone.

Second, the extinction is restricted to the range $A_0 \in [0, 10]$ and within this range we adopt an ad hoc extinction prior of exponential form, $P(A_0|d) \propto e^{-A_0/\mu}$, where the mean value $\mu$ depends on Galactic latitude $b$ and distance $d$,

$$\mu = \frac{1 + 9\sin b}{1000 \cdot (1 + \exp[-(d - 100)/10])}. \quad (6)$$

This specific functional form and the choice of coefficients is the result of several test runs, reducing the occurrence of spuriously large extinctions in the validation sample.

Third, the distance was restricted to the range from 1 pc to 100 kpc. Furthermore, we adopt a distance prior of the form $P(d) \propto d^2 e^{-d/L}$ (as introduced in Bailer-Jones 2015) where the length scale $L$ depends on Galactic coordinates and has been mapped from the Gaia Early Data Release 3 (Gaia EDR3) mock catalog of Rybizki et al. (2020) excluding the Large Magellanic Cloud (LMC) and the Small Magellanic Cloud (SMC). To this end, we binned the mock data in Galactic coordinates and computed the mean distance $\langle d \rangle$ in each bin which is an estimator of the length scale $L = \frac{1}{3}\langle d \rangle$ under the assumed prior distribution $P(d) \propto d^2 e^{-d/L}$. The length scale is then interpolated over the grid in Galactic coordinates in order to provide a smooth distance–prior variation over the sky. However, we set the length scale of the prior to be a factor of ten smaller than the result from the Gaia EDR3 mock catalogue in an attempt to suppress outliers with unreasonably large distances. This also improved the comparison of temperature estimates to literature values, for example. Unfortunately, our validation sample lacked sources at large distances and therefore failed to show that this leads to a systematic underestimation of distances by GSP-Phot (see Sect. 3.7).

Fourth, Eq. (A.2) contains the factor $P([M/H], T_{\rm eff}, \log g, M_G, \log_{10}\tau, \log_{10}\mathcal{M})$, where there is an inter-dependency between the six components. Following Bailer-Jones (2011), we adopt a Hertzsprung-Russell-diagram prior, that is, we approximate the last factor as $P([M/H], T_{\rm eff}, \log g, M_G)$. Our specific Hertzsprung-Russell-diagram prior was constructed from the Gaia Universe Model Snapshot (Robin et al. 2012). We note that the PARSEC isochrones used by GSP-Phot and the Hertzsprung-Russell-diagram prior derived from GUMS are not always consistent. This may cause discrepancies, for example for low-mass dwarfs (see discussion of results on the Local Bubble in Babusiaux et al. 2022).

Finally, by adopting the forward isochrone modelling (see Sect 2.3), we restrict the parameters $T_{\rm eff}$, $\log g$, $M_G$, and radius, which can only populate regions that are reached by isochrones. Although this is formally part of the likelihood function, using isochrones in this way introduces a significant amount of prior astrophysical information.

We note that some of the priors are on derived parameters (Sect. 2.4) instead of fit parameters. This is somewhat uncommon but still formally correct in a Bayesian sense (see Appendix A.1). Furthermore, we note that we mainly employ priors as regularisation in order to suppress spuriously large extinctions and distances. We could have used other priors that are more motivated by astrophysics, for example an initial mass function, but we find that such priors are too weak to compete with the likelihood and so cannot confine the parameters to plausible regions of the parameter space.

### 2.6. Multi-library approach

GSP-Phot not only employs isochrone models but also requires model grids of synthetic spectral energy distributions (SEDs) of stellar atmosphere models. As introduced in Bailer-Jones et al. (2013), GSP-Phot uses four different such libraries of synthetic SEDs: MARCS for $T_{\rm eff}$ between 2500 and 8000 K, PHOENIX for $T_{\rm eff}$ between 3000 and 10 000 K, A-stars for $T_{\rm eff}$ between 6000 and 20 000 K, and OB for $T_{\rm eff}$ between 15 000 and 55 000 K. More details about these model libraries are provided in Creevey et al. (2022).

Results from each library are reported individually in Gaia DR3,[2] in case users have preferences for one particular library. We do *not combine* the different estimates. Instead, we recommend a best library for users who prefer a single result per star.[3]

We identify the best library from the log-posterior probabilities of the MCMC samples. Let $\theta$ denote the GSP-Phot parameters (temperature, extinction, distance, etc.) and $\boldsymbol{x}$ the BP/RP spectra, $\varpi$ the measured parallax, and $G$ the apparent magnitude. Then, $p(\theta_s|\boldsymbol{x}, \varpi, G)$ denotes the posterior probability of the $s$th MCMC sample where $s = 1, 2, \ldots, S$ and $S$ is the number of MCMC samples (same for all sources). We tested various different measures of goodness-of-fit in order to identify the best library, such as the maximum posterior value in the MCMC or the Bayesian evidence estimated by the harmonic mean (e.g. Wolpert & Schmidler 2012). In the end, we obtained the best results when identifying the best library as the one having the highest mean log-posterior value averaged over the MCMC samples:

$$\langle \log p(\theta|\boldsymbol{x}, \varpi, G) \rangle = \frac{1}{S} \sum_{s=1}^{S} \log p(\theta_s|\boldsymbol{x}, \varpi, G). \quad (7)$$

We note that Eq. (7) corresponds to a Monte-Carlo estimate of the differential entropy,

$$h = -\int p(\theta|\boldsymbol{x}, \varpi, G) \log p(\theta|\boldsymbol{x}, \varpi, G)\, d\theta, \quad (8)$$

which means that $h \approx -\langle \log p(\theta|\boldsymbol{x}, \varpi, G) \rangle$. In other words, the best library is chosen to be the library whose posterior distribution has the lowest differential entropy, that is, it provides the most information about the source from the point of view of information theory. This identification scheme of the best library is adequate but not perfect (see Sect. 3.5).

### 2.7. Computational cost

The objective of GSP-Phot is to provide stellar parameter estimates for hundreds of millions of stars in Gaia DR3 (and ultimately all 1.8 billion sources expected in Gaia DR4). However, given the Gaia data release planning, we only have a limited amount of time available for processing. In order to comply with

[2] The results from individual libraries are provided in the Gaia archive table named `astrophysical_parameters_supp`.

[3] The best-library results are provided in the Gaia archive tables named `gaia_source` and `astrophysical_parameters`.

these limited resources, GSP-Phot can only process sources with $G < 19$ (see Creevey et al. 2022, Table 1 therein). The actual processing of GSP-Phot in producing the results for Gaia DR3 took approximately 360 000 CPU hours on about 1400 cores, which equates to 257 hours.

One consequence of limited computational resources is that we cannot afford long convergence phases in our MCMC sampling. We therefore need a good initial guess in order to accelerate convergence. This initial guess is provided in a two-step process, starting with a machine-learning algorithm called Extremely Randomised Trees (Geurts et al. 2006), which is one of the support algorithms within GSP-Phot mentioned in Sect. 2.1.[4] As in Liu et al. (2012), the resulting initial guess is then further refined by a Newton-Raphson algorithm (Bailer-Jones 2010), which is the second support algorithm in GSP-Phot. These two previous steps are computationally inexpensive and allow us to get away with very short convergence and relaxation phases in our MCMC.

Furthermore, while Liu et al. (2012) employed a classic Metropolis-Hastings MCMC, we changed this to the `emcee` (Foreman-Mackey et al. 2013). The reason is that it is impossible to configure the step size of the proposal distribution in the Metropolis-Hastings MCMC such that it works well when all sources have different signal-to-noise ratios. Conversely, the `emcee` is an ensemble MCMC that does not require any proposal distribution to be fine-tuned, and can provide very efficient sampling for all sources. Appendix A.2 gives further details on the MCMC configuration.

Last but not least, a key choice is that GSP-Phot uses multilinear interpolation over rectangular model grids. Among all possible interpolation schemes, we found this to be computationally highly efficient.[5] First, there is a 3D linear interpolation over isochrones with a rectangular grid in age, [M/H] and initial mass. Second, there is a 4D linear interpolation over rectangular grids of $T_{\rm eff}$, $\log g$, [M/H], and $A_0$. In contrast, Liu et al. (2012) used a thin-plate-spline smoothing, which has a computational cost that is about two to three orders of magnitude more expensive. Likewise, GSP-Phot cannot afford the computational cost of propagating interpolation errors using a Gaussian process as in Czekala et al. (2015). This inevitably causes an underestimation of the uncertainties.

### 2.8. Values, uncertainties, and MCMC chains provided in Gaia DR3

The parameter values and lower and upper confidence levels that are provided in Gaia DR3 are the median and 16th and 84th percentiles of the MCMC samples, respectively. We choose the median value because the mean (or mode) values can lie outside the confidence interval, especially (but not exclusively) in the presence of outliers, and inferring the mode from the MCMC samples requires additional computational cost.

We emphasise that we only provide one-dimensional confidence levels, and not correlations between parameters. The reason is that a correlation matrix implies that the posterior probability is Gaussian, which is not the case. Instead, for most sources, the MCMC samples exhibit clear evidence of non-Gaussianity (e.g. asymmetry, curved contours, heavy tails). We therefore provide MCMC samples themselves to enable the user to correctly propagate uncertainties through any subsequent analysis. For reasons of data volume, it is not possible to provide full MCMC chains for all sources. Therefore, MCMC chains are provided only for the best-library results. Even here, we provide a reduced sample comprising only the last 100 MCMC samples for most sources.[6] The full MCMC chain with all 2000 samples (see Appendix A.2) is provided only for sources brighter than $G < 12$ and for a random subset of 1% of sources fainter than that. Moreover, due to the filtering described in Sect. 3.3, around 10% of best-library results have no MCMC samples, because the originally best library was filtered out together with its MCMC and another library stepped up to fill the role of best library. This can happen, for example, when the originally best library provided a good fit to the BP/RP spectra, with their numerous pixels dominating the likelihood function, but otherwise poorly predicted the parallax or apparent $G$ magnitude.

## 3. Application to Gaia data

In this section, we show some basic validation results from GSP-Phot. We first discuss aspects of MCMC convergence and filtering of results. We then verify the results at the level of distributions, for example, via a Hertzsprung-Russell diagram. Finally, we compare GSP-Phot results to literature values. Further complementary validation results are presented by Creevey et al. (2022), Fouesneau et al. (2022b), Delchambre et al. (2022), Babusiaux et al. (2022), Gaia Collaboration, Creevey et al. (2022), and Gaia Collaboration, Schultheis et al. (2022).

### 3.1. Mismatch between models and observed BP/RP spectra

The forward modelling of BP/RP spectra by GSP-Phot relies heavily on the agreement between observed BP/RP spectra and models thereof. Unfortunately, this agreement is not perfect. In order to illustrate this, we take solar twins from Galarza et al. (2021) and select 18 twins with $A_0 < 0.001$mag. For each of these 18 solar twins, we rescale their observed BP/RP spectra to $G = 15$ from their actual apparent $G$ magnitude in order to make their flux levels comparable to each other and to model spectra from PHOENIX and MARCS (see Creevey et al. 2022) as well as to the model spectrum `sun_model_001` from the CALSPEC library[7] (Bohlin et al. 1995, 2014, 2020) and to the 3D Stagger model at $T_{\rm eff} = 5787$K, $\log g = 4.44$, [Fe/H]=0 (Palacios et al. 2010).[8] In Fig. 1a and b, we compare these models to the observed BP/RP spectra, which shows good agreement to first order. However, if we inspect the differences between models and observed BP/RP spectra in Fig. 1c and d, we note that PHOENIX poorly matches BP, whereas MARCS poorly matches RP. CALSPEC `sun_model_001` is a poor match to both BP and RP. The 3D Stagger model matches BP reasonably well but not RP. The flux differences are as large as 10% in BP and 4% in RP per pixel, which are significant compared to the typical flux uncertainties of 2% and below in BP and about 0.5% in RP for sources

[4] Extremely Randomised Trees replace the Support Vector Regression previously used in Liu et al. (2012) because they are much easier to train.

[5] There are smoothing algorithms that are computationally faster but make additional approximations, e.g. those used in Fouesneau et al. (2022a).

[6] Note that the reported values and confidence levels are *always* estimated from the full MCMC having 2000 samples, even if the reported MCMC is reduced to only 100 samples.

[7] `https://www.stsci.edu/hst/instrumentation/reference-data-for-calibration-and-tools/astronomical-catalogs/calspec`

[8] `http://npollux.lupm.univ-montp2.fr/DBPollux/PolluxAccesDB`

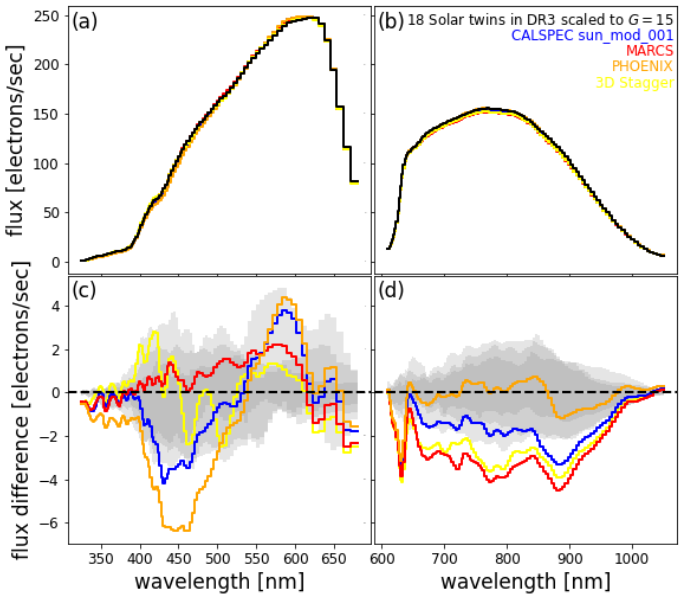


**Fig. 1.** Mismatch between observed BP/RP spectra of 18 solar twins from Galarza et al. (2021) with $A_0 < 0.001$mag (black lines and grey contours) and model BP/RP spectra from PHOENIX (orange lines), MARCS (red lines), CALSPEC `sun_model_001` (blue lines), and 3D Stagger (yellow lines). Panel (a): All BP spectra scaled to $G = 15$. Panel (b): Same as panel (a) but for RP. Panel (c): Differences from median observed BP spectrum in panel (a). Panel (d): Same as panel (c) but for RP. Grey contours in panels (c) and (d) indicate the pixel-wise central 68%, 90%, and min/max intervals given the 18 solar twins.

in this apparent magnitude range, that is, between $G = 7.4$ and $G = 8.7$. We also emphasise that while we use Sun-like stars to illustrate this mismatch in Fig. 1, it is very likely that also other types of stars suffer from similar mismatches.

This mismatch can only be partially ascribed to imperfections in the CU5 instrument model, namely where all models agree with each other but disagree with the observations (in the steep RP cutoff 620-650nm and at the blue end of BP 320-400nm, see Montegriffo et al. 2022). Nevertheless, over wide ranges, the various model spectra differ significantly not only from the observed spectra but also from each other (400-650nm in BP, 680-950nm in RP). This mismatch can only originate from a genuine difference between the various model SEDs, as all models have their BP/RP spectra simulated with the exact same CU5 instrument model from Montegriffo et al. (2022). This systematic disagreement between models such as MARCS and PHOENIX most likely originates from different opacities which lead to different degrees of flux redistribution. Spectral lines are mostly invisible in low-resolution BP/RP spectra, such that the continuum shape is very important.

As a result of the systematic mismatch between models and observed BP/RP spectra, GSP-Phot results for parameters such as temperature and extinction often tend to cluster at the grid points of the model grids used for multilinear interpolation. This is visible as stripes when plotting GSP-Phot parameters. The reason is that the parameter optimisation struggles to make the model fit the observed BP/RP spectrum, which in the presence of systematic mismatches does not work perfectly. The closest fit can often only be achieved by letting the pixel fluxes of the model take a maximal or minimal value. As linear interpolation is a form of *monotonic interpolation*, maxima or minima can only be acquired at grid points, but not in between grid points.

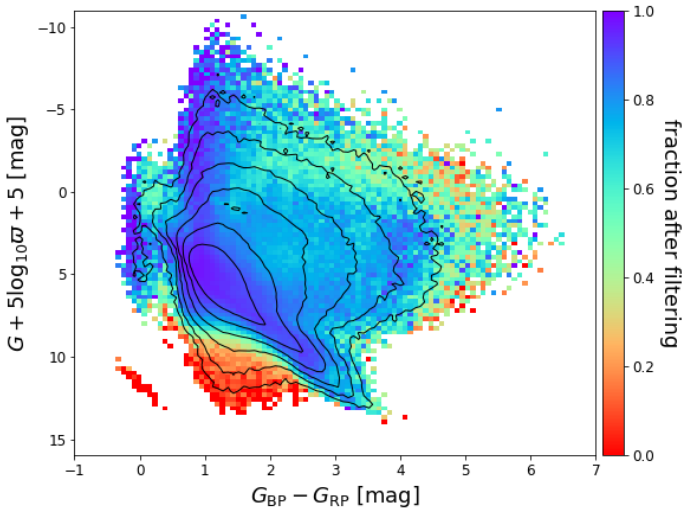


**Fig. 2.** Fraction of sources surviving the filtering described in Sect. 3.3 in the observed CMD. Contours indicate density of all sources with $G \leq 19$. This figure used a random subset of 2 815 418 sources drawn from the main catalogue. This plot does not include sources without parallax measurement or with negative parallax.

In that sense, the presence of such stripes can be interpreted as an indicator of a mismatch between models and data.

### 3.2. MCMC convergence

As explained in Sect. 2.7 and Appendix A.2, we have to use a fixed number of MCMC iterations in order to comply with the limited computational resources. This leads to the possibility of non-convergence. Visual inspection of 500 randomly chosen MCMC chains suggests that about 50% of cases show at least minor evidence of non-convergence (e.g. drift in at least one parameter). However, those cases do not appear to correspond to outliers in scientific validation because test runs with longer MCMC chains and better convergence did not yield better scientific results (e.g. in terms of lower differences to literature values). Instead, scientific outliers appear to be cases where the MCMC got stuck in a local optimum, that is, converged to a bad solution.

### 3.3. Filtering of results

As mentioned in Sect. 2.7 and further explained in Creevey et al. (2022), GSP-Phot has only processed sources with $G < 19$ due to limited computational resources.[9] However, while all sources with $G < 19$ have been processed, not all the GSP-Phot results are published in Gaia DR3. Instead, based on initial validation work, we filter out results when one or more of the following conditions apply:

i There is no parallax available. In such a case, the distance estimate is usually unreliable and most of these sources would end up with improper solutions, for example in the colour–magnitude diagram.

[9] Gaia sources with $G > 19$ (about two-thirds of the entire Gaia sample) have therefore never been processed with GSP-Phot. This is not due to a lack of data quality but simply due to a lack of computational resources.

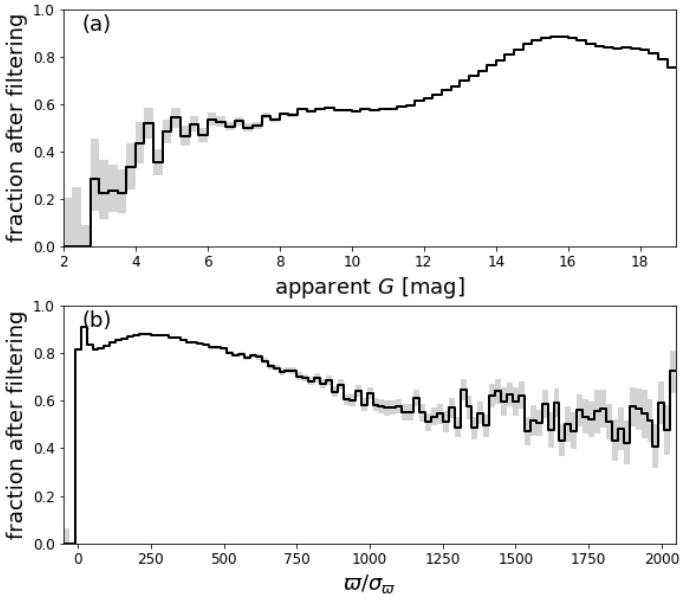


**Fig. 3.** Fraction of sources surviving the filtering described in Sect. 3.3 as a function of apparent $G$ magnitude (panel a) and parallax signal-to-noise ratio (panel b). Both panels use a random subset of about 100 million sources with $G < 19$. Grey intervals indicate the horizontally decreasing intervals of 68% confidence assuming a beta distribution in each bin.

ii The number of transits in the BP or RP spectrum is below 10 or 15, respectively. Such spectra are not of sufficient quality for the GSP-Phot analysis.

iii The observed apparent $G$ magnitude is poorly predicted, such that $M_G + A_G + 5\log_{10} d - 5$ differs from $G$ by more than 0.1mag.

iv The inverse distance differs from the measured parallax by more than ten times the parallax error.

v The MCMC acceptance rate is below 10%, suggesting that the initial guess was poor and did not allow the MCMC to properly explore the parameter space (see Sect. 3.2).

Of the original 575.9 million sources with $G < 19$, 471 million survive the filtering process (~81.7%).[10] Figure 2 shows where the sources lost due to filtering reside in the CMD. As GSP-Phot has no specific models for white dwarfs, almost all white dwarfs are lost. Otherwise, Fig. 2 shows that no other population is selectively affected by the filtering process. Furthermore, Fig. 3a shows that the completeness is around 0.8 for apparent $G$ magnitudes of between 16 and 18, whereas for magnitudes brighter than $G = 12$ the completeness is around 0.5, falling below 0.5 for the very brightest sources at $G < 5$. Intuitively, one may expect higher completeness at the bright end due to better data quality. However, as is evident from Fig. 3b, the high parallax quality at the bright end is exactly where the GSP-Phot completeness is lowest. The main reason for this behaviour is the filter (iv) requiring that the inverse GSP-Phot distance agrees to the parallax within ten times the parallax measurement error. As GSP-Phot infers the distance from the amplitude of the BP/RP spectra (see Sect. 2.4), the noise on the BP/RP amplitude makes it increasingly difficult to agree with the parallax to within $10\sigma$ as the parallax measurement error decreases at the bright end.

[10] The vast majority of filtered cases for $G > 12$ are due to a mismatch of observed and predicted apparent $G$ magnitudes.

**Table 1.** Comparison of absolute differences between best $T_{\rm eff}$ estimates (in K) from GSP-Phot and literature values. The columns indicate, from left to right, the median absolute difference (MedAD), the mean absolute difference (MAD), the root-mean-square difference (RMSD), the absolute difference not exceeded by 75% of sources (AD 75%), and the absolute difference not exceeded by 90% of sources (AD 90%).

| catalogue | MedAD | MAD | RMSD | AD 75% | AD 90% |
|---|---|---|---|---|---|
| APOGEE | 169 | 418 | 1294 | 440 | 824 |
| GALAH | 110 | 150 | 228 | 198 | 315 |
| LAMOST | 110 | 156 | 253 | 198 | 327 |
| RAVE | 160 | 227 | 390 | 296 | 483 |

### 3.4. CMD, HRD, and $T_{eff}$-log $g$ diagrams

The goal of GSP-Phot is to characterise all single stars in the Gaia catalogue. In Fig. 4, we demonstrate that the reddening and the absolute magnitude estimated by GSP-Phot indeed produce a de-reddened CMD that appears astrophysically plausible. Likewise, the Hertzsprung-Russell diagram and $T_{\rm eff}$-log $g$ diagram shown in Fig. 5 appear plausible. Nevertheless, we do see a prominent vertical stripe at $T_{\rm eff} = 15\,000$K in both panels of Fig. 5, which is a pile-up effect at the lower boundary of the OB library. We also see some minor vertical stripes at various temperatures, which are most likely a result of multilinear interpolation struggling in the presence of the mismatch between models and real BP/RP spectra (see discussion in Sect. 3.1). The Hertzsprung-Russell diagram in Fig. 5a also shows a hook at around 4000K protruding out of the giant population towards fainter magnitudes.

As the GSP-Phot forward model uses isochrones (see Sect. 2.3), we see in Fig. 4b and Fig. 5 that only the regions covered by these isochrones are populated, that is, there is no extrapolation.

### 3.5. Comparison to literature values

We compare our parameter estimates to those from the literature, specifically 256 967 stars from APOGEE DR16 (Jönsson et al. 2020), 169 825 stars from GALAH DR3 (Buder et al. 2021), 513 669 stars from LAMOST DR4 (Wu et al. 2011, 2014), and 153 284 stars from RAVE DR6 (Steinmetz et al. 2020). We note that these literature values were estimated from spectra with resolutions ranging from 1000 (LAMOST) to 28 000 (GALAH). These spectral resolutions are significantly higher than those of the BP/RP spectra (20-60 for BP, 30-50 for RP, Montegriffo et al. 2022).

#### 3.5.1. Effective temperature

A major hurdle for GSP-Phot parameter estimates in general and effective temperatures in particular is the temperature–extinction degeneracy. This latter originates from the fact that a red star could be genuinely cool or have a higher temperature but is subject to notable line-of-sight dust attenuation and according reddening. Using only the low-resolution optical BP/RP spectra, it is very difficult to distinguish between these two cases. Employing the parallax and apparent magnitude can mitigate but not fully break this degeneracy (Bailer-Jones 2011). Nevertheless, the GSP-Phot temperatures are affected to some degree. Conversely, the effective temperatures from, for example, APOGEE, GALAH, LAMOST, or RAVE are derived from absorption lines in spectra of much higher resolution and are therefore unaffected by extinction.

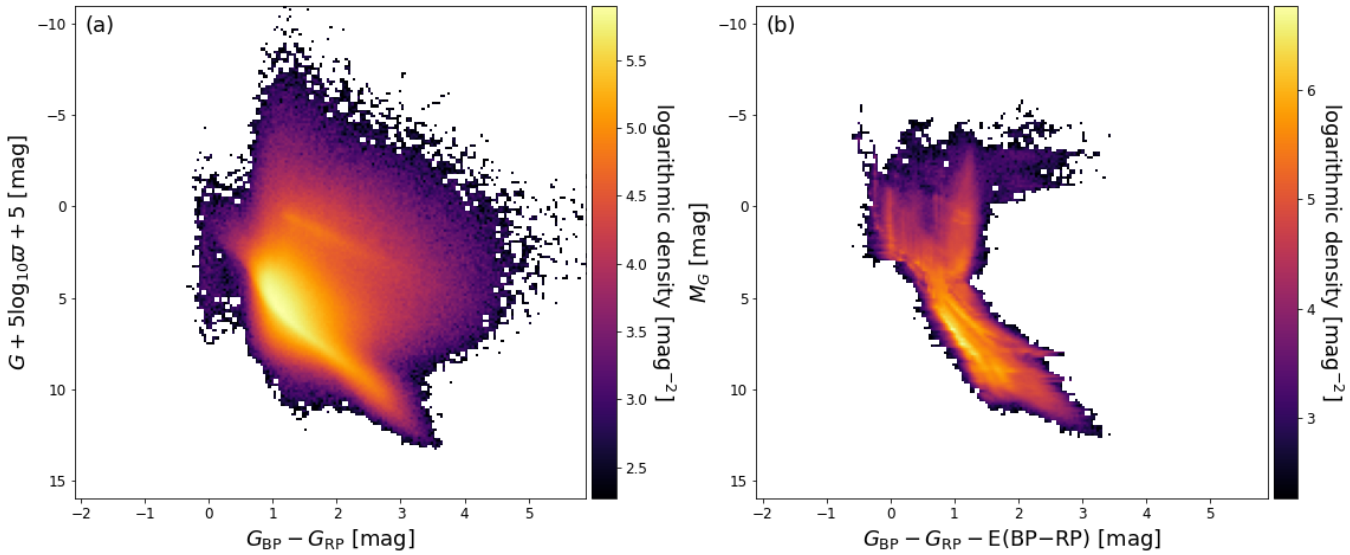


**Fig. 4.** Observed colour–magnitude diagram (panel a) and de-reddened colour–magnitude diagram (panel b). Both panels use the same sample of 2 598 519 stars that has been randomly selected from the main catalogue (see Appendix B for the ADQL query).

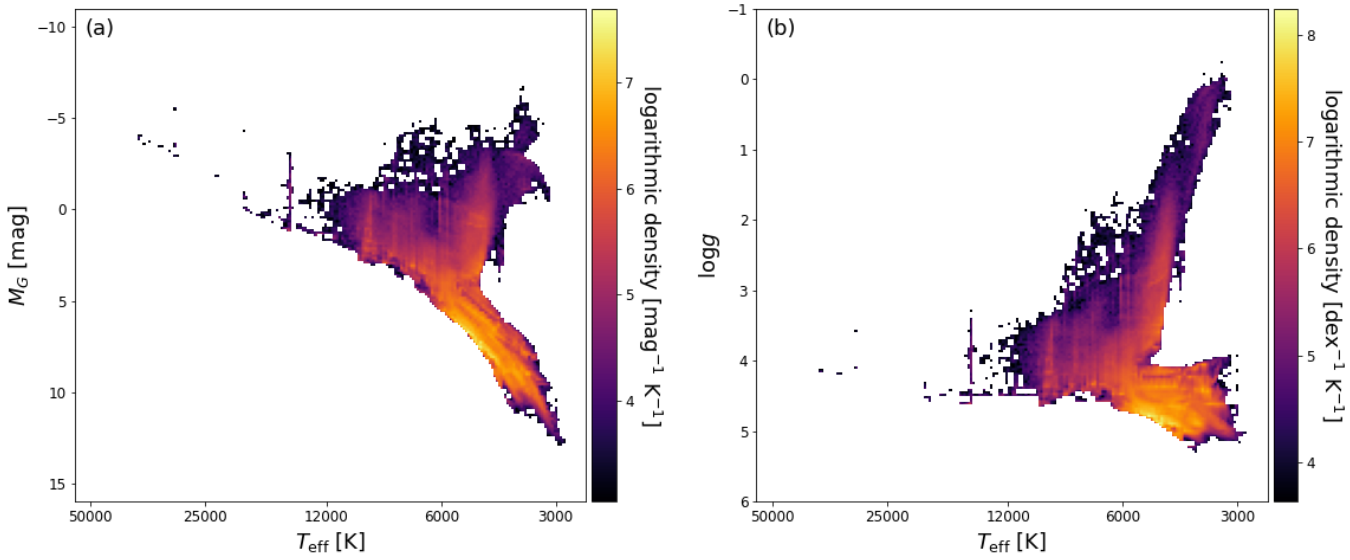


**Fig. 5.** Hertzsprung-Russell diagram (panel a) and $T_{\rm eff}$-log $g$ diagram (panel b). The same sample as in Fig. 4 is used. The ADQL query for this plot can be found in Appendix B.

Table 1 compares the absolute differences between our $T_{\rm eff}$ estimates and literature values. We can see that there is only a mild dependence on the reference catalogue and that overall the median absolute difference shows that half of our results agree with literature values to within ∼170 K. For GALAH DR3 and LAMOST DR4 in particular, half of our results agree to within 110 K. The deviations are larger for APOGEE, which includes a significant fraction of red giants with relatively high extinction. For these stars, the temperature–extinction degeneracy is particularly difficult to break for GSP-Phot, resulting in hot stars being favoured and the OB library being labelled as best library. If we exclude all results from the OB library, the RMS difference reduces to 662 K, which is still higher than for the other catalogues.

In order to exclude the possibility that APOGEE could have any internal inconsistency in itself, we also compare the results for the subset of 4015 stars shared by APOGEE DR16, GALAH DR3, and our results. On this specific subset, the RMS difference between GSP-Phot $T_{\rm eff}$ estimates and APOGEE and GALAH values is 269K and 263K, respectively, whereas the RMS difference between APOGEE and GALAH values is only 116K. This suggests that GALAH DR3 and APOGEE DR16 are in good mutual agreement and that we get a genuine overestimation of $T_{\rm eff}$ in GSP-Phot for stars with line-of-sight extinctions $A_0 \gtrsim 2$. Figure 6a reveals that the largest temperature differences occur for red giant stars, whereas GSP-Phot estimates appear to be consistent for main sequence stars. However, for stars with high-quality parallaxes, the GSP-Phot temperature estimates are much better and still usable in the red giant branch, as is evident from Fig. 6b. If we impose a parallax quality cut of $\frac{\varpi}{\sigma_\varpi} > 20$, the median absolute deviation and the RMS deviation drop from 169K and 1294K (Table 1) to 105K and 369K, respectively, although

the number of stars also decreases by nearly a factor of two. Therefore, this appears to be not only due to the temperature–extinction degeneracy, but also in part to the systematic underestimation of distances caused by an overly harsh distance prior (Fouesneau et al. 2022b).

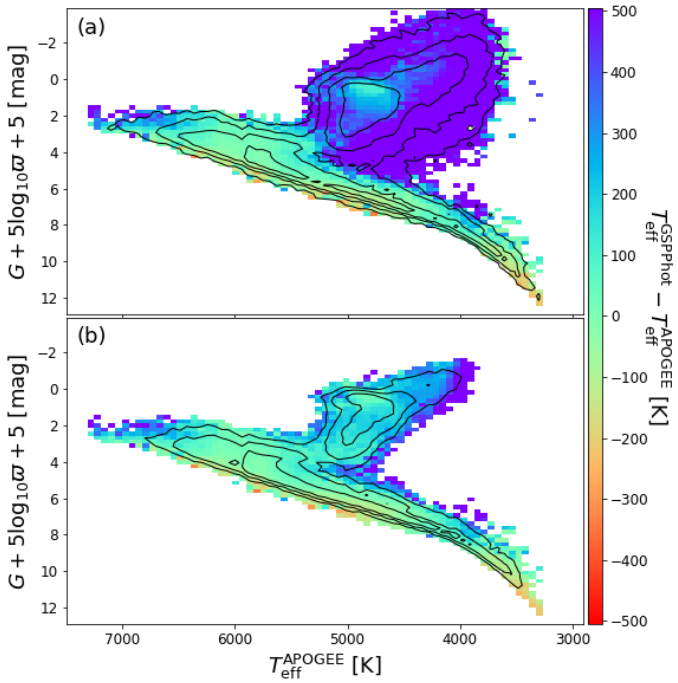


**Fig. 6.** Differences in $T_{\rm eff}$ between results and literature values from APOGEE DR16 (Jönsson et al. 2020) in the Hertzsprung-Russell diagram. Panel (a): All sources. Panel (b): Sources with $\varpi/\sigma_\varpi > 20$. Contours indicate source density decreasing by factors of 3.

Considering the subset of 6560 stars shared by GALAH DR3, RAVE DR6, and our results, the RMS difference between GSP-Phot $T_{\rm eff}$ estimates and those of GALAH and RAVE is 230 K and 281 K, respectively, whereas the RMS difference between GALAH and RAVE is 214 K. This suggests that, for this subset, the GSP-Phot results are fully consistent with the typical uncertainties in the literature values.

Figure 7a reveals the temperature–extinction degeneracy very clearly. GSP-Phot overestimates $T_{\rm eff}$ with rising $A_0$ estimate, because an overestimated extinction can compensate for an overestimated temperature. This is hardly surprising given that GSP-Phot is using only optical data. However, for $A_0 \lesssim 4$, the 84th percentile does not exceed 1000K, that is, in that regime our temperature estimates are reasonably stable. Furthermore, Fig. 7b shows that GSP-Phot estimates a low extinction for the vast majority of stars. Here, we emphasise that, in principle, the temperature–extinction degeneracy works both ways: it can also cause a simultaneous underestimation of $T_{\rm eff}$ and $A_0$. However, the majority of our validation targets with literature values usually have very low extinction.[11] If the true extinction is already close to zero, then there will be no room to significantly underestimate $A_0$ because of GSP-Phot's non-negativity constraint on $A_0$. In Sect. 3.10, we use a sample that is not restricted to low-extinction stars and it indeed shows the temperature–extinction degeneracy working both ways.

[11] This is a consequence of ground-based spectroscopic surveys preferentially targeting relatively bright sources.

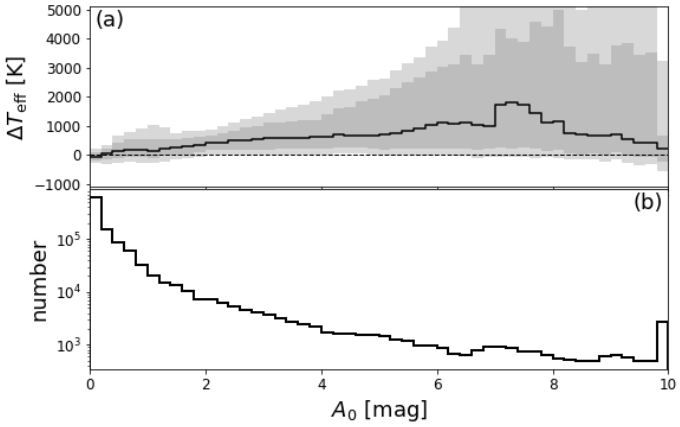


**Fig. 7.** Impact of extinction onto differences in $T_{\rm eff}$ between results and literature values. Panel (a): Differences in $T_{\rm eff}$ between our GSP-Phot results and literature values as function of estimated extinction $A_0$ for the joint samples from APOGEE DR16 (Jönsson et al. 2020), GALAH DR3 (Buder et al. 2021), LAMOST DR4 (Wu et al. 2011, 2014), and RAVE DR6 (Steinmetz et al. 2020). If the same star appears in more than one literature resource, it enters this plot multiple times with identical GSP-Phot results but different literature values. The data are binned into $A_0$ ranges of 0.2mag, where the solid black line indicates the median residual and the two shaded regions show the 5th-to-95th percentiles and the 16th-to-84th percentiles, respectively. Panel (b): Distribution of $A_0$ estimates in this sample.

**Table 2.** Comparison of best log $g$ estimate from GSP-Phot to literature values. Columns as in Table 1.

| catalogue | MedAD | MAD | RMSD | AD 75% | AD 90% |
|---|---|---|---|---|---|
| APOGEE | 0.218 | 0.406 | 0.626 | 0.570 | 1.054 |
| GALAH | 0.059 | 0.102 | 0.163 | 0.119 | 0.255 |
| LAMOST | 0.104 | 0.154 | 0.236 | 0.197 | 0.332 |
| RAVE | 0.252 | 0.335 | 0.465 | 0.451 | 0.709 |

The skymaps of temperature differences in Fig. 8 also show that GSP-Phot typically overestimates $T_{\rm eff}$ for APOGEE targets, whereas the results are typically much better for targets from other catalogues. Nevertheless, all literature catalogues suggest that we tend to overestimate $T_{\rm eff}$ in the Galactic plane. Again, this is a consequence of the temperature–extinction degeneracy.

### 3.5.2. Surface gravity

Table 2 demonstrates that our log $g$ estimates are also very good: as is evident from the median absolute differences, half of our sources agree with literature values of log $g$ to within 0.25 dex. For GALAH DR3 in particular, 75% of our results agree to within 0.12 dex. Given the low resolution of BP/RP, this may appear somewhat surprising, but the main constraint on log $g$ is actually imposed by the prediction of the apparent $G$ magnitude together with the measured parallax that provide a rather tight constraint on the luminosity and absolute magnitude of the star. For sources with poor parallax measurements ($\frac{\varpi}{\sigma_\varpi} < 10$), Fig. 9 reveals that GSP-Phot tends to systematically overestimate log $g$. This is related to an overly harsh distance prior, as is discussed in Sect. 3.7.

The excellent quality of our log $g$ estimates for sources with good parallaxes is demonstrated further by the comparison to asteroseismic values in Fig. 10a. There is only a mild systematic overestimation by ~0.25 dex for giants with log $g < 2.5$ dex. In particular, the median absolute difference quoted in Fig. 10a

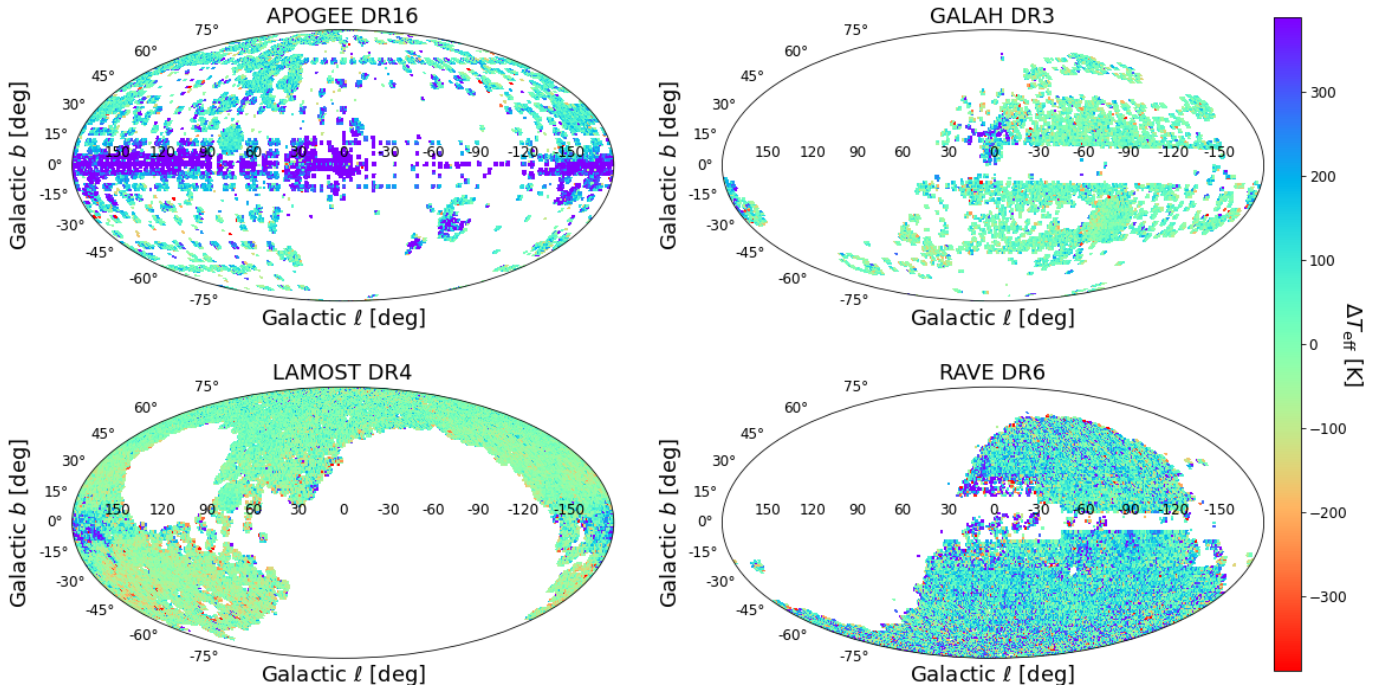


**Fig. 8.** Skymaps of differences between our results and literature $T_{\rm eff}$ for APOGEE DR16 (Jönsson et al. 2020), GALAH DR3 (Buder et al. 2021), LAMOST DR4 (Wu et al. 2011, 2014), and RAVE DR6 (Steinmetz et al. 2020). All skymaps use the Mollweide projection where lines of constant latitude are horizontal straight lines parallel to the equator.

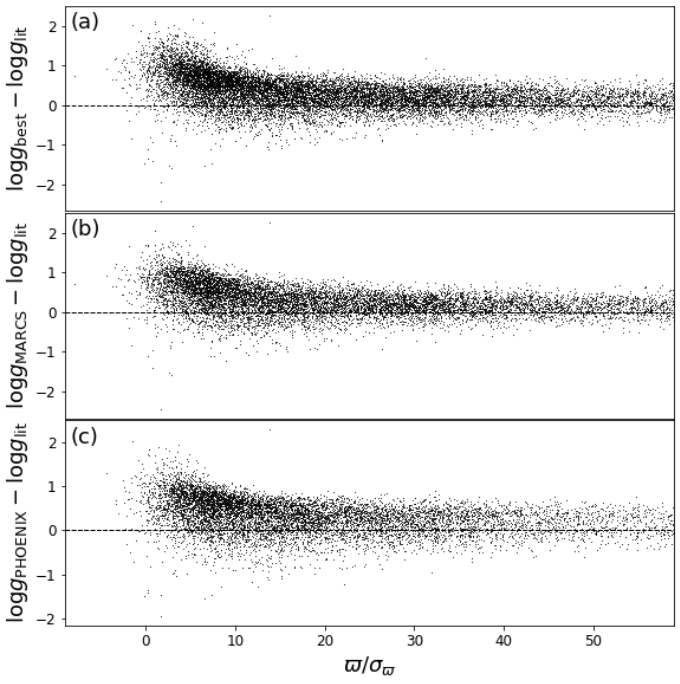


**Fig. 9.** Bias of surface gravity estimates for stars with low parallax signal-to-noise ratio, $\varpi/\sigma_\varpi$, for 25 169 red clump stars (Jönsson et al. 2020; Bovy et al. 2014). Panel (a): GSP-Phot best-library results. Panel (b): GSP-Phot MARCS library results. Panel (c): GSP-Phot PHOENIX library results.

shows that half of our $\log g$ values agree with asteroseismic values to within 0.2 dex or better.

Given that asteroseismic estimates of $\log g$ typically have very low uncertainties (Creevey et al. 2013), the deviations in Fig. 10 should be fully explained by GSP-Phot's own uncertainties. If we therefore normalise the differences by our uncertainty estimates, we can look for a distribution similar to a unit Gaussian. More specifically, we normalise by our lower confidence interval if our $\log g$ estimate is above the asteroseismic value and we normalise by our upper confidence interval if it is below. If we denote the normalised differences as $d$ and the asteroseismic reference value as $\log g_{\rm AS}$, we have

$$d = \begin{cases} \dfrac{\text{logg\_gspphot} - \log g_{\rm AS}}{\text{logg\_gspphot\_upper} - \text{logg\_gspphot}} & \Leftrightarrow \text{logg\_gspphot} < \log g_{\rm AS} \\ \dfrac{\text{logg\_gspphot} - \log g_{\rm AS}}{\text{logg\_gspphot} - \text{logg\_gspphot\_lower}} & \Leftrightarrow \text{logg\_gspphot} > \log g_{\rm AS} \end{cases} . \quad (9)$$

We emphasise that the lower and upper confidence levels are 16th and 84th percentiles of the MCMC samples and the reported value is the median (see Sect. 2.8) such that

logg_gspphot_lower ≤ logg_gspphot ≤ logg_gspphot_upper,

and logg_gspphot_upper − logg_gspphot ≥ 0 is the upper error while logg_gspphot − logg_gspphot_lower ≥ 0 is the lower error. We exclude sources for which the asteroseismic reference value is below 2.5 dex, because our $\log g$ estimates are biased in that regime. This systematic error would compromise the validation of our uncertainty estimates that are meant to account for random errors only. Unfortunately, Fig. 10b shows that while the distribution of normalised differences is indeed centred on zero, it is very far from a unit Gaussian. Instead, Fig. 10b suggests that our uncertainties are underestimated by a factor of ∼10. This value of 10 has been estimated by a maximum-likelihood fit of the standard deviation given the normalised residuals. Indeed, the distribution appears to be not even Gaussian at all, exhibiting a sharper peak and heavier tails than expected from a Gaussian. We return to this issue in Sect. 3.9.

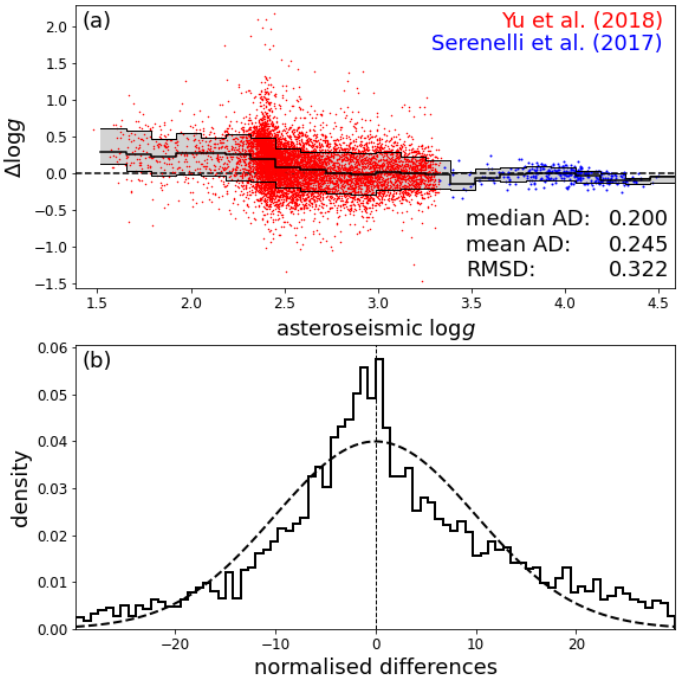


**Fig. 10.** Comparison to asteroseismic surface gravities. Panel (a): Comparison between our best-library log $g$ estimates and asteroseismic values from Serenelli et al. (2017) (blue points) and Yu et al. (2018) (red points). The solid black line shows the median difference and the grey region the central 68% interval. Panel (b): Distribution of normalised differences defined in Eq. (9) for asteroseismic log $g > 2.5$ compared to a Gaussian with zero mean and $\sigma = 10$ (dashed black line).

**Table 3.** Comparison of the best [M/H] estimate from GSP-Phot to literature values. Same as Table 1.

| catalogue | MedAD | MAD | RMSD | AD 75% | AD 90% |
|---|---|---|---|---|---|
| APOGEE | 0.210 | 0.303 | 0.450 | 0.384 | 0.644 |
| GALAH | 0.210 | 0.238 | 0.295 | 0.326 | 0.452 |
| LAMOST | 0.204 | 0.248 | 0.330 | 0.328 | 0.477 |
| RAVE | 0.198 | 0.254 | 0.350 | 0.344 | 0.524 |

#### 3.5.3. Metallicity

Table 3 compares the GSP-Phot metallicity estimates to literature values. Despite the low resolution of BP/RP spectra, half of the sources agree with literature values of [M/H] to within 0.21. However, we caution that the [M/H] provided by GSP-Phot are systematically too low, which is not obvious from Table 3. While we would expect a systematic underestimation of [M/H] for sources with overestimated extinction due to the degeneracy between these parameters, we also observe that the GSP-Phot [M/H] are too low when the extinctions are adequately estimated. Consequently, the GSP-Phot [M/H] values should only be used with due caution. However, we find that GSP-Phot [M/H] estimates can be empirically calibrated to the [Fe/H] scale of LAMOST DR6[12] (Zhao et al. 2012; Deng et al. 2012; Liu et al. 2015) see also Sect. 11.3.3.6 in the online documentation for details.

We briefly outline this empirical metallicity calibration procedure here: Our objective is to use a multivariate adaptive regression spline (hereafter MARS, Friedman 1991) in order to learn a mapping from GSP-Phot's biased [M/H] to some well-established metallicity estimates. We considered various literature catalogues as possible training samples and eventually opted for LAMOST DR6 because it provides a broad range of metallicity values but does not probe too deeply into high-extinction regions in the Galactic disk.[13] As LAMOST provides [Fe/H] estimates, our MARS model not only needs to remove the systematic errors from the GSP-Phot [M/H] but also to translate from [M/H] to [Fe/H]. As the metallicity bias in GSP-Phot also depends on stellar parameters, the input features of the MARS model include the effective temperature, surface gravity, the biased [M/H] value itself, and the extinction and reddening. It also includes Galactic latitude, which helps with the translation from [M/H] to [Fe/H].[14] The trained MARS model then provides the calibrated [Fe/H].

We test this calibration with FGK members ($T_{\rm eff}$ between 4000 and 6500 K) of open clusters with known metallicities after rejection of stars with poor parallax measurements ($\frac{\varpi}{\sigma_\varpi} < 10$). Cluster members and mean cluster metallicities were taken from Soubiran et al. (2021). Figure 11a shows the individual metallicities [M/H] minus reference [Fe/H] of the parent cluster as a function of the temperature of the star. This test involves nearly 56 000 stars in 187 open clusters. Figure 11b shows that there is a net improvement of the metallicities when the calibration is applied —in both the offset and the dispersion— over the temperature range. However, the systematic errors are not completely removed by the calibration and still depend on the surface gravity (despite the MARS calibration model taking log $g$ as an input feature), as is evident from Fig. 11c. Therefore, we emphasise that our empirical calibration is simply an illustration. The users are explicitly encouraged to find better calibration procedures of their own. Nevertheless, as mentioned in Gaia Collaboration, Creevey et al. (2022), the [M/H] provided by GSP-Phot , together with its estimates of temperature and gravity, can still be used to select solar-like stars whose RVS spectra are in close agreement with those of known solar analogues. Therefore, GSP-Phot [M/H] estimates, in spite of their large systematic errors, still contain some exploitable information about the actual metallicity of the star, which is why they were not removed from the Gaia DR3 release.

This issue of [M/H] discrepancies is likely due to the mismatch between observed BP/RP spectra and the models employed by GSP-Phot (see Sect. 3.1). Given the large differences between observed and model BP/RP spectra shown in Fig. 1c and d, it is not surprising that the metallicity estimates are of poor quality. Metallicity is the weakest parameter in GSP-Phot, in the sense that it has the lowest impact on the shape of BP/RP spectra and so is most affected by model–data mismatches. In particular, the metallicity information is largely encoded at the blue end of the BP spectrum, where Fig. 1c shows large discrepancies between observations and models.

### *3.6. Extinction*

GSP-Phot extinction estimates are validated in various places. For example, Gaia Collaboration, Creevey et al. (2022) select solar-like stars from GSP-Spec results (based on RVS spectra,

[12] `http://dr6.lamost.org/v2/catalogue`

[13] Our preferred solution would have been to train an [M/H] calibration based on GSP-Spec results (Recio-Blanco et al. 2022). Unfortunately, due to GSP-Phot filtering and GSP-Spec flagging, the overlap of both Apsis modules has an insufficient number of low-metallicity stars.

[14] The conversion from [M/H] to [Fe/H] strictly requires knowledge of [$\alpha$/Fe]. While [$\alpha$/Fe] is not available from GSP-Phot, [$\alpha$/Fe] varies between the Galactic plane and high latitudes such that the MARS model can infer an approximate conversion from the Galactic latitude.

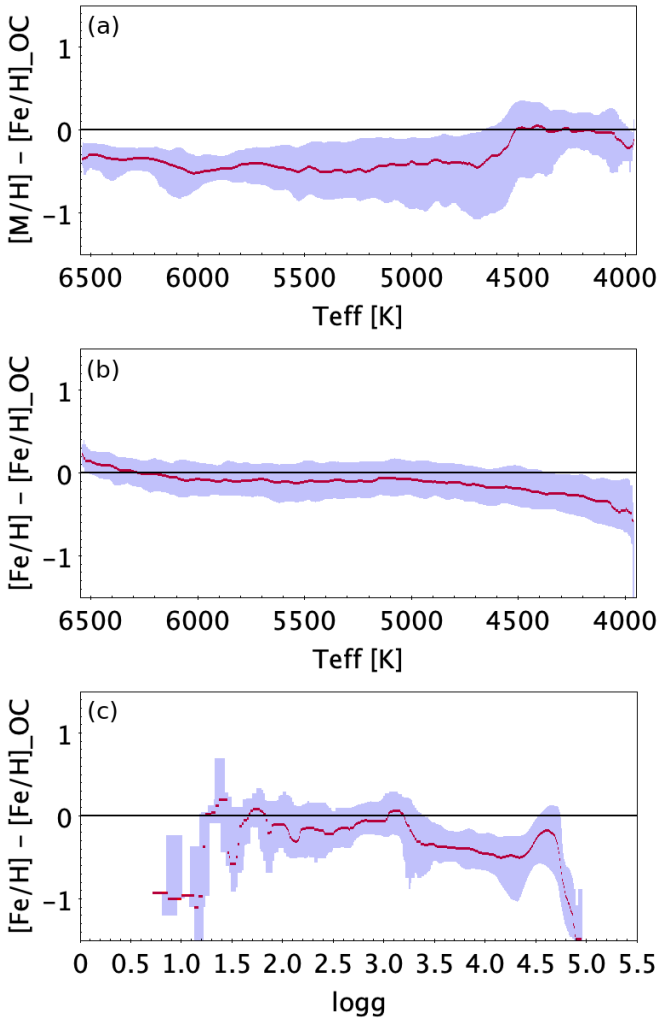


**Fig. 11.** Difference between GSP-Phot metallicities of individual FGK members and the mean [Fe/H] of the parent open cluster for stars with $\frac{\varpi}{\sigma_\varpi} \geq 10$ and MARCS models, before and after the calibration. We note that the calibration on LAMOST [Fe/H] values translates [M/H] to [Fe/H]. The red line is the median of the distribution. The blue area is delimited by the 16th and 84th quantiles. Panel (a): Residuals versus $T_{\rm eff}$ before the calibration. Panel (b): Residuals versus $T_{\rm eff}$ after the calibration. Panel (c): Residuals versus $\log g$ after the calibration.

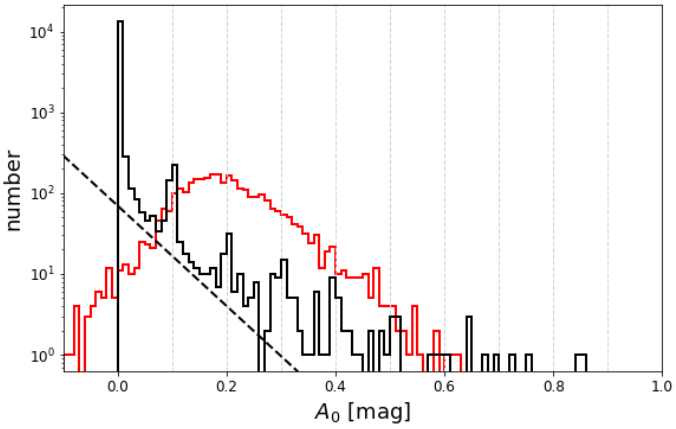


**Fig. 12.** Distribution of $A_0$ estimates for stars in the Local Bubble ($\varpi > 20$ mas). The dashed black line indicates the slope of an exponential with 0.07 mag scale length. The vertical dashed grey lines indicate grid points of $A_0$ from multilinear interpolation of model spectra. (See Appendix B for the ADQL query.) The red histogram shows `av50` extinction estimates from StarHorse2021 (Anders et al. 2022) with `sh_outflag`=0000.

Recio-Blanco et al. 2022) and show that the $G_{\rm BP} - W_2$ colour of those solar-like stars is in close agreement with the linear trend with the GSP-Phot $A_{\rm BP}$ estimate to within 0.087 mag RMS scatter. Gaia Collaboration, Schultheis et al. (2022) find good agreement between GSP-Phot's $E(G_{\rm BP} - G_{\rm RP})$ reddening and the equivalent widths of diffuse interstellar bands (DIBs) measured from Gaia RVS spectra. GSP-Phot extinction estimates are also used to estimate a map of total Galactic extinction and Delchambre et al. (2022) report that this map agrees very well with Planck data for $A_0 < 4$ mag and is also in good agreement with the Schlegel map (Schlegel et al. 1998). In this section, we complement the aforementioned findings with some additional validation results.

### 3.6.1. Local Bubble and non-negativity

GSP-Phot imposes the constraint $A_0 \geq 0$, which reflects the fact that extinctions cannot be negative. This causes GSP-Phot to systematically overestimate $A_0$ in regimes where the actual extinction is very low (also see TGE results in Delchambre et al. 2022). We illustrate this effect with the Local Bubble: Gaia DR3 contains 51 983 sources with parallaxes larger than 20 mas (i.e. closer than 50 pc), which should have very low extinction. Of these, 14 862 have GSP-Phot results[15] and the $A_0$ distribution is shown in Fig. 12. While the average extinction in this sample is $A_0 = 0.1$ mag, the values can be significantly larger. This is expected given that $A_0$ is subject to measurement noise. For a non-negative random variate with a true value of zero, we expect an exponential distribution given that this is the maximum-entropy distribution for such a random variate (e.g. Dowson & Wragg 1973). Indeed, Fig. 12 shows that the $A_0$ distribution is roughly matched by an exponential with a scale length of 0.07 mag. Similarly to Fig. 12, we obtain exponentials with scale lengths of 0.07 mag for $A_{\rm BP}$, 0.06 mag for $A_G$, and 0.05 mag for $A_{\rm RP}$. These also provide rough error estimates, at least for bright sources in the Local Bubble. In particular, the 0.07 mag for $A_{\rm BP}$ is consistent with the 0.087 mag RMS scatter in $A_{\rm BP}$ reported for solar-like stars in Gaia Collaboration, Creevey et al. (2022). However, we note that this is not purely random but includes some systematic errors. Figure 12 shows peaks at the grid points used for the multilinear interpolation of model spectra, in particular around $A_0$ values of 0, 0.1, and 0.2. As discussed in Sect. 3.1, this is most likely a result of the mismatch between models and real BP/RP spectra. Furthermore, the extinction overestimation tends to affect some parts of the CMD more than others (see Babusiaux et al. 2022). In particular, low-mass dwarfs tend to have their extinction overestimated, which may be related to our Hertzsprung-Russell diagram prior using isochrones that do not coincide with the PARSEC isochrones employed by GSP-Phot in this regime (see Sect. 2.5).

[15] The high-quality parallaxes in the Local Bubble make it difficult for GSP-Phot to match the inverse distance within $10\sigma$ of the parallax measurement (see Sect. 3.3), which means many sources are filtered out.

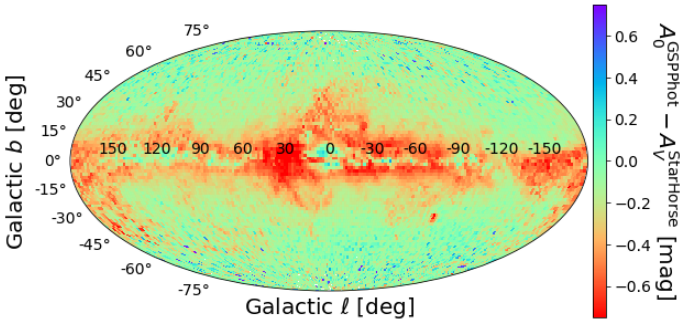


**Fig. 13.** Difference between the GSP-Phot $A_0$ and the `av50` provided by StarHorse2021 with `sh_outflag`=0000 (Anders et al. 2022) on the sky. This skymap uses the Mollweide projection where lines of constant latitude are horizontal straight lines parallel to the equator.

#### 3.6.2. Comparison to StarHorse2021

We briefly compare the GSP-Phot $A_0$ estimate to the `av50` estimate from StarHorse2021 (Anders et al. 2022). Returning to the Local Bubble ($\varpi > 20$ mas), the red histogram in Fig. 12 shows that StarHorse2021 overestimates extinction with a mean `av50` of 0.16 mag, which is about twice as large as the value from GSP-Phot. In particular, despite StarHorse2021 allowing for slightly negative extinctions, the `av50` does not peak near zero.

We also compare extinctions for a random subset of one million stars. On average, `av50` estimated by StarHorse2021 is 0.36 mag higher than the $A_0$ estimated by GSP-Phot. As is evident from Fig. 13, this difference is mainly driven by high-extinction regions. Outside such regions, StarHorse2021 `av50` appears to be about 0.1 mag higher than GSP-Phot $A_0$, which is very similar to the difference we find in the Local Bubble. Anders et al. (2022) report in their Fig. 15 a systematic overestimation of `av50` in open clusters from Cantat-Gaudin et al. (2020) that is consistent with the 0.1 mag difference to the GSP-Phot $A_0$. However, in high-extinction regions, the differences can easily reach 0.7 mag or higher (Fig. 13). This systematic difference cannot be understood as `av50` being the Johnson $V$ band extinction and $A_0$ being the monochromatic extinction at 541.4 nm. Even though the difference between these two extinction concepts becomes more pronounced as the extinction increases, the effect goes in the opposite direction, that is, `av50` should become increasingly smaller than $A_0$, not larger (see Sect. 11.2.3.1.4 in the online documentation for details).

#### 3.6.3. Comparison to Bayesstar19

For further validation, we compare the GSP-Phot $A_0$ extinctions to the $A_V$ extinctions derived from the Bayestar19 3D extinction map (Green et al. 2019). As mentioned above, the GSP-Phot $A_0$ is the monochromatic extinction at 541.4 nm and the parameter in our extinction law (Fitzpatrick 1999). From Bayestar19, $A_V$ is conceptually the closest to GSP-Phot $A_0$. We sample the 3D extinction map from Bayestar19 using the GSP-Phot distance to each star. We define two subsamples for comparison: (1) a randomly selected sample of 1 million sources spread throughout the sky, and (2) all the sources in the direction of the Cygnus X star-formation region, $73 \leq l \leq 87, -4 \leq b \leq 6$; these samples are shown in Fig. 14 and Fig. 15. In both samples, GSP-Phot predicts a higher extinction than Bayestar19, as is evident from the all-sky sample in Fig. 14c as well as Cygnus X in Fig. 15c. The differences between GSP-Phot and Bayestar19 extinctions become larger as the extinctions increase. However, as reported in Delchambre et al. (2022), Bayesstar19 appears to also estimate lower extinctions compared to data from Planck and Schlegel. We note that a comparison between Fig. 14a and b shows that the GSP-Phot map shows finer structures with larger contrast (i.e. higher extinctions). We put forward a possible explanation for this below.

Looking at the Cygnus X star-formation region, GSP-Phot appears to more faithfully recover extinctions towards compact high-density regions when compared to Bayestar19. This is evident from comparing Fig. 15a and b, where regions of significantly higher extinction are visible in GSP-Phot results, tracing the structure of the dense regions of ongoing star formation. In Fig. 15c, we directly compare the extinctions from GSP-Phot and Bayestar19. We see two clear populations of sources: the majority of sources have similar extinctions in the two catalogues, although GSP-Phot is systematically higher, while a smaller population have large extinctions in GSP-Phot but negligible extinction predicted by Bayestar19. The first of these populations can be tied to stars in diffuse regions. The second population, where GSP-Phot predicts large extinctions but Bayestar19 does not, is only seen in regions that have high ISM densities, meaning GSP-Phot successfully recovers stars in regions with large dust density and active star formation while Bayestar19 does not.[16]

We notice that GSP-Phot maps show finer structures and larger extinction values than maps from Bayesstar19 in Fig. 14a and b and Fig. 15a and b. We speculate that while Bayestar19 is capable of detecting high-extinction sources, the Gaussian process model the authors applied essentially smoothes out the map and averages over many lines of sight in its grid pixel. Therefore, a high-extinction line of sight may become averaged down because it occupies a small volume and therefore only affects a small fraction of sources. GSP-Phot on the other hand is not biased by this and recovers an extinction for each source individually, leaving the high-extinction sources as high extinction. While small-scale structures also exist in low-extinction regions, the bias from averaging will be more visible in high-dust-density, active star-formation regions. Those small-scale structures may be washed out by the Gaussian process model employed by Bayestar19.

As an aside, we mention that both Fig. 14c and Fig. 15c show horizontal stripes in the GSP-Phot $A_0$ estimates. These are the same linear interpolation issues that we already observed in the Local Bubble (Fig. 12) and which are most likely caused by the mismatch between models and real BP/RP spectra (see Sect. 3.1).

### 3.7. Systematic underestimation of distances

Fouesneau et al. (2022b) show that GSP-Phot distances of cluster member stars are consistent with the cluster distances from Cantat-Gaudin et al. (2020) only out to 2-3 kpc, and that GSP-Phot distances become systematically too low beyond 3 kpc. Similar results are reported for stars with asteroseismic distances by Huber et al. (2017) and Anders et al. (2017). However, Fouesneau et al. (2022b) also show that GSP-Phot distances are reli-

[16] We double-checked that these stars driving the high extinction values are not low-mass dwarfs at the faint/cool end of the main sequence which can sometimes exhibit spuriously large extinctions, e.g. in the Local Bubble (Babusiaux et al. 2022). Instead, these high-extinction values in Cygnus X1 appear to be driven by red giant stars.

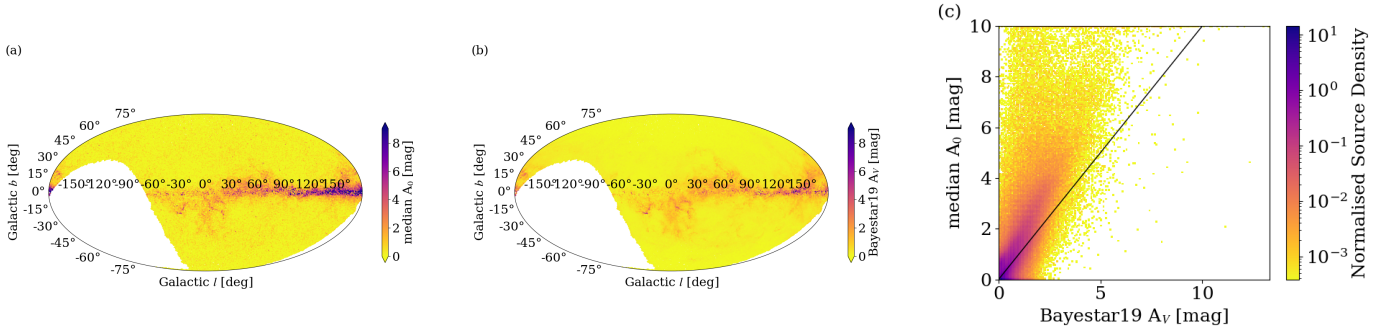


**Fig. 14.** Extinction comparison between GSP-Phot and Bayestar19 for a randomly selected sample of 1 million sources. Panel (a): Skymap of $A_0$ provided by GSP-Phot, taking a pixel-wise median value. Panel (b): Skymap of $A_V$ provided by Bayestar19, taking a pixel-wise median value. Panel (c): One-to-one comparison of the GSP-Phot median $A_0$ and the Bayestar19 $A_V$. Both skymaps use the Mollweide projection where lines of constant latitude are horizontal straight lines parallel to the equator. All panels show the identical sample of stars. The missing data in panels (a) and (b) are due to the footprint of Bayestar19.

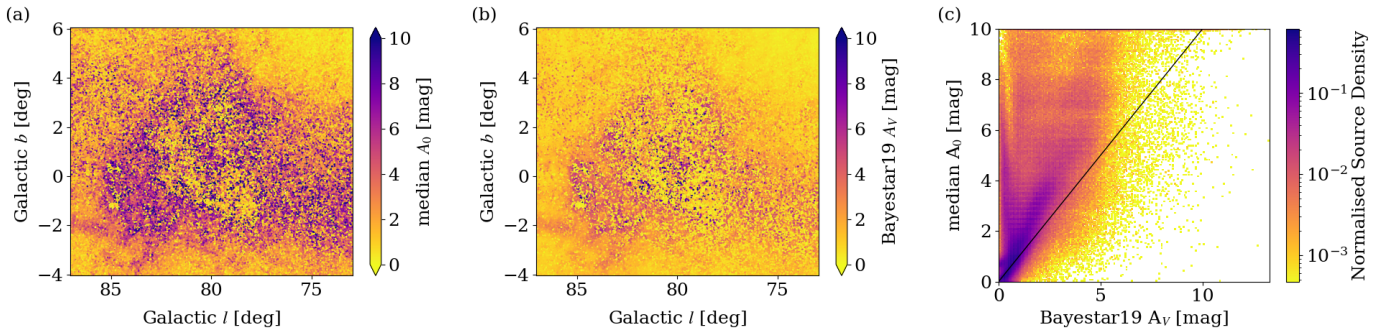


**Fig. 15.** Same as Fig. 14 but for the dense star-formation region Cygnus X1.

able out to 10 kpc for stars with high-quality parallax measurements ($\varpi/\sigma_\varpi \geq 10$). This dependence on parallax quality suggests that the systematic underestimation of distances may be related to the distance prior. As we note in Sect. 2.5, the length scale of the distance prior is set to one-tenth of the length scale that we compute from the Gaia EDR3 mock catalogue of Rybizki et al. (2020). The objective to do so was to reduce the differences to literature values, for example, for effective temperatures. Unfortunately, this distance prior is overly harsh, resulting in a systematic underestimation of distances by GSP-Phot for sources with low parallax quality. This may also compromise other parameters too, such as the $\log g$ estimates of red clump stars in Fig. 9. This view is also supported by Fig. 6, which demonstrates that restricting to high parallax quality stabilises GSP-Phot and, in that case, also improves the temperature estimates.

In order to confirm this interpretation, we locally reprocess 5 million sources with the length scale of the distance prior relaxed by a factor of ten, that is, restoring the value we compute from the Gaia EDR3 mock catalogue of Rybizki et al. (2020). Unfortunately, these 5 million sources are not representative of the sample as a whole, but other sources are not available for this exercise for various reasons. Figure 16 shows that while the GSP-Phot distances in Gaia DR3 do not follow the inverse parallax distribution very well, the situation clearly improves when we relax the distance prior. We caution though that inverting parallaxes is not recommended (e.g. Bailer-Jones 2015), in particular given that many sources in this sample may have very noisy parallax measurements. Nevertheless, given this systematic underestimation of distances in Gaia DR3 for stars with low parallax quality, GSP-Phot distances cannot be used to map the Milky Way spiral arms (Gaia Collaboration, Drimmel et al. 2022), the spatial distributions of the diffuse-interstellar-band absorption (Gaia Collaboration, Schultheis et al. 2022), or chemical cartography (Gaia Collaboration, Recio-Blanco et al. 2022).

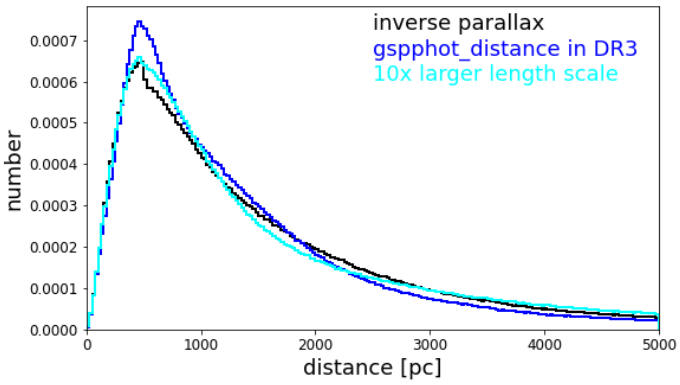


**Fig. 16.** Distributions of inverse parallax (black histogram), GSP-Phot distances in Gaia DR3 with overly harsh distance prior (blue histogram), and GSP-Phot distances after reprocessing with a relaxed distance prior (cyan histogram). The parallaxes shown here and also used during the GSP-Phot processing all include the zero-point correction from Lindegren et al. (2021a).

### 3.8. Further validation results

As mentioned above, GSP-Phot results are validated in various publications accompanying Gaia DR3. Here, we want to briefly highlight some of these findings: In Fouesneau et al. (2022b), we show that using the radius and distance from GSP-Phot, we can predict angular diameters that are in excellent agreement with measurements from ground-based interferometry. In Creevey et al. (2022), we show that results from GSP-Phot and FLAME are in very good agreement for radii, luminosity, and bolometric correction. We also report relatively good agreement in terms of effective temperatures and extinctions between GSP-Phot and ESP-HS for hot stars ($T_{\rm eff} > 7500$K), which is used for the OB sample definitions in Gaia Collaboration, Drimmel et al. (2022) and Gaia Collaboration, Creevey et al. (2022). Recio-Blanco et al. (2022) report good agreement between results from GSP-Phot (low-resolution BP/RP spectra) and GSP-Spec (RVS spectra) for effective temperatures and surface gravities. Gaia Collaboration, Creevey et al. (2022) demonstrate that when selecting solar-like stars from GSP-Spec results, the colours of the resulting candidates are in good agreement with those of known solar twins for stars where $A_0 < 0.001$mag according to GSP-Phot. Furthermore, BP/RP spectra of solar-like stars exhibit a clear dimming and reddening trend with increasing $A_0$. Gaia Collaboration, De Ridder et al. (2022) find that temperature uncertainties from GSP-Phot are too small by a factor of approximately 4 for $\delta$Scuti and $\gamma$Doradus stars (spectral type early-F to mid-A), as well as for hotter variable stars, such as SPB or $\beta$Cephei (spectral type B9 or hotter). Finally, Babusiaux et al. (2022) provide an overview of the main issues identified in Gaia DR3 data, including GSP-Phot results.

### 3.9. Uncertainty validation

It is conceptually very challenging to validate uncertainty estimates because we not only need reliable reference values but also reliable uncertainties on these reference values. In Fig. 10 we can validate our $\log g$ uncertainties assuming that measurement errors in asteroseismic gravities are negligible compared to GSP-Phot uncertainties (e.g. Creevey et al. 2013).

We further validate our uncertainty estimates by employing the BP/RP split-epoch validation dataset introduced in De Angeli et al. (2022): first, we go back to the epoch BP/RP spectra of each source[17] and randomly group them into two sets; second, for both sets of epoch spectra we compute a mean BP/RP spectrum. This procedure provides two statistically independent BP/RP spectra for each source. Each BP/RP spectrum now only has half of the epochs of the actual source, and so this procedure produces spectra with slightly lower signal-to-noise ratio. As both spectra belong to the same source, the parameters obtained from processing both BP/RP spectra with GSP-Phot must be consistent with each other within their respective uncertainties. This should even be the case for intrinsically variable sources because the splitting of epoch spectra is done randomly. We note that we do not need to know the true parameters of each source; it is sufficient to know that the two randomly split spectra belong to the same source. As it is too time-consuming to perform this test for all sources, we can only do it for a small sample of 17 994 sources for which the necessary epoch BP/RP spectra were still available. This sample is not representative but still covers the apparent $G$ magnitude range reasonably well. We also require that GSP-Phot results for both components pass the filters described in Sect. 3.3.

[17] In Gaia DR3, each source typically has ∼40 epoch BP/RP spectra, but this can vary between 10 and over 150.

**Table 4.** Inflation factors necessary to make the uncertainty intervals of randomly split pairs overlap in about 84% of cases inferred from Fig. 18.

| parameter | $T_{\rm eff}$ | $A_0$ | $\log g$ | [M/H] | distance | $M_G$ |
|---|---|---|---|---|---|---|
| factor | 2.0 | 1.8 | 2.5 | 3.4 | 1.6 | 1.9 |

As a first test, Fig. 17 simply compares the estimates within each pair. The median absolute differences are 41 K for effective temperature, 0.041 mag for extinction $A_0$, 0.031 for gravity, 0.058 for metallicity, 0.081 mag for absolute $M_G$ magnitude, and 2.7% for distance. These differences are purely due to random noise and are much smaller than the differences with respect to literature values (e.g. the median absolute differences between temperatures and literature values range from 110 to 170 K in Table 1). This suggests that the differences from literature values are not driven by random errors but are rather dominated by systematic errors such as different temperature scales in GSP-Phot and other surveys.

If we compare the differences between the parameters to their uncertainty estimates, we find that, for 30%-60% of randomly split pairs, the parameters of the two components are outside each other's 68% confidence intervals. In reality, this should only happen in 16% of cases, which suggests that our uncertainties are systematically underestimated.

In order to assess by how much our uncertainties are underestimated, Fig. 18 investigates how the fraction of non-overlapping uncertainty intervals from GSP-Phot between the pairs decreases as we inflate the uncertainty intervals. The inflation is done by applying the same factor to the asymmetric intervals on both sides of each parameter. Table 4 suggests that the uncertainties are systematically too small by factors ranging from 1.6 for distance to 3.4 for metallicity. However, we have to caution that the 17 994 sources from the BP/RP split-epoch validation dataset may not be large enough or representative enough, meaning that the values in Table 4 can only provide a rough indication. In particular, for asteroseismic gravities, Fig. 10 suggests that our uncertainties are too small by a factor of 10. Given this somewhat unclear situation, we do not apply any correction to the GSP-Phot uncertainties in the published Gaia DR3 data.

### 3.10. Temperature–extinction degeneracy

As a final use case of the BP/RP split-epoch validation dataset, Fig. 19 clearly illustrates the degeneracy between effective temperature and line-of-sight extinction by comparing the differences in parameters for each pair. This degeneracy originates from the fact that low-resolution, optical BP/RP spectra are very similarly affected by both parameters. Figure 19 also shows that the temperature–extinction degeneracy affects dwarfs and giants in different ways: while dwarfs can exhibit temperature variations as large as several hundred Kelvin even for small extinction variations, giants usually exhibit smaller temperature variations that are accompanied with much larger extinction variations. We note in particular that the temperature–extinction degeneracy works both ways in Fig. 19, causing simultaneous underestimation of $T_{\rm eff}$ and $A_0$ just as frequently as a simultaneous overestimation.

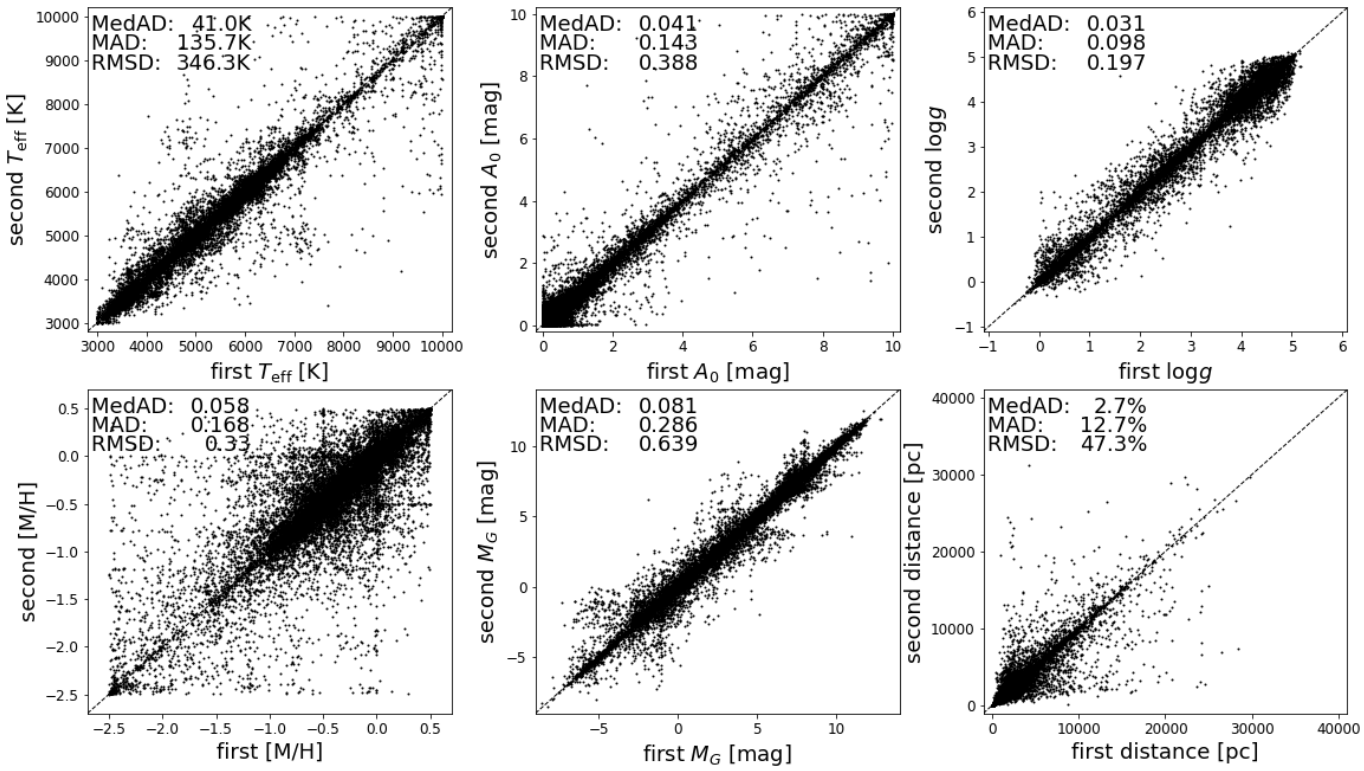


**Fig. 17.** Comparison of parameters between two components of 17 994 sources from the BP/RP split-epoch validation dataset. Quoted numbers summarise the median absolute difference (MedAD), the mean absolute difference (MAD), and the root-mean-square difference (RMSD).

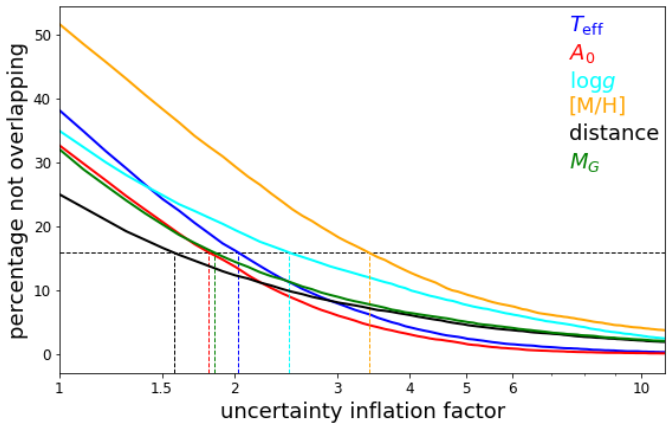


**Fig. 18.** Percentage of parameter pairs (in BP/RP split-epoch validation dataset) outside each other's 68% confidence interval as a function of uncertainty inflation factor. The horizontal black dashed line at 16% shows the expected percentage if uncertainties were correctly estimated. Vertical colour dashed lines indicate the inflation factors necessary for each parameter in order to bring the percentage of non-overlapping intervals to the expected 16% level.

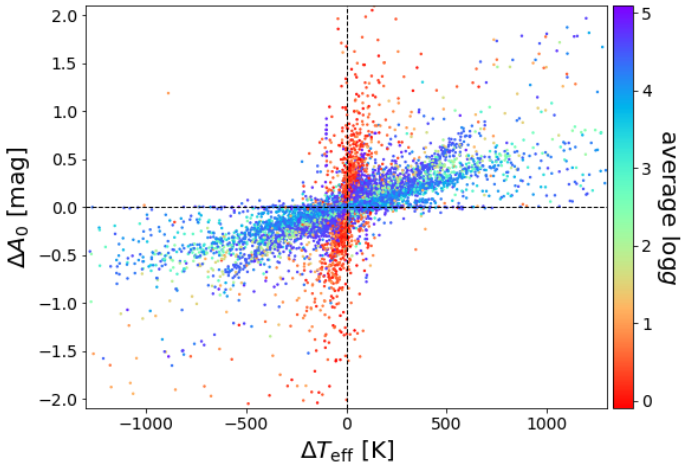


**Fig. 19.** Illustration of the temperature–extinction degeneracy using the parameter differences between pairs from the BP/RP split-epoch validation dataset (c.f. Sect. 3.9). Colour-coding is done by the average surface gravity of both components.

## 4. Summary

In Gaia DR3, one of the major new data products is a collection of 220 million low-resolution BP/RP spectra. In this paper, we explain how the CU8 Apsis module GSP-Phot provides a homogeneous catalogue of stellar-parameter estimates for 471 million sources with $G < 19$ based on these BP/RP spectra, parallax, and integrated photometry. We emphasise that GSP-Phot assumes that each source is a single star and that using combined BP/RP spectra implies that any intrinsic time variability is lost (De Angeli et al. 2022). One of the main design features of GSP-Phot is to not normalise the BP/RP spectra but instead to exploit the apparent flux level of the BP/RP spectra as an observational constraint on the radius and distance of the star (see Eq. (2)). GSP-Phot also employs PARSEC isochrones in order to guarantee astrophysically self-consistent stellar temperatures, gravities, metallicities, radii, and absolute magnitudes (see Sect. 2.3).

However, in Gaia DR3, GSP-Phot does not directly account for parallax and apparent $G$ magnitude when solving for the dis-

tance from the amplitude of BP/RP spectra (parallax and apparent magnitude still enter the GSP-Phot likelihood function in Eq. (5)). While this will be fixed in the future, in Gaia DR3 this requires filtering out GSP-Phot results where the distance is inconsistent with the measured parallax or the apparent $G$ magnitude (see Sect. 3.3). In particular, this filtering removes GSP-Phot results for virtually all white dwarfs. Furthermore, GSP-Phot results have largely been filtered out for sources with negative parallaxes. Otherwise, this filtering does not appear to specifically affect any particular stellar population (see Fig. 2) and GSP-Phot results are usually complete at the 80% level except for sources brighter than $G < 13$ or sources with extremely high parallax qualities (see Fig. 3).

Despite the filtering, the GSP-Phot results remaining in Gaia DR3 can still suffer from several systematic effects (an overview of the main issues identified in the Gaia DR3 data is given by Babusiaux et al. 2022):

1. GSP-Phot systematically underestimates distances for sources with low parallax quality, which applies to most sources in Gaia DR3. As discussed in Sect. 3.7, this is due to an overly harsh distance prior employed by GSP-Phot. Sources with high-quality parallax measurements ($\varpi/\sigma_\varpi > 20$) should have reliable distances.
2. GSP-Phot imposes a non-negativity constraint on extinction. As a result, in low-extinction regions, GSP-Phot tends to overestimate extinction (e.g. in the Local Bubble as shown in Fig. 12).
3. The [M/H] estimates from GSP-Phot are dominated by large systematic errors which reduce them to the level of qualitative information. We therefore advise against using them. However, the [M/H] estimates provided by GSP-Phot are still sufficiently informative that they can be empirically calibrated onto the LAMOST DR6 [Fe/H] scale, as we illustrate in Fig. 11.
4. Given that BP/RP spectra have very low resolution and only cover the wavelength range from 320 to 1050 nm, there is a degeneracy between effective temperature and line-of-sight extinction: Increasing the star's effective temperature can be compensated by simultaneously increasing the line-of-sight extinction, thereby producing very similar BP/RP spectra. The strength of this temperature–extinction degeneracy varies with stellar population: while main sequence dwarfs can exhibit temperature variations as large as several hundred Kelvin even for small extinction variations, red giant stars usually exhibit smaller temperature variations that are accompanied with much larger extinction variations (see Fig. 19). This is a major limitation for GSP-Phot. We also point out that the temperature–extinction degeneracy, in principle, works both ways, but in low-extinction regimes tends to primarily cause an overestimation of temperatures and extinctions simply because the non-negativity constraint leaves no room to underestimate extinction.

Another fundamental limitation of GSP-Phot in Gaia DR3 is a mismatch between observed BP/RP spectra and models. We illustrate this using solar-like stars in Fig. 1. This mismatch is likely responsible for the poor quality of GSP-Phot [M/H] estimates, given that metallicity has the weakest impact on the shape of BP/RP spectra and is therefore easiest to compromise. At the aesthetic level, this mismatch also causes stripes in GSP-Phot results (e.g. Fig. 15c). As we discuss in Sect. 3.1, this mismatch is unlikely to originate from the CU5 instrument model, but rather different solar model SEDs result in BP/RP spectra whose differences are easily measurable from Gaia DR3 data. While this is unfortunate for GSP-Phot, it will allow the community to further refine stellar atmospheric models.

Given all the aforementioned limitations, GSP-Phot results still compare well to expected values: In a comparison with GALAH DR3 and LAMOST DR4, half of the stars have temperatures that deviate by less than 110 K from the literature values (see Table 1). The differences are larger for APOGEE DR16, which probes deeper into distant stars in the high-extinction regimes of the Galactic disk. If we restrict the comparison to APOGEE values to high-quality parallaxes ($\varpi/\sigma_\varpi > 20$), we obtain results that are just as good as for GALAH DR3 or LAMOST DR4. Concerning surface gravities, half of the stars deviate by less than 0.25 from literature values (see Table 2). A comparison to asteroseismic gravities confirms a median absolute difference of 0.2 (see Fig. 10a). Concerning extinctions, the Local Bubble suggests typical uncertainties of 0.07 mag in $A_0$ and $A_{\rm BP}$, and slightly lower uncertainties of 0.06 mag in $A_G$ and 0.05 mag in $A_{\rm RP}$, reflecting the different susceptibilities to extinction of each band. This also agrees with the scatter of 0.087 mag in $A_{\rm BP}$ that we find in solar-like stars in Gaia Collaboration, Creevey et al. (2022). A comparison to StarHorse2021 shows that, firstly, there is a global offset of `av50` that is about 0.1 mag larger than the GSP-Phot $A_0$ (e.g. at high latitudes and in the Local Bubble). This offset is likely due to a systematic overestimation of `av50`, which is also evident from the comparison to open clusters made by Anders et al. (2022). Secondly, in high-extinction regions, the StarHorse2021 `av50` can be substantially larger than the GSP-Phot $A_0$ (see Fig. 13). This effect cannot be explained by the different definitions of `av50` and $A_0$ (which should work in the exact opposite direction). The systematic differences between GSP-Phot and StarHorse2021 extinction estimates are currently not understood.

We caution that the uncertainty estimates from GSP-Phot tend to be much smaller than the typical differences from reference values. Validation of uncertainties is very difficult. In particular, a simple comparison to literature values is insufficient because the literature values also often have underestimated uncertainties. In this work, we circumvent this problem by producing two statistically independent incarnations of a limited sample of stars (the BP/RP split-epoch validation dataset discussed in Sect. 3.9). For these, we do not need to know their true stellar parameters. Instead, it is sufficient to know that both incarnations represent the exact same star, such that the GSP-Phot results for both should be consistent within their respective uncertainties. Unfortunately, we find that this is not the case, that is, GSP-Phot uncertainties are systematically underestimated by factors ranging from 1.6 to 3.4 (see Table 4). Likewise, uncertainties on surface gravities appear to be underestimated by a factor of ∼10, as is evident from a comparison to asteroseismic values (see Fig. 10b). There are multiple reasons why GSP-Phot underestimates uncertainties: Firstly, as mentioned above, some priors are too harsh and thereby may overly restrict the fit procedure. Secondly, the CU8 Apsis chain ignores correlations between pixels even though they exist (e.g. Creevey et al. 2022). Finally, the aforementioned mismatch between observed BP/RP spectra and models not only causes systematic errors, but when the fit struggles to make the models match the observed data, it usually also leads to an underestimation of uncertainties.

*Acknowledgements.* This work presents results from the European Space Agency (ESA) space mission Gaia. Gaia data are being processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the Gaia MultiLateral Agreement (MLA). The Gaia mission website is

https://www.cosmos.esa.int/gaia. The Gaia archive website is https://archives.esac.esa.int/gaia. Acknowledgements are given in Appendix C

## References


Anders, F., Chiappini, C., Rodrigues, T. S., et al. 2017, A&A, 597, A30
Anders, F., Khalatyan, A., Queiroz, A. B. A., et al. 2022, A&A, 658, A91
Babusiaux, C., Fabricius, C., & et al. 2022, A&A, submitted
Bailer-Jones, C. A. L. 2010, MNRAS, 403, 96
Bailer-Jones, C. A. L. 2011, MNRAS, 411, 435
Bailer-Jones, C. A. L. 2015, PASP, 127, 994
Bailer-Jones, C. A. L., Andrae, R., Arcay, B., et al. 2013, A&A, 559, A74
Bohlin, R. C., Colina, L., & Finley, D. S. 1995, AJ, 110, 1316
Bohlin, R. C., Gordon, K. D., & Tremblay, P. E. 2014, PASP, 126, 711
Bohlin, R. C., Hubeny, I., & Rauch, T. 2020, AJ, 160, 21
Bovy, J., Nidever, D. L., Rix, H.-W., et al. 2014, ApJ, 790, 127
Buder, S., Sharma, S., Kos, J., et al. 2021, MNRAS[arXiv:2011.02505]
Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. 2020, A&A, 640, A1
Carrasco, J. M., Weiler, M., Jordi, C., et al. 2021, A&A, 652, A86
Chen, Y., Bressan, A., Girardi, L., et al. 2015, MNRAS, 452, 1068
Creevey, O., Sordo, R., Pailler, F., et al. 2022, A&A, submitted
Creevey, O. L., Thévenin, F., Basu, S., et al. 2013, MNRAS, 431, 2419
Czekala, I., Andrews, S. M., Mandel, K. S., Hogg, D. W., & Green, G. M. 2015, ApJ, 812, 128
De Angeli, F., Weiler, M., Montegriffo, P., & et al. 2022, A&A, submitted
Delchambre, L., Bailer-Jones, C., Bellas-Velidis, I., Drimmel, R., & et al. 2022, A&A, submitted
Deng, L.-C., Newberg, H. J., Liu, C., et al. 2012, Research in Astronomy and Astrophysics, 12, 735
Dowson, D. & Wragg, A. 1973, IEEE Transactions on Information Theory, 19, 689
Fitzpatrick, E. L. 1999, PASP, 111, 63
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Fouesneau, M., Andrae, R., Dharmawardena, T., et al. 2022a, arXiv e-prints, arXiv:2201.03252
Fouesneau, M., Frémat, Y., Andrae, R., Korn, A., & et al. 2022b, A&A, in prep.
Friedman, J. H. 1991, The Annals of Statistics, 19, 1
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
Gaia Collaboration, Creevey, O., Sarro, L., Lobel, A., & et al. 2022, A&A, submitted
Gaia Collaboration, De Ridder, J., Ripepi, V., Aerts, C., & et al. 2022, A&A, submitted
Gaia Collaboration, Drimmel, R., Romero-Gomez, M., Chemin, L., & et al. 2022, A&A, in prep.
Gaia Collaboration, Recio-Blanco, A., Kordopatis, G., de Laverny, P., & et al. 2022, A&A, submitted
Gaia Collaboration, Schultheis, M., Zhao, H., Zwitter, T., & et al. 2022, A&A, accepted
Gaia Collaboration, Vallenari, A., Brown, A., Prusti, T., & et al. 2022, A&A, in prep.
Galarza, J. Y., López-Valdivia, R., Lorenzo-Oliveira, D., et al. 2021, MNRAS, 504, 1873
Geurts, P., Ernst, D., & Wehenkel, L. 2006, Machine Learning, 63, 3
Green, G. M., Schlafly, E., Zucker, C., Speagle, J. S., & Finkbeiner, D. 2019, ApJ, 887, 93
Huber, D., Zinn, J., Bojsen-Hansen, M., et al. 2017, ApJ, 844, 102
Jönsson, H., Holtzman, J. A., Allende Prieto, C., et al. 2020, AJ, 160, 120
Lindegren, L., Bastian, U., Biermann, M., et al. 2021a, A&A, 649, A4
Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021b, A&A, 649, A2
Liu, C., Bailer-Jones, C. A. L., Sordo, R., et al. 2012, MNRAS, 426, 2463
Liu, X.-W., Zhao, G., & Hou, J.-L. 2015, Research in Astronomy and Astrophysics, 15, 1089
Montegriffo, P., De Angeli, F., Andrae, R., Riello, M., & et al. 2022, A&A, submitted
Palacios, A., Gebran, M., Josselin, E., et al. 2010, A&A, 516, A13
Pastorelli, G., Marigo, P., Girardi, L., et al. 2020, MNRAS, 498, 3283
Recio-Blanco, A., de Laverny, P., Palicio, P., Kordopatis, G., & et al. 2022, A&A submitted
Riello, M., De Angeli, F., Evans, D. W., et al. 2021, A&A, 649, A3
Robin, A. C., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100
Rybizki, J., Demleitner, M., Bailer-Jones, C., et al. 2020, PASP, 132, 074501
Sartoretti, P., Marchal, O., Babusiaux, C., Jordi, C., & et al. 2022, A&A, submitted
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Schönrich, R. & Bergemann, M. 2014, MNRAS, 443, 698
Seabroke, G., Sartoretti, P., & et al. 2022, A&A, in prep.
Serenelli, A., Johnson, J., Huber, D., et al. 2017, ApJS, 233, 23
Soubiran, C., Brouillet, N., & Casamiquela, L. 2021, arXiv e-prints, arXiv:2112.07545
Steinmetz, M., Guiglion, G., McMillan, P. J., et al. 2020, AJ, 160, 83
Tang, J., Bressan, A., Rosenfield, P., et al. 2014, MNRAS, 445, 4287
Wolpert, R. L. & Schmidler, S. C. 2012, Statistica Sinica, 22, 1233
Wu, Y., Du, B., Luo, A., Zhao, Y., & Yuan, H. 2014, in IAU Symposium, Vol. 306, Statistical Challenges in 21st Century Cosmology, ed. A. Heavens, J.-L. Starck, & A. Krone-Martins, 340–342
Wu, Y., Luo, A.-L., Li, H.-N., et al. 2011, Research in Astronomy and Astrophysics, 11, 924
Yu, J., Huber, D., Bedding, T. R., et al. 2018, ApJS, 236, 42
Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, Research in Astronomy and Astrophysics, 12, 723



[1] Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
[2] INAF - Osservatorio astronomico di Padova, Vicolo Osservatorio 5, 35122 Padova, Italy
[3] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom
[4] INAF - Osservatorio Astrofisico di Torino, via Osservatorio 20, 10025 Pino Torinese (TO), Italy
[5] Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen Ø, Denmark
[6] DXC Technology, Retortvej 8, 2500 Valby, Denmark
[7] IRAP, Université de Toulouse, CNRS, UPS, CNES, 9 Av. colonel Roche, BP 44346, 31028 Toulouse Cedex 4, France
[8] Observational Astrophysics, Division of Astronomy and Space Physics, Department of Physics and Astronomy, Uppsala University, Box 516, 751 20 Uppsala, Sweden
[9] Laboratoire d'astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy Saint-Hilaire, 33615 Pessac, France
[10] GEPI, Observatoire de Paris, Université PSL, CNRS, 5 Place Jules Janssen, 92190 Meudon, France
[11] ATG Europe for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[12] CIGUS CITIC - Department of Computer Science and Information Technologies, University of A Coruña, Campus de Elviña s/n, A Coruña, 15071, Spain
[13] National Observatory of Athens, I. Metaxa and Vas. Pavlou, Palaia Penteli, 15236 Athens, Greece
[14] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France
[15] INAF - Osservatorio Astrofisico di Catania, via S. Sofia 78, 95123 Catania, Italy
[16] Telespazio for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[17] Dpto. de Matemática Aplicada y Ciencias de la Computación, Univ. de Cantabria, ETS Ingenieros de Caminos, Canales y Puertos, Avda. de los Castros s/n, 39005 Santander, Spain
[18] CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[19] Centre for Astrophysics Research, University of Hertfordshire, College Lane, AL10 9AB, Hatfield, United Kingdom
[20] Institut d'Astrophysique et de Géophysique, Université de Liège, 19c, Allée du 6 Août, B-4000 Liège, Belgium
[21] APAVE SUDEUROPE SAS for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[22] Theoretical Astrophysics, Division of Astronomy and Space Physics, Department of Physics and Astronomy, Uppsala University, Box 516, 751 20 Uppsala, Sweden
[23] Royal Observatory of Belgium, Ringlaan 3, 1180 Brussels, Belgium
[24] European Space Agency (ESA), European Space Astronomy Centre (ESAC), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[25] Data Science and Big Data Lab, Pablo de Olavide University, 41013, Seville, Spain

[26] Department of Astrophysics, Astronomy and Mechanics, National and Kapodistrian University of Athens, Panepistimiopolis, Zografos, 15783 Athens, Greece

[27] Dipartimento di Fisica e Astronomia ""Ettore Majorana"", Università di Catania, Via S. Sofia 64, 95123 Catania, Italy

[28] LESIA, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Université de Paris, 5 Place Jules Janssen, 92190 Meudon, France

[29] Université Rennes, CNRS, IPR (Institut de Physique de Rennes) - UMR 6251, 35000 Rennes, France

[30] Aurora Technology for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain

[31] CIGUS CITIC, Department of Nautical Sciences and Marine Engineering, University of A Coruña, Paseo de Ronda 51, 15071, A Coruña, Spain

[32] IPAC, Mail Code 100-22, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA

[33] Thales Services for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France

[34] Dpto. de Inteligencia Artificial, UNED, c/ Juan del Rosal 16, 28040 Madrid, Spain

[35] Institute of Global Health, University of Geneva

[36] Applied Physics Department, Universidade de Vigo, 36310 Vigo, Spain

[37] Sorbonne Université, CNRS, UMR7095, Institut d'Astrophysique de Paris, 98bis bd. Arago, 75014 Paris, France

[38] European Space Agency (ESA), European Space Astronomy Centre (ESAC), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain

## Appendix A: Technical details

### Appendix A.1: Deriving the prior distributions

In this Appendix, we derive the factorised priors step by step. Again, we emphasise that while priors are usually defined for actual fit parameters, they can just as well be defined for derived parameters instead. This may be uncommon but we find it easier to impose a prior in the Hertzsprung-Russell diagram (temperature vs. absolute magnitude, both derived parameters) than over the fit parameters of initial mass and age. For a given BP/RP spectrum $\boldsymbol{s}$, apparent $G$ magnitude, and parallax $\varpi$, the following posterior distribution is provided in Gaia DR3:

$$
\begin{aligned}
&P(T_{\rm eff}, \log g, [M/H], A_0, A_G, d, R, M_G|\boldsymbol{s}, G, \varpi) \\
&= \int d\log_{10}\tau \int d\log_{10}\mathcal{M} \\
&\quad P(\log_{10}\tau, \log_{10}\mathcal{M}, A_G, T_{\rm eff}, \log g, [M/H], A_0, A_G, d, R, M_G|\boldsymbol{s}, G, \varpi).
\end{aligned} \tag{A.1}
$$

We note that while the MCMC sampling itself makes use of log-age, $\log_{10}\tau$, and log-initial mass, $\log_{10}\mathcal{M}$, these values are not actually provided in Gaia DR3, which is why they are marginalised out from the user perspective. Nevertheless, these parameters are still necessary for the MCMC in order to establish astrophysically consistent relations, for example between temperature and radius. Applying Bayes' theorem, we obtain:

$$
\begin{aligned}
&P(T_{\rm eff}, \log g, [M/H], A_0, A_G, d, R, M_G|\boldsymbol{s}, G, \varpi) \\
&\propto \int d\log_{10}\tau \int d\log_{10}\mathcal{M} \\
&\quad P(\boldsymbol{s}, G, \varpi|\log_{10}\tau, \log_{10}\mathcal{M}, A_G, T_{\rm eff}, \log g, [M/H], A_0, d, R, M_G) \\
&\quad \cdot P(\log_{10}\tau, \log_{10}\mathcal{M}, A_G, T_{\rm eff}, \log g, [M/H], A_0, d, R, M_G) \\
&= \int d\log_{10}\tau \int d\log_{10}\mathcal{M} \\
&\quad P(\boldsymbol{s}, G, \varpi|\log_{10}\tau, \log_{10}\mathcal{M}, A_G, T_{\rm eff}, \log g, [M/H], A_0, d, R, M_G) \\
&\quad \cdot P(A_G|\log_{10}\tau, \log_{10}\mathcal{M}, T_{\rm eff}, \log g, [M/H], A_0, d, R, M_G) \\
&\quad \cdot P(A_0|\log_{10}\tau, \log_{10}\mathcal{M}, T_{\rm eff}, \log g, [M/H], d, R, M_G) \\
&\quad \cdot P(d|\log_{10}\tau, \log_{10}\mathcal{M}, T_{\rm eff}, \log g, [M/H], R, M_G) \\
&\quad \cdot P(R|\log_{10}\tau, \log_{10}\mathcal{M}, [M/H], T_{\rm eff}, \log g, M_G) \\
&\quad \cdot P(\log_{10}\tau, \log_{10}\mathcal{M}, [M/H], T_{\rm eff}, \log g, M_G).
\end{aligned}
$$

The first factor is the likelihood of the observables. We can safely assume that the observables $\boldsymbol{s}$, $G$, and $\varpi$ are statistically independent measurements by the Gaia satellite, such that their likelihoods factorise. Dropping all irrelevant dependencies, we therefore obtain:

$$
\begin{aligned}
&P(T_{\rm eff}, \log g, [M/H], A_0, A_G, d, R, M_G|\boldsymbol{s}, G, \varpi) \\
&\propto \int d\log_{10}\tau \int d\log_{10}\mathcal{M} \\
&\quad P(\boldsymbol{s}|T_{\rm eff}, \log g, [M/H], A_0, d, R) \cdot P(G|A_G, d, M_G) \cdot P(\varpi|d) \\
&\quad \cdot P(A_G|T_{\rm eff}, \log g, [M/H], A_0) \\
&\quad \cdot P(A_0|d) \\
&\quad \cdot P(d) \\
&\quad \cdot P(R|\log_{10}\tau, \log_{10}\mathcal{M}, [M/H]) \\
&\quad \cdot P([M/H], T_{\rm eff}, \log g, M_G, \log_{10}\tau, \log_{10}\mathcal{M}).
\end{aligned} \tag{A.2}
$$

This is the fully simplified posterior that is optimised by the Aeneas MCMC.

### Appendix A.2: MCMC configuration

As explained in Sect. 2.7, GSP-Phot makes use of the `emcee` algorithm (Foreman-Mackey et al. 2013). For the ensemble size, we choose 100 walkers in order to explore the 4D parameter space of age, initial mass, metallicity, and $A_0$. We then set up a procedure which we find minimises the risk of the `emcee` getting stuck in the next best local optimum: First, we initialise the `emcee` ensemble in a small ball around the initial guess and let it expand for 50 iterations. After these initial 50 iterations we repeat the following procedure five times:

1. From all previous samples (not only the last ensemble state), identify the 100 best samples, having the highest posterior probability (without repetition of samples).
2. Re-initialise the EMCEE ensemble with these 100 best walkers. Erase previous `emcee` history.
3. Run for 25 iterations.

After this procedure, we assume that the `emcee` ensemble has converged and we run it for another 145 iterations. From this final phase, we start from the last ensemble state (100 samples) and work backwards through the MCMC chain taking an ensemble snapshot every 7th iteration until we have gathered a total of 2000 samples. These 2000 samples are then used to estimate the reported median values and confidence intervals.

We note that a thin-out factor of 7 is most likely insufficient to guarantee absence of autocorrelations in the samples. Likewise, after the fifth and last clipping of the `emcee` ensemble we only have 36 iterations before taking the first ensemble snapshot for inference, which is not always sufficient to guarantee relaxation. As explained in Sect. 2.7, these choices are the results of limited computational resources. Experiments with longer MCMC chains and more ensemble walkers only show a mild improvement of scientific results, i.e. this is no major limitation.

## Appendix B: Example ADQL queries

The following query produces the random sample used in Fig. 4 and Fig. 5.

```
SELECT
gaia.source_id,
gaia.parallax,
gaia.parallax_error,
gaia.phot_g_mean_mag,
gaia.phot_bp_mean_mag,
gaia.phot_rp_mean_mag,
gaia.teff_gspphot,
gaia.logg_gspphot,
gaia.ebpminrp_gspphot,
apsis.mg_gspphot
FROM (
SELECT
source_id,parallax,parallax_error,
phot_g_mean_mag,phot_bp_mean_mag,phot_rp_mean_mag,
teff_gspphot,logg_gspphot,ebpminrp_gspphot
FROM user_dr3int5.gaia_source
WHERE random_index<10000000
AND teff_gspphot IS NOT NULL
) AS gaia
JOIN user_dr3int5.astrophysical_parameters AS apsis
ON gaia.source_id=apsis.source_id
```

The query below produces the sample of the Local Bubble used in Fig. 12.

```
SELECT
gaia.source_id,
gaia.parallax,
gaia.parallax_error,
gaia.phot_g_mean_mag,
gaia.phot_bp_mean_mag,
gaia.phot_rp_mean_mag,
apsis.azero_gspphot,
apsis.ag_gspphot,
apsis.abp_gspphot,
apsis.arp_gspphot
FROM (
    SELECT
    source_id,parallax,parallax_error,
    phot_g_mean_mag,
    phot_bp_mean_mag,phot_rp_mean_mag
    FROM user_dr3int5.gaia_source
    WHERE parallax>20 AND teff_gspphot IS NOT NULL
) AS gaia
JOIN user_dr3int5.astrophysical_parameters AS apsis
ON gaia.source_id=apsis.source_id
```

## Appendix C: Acknowledgements

We thank our DPAC colleagues from CU5, Paolo Montegriffo, Dafydd Wyn Evans, Michael Weiler, Carme Jordi, Elena Pancino and Carla Cacciari, who have continuously supported us with their expertise on BP/RP spectra, their instrument characteristics and calibration. We also thank our DPAC colleagues from CU9, Carine Babusiaux, Mercè Romero-Gómez and Francesca Figueras, for their validation work and valuable feedback. Last but not least, we thank our former colleagues Tri Astraadmadja, Dae-Won Kim, Kester Smith, Paravskevi Tsalmantza, Rainer Klement, and Carola Tiede. This research was achieved using the POLLUX database (`http://pollux.oreme.org`) operated at LUPM (Université Montpellier - CNRS, France with the support of the PNPS and INSU).

This work presents results from the European Space Agency (ESA) space mission Gaia. Gaia data are being processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the Gaia MultiLateral Agreement (MLA). The Gaia mission website is `https://www.cosmos.esa.int/gaia`. The Gaia archive website is `https://archives.esac.esa.int/gaia`.

The Gaia mission and data processing have financially been supported by, in alphabetical order by country:

- the Algerian Centre de Recherche en Astronomie, Astrophysique et Géophysique of Bouzareah Observatory;
- the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF) Hertha Firnberg Programme through grants T359, P20046, and P23737;
- the BELgian federal Science Policy Office (BELSPO) through various PROgramme de Développement d'Expériences scientifiques (PRODEX) grants and the Polish Academy of Sciences - Fonds Wetenschappelijk Onderzoek through grant VS.091.16N, and the Fonds de la Recherche Scientifique (FNRS), and the Research Council of Katholieke Universiteit (KU) Leuven through grant C16/18/005 (Pushing AsteRoseismology to the next level with TESS, GaiA, and the Sloan DIgital Sky SurvEy – PARADISE);
- the Brazil-France exchange programmes Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) - Comité Français d'Evaluation de la Coopération Universitaire et Scientifique avec le Brésil (COFECUB);
- the Chilean Agencia Nacional de Investigación y Desarrollo (ANID) through Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) Regular Project 1210992 (L. Chemin);
- the National Natural Science Foundation of China (NSFC) through grants 11573054, 11703065, and 12173069, the China Scholarship Council through grant 201806040200, and the Natural Science Foundation of Shanghai through grant 21ZR1474100;
- the Tenure Track Pilot Programme of the Croatian Science Foundation and the École Polytechnique Fédérale de Lausanne and the project TTP-2018-07-1171 'Mining the Variable Sky', with the funds of the Croatian-Swiss Research Programme;
- the Czech-Republic Ministry of Education, Youth, and Sports through grant LG 15010 and INTER-EXCELLENCE grant LTAUSA18093, and the Czech Space Office through ESA PECS contract 98058;
- the Danish Ministry of Science;
- the Estonian Ministry of Education and Research through grant IUT40-1;
- the European Commission's Sixth Framework Programme through the European Leadership in Space Astrometry (ELSA) Marie Curie Research Training Network (MRTN-CT-2006-033481), through Marie Curie project PIOF-GA-2009-255267 (Space AsteroSeismology & RR Lyrae stars, SAS-RRL), and through a Marie Curie Transfer-of-Knowledge (ToK) fellowship (MTKD-CT-2004-014188); the European Commission's Seventh Framework Programme through grant FP7-606740 (FP7-SPACE-2013-1) for the Gaia European Network for Improved data User Services (GENIUS) and through grant 264895 for the Gaia Research for European Astronomy Training (GREAT-ITN) network;
- the European Cooperation in Science and Technology (COST) through COST Action CA18104 'Revealing the Milky Way with Gaia(MW-Gaia)';
- the European Research Council (ERC) through grants 320360, 647208, and 834148 and through the European Union's Horizon 2020 research and innovation and excellent science programmes through Marie Skłodowska-Curie grant 745617 (Our Galaxy at full HD – Gal-HD) and 895174 (The build-up and fate of self-gravitating systems in the Universe) as well as grants 687378 (Small Bodies: Near and Far), 682115 (Using the Magellanic Clouds to Understand the Interaction of Galaxies), 695099 (A sub-percent distance scale from binaries and Cepheids – CepBin), 716155 (Structured ACCREtion Disks – SACCRED), 951549 (Sub-percent calibration of the extragalactic distance scale in the era of big surveys – UniverScale), and 101004214 (Innovative Scientific Data Exploration and Exploitation Applications for Space Sciences – EXPLORE);
- the European Science Foundation (ESF), in the framework of the Gaia Research for European Astronomy Training Research Network Programme (GREAT-ESF);
- the European Space Agency (ESA) in the framework of the Gaia project, through the Plan for European Cooperating States (PECS) programme through contracts C98090

and 4000106398/12/NL/KML for Hungary, through contract 4000115263/15/NL/IB for Germany, and through PROgramme de Développement d'Expériences scientifiques (PRODEX) grant 4000127986 for Slovenia;
- the Academy of Finland through grants 299543, 307157, 325805, 328654, 336546, and 345115 and the Magnus Ehrnrooth Foundation;
- the French Centre National d'Études Spatiales (CNES), the Agence Nationale de la Recherche (ANR) through grant ANR-10-IDEX-0001-02 for the 'Investissements d'avenir' programme, through grant ANR-15-CE31-0007 for project 'Modelling the Milky Way in the Gaiaera' (MOD4Gaia), through grant ANR-14-CE33-0014-01 for project 'The Milky Way disc formation in the Gaiaera' (ARCHEOGAL), through grant ANR-15-CE31-0012-01 for project 'Unlocking the potential of Cepheids as primary distance calibrators' (UnlockCepheids), through grant ANR-19-CE31-0017 for project 'Secular evolution of galxies' (SEGAL), and through grant ANR-18-CE31-0006 for project 'Galactic Dark Matter' (GaDaMa), the Centre National de la Recherche Scientifique (CNRS) and its SNO Gaiaof the Institut des Sciences de l'Univers (INSU), its Programmes Nationaux: Cosmologie et Galaxies (PNCG), Gravitation Références Astronomie Métrologie (PNGRAM), Planétologie (PNP), Physique et Chimie du Milieu Interstellaire (PCMI), and Physique Stellaire (PNPS), the 'Action Fédératrice Gaia' of the Observatoire de Paris, the Région de Franche-Comté, the Institut National Polytechnique (INP) and the Institut National de Physique nucléaire et de Physique des Particules (IN2P3) co-funded by CNES;
- the German Aerospace Agency (Deutsches Zentrum für Luft- und Raumfahrt e.V., DLR) through grants 50QG0501, 50QG0601, 50QG0602, 50QG0701, 50QG0901, 50QG1001, 50QG1101, 50QG1401, 50QG1402, 50QG1403, 50QG1404, 50QG1904, 50QG2101, 50QG2102, and 50QG2202, and the Centre for Information Services and High Performance Computing (ZIH) at the Technische Universität Dresden for generous allocations of computer time;
- the Hungarian Academy of Sciences through the Lendület Programme grants LP2014-17 and LP2018-7 and the Hungarian National Research, Development, and Innovation Office (NKFIH) through grant KKP-137523 ('SeismoLab');
- the Science Foundation Ireland (SFI) through a Royal Society - SFI University Research Fellowship (M. Fraser);
- the Israel Ministry of Science and Technology through grant 3-18143 and the Tel Aviv University Center for Artificial Intelligence and Data Science (TAD) through a grant;
- the Agenzia Spaziale Italiana (ASI) through contracts I/037/08/0, I/058/10/0, 2014-025-R.0, 2014-025-R.1.2015, and 2018-24-HH.0 to the Italian Istituto Nazionale di Astrofisica (INAF), contract 2014-049-R.0/1/2 to INAF for the Space Science Data Centre (SSDC, formerly known as the ASI Science Data Center, ASDC), contracts I/008/10/0, 2013/030/I.0, 2013-030-I.0.1-2015, and 2016-17-I.0 to the Aerospace Logistics Technology Engineering Company (ALTEC S.p.A.), INAF, and the Italian Ministry of Education, University, and Research (Ministero dell'Istruzione, dell'Università e della Ricerca) through the Premiale project 'MIning The Cosmos Big Data and Innovative Italian Technology for Frontier Astrophysics and Cosmology' (MITiC);
- the Netherlands Organisation for Scientific Research (NWO) through grant NWO-M-614.061.414, through a VICI grant (A. Helmi), and through a Spinoza prize (A. Helmi), and the Netherlands Research School for Astronomy (NOVA);
- the Polish National Science Centre through HARMONIA grant 2018/30/M/ST9/00311 and DAINA grant 2017/27/L/ST9/03221 and the Ministry of Science and Higher Education (MNiSW) through grant DIR/WK/2018/12;
- the Portuguese Fundação para a Ciência e a Tecnologia (FCT) through national funds, grants SFRH/BD/128840/2017 and PTDC/FIS-AST/30389/2017, and work contract DL 57/2016/CP1364/CT0006, the Fundo Europeu de Desenvolvimento Regional (FEDER) through grant POCI-01-0145-FEDER-030389 and its Programa Operacional Competitividade e Internacionalização (COMPETE2020) through grants UIDB/04434/2020 and UIDP/04434/2020, and the Strategic Programme UIDB/00099/2020 for the Centro de Astrofísica e Gravitação (CENTRA);
- the Slovenian Research Agency through grant P1-0188;
- the Spanish Ministry of Economy (MINECO/FEDER, UE), the Spanish Ministry of Science and Innovation (MICIN), the Spanish Ministry of Education, Culture, and Sports, and the Spanish Government through grants BES-2016-078499, BES-2017-083126, BES-C-2017-0085, ESP2016-80079-C2-1-R, ESP2016-80079-C2-2-R, FPU16/03827, PDC2021-121059-C22, RTI2018-095076-B-C22, and TIN2015-65316-P ('Computación de Altas Prestaciones VII'), the Juan de la Cierva Incorporación Programme (FJCI-2015-2671 and IJC2019-04862-I for F. Anders), the Severo Ochoa Centre of Excellence Programme (SEV2015-0493), and MICIN/AEI/10.13039/501100011033 (and the European Union through European Regional Development Fund 'A way of making Europe') through grant RTI2018-095076-B-C21, the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia 'María de Maeztu') through grant CEX2019-000918-M, the University of Barcelona's official doctoral programme for the development of an R+D+i project through an Ajuts de Personal Investigador en Formació (APIF) grant, the Spanish Virtual Observatory through project AyA2017-84089, the Galician Regional Government, Xunta de Galicia, through grants ED431B-2021/36, ED481A-2019/155, and ED481A-2021/296, the Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), funded by the Xunta de Galicia and the European Union (European Regional Development Fund – Galicia 2014-2020 Programme), through grant ED431G-2019/01, the Red Española de Supercomputación (RES) computer resources at MareNostrum, the Barcelona Supercomputing Centre - Centro Nacional de Supercomputación (BSC-CNS) through activities AECT-2017-2-0002, AECT-2017-3-0006, AECT-2018-1-0017, AECT-2018-2-0013, AECT-2018-3-0011, AECT-2019-1-0010, AECT-2019-2-0014, AECT-2019-3-0003, AECT-2020-1-0004, and DATA-2020-1-0010, the Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya through grant 2014-SGR-1051 for project 'Models de Programació i Entorns d'Execució Parallels' (MPEXPAR), and Ramon y Cajal Fellowship RYC2018-025968-I funded by MICIN/AEI/10.13039/501100011033 and the European Science Foundation ('Investing in your future');
- the Swedish National Space Agency (SNSA/Rymdstyrelsen);
- the Swiss State Secretariat for Education, Research, and Innovation through the Swiss Activités Nationales Complémentaires and the Swiss National Science Founda-

tion through an Eccellenza Professorial Fellowship (award PCEFP2_194638 for R. Anderson);
- the United Kingdom Particle Physics and Astronomy Research Council (PPARC), the United Kingdom Science and Technology Facilities Council (STFC), and the United Kingdom Space Agency (UKSA) through the following grants to the University of Bristol, the University of Cambridge, the University of Edinburgh, the University of Leicester, the Mullard Space Sciences Laboratory of University College London, and the United Kingdom Rutherford Appleton Laboratory (RAL): PP/D006511/1, PP/D006546/1, PP/D006570/1, ST/I000852/1, ST/J005045/1, ST/K00056X/1, ST/K000209/1, ST/K000756/1, ST/L006561/1, ST/N000595/1, ST/N000641/1, ST/N000978/1, ST/N001117/1, ST/S000089/1, ST/S000976/1, ST/S000984/1, ST/S001123/1, ST/S001948/1, ST/S001980/1, ST/S002103/1, ST/V000969/1, ST/W002469/1, ST/W002493/1, ST/W002671/1, ST/W002809/1, and EP/V520342/1.

The GBOT programme uses observations collected at (i) the European Organisation for Astronomical Research in the Southern Hemisphere (ESO) with the VLT Survey Telescope (VST), under ESO programmes 092.B-0165, 093.B-0236, 094.B-0181, 095.B-0046, 096.B-0162, 097.B-0304, 098.B-0030, 099.B-0034, 0100.B-0131, 0101.B-0156, 0102.B-0174, and 0103.B-0165; and (ii) the Liverpool Telescope, which is operated on the island of La Palma by Liverpool John Moores University in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias with financial support from the United Kingdom Science and Technology Facilities Council, and (iii) telescopes of the Las Cumbres Observatory Global Telescope Network.