Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **521,** 2745–2764 (2023) https://doi.org/10.1093/mnras/stad574
Advance Access publication 2023 March 3

Hunting for C-rich long-period variable stars in the Milky Way's bar-bulge using unsupervised classification of *Gaia* BP/RP spectra

Jason L. Sanders [1]★ and Noriyuki Matsunaga[2]

[1]*Department of Physics and Astronomy, University College London, London WC1E 6BT, UK*
[2]*Department of Astronomy, School of Science, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

Accepted 2023 February 16. Received 2023 February 14; in original form 2022 December 16

ABSTRACT

The separation of oxygen- and carbon-rich asymptotic giant branch sources is crucial for their accurate use as local and cosmological distance and age/metallicity indicators. We investigate the use of unsupervised learning algorithms for classifying the chemistry of long-period variables from Gaia DR3's BP/RP spectra. Even in the presence of significant interstellar dust, the spectra separate into two groups attributable to O-rich and C-rich sources. Given these classifications, we utilize a supervised approach to separate O-rich and C-rich sources without blue and red photometers (BP/RP) spectra but instead given broadband optical and infrared photometry finding a purity of our C-rich classifications of around 95 per cent. We test and validate the classifications against other advocated colour–colour separations based on photometry. Furthermore, we demonstrate the potential of BP/RP spectra for finding S-type stars or those possibly symbiotic sources with strong emission lines. Although our classification suggests the Galactic bar-bulge is host to very few C-rich long-period variable stars, we do find a small fraction of C-rich stars with periods > 250 day that are spatially and kinematically consistent with bar-bulge membership. We argue the combination of the observed number, the spatial alignment, the kinematics, and the period distribution disfavour young metal-poor star formation scenarios either *in situ* or in an accreted host, and instead, these stars are highly likely to be the result of binary evolution and the evolved versions of blue straggler stars already observed in the bar-bulge.

Key words: stars: AGB – stars: variables: general – Galaxy: bulge.

1 INTRODUCTION

Stars of masses between $\sim$0.8 and $\sim 8 M_{\odot}$ will eventually pass through the asymptotic giant branch (AGB) phase, the final nuclear burning stage, characterized by high luminosity and large cool convective envelopes (Herwig 2005). AGB stars are significant components of a galaxy's stellar population not only in terms of luminosity but also in terms of their contributions to the nucleosynthetic build-up of elements in their galaxy (Karakas & Lattanzio 2014; Kobayashi, Karakas & Lugaro 2020). They are also typically long-period variables (LPV) with the Mira variables representing the highest amplitude AGB pulsators typically attributed to fundamental mode pulsations. During these pulsation phases, AGB stars lose significant quantities of their mass producing large amounts of dust (Höfner & Olofsson 2018). The chemistry of the AGB envelope and the properties of the circumstellar dust are determined by the carbon to oxygen ratio, C/O, which in turn affects the subsequent evolutionary properties of the star (see Wallerstein & Knapp 1998; Lloyd Evans 2010, for thorough reviews of carbon stars). The subdominant element will be bound up almost entirely in CO leaving the dominant element to form carbon- or oxygen-rich molecules, such as C_2 and CN, and TiO and SiO ,respectively. When C/O $\approx$1 the AGB star will be an S-type star exhibiting intermediate chemistry with a mixture of carbon- and oxygen-rich species and characteristically ZrO.

The C/O ratio of an AGB star is altered by the third dredge-up bringing newly-synthesized carbon to the outer atmosphere. The strength of the dredge-up and its subsequent impact on the C/O ratio is a function of both the metallicity and age/mass of the AGB star (Karakas 2014; Höfner & Olofsson 2018). Loosely, higher mass stars have stronger dredge-up episodes and lower metallicity stars have lower levels of oxygen in their atmosphere so are more readily diluted by dredged-up carbon. In this way, the fraction of C-rich to O-rich AGB stars can be a useful indicator of the age and/or metallicity of a stellar population (Brewer, Richer & Crabtree 1995; Boyer et al. 2013). For example, in the Milky Way disc, the fraction of C-rich to O-rich AGB stars increases with radius (Blanco, McCarthy & Blanco 1984; Ishihara et al. 2011), which could be a reflection of the metallicity gradient and/or the inside-out growth of the disc. Due to its predominantly old, metal-rich population (Bensby et al. 2013; Catchpole et al. 2016; Bovy et al. 2019; Hasselquist et al. 2021), the Milky Way's bar-bulge is expected, and largely observed, to host relatively few C-rich LPV stars. Matsunaga et al. (2017) discovered five C-rich Mira variable stars with magnitudes consistent with bar-bulge membership and argued that the previously-discovered symbiotic C-rich Mira variable from Miszalski, Mikołajewska & Udalski (2013) was more consistent with foreground disc membership because the 2MASS data were taken around the light curve minimum. Although Matsunaga et al.

★ E-mail: jason.sanders@ucl.ac.uk

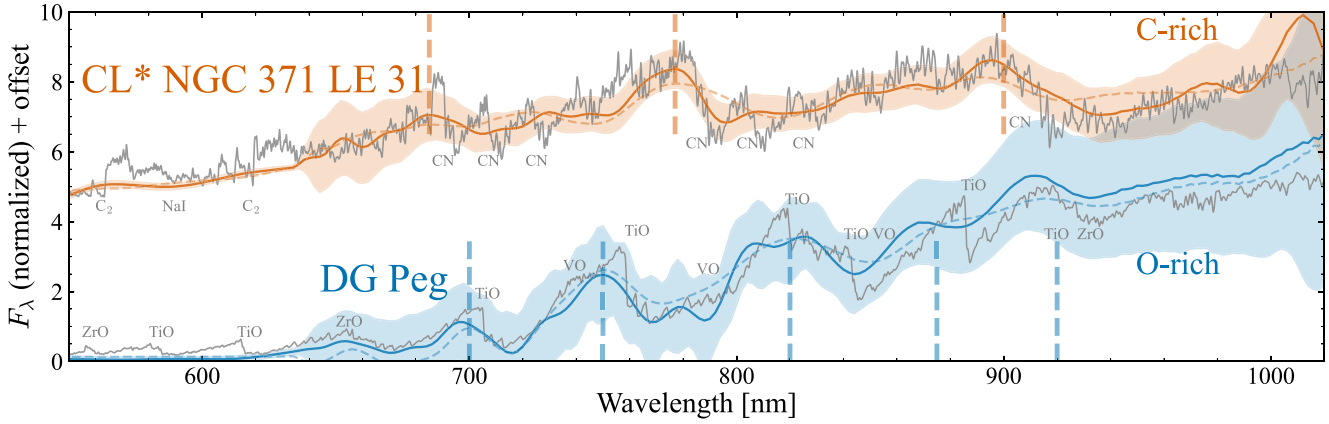

Figure 1. Two example long-period variable star blue and red photometers (BP/RP) spectra (with the $\pm 1\sigma$ bracket) compared to spectra from the X-Shooter library (XSL, smoothed to 0.36 nm, Chen et al. 2014; Gonneau et al. 2020; Verro et al. 2022). The dashed line shows the truncated BP/RP spectra. All spectra have been normalized by the integral of the XSL spectra from 350 to 1020 nm and the C-rich spectrum has been offset vertically for clarity. DG Peg is an oxygen-rich Mira variable star (spectral type M4e) while CL* NGC 371 LE 31 is a carbon-rich semiregular variable in the SMC cluster NGC 371. We have marked the peaks arising from the molecular band heads (as labelled), which are easily distinguishable in the BP/RP spectra (primarily TiO for O-rich, CN for C-rich).

(2017) favour binary evolution as the production mechanism for these C-rich variables, it is also possible they are the result of single star evolution and arise from either accretion or recent *in situ* metal-poor star formation. The occurrence of dwarf carbon stars (e.g. Whitehouse et al. 2018), CH stars, and carbon-enhanced metal-poor stars rich in s-process elements (Koch et al. 2016; Arentsen et al. 2021) demonstrates that binary evolution can significantly alter the C/O ratio of some stars, particularly those that are metal-poor (De Marco & Izzard 2017). Indeed Azzopardi, Lequeux & Rebeirot (1988) and Azzopardi et al. (1991) found a population of C-rich giants towards the Galactic bulge that are too faint to be AGB stars in the bulge but could instead be lower luminosity red giant branch stars in the bulge that are the result of binary evolution (or plausibly associated with the Sagittarius dwarf spheroidal galaxy, Ng 1997, 1998). Hunting for C-rich stars in the Galactic bulge is a useful archaeological probe for discovering accreted or young *in situ* populations, but disentangling the different formation channels to reveal details of the evolutionary history of the Galaxy requires the complement of spectroscopic and kinematic analysis now possible with *Gaia*.

Observations in the Large Magellanic Cloud (LMC, e.g. Glass & Evans 1981; Iwanek, Soszyński & Kozłowski 2021) have demonstrated that LPVs reside along a series of period–luminosity relations believed to be associated with different pulsation modes (Wood 2000). In particular, the high-amplitude Mira variables lie along a tight sequence in the dust-insensitive Wesenheit indices versus period (Soszyński et al. 2009). For this reason, Mira variables are increasingly seeing use as distance tracers for structure within the Milky Way and Local Group (Catchpole et al. 2016; Deason et al. 2017; Grady, Belokurov & Evans 2020; Semczuk et al. 2022), and also as an alternative distance ladder calibrator for measurements of H_0 (Huang et al. 2018, 2020). However, in general, O-rich and C-rich Mira variables are known to lie along separate period–luminosity sequences with the O-rich Mira variables typically obeying tighter period–luminosity relations (Glass & Evans 1981; Feast et al. 1989; Ita et al. 2004; Groenewegen 2004; Fraser, Hawley & Cook 2008; Riebel et al. 2010; Ita & Matsunaga 2011; Yuan et al. 2017a, b; Bhardwaj et al. 2019; Iwanek et al. 2021). This appears to be in large part due to the differing amounts of circumstellar dust (Ita & Matsunaga 2011). Careful characterization and separation of the two types are essential for precision work both in the cosmological and local group setting.

Traditionally, carbon stars have been identified by their C_2 Swan bands and CN bands in objective-prism plates (Secchi 1868; Nassau & Velghe 1964; MacConnell 1988; Aaronson et al. 1990). In more recent years, the separation of O-rich and C-rich AGB stars has been performed on large samples using infrared photometry taking advantage of the silicate bands in the O-rich spectra at ∼9 and ∼ 18 μm compared to the SiC band at ∼ 11 μm e.g. AKARI (Ishihara et al. 2011), WISE (Lian et al. 2014; Nikutta et al. 2014; Suh & Hong 2017), Spitzer (Kastner et al. 2008; Groenewegen & Sloan 2018), and MSX (Lewis et al. 2020a, b) guided by the more detailed infrared spectroscopic view from the IRAS LRS, the ISO SWS, and the Spitzer IRS instruments (Olnon et al. 1986; Kraemer et al. 2002). Soszyński et al. (2009) separated O-rich and C-rich sources in the LMC using the Wesenheit $W_{I,V-I}$ versus $W_{Ks,J-Ks}$ diagram where the two Wesenheit indices are $W_{I,V-I} = I - 1.55(V - I)$ and $W_{Ks,J-Ks} = K_s - 0.686(J - K_s)$. In a similar vein, Lebzelter et al. (2018) proposed a scheme for separating O-rich and C-rich sources in the LMC using a combination of *Gaia* and 2MASS broadband photometry. This is highly desirable due to the all-sky availability of this photometry. These authors advocated for a separation in the 'colour–magnitude' diagram of $W_{\mathrm{RP,\,BP\text{-}RP}} - W_{Ks,J-Ks}$ versus K_s where $W_{\mathrm{RP,\,BP\text{-}RP}} = G_{\mathrm{RP}} - 1.3(G_{\mathrm{BP}} - G_{\mathrm{RP}})$. There is relatively weak curvature of the O-rich/C-rich separation line in this space meaning an analogue of this separation using only $W_{\mathrm{RP,\,BP\text{-}RP}} - W_{Ks,J-Ks}$ can also in theory be used for non-LMC stars. These Wesenheit diagrams are effective as, in the optical, oxygen-rich AGB stars have a set of TiO bands (also ZrO and VO, see Fig. 4 of Lançon & Wood 2000, for identification of the lines, see also Van Eck et al. 2017 and Yao et al. 2017), whereas the carbon-rich stars have a set of Swan C_2 bands (at < 600 nm) and CN bands (at > 700 nm) (e.g. fig. 7 of Lançon & Mouhcine 2002). Representative O-rich and C-rich LPV spectra from the X-Shooter Library (Chen et al. 2014; Gonneau et al. 2020; Verro et al. 2022) are shown in Fig. 1 where these bands are clearly visible. Both the locations and separations between the bands are distinct for the two types of star leading Lebzelter et al. (2022) to develop a classification on the basis

of peak separation in pseudo-wavelength (`median_delta_wl_rp`) in the Gaia DR3 RP spectra. However, Lebzelter et al. (2022) report that the classification scheme performs poorly for low signal-to-noise spectra and/or highly-extincted sources and suggest only trusting the C-rich classification if $7 <$ `median_delta_wl_rp` < 11 and $G_{\rm BP} < 19$.

Here, we investigate the performance of an unsupervised O-rich/C-rich classification scheme using the *Gaia* third data release (DR3) BP/RP spectra (also called XP spectra, Carrasco et al. 2021; De Angeli et al. 2022; Montegriffo et al. 2022). Lucey et al. (2022) has already demonstrated the power of utilising a supervised classification on the BP/RP spectra for the identification of carbon-enhanced metal-poor stars. Unsupervised classification seeks to find similarities between presented training examples such that clusters of 'similar' data can be assigned labels. Numerous clustering algorithms exist but often the challenge is representing the data set in a space that is amenable to clustering. In many cases, the data are of high dimensionality making clustering algorithms very computationally expensive, or the natural clustering is along complex surfaces in the high-dimensional space. One solution to both of these problems is to project the data to a lower dimensional space that encodes as much information from the higher dimensional space as possible. For example, principal component analysis (PCA) seeks to find the most informative linear combinations of the higher dimensional space. However, it is inappropriate for significantly reducing the dimensionality of the data as it is limited to only considering linear combinations of the input data dimensions whilst often clusters lie along highly non-linear surfaces.

Several algorithms have sought to circumvent this limitation. For example, t-SNE (t-stochastic neighbour embedding, van der Maaten & Hinton 2008) first finds the similarities between the data points in the high dimensional space $\boldsymbol{x}$ using a Gaussian kernel and then attempts to find the low dimensional projection $\boldsymbol{y}$ that minimizes the Kullback–Leibler (KL) divergence between the higher dimensional similarity distribution and that of the lower dimensional representation assuming the lower dimensional similarity distribution follows a Student t distribution. t-SNE has found considerable use for astrophysics applications, in particular in the analysis of spectroscopic data sets (e.g. Traven et al. 2017; Anders et al. 2018). One disadvantage of the t-SNE algorithm is that it only preserves a sense of distance (metric) between local points in the lower dimension space. Additionally, the original implementation was computationally expensive for large data sets (mostly due to having to construct the N by N high-dimensional similarity distribution) and in practical applications is often combined with an initial PCA to an intermediate dimensionality data set. Uniform Manifold Approximation and Projection (UMAP, McInnes, Healy & Melville 2018) was designed to solve both of these issues with t-SNE. Instead of utilizing the KL divergence, UMAP uses the cross-entropy which ensures $|\boldsymbol{y}_i - \boldsymbol{y}_j| \rightarrow \infty$ as $|\boldsymbol{x}_i - \boldsymbol{x}_j| \rightarrow \infty$ (whilst for the KL divergence $|\boldsymbol{y}_i - \boldsymbol{y}_j|$ can take any value as $|\boldsymbol{x}_i - \boldsymbol{x}_j| \rightarrow \infty$). Other computational/algorithmic improvements have enabled significant speed-ups for UMAP compared to t-SNE although in essence, the algorithms share significant similarities and the t-SNE implementation from Poličar, Stražar & Zupan (2019) is competitive with UMAP in terms of computational time. Despite its relatively recent creation, the UMAP algorithm has already found use in astrophysics applications (Reis et al. 2019; Kim et al. 2022; Grondin et al. 2023).

Here, we investigate the use of these unsupervised classification algorithms (UMAP and t-SNE) for determining the chemistry of AGB stars using their *Gaia* DR3 BP/RP spectra. We begin by describing the BP/RP spectra and our chosen unsupervised classification scheme in Section 2. BP/RP spectra are unavailable for stars with $G > 17.65$ although there are many identified LPVs from *Gaia* fainter than this limit. We, therefore, extend our unsupervised classification to the fainter objects with a supervised scheme using *Gaia* and 2MASS photometry in Section 2.6. We validate our results by inspecting previously employed colour–colour diagrams for separating C-rich and O-rich sources in Section 3. Finally, we use the new classification results to search for C-rich variables in the Galactic bulge in Section 4.

2 UNSUPERVISED CLASSIFICATION OF O-RICH/C-RICH LONG-PERIOD VARIABLES

Our primary data source is the *Gaia* DR3 long-period variable candidate catalogue (Lebzelter et al. 2022) as part of the full third *Gaia* data release (Gaia Collaboration 2016, 2022b) that complements the astrometric results from the early third *Gaia* data release (Gaia Collaboration 2021). This is the second version of the long-period candidate table from *Gaia* after the first catalogue of Mowlavi et al. (2018) based on *Gaia* DR2 photometry. *Gaia*'s long-period variable processing consists of two pipelines: a generic classification pipeline for assigning classes to all variables (Rimoldini et al. 2019) and then the specific object study (SOS) for the stars classified as long-period variables by the first stage (in addition to a very low number of additional likely LPVs largely classified as symbiotic stars in the classification pipeline). Additional quality cuts are also performed on the stars processed by the SOS on the basis of colour, high G-band variability, number of epochs, and valid astrometry. In total there are 1 720 588 stars in the *Gaia* DR3 long-period variable candidate catalogue of which 392 240 have reported periods.

The BP/RP spectra from *Gaia* DR3 (Carrasco et al. 2021; De Angeli et al. 2022; Montegriffo et al. 2022) are low resolution ($R = \lambda/\delta\lambda \approx 25-100$) and together the BP/RP cover the optical range from 330 to 1050 nm. The spectra are provided for ∼220 million stars in *Gaia* with $G < 17.65$. Despite their low resolution, several studies have demonstrated that the information content is sufficient to measure bulk spectroscopic parameters ($T_{\rm eff}$, $\log g$, [M/H], Liu et al. 2012; Witten et al. 2022; Xylakis-Dornbusch et al. 2022; Andrae et al. 2022; Fouesneau et al. 2022; Creevey et al. 2022; Belokurov et al. 2022; Rix et al. 2022) although detailed chemical abundances are likely too challenging (Gavel et al. 2021). However, particularly in cooler stars, the presence of molecular features in these spectra is detectable (Lucey et al. 2022) enabling more accurate metallicity determinations and detailed carbon abundances. For each star, the BP/RP spectra are provided as two sets of coefficients (with associated covariance matrices) from which both the internally- or externally-calibrated spectra can be reconstructed on a wavelength grid of choice through multiplication with a set of basis functions (times a mixing term for connecting the BP and RP spectra at their interface). This can be done using the GAIAXPY python package[1] (Montegriffo et al. 2022). The basis functions are linear combinations of Hermite polynomials chosen through a PCA/SVD on a set of BP/RP calibrators. In this way, the first coefficient contains the most information for the average calibrator star and higher-order coefficients provide ever-decreasing corrections to this average calibrator star's spectrum. It is expected that stars that are significantly different to the average calibrator star in the full data set will not necessarily follow this hierarchy. The *Gaia* DR3 data also

[1] Available from https://gaia-dpci.github.io/GaiaXPy-website/. We use v1.1.2 10.5281/zenodo.6642313.

provides a truncation order (`xp_n_relevant_bases`) for each star indicating the last coefficient that is significant compared to the noise. Of the 1 720 588 Gaia DR3 LPVs, 1 205 121 have BP/RP spectra.

In Fig. 1 we display two spectra from the X-Shooter Library (Chen et al. 2014; Gonneau et al. 2020; Verro et al. 2022): DG Peg is an O-rich Mira variable star (*Gaia* DR3 1768290812321736576 with a 147 d period and $G = 11.1$ mag) while CL∗ NGC 371 LE 31 is a semiregular variable star (although with a high I amplitude of ≈ 0.69 mag Soszyński et al. 2009) in the Small Magellanic Cloud open cluster NGC 371 (*Gaia* DR3 4690506059969811968 with a 301 d period and $G = 16$ mag). We have normalized the spectra by the integral $\int \mathrm{d}\lambda\, \lambda F_\lambda$ over 350 to 1020 nm and the C-rich spectrum is offset vertically for clarity. We also display the *Gaia* DR3 BP/RP spectra constructed from the coefficients on a 2 nm grid (using GaiaXPy) along with the $\pm 1\sigma$ band implied by the covariance matrix for the coefficients. The BP/RP spectra have been divided by the same normalization factor as the X-Shooter spectra. There is a very good correspondence between the BP/RP spectra and the X-Shooter spectra showing that both are well absolutely calibrated. We further display the BP/RP spectra constructed by truncating the spectra as dashed lines (the number of relevant RP bases are 2 and 4 for DG Peg and CL∗ NGC 371 LE 31, respectively) We see in these cases the truncated spectra largely do a good job of capturing the features we are interested in here. From Fig. 1 it is clear that the O-rich and C-rich spectra display very different sets of features that are well captured in the BP/RP spectra. The O-rich DG Peg spectrum shows a series of TiO bands at 705.4, 758.9, 819.4843.2, 885.9, and 920.9 nm (Bobrovnikoff 1933; Sharpless 1956; Mürset & Schmid 1999) that give rise to characteristic peaks at 700, 750, 820, and 920 nm with the band head at 885.9 nm giving rise to a peak around 875 nm that often blends with the 920 nm peak in the BP/RP spectra. There are a series of other band heads bluer than 700 nm also due to TiO but in general, these are hard to identify in the BP/RP spectra due to the lower flux. Figure 4 of Lançon & Wood (2000) labels the TiO transitions responsible for these characteristic peaks/troughs in a cool and a warm Mira spectrum. In the cooler spectra, VO features appear at ∼740, ∼790, and ∼860 nm broadening the neighbouring TiO features (there is the suggestion of this in the double minimum structure at 780 nm where the right minimum is due to VO absorption). The C-rich spectrum however has a set of three peaks at 683, 778, and 900 nm due to a series of CN features (see e.g. Gonneau et al. 2016). For distinguishing O-rich and C-rich stars, the region between 750 and 850 nm is particularly clear where the CN band head sits at the minimum between two TiO band heads. It is quite clear from this example that BP/RP spectra have the capability to distinguish between O-rich and C-rich stars (as already evidenced by Lebzelter et al. 2022).

As we are dealing with solely variable stars, it is worth considering how variability alters our interpretation of the *Gaia* data. The BP/RP spectra are the average of *Gaia*'s observations of each source over many transits (typically 20−40). For variable stars, we therefore approximately observe the properties of the stars averaged over period (*Gaia* DR3's observing window is 34 months so most LPVs have at least one period of observations). During the pulsation of long-period variables, the temperature of their envelopes varies giving rise to different balances of molecular species (as well as varying emission line ratios e.g. Yao et al. 2017 although these are less important for our work). In O-rich Mira variables, TiO is only present at low levels at maximum brightness (temperature) but its production strongly increases towards lower temperatures giving rise to the high amplitudes in the visual bands (Reid & Goldston 2002). In C-rich Mira variables, C_2 and CN are formed at higher temperatures deeper in the atmosphere and so their relative abundance does not change significantly during the pulsation cycle (Lançon & Wood 2000). Fig. 3 of Lebzelter et al. (2022) shows the RP spectra of the O-rich star T Aqr and the C-rich star RU Vir at the individual observing epochs where it is clear that for the O-rich star the TiO bands reduce in depth at peak brightness (particularly the band at ∼ 900 nm or ∼40 in pseudo-wavelength units) while the depth of the C-rich bands remains quite similar across all epochs. Aliasing may arise as an issue for stars with periods around 190 or 380 d (awkward periods for *Gaia*'s scanning law) where the BP/RP spectra may only be averaging over similar phases.

2.1 High-amplitude variable selection

Our primary interest is the high-amplitude Mira variable stars in *Gaia*'s LPV candidate catalogue. For this reason, we limit our analysis in the main body of this work to a high-amplitude subset of the *Gaia* DR3 LPVs. In Appendix A we show the results of running the approach on all *Gaia* LPVs. It should be noted that sources in the *Gaia* DR3 LPV catalogue are not assigned periods and amplitudes if the photometry is likely contaminated. In turn, this possibly means the BP/RP spectra are contaminated. Many of these sources are faint and in the crowded midplane and it is more of an issue for lower amplitude variables. In our analysis, we have opted to process all the stars without consideration of any data quality cuts under the proviso that any subsequent analysis will consider data quality more explicitly. For example, typical fields used for selecting high-quality data are the BP/RP photometric excess (although as discussed by Riello et al. 2021, the BP/RP photometric excess is a less useful quality cut for red variable sources) and the fraction of contaminated transits. A further consideration is that some fraction of the classified LPVs may be contaminating young stellar objects (YSOs, see the discussion in Mowlavi et al. 2018, for the Gaia DR2 LPV catalogue which likely still applies for the DR3 data). However, some basic parallax cuts can approximately remove this contaminant.

We select all LPVs with peak-to-peak semi-amplitudes,[2] `amplitude`, > 0.32 mag from the *Gaia* DR3 LPV candidate catalogue (Lebzelter et al. 2022). In Appendix B we compare the different amplitude measures for these stars from *Gaia* DR3. This amplitude cut includes Mira variable stars which are typically defined as having $\Delta V > 1.25$ and for which Grady, Belokurov & Evans (2019) advocate a cut of $\Delta G > 0.433$ mag. At the lower amplitude end, there will also be some semiregular variables. Furthermore, around 190 and 380 d periods `amplitude` can significantly overestimate the true amplitude due to poor phase coverage related to *Gaia*'s scanning law. However, for the majority of stars, `amplitude` agrees with amplitude estimates from the scatter of the photometric data points. There are 99 212 stars satisfying this amplitude cut of which 79 944 have BP/RP spectra.

2.2 Default procedure

We use the BP/RP coefficients normalized to the first RP coefficient as our input data (note we are here ignoring uncertainties in the coefficients due to the limitations of the UMAP and t-SNE approaches). Our default setup is to run UMAP on the entire $(2 \times 55 - 1 = 109)$ set of normalized coefficients to reduce it down to a two-dimensional projection. For UMAP there are two key

[2]We use Δ to denote the semi-amplitude throughout this work.

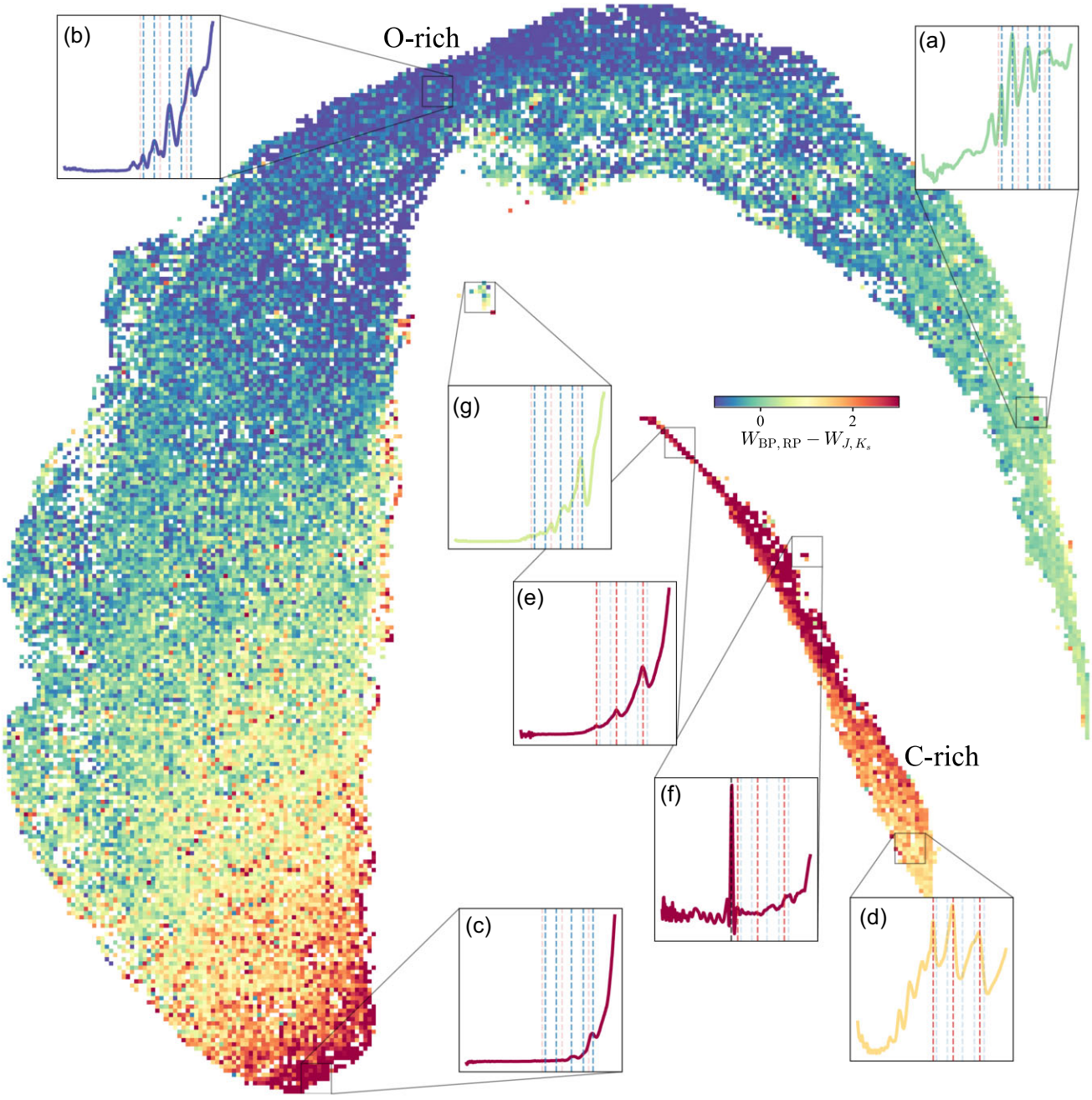

Figure 2. Two-dimensional UMAP projection of the BP/RP spectra coefficients for *Gaia* DR3 high-amplitude long-period variables. Each pixel is coloured by the mean Wesenheit colour $W_{\mathrm{RP,\,BP\text{-}RP}} - W_{Ks,\,J-Ks}$ advocated as a good metric for separating O-rich and C-rich stars by Lebzelter et al. (2018). The upper crescent is populated by O-rich stars while the lower spur is populated by C-rich stars. The insets show the median $\lambda^{-2}F_{\lambda}$ BP/RP spectrum between 336 and 1020 nm in each marked UMAP bin with the features from Fig. 1 marked as dashed vertical lines (blue for O-rich, red for C-rich). As we move along the crescent from right to left [from (a) to (c)], the O-rich stars broadly become redder due to a combination of effective temperature and extinction variation. Moving vertically along the C-rich sequence from (d) to (e), the stars get redder. The small island (f) to the right of the C-rich sequence has significant Hα emission (marked as black dashed) whilst the small island (g) off the centre of the crescent has a mixture of O-rich and C-rich features attributable to S-type stars.

hyperparameters, `min_dist` and `n_neighbors`, setting the minimum distance between points in the low-dimensional space and the size of the neighbourhood used for finding the local manifold, respectively. We have found `min_dist` =0.05 and `n_neighbors` =15 produces results that cleanly separate the dataset into two distinct groupings. After projection, we z-score the output and then run DBSCAN (Ester et al. 1996) to cluster the data with a standard maximum distance $\epsilon = 0.09$ (although simpler linear cuts would also suffice).

The results of our procedure are shown in Fig. 2.[3] We see that in the two-dimensional UMAP space the high-amplitude variables split into two clear populations: a crescent and a spur. Fig. 2 is coloured by the mean Wesenheit colour $W_{\mathrm{RP,\,BP\text{-}RP}} - W_{Ks,\,J-Ks}$ which was advocated by Lebzelter et al. (2018) as a good space to separate

[3]Note whenever we display a UMAP (or t-SNE) projection we have opted to drop the tick labels as the absolute values are unimportant.

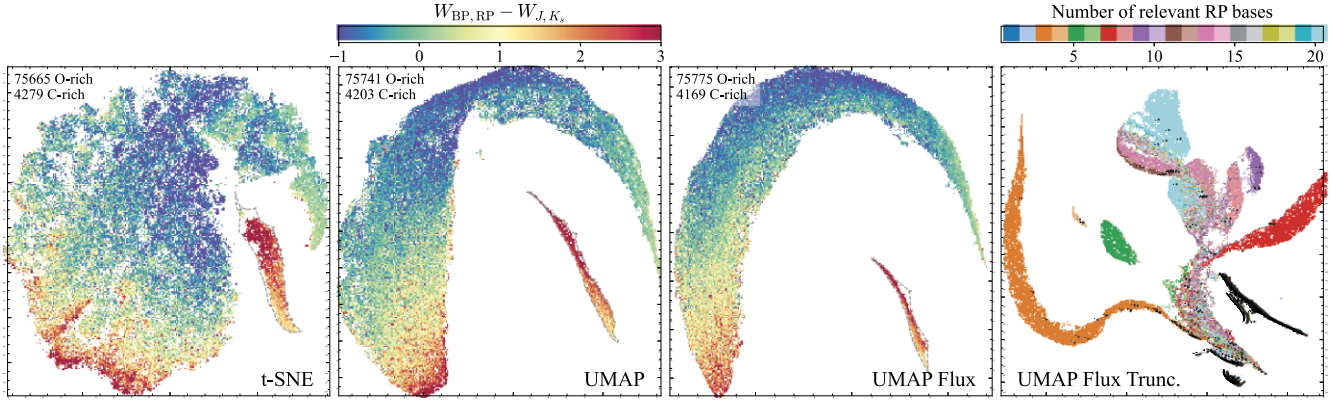

Figure 3. Variants of our unsupervised learning approach – the left three panels show the t-SNE projection, the UMAP projection of the BP/RP coefficients (identical to Fig. 2) and the UMAP projection of the binned spectra respectively, all coloured by the Wesenheit colour $W_{\rm RP,\,BP\text{-}RP} - W_{Ks,\,J-Ks}$ The grey contour shows the group identified by DBSCAN that we label as C-rich stars. The right panel shows the UMAP projection of the binned spectra computed from the truncated spectra coloured by the number of relevant RP bases. The black points are C-rich stars identified from the UMAP projection of the coefficients. Clearly, the stars have been separated more on the basis of their truncation order than on any intrinsic similarities in the data.

O-rich and C-rich stars in the LMC. We see indeed that the spur has mostly $W_{\rm RP,\,BP\text{-}RP} - W_{Ks,\,J-Ks} > 1$ so we identify it with C-rich sources whilst the crescent has mostly $W_{\rm RP,\,BP\text{-}RP} - W_{Ks,\,J-Ks} < 1$ so we identify it with O-rich sources. Note however that at the far left end of the crescent the stars are redder and as red as the C-rich spur.

To further validate our assignment of O- and C-rich to the crescent and spur, we have plotted a series of median BP/RP spectra in the insets (we divide by λ^2 so the structure of the peaks is more visible). Spectrum (a) shows a typical low-extinction O-rich spectrum where the set of peaks identified in Fig. 1 are clearly seen. As we move around the crescent to spectrum (b) and onto spectrum (c) the set of O-rich peaks is still visible but the spectra are increasingly reddened. There are also less obvious variations in the shapes of the absorption features (which are modulated by the varying extinction) with the final peak at ~ 910 nm broader and boxier in spectrum (a) than in spectrum (b). This is a reflection of varying effective temperatures producing varying depths of the TiO$\delta(\Delta\nu = 0)$ absorption feature. A similar sequence is seen moving up the spur from spectrum (d) which shows the three peaks of a C-rich spectrum seen in Fig. 1 through to spectrum (e) which is distinctly more reddened but the peak structure is still visible. We attribute the relative narrowness of the C-rich spur to degenerate changes in both effective temperature and extinction rather than any change in the surface chemistry as the CN features are not strong functions of effective temperature. However, the broadness of the O-rich crescent is because both extinction and effective temperature (through the relative depths of TiO features) control the spectrum shape and can be distinguished. There are a couple of exceptions to these broad trends due to effective temperature and extinction variation which are apparent as small islands in Fig. 2. The very small island (f) just off the C-rich spur has a median spectrum showing clear Hα emission characteristic of a symbiotic system (this island contains the source identified by Miszalski et al. 2013). Mira variables also show strong Balmer series emission lines that vary with the pulsation phase but typically the line-to-continuum flux ratio is at most $\sim$2 (Yao et al. 2017) while the symbiotic carbon star from Miszalski et al. (2013) has a line-to-continuum ratio of $\gg$5. It is likely Mira variable emission lines cannot be detected in BP/RP spectra. Finally, island (g) sits partway between the crescent and spur and its median spectrum shows evidence of both O-rich and C-rich behaviour. This is characteristic of an S-star that has intermediate chemistry C/O$\sim$1. We will discuss these stars further below.

2.3 Model variants

Before discussing in more detail the structure of the UMAP projection we briefly mention other approaches for separating the spectra. As a comparison, we have run some variants of our algorithm. First, we use t-SNE. t-SNE differs from UMAP primarily in the properties of the high-dimensional similarity distribution that are retained in the lower-dimensional projection. The cross-entropy used in UMAP preserves both a sense of local and global structure, while the t-SNE KL divergence preserves only local structure (as discussed in the introduction). We use the open t-SNE implementation from Poličar et al. (2019) on the full $2 \times 55 - 1$ dimensional space. A high perplexity (of $\sim$100) seems a good choice to maximise the separation between the two clusters. We show the results in Fig. 3 and see that there is a clear island that we can associate with C-rich stars. However, it is not as clearly separated from the bulk of the O-rich spectra. Furthermore, there is significant clumping within the O-rich region which we believe is more a reflection of the high perplexity choice than any sets of stars within the expected smooth continuum. Also, the S-star island is associated with the C-rich island while the more global metric preserving UMAP places it closer to the O-rich island.

We further experiment by using as inputs to UMAP the externally-calibrated spectra on a wavelength grid (336– 1020 nm with 2 nm spacing). We first normalize the spectra by dividing them by the sum of the binned flux values before performing the same procedure as applied to the coefficients. We show the results in the third panel of Fig. 3. We have found in general there is a very similar degree of separation using both the coefficient and the flux space. The dimensionality of the coefficient space is smaller (although our wavelength grid choice here is arbitrary) and the dimensions are anticipated to be less correlated than for the sampled flux space so we prefer the coefficient space.

Finally, we investigate truncating the BP/RP coefficients using the `xp_n_relevant_bases`. For each star, we set all higher-order coefficients to zero and construct the normalized externally-calibrated spectra on the previously mentioned wavelength grid. Running UMAP on this set produces the right panel of Fig. 3. This clearly has significantly more structure than when using the full spectra without truncation and colouring by truncation order in RP makes it clear that each cluster is linked to a specific truncation

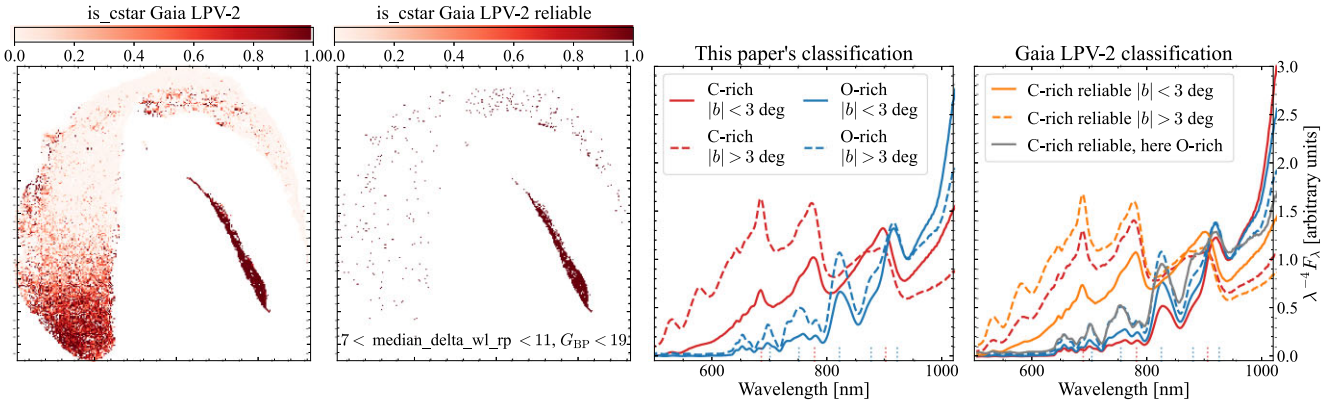

Figure 4. Comparison with the *Gaia* DR3 2nd LPV catalogue classifications. The left panel shows the UMAP projection of the high-amplitude sample coloured by the `is_cstar` flag from the Gaia catalogue. The second panel restricts to the reliable flags with $7 <$ `median_delta_wl_rp` < 11 and $G_{\rm BP} < 19$. In the right two panels, we show the median normalized BP/RP spectra classified using this paper's methodology (third panel) and the Gaia DR3 methodology (fourth panel). We split by those classified as C-rich (red) and O-rich (blue) at high ($|b| > 3$ deg, dashed) and low ($|b| < 3$ deg, solid) latitudes, respectively. In the fourth panel, the orange lines show the median of the reliable C-rich classifications from *Gaia*. Finally, the grey line in the right panel shows the median of the spectra classified here as O-rich but as reliable C-rich in the *Gaia* DR3 release. From this, it is evident that the reportedly reliable classifications from *Gaia* DR3 are on the whole robust for separating O-rich and C-rich, but the suspected unreliable low-latitude C-rich classifications are primarily O-rich stars.

order. Furthermore, the stars identified as C-rich in the UMAP projection of the coefficients are split and not clearly identifiable as an association. It, therefore, seems, at least in this case, that truncation is significantly deteriorating the results. Our case may be special as we are considering very red (both intrinsically and due to interstellar dust) and variable objects. There is clearly significant information in the higher-order terms that is being neglected by a hard cut. As we will see, the basis expansion also appears sub-optimal for the C-rich stars as the chi-squared/standard deviation of the fit for these stars is significantly larger than for the O-rich.

2.4 Comparison with *Gaia* DR3 classifications

In the *Gaia* DR3 LPV candidate catalogue (Lebzelter et al. 2022), there is already a flag for whether a source is O-rich or C-rich on the basis of the BP/RP spectra. This identification is based upon the separation of the most prominent peaks in the (internally-calibrated) RP spectra with C-rich spectra having more separated peaks than O-rich stars. As already noted by Lebzelter et al. (2022), the *Gaia* DR3 LPV `is_cstar` flag can produce unreliable results. Lebzelter et al. (2022) advises only using the classification if $7 <$ `median_delta_wl_rp` < 11 and $G_{\rm BP} < 19$. Without this cut, there is a high number of C-rich LPVs in the Galactic midplane and in particular in the Galactic bulge, somewhat in tension with the previous work discussed in the introduction. It appears that in the presence of significant interstellar dust the identification of the peaks fails possibly because the spectrum is so skewed to the red that the red edge of the RP bandpass is being identified as a peak associated with an absorption feature. In the lower panels of Fig. 4, we show the UMAP diagram coloured by the `is_cstar` flag both for the entire set and restricting to the reliably flagged stars. This demonstrates that the main cluster we have identified as C-rich is populated by stars reliably flagged as C-rich through the method of Lebzelter et al. (2022). However, there is also a clear overabundance of stars with non-zero `is_cstar` in the high-extinction end of the O-rich cluster. Even of those stars reliably flagged as C-rich, there are many that reside in the main O-rich cluster.

In the right panels of Fig. 4, we show the median of the normalized spectra for those stars we identify as O or C-rich split between a low- and high-latitude sample at $|b| = 3$ deg. We see the common features in both the low and high-latitude samples for each set. In the rightmost panel, we display similar using the `is_cstar` flag (not restricting to the reliable subset). Clearly, the low-latitude C-rich sample very closely resembles the low-latitude O-rich sample. Restricting to the reliable `is_cstar` subset, we see the high- and low-latitude C-rich samples have the characteristic C-rich features. Finally, we show the median spectrum of the reliable `is_cstar` stars that we have identified as O-rich. Again, this has evidence of the O-rich peak structure so we believe our assignment is more robust. Performing the same tests but separating the sample instead on A_0 from the *Gaia* DR3 optimized total galactic extinction map yields near-identical conclusions. In this case, however, even the reliable C-rich Gaia LPV-2 classifications for the high-extinction sample are clearly predominantly O-rich. In conclusion, we have found that our method offers an improvement over the *Gaia* DR3 LPV catalogue, especially for the reddest most extincted sources in *Gaia*. This is perhaps not surprising as we have used all the information from the spectra.

In the *Gaia* DR3 data release, there is a table of 'golden carbon stars' that have been selected from an initial list classified from the BP/RP spectra using a random forest classifier trained on synthetic data and *Gaia* data for known Galactic carbon stars (Gaia Collaboration et al. 2022a). The initial list of 386 936 candidates was filtered using the strength of the right two peaks marked on the C-rich spectrum of Fig. 1 to reduce the sample down to 15 740 'bona fide' carbon stars. 13 513 of the golden carbon stars are classified as LPVs in *Gaia* DR3 of which we classify 13 239 as C-rich using the classification of the full set in Appendix A. A total of 61 are classified as O-rich and 213 do not have BP/RP spectra in the *Gaia* DR3 data release. We have a total of 23 737 C-rich classifications based on the BP/RP spectra. For the high-amplitude subset described in the body of this paper, there are 1835 in the golden carbon star list of which only four are classified as O-rich by our algorithm. We find a total of 4203 C-rich LPV stars with BP/RP spectra.

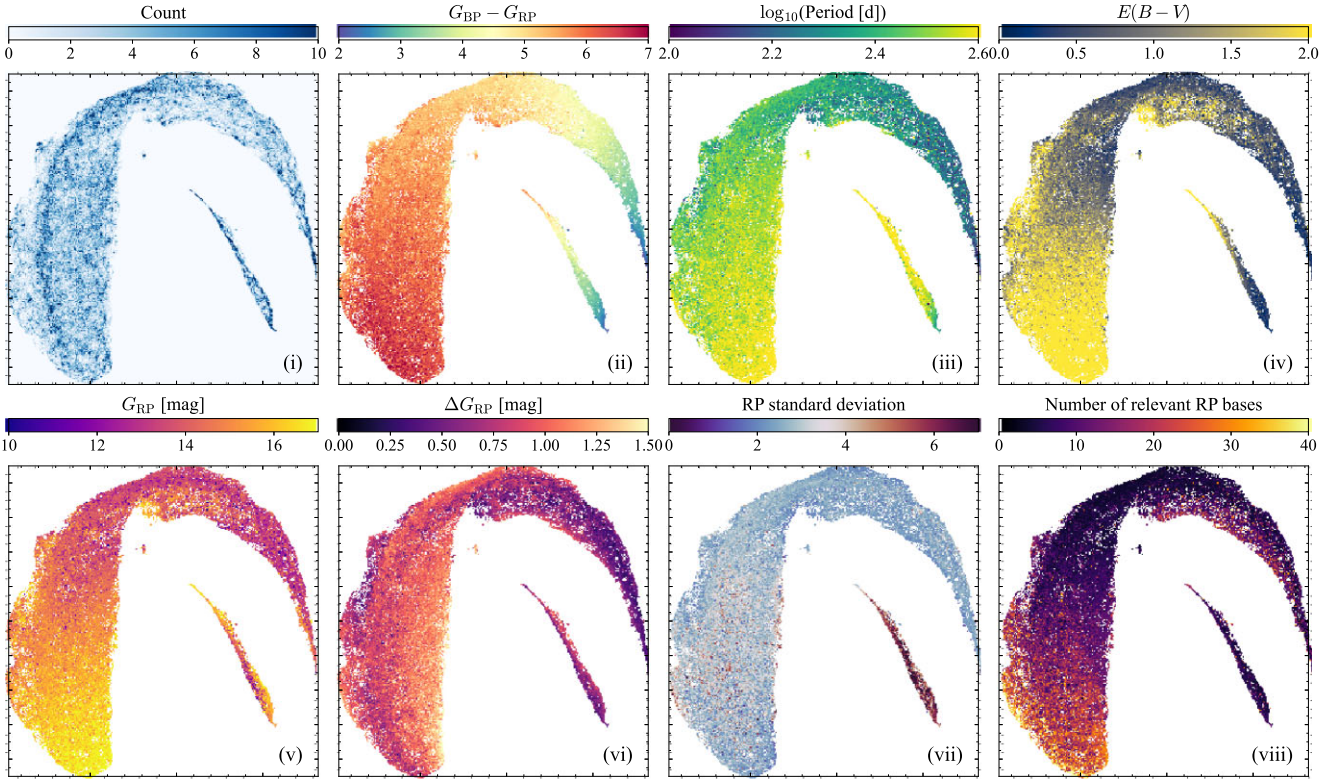

Figure 5. Two-dimensional UMAP projections coloured by the means of different binned quantities. We show the UMAP projection of the coefficients coloured by (i) the count per bin, (ii) the $G_{BP} - G_{RP}$ colour, (iii) the logarithm of the period, (iv) the Schlegel, Finkbeiner & Davis (1998) interstellar extinction, (v) the G_{RP} magnitude, (vi) the G_{RP} amplitude, (vii) the standard deviation of the RP spectrum fit, and (viii) the number of relevant RP basis functions.

2.5 The structure of the uniform manifold approximation and projection

We now investigate further the structure of the UMAP projection shown in Fig. 2, in particular focusing on the internal structure of the crescent and spur. Fig. 5 shows the UMAP projection coloured by various properties. From the UMAP diagrams, it is clear that both the O-rich and C-rich stars form approximately one-dimensional sequences with the O-rich sequence having a more significant width perpendicular to this. However, even along the O-rich crescent, there is a ridge line. The one-dimensional sequences are largely arising due to broad ($G_{BP} - G_{RP}$) variation as evidenced by panel (ii). This can arise either through effective temperature or extinction variation. For general stars, it is quite difficult to separate effective temperature and extinction variation using broadband photometry as the extinction vector typically lies near parallel to the stellar sequence. However, absorption line structure from spectroscopy is sensitive to the effective temperature of the stars and so allows for extinction-independent measurements. As seen in panel (iv), the interstellar $E(B - V)$ from Schlegel et al. (1998) approximately increases along the crescent. Using A_0 from the *Gaia* DR3 optimized total galactic extinction map produces near identical trends for the 75 per cent of our sample that fall in the on-sky region covered by the map. This trend with extinction suggests the direction perpendicular to the extinction gradient is due to detected effective temperature variation. This is somewhat evidenced by the amplitude variation, ΔG_{RP}, across the sequence shown in panel (vi). The period colouring in panel (iii) shows that period is also increasing around the sequence. This could be arising from intrinsic variations with longer-period cooler stars displaying distinct spectral features, perhaps arising from more circumstellar extinction through higher mass loss, or it could be because longer-period stars are associated with younger populations and so more confined to the higher extinction midplane. As already described, the relative abundance of molecular species gives a probe of effective temperature. As CN forms deeper in the atmosphere and is largely independent of the surface temperature for these stars, there is little effective temperature variation in the observed BP/RP spectra for C-rich stars and they lie along a narrow sequence in the UMAP diagram due almost entirely to extinction. On the contrary, in O-rich stars, it is expected that the abundance of TiO is a strong function of effective temperature (Reid & Goldston 2002). We define the relative depth of the TiO features as

$$\delta\mathrm{TiO} = \frac{\mathrm{TiO}\epsilon(\Delta\nu = -1) - \mathrm{TiO}\epsilon(\Delta\nu = 0)}{\mathrm{TiO}\epsilon(\Delta\nu = 0) - \mathrm{TiO}\delta(\Delta\nu = -1)}, \quad (1)$$

where

$$\mathrm{TiO}j = \int_{\lambda_j - \delta\lambda}^{\lambda_j + \delta\lambda} \mathrm{d}\lambda\, F_\lambda, \quad (2)$$

and $(\lambda_{\delta(\Delta\nu=-1)}, \lambda_{\epsilon(\Delta\nu=0)}, \lambda_{\epsilon(\Delta\nu=-1)}) = (779, 853, 940)$ nm are the locations of three prominent TiO bands as identified in Fig. 1. We choose $\delta\lambda = 15$ nm. δTiO is defined to be less sensitive to the broad spectrum shape which is sensitive to both effective temperature and extinction. Colouring the UMAP diagram by this feature as shown in Fig. 6, it is clear that the TiO band depth (or effective temperature) is varying across the O-rich UMAP sequence. This is particularly evident in the top half of the spur. In Fig. 4 of Lançon & Wood (2000), in warmer stars the depths of $\mathrm{TiO}\epsilon(\Delta\nu = 0)$ and $\mathrm{TiO}\delta(\Delta\nu = -1)$ are comparable but $\mathrm{TiO}\epsilon(\Delta\nu = -1)$ is weak. For the cooler stars, the $\mathrm{TiO}\epsilon(\Delta\nu = -1)$ depth increases to be similar to the relative depths

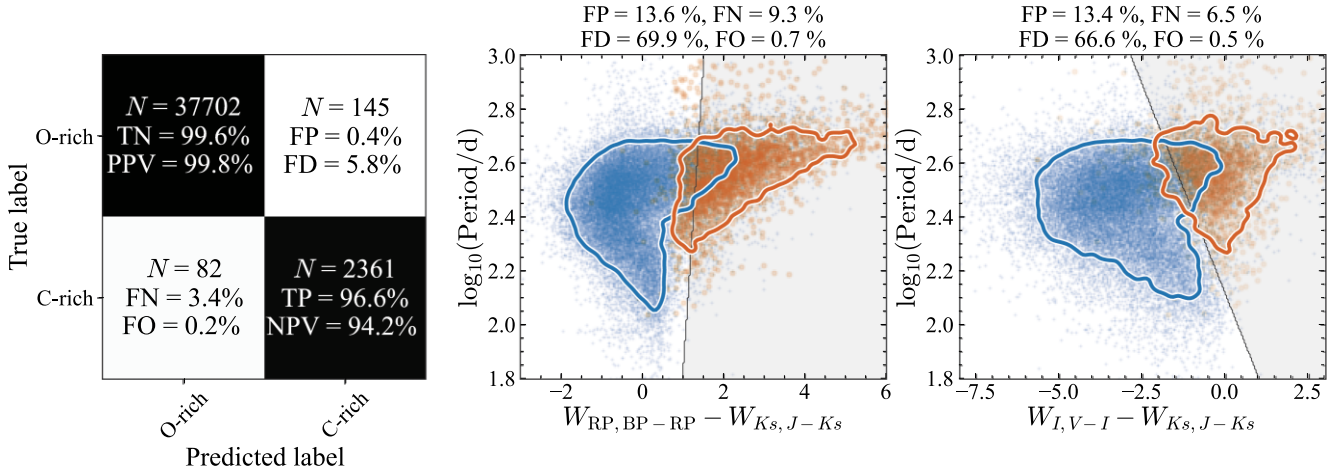

Figure 6. Comparison with literature identifications of C-stars (orange) and S-stars (green). The top panel shows the UMAP projection coloured by the TiO ratio constructed in equation (1). The SC stars are triangles and the MS stars are squares. Any stars with emission lines are outlined in black. The lower panel shows the median spectra of O-rich stars (M, thin blue), MS stars (green thicker), SC stars (orange short-dash), and those stars classified as C-rich on SIMBAD but classified as O-rich here (grey long-dashed).

of TiO$\epsilon(\Delta\nu = 0)$ and TiO$\delta(\Delta\nu = -1)$. Therefore, δTiO decreases as the star gets cooler.

Panel (vii) of Fig. 5 displays the standard deviation of the *Gaia* RP spectrum fit, i.e. the chi-squared per degree of freedom, as provided in the *Gaia* DR3 catalogue. Interestingly this displays a very clear separation between O-rich and C-rich sources with the C-rich sources having much poorer fits. This does not appear to be linked to their typically slightly fainter magnitudes (panel (v) shows there are many O-rich stars of similar magnitudes with similar quality fits) nor is it linked to their amplitudes as C-rich stars actually typically have lower amplitudes in the visual bands than the O-rich stars. Our explanation is that the RP basis function choice is optimized for a set of calibrators and C-rich stars are probably a minority population in this set (if present at all). As the C-rich spectral features are very distinct from the more typical O-rich stars, the RP basis functions do not completely capture the behaviour of C-rich stars. It is interesting then that more generically the RP standard deviation could be utilized to identify C-rich stars (although note that some O-rich stars can also have poorly fitting solutions for other reasons). Despite the poorer fits of the C-rich stars, panel (viii) shows that they have a low number of relevant RP bases. This is likely again because the ordering of the bases has been based upon typical calibrator stars, while C-rich stars probably have insignificant information in intermediate terms (that may capture the molecular features in an O-rich spectrum) and more significant information in the higher order terms. Note as well that higher extinction O-rich stars require more terms to capture their behaviour.

2.5.1 *S-stars and comparison with literature classifications*

S-stars have intermediate C/O ∼1 and subsequently chemistry that shares characteristics with both O-rich and C-rich stars. One identifying feature is the presence of ZrO molecular features. We already observed from Fig. 2 that a small clump of stars near the O-rich crescent had median BP/RP spectra indicative of S-stars. We have taken all S-stars from SIMBAD (Wenger et al. 2000) and matched 240 to our sample. We show their locations on the UMAP plane in Fig. 6. We also separate out MS type (those S-stars with more O-rich chemistry) and SC type (those with more C-rich chemistry, see Yao et al. 2017, for a clear illustration of the progression). The SC stars all lie on the C-rich spur while the MS stars lie in the O-rich crescent. Although they are distributed across the entire crescent, there are a number of overdensities, particularly on the underside of the crescent, and also crucially on the small island identified in Fig. 2. We reason that the overdensity of S-stars along the underside of the crescent is due to ZrO in these stars which lies at the same location in the spectrum as TiO$\epsilon(\Delta\nu = -1)$. Van Eck et al. (2017) show a series of spectra with increasing C/O where the TiO features weaken but the absorption at 940 nm stays quite constant due to the increasing contribution of ZrO. We have also indicated those stars classified in SIMBAD as emission line objects (presumed mostly from the identification of Hα and other Balmer lines, but this is a heterogeneous set) but these appear to be indistinguishable from the bulk of the stars.

Furthermore, we have obtained 10 000 C stars from SIMBAD (Wenger et al. 2000) and found 805 matches in our data set. These are also displayed in Fig. 6. The majority live along the C-rich spur with a few in the O-rich crescent. The median spectrum of the objects 'misclassified' by us as O-rich is shown in the lower panel of Fig. 6. It is not clear what the exact nature of these misclassified stars is but one can clearly see the TiO absorption at ∼ 850 nm so we are inclined to classify them as O-rich.

2.6 Supervised classification

As BP/RP spectra are only available for stars with $G < 17.65$, utilizing the BP/RP classifications alone would remove many highly-extincted stars. From the work of Lebzelter et al. (2018) it is clear broadband optical and infrared photometry can be used effectively to separate O-rich and C-rich sources. We use the previous classifications to train a gradient-boosted random forest classifier (XGBoost, Chen & Guestrin 2016). Due to the imbalance of the data set, we use weights inversely proportional to the number of each class in the data set. We have found the best performance is obtained using ($J - K_s$, $G_{BP} - G_{RP}$, $G - G_{RP}$, Period, ΔG_{RP}), where ΔG_{RP} is computed from `std_dev_mag_rp` (see Appendix B). We limit ourselves to considering stars with high-quality 2MASS photometry (`ph_qual` = 'A')

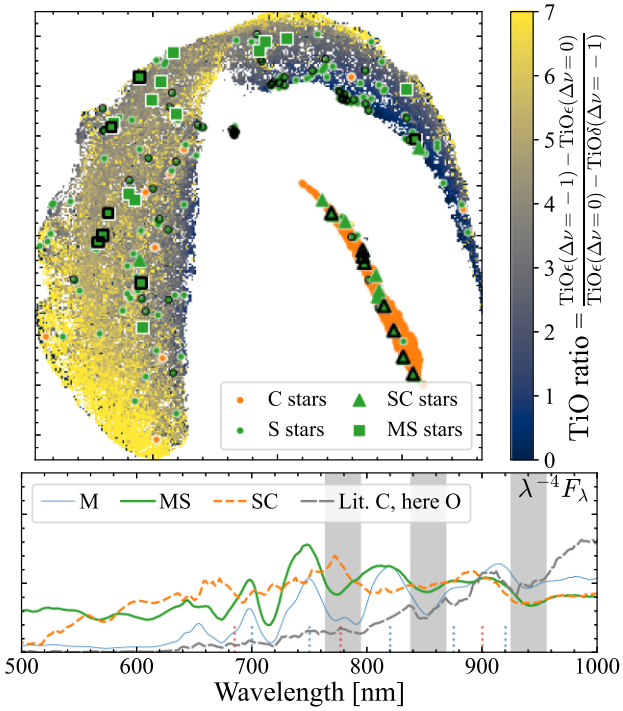

Figure 7. Supervised classification schemes: the left panel shows the confusion matrix for the application of XGBoost to the classification of O/C-rich stars from photometric colours, periods, and amplitude. We report the number and other statistics described in the text. The central panel shows two other optical and infra-red photometric spaces for separating O/C-rich stars. The blue and orange clouds show the O-rich and C-rich spectroscopic classifications with the contours containing 80 per cent of each data set. The grey line shows the best linear separation between the two classes.

and a low fraction of blended/contaminated BP and RP observations (`phot_bp/rp_n_contaminated/blended_transits/phot_bp/rp_n_obs`≤0.1) (Riello et al. 2021). The resulting feature importance is (0.19, 0.52, 0.13, 0.14, and 0.03). We also store the classification probabilities from XGBoost. We give the resulting confusion matrix in the left panel of Fig. 7 where we quote the total numbers of correct and incorrect classifications, the true positive, true negative, false positive, and negative rates (all normalized by the number of true classifications) and the positive predictive value (PPV, positive predicted as positive), the negative predictive value (NPV, negative predicted as negative), the false discovery (FD) rate (negative predicted as positive), and the false omission (FO) rate (positive predicted as negative, all normalized by the number of predictions made). Here 'positive' is a C-rich classification. For the identification of C-rich stars, the false discovery rate (related to the purity of the sample) of the C-rich predictions is the most important. Here, we find 5.8 per cent. We further only lose FN = 3.4 per cent of genuine C-rich stars so the completeness is also high. As the number of O-rich stars overwhelms C-rich stars, the purity of the O-rich sample (PPV) is very high (99.8 per cent). The classifier metrics are weak functions of G making their extrapolation to the fainter stars without BP/RP spectra valid.

We further inspect previously employed classification schemes based on optical and near-infrared data. As already evidenced in Fig. 2, the Wesenheit colour–colour indicator from Lebzelter et al. (2018) performs well to separate O-rich and C-rich even for non-LMC stars. We display the projection in the central panel of Fig. 7. We perform a simple linear support vector machine classification in the $W_{\mathrm{RP,\,BP\text{-}RP}} - W_{Ks,\,J-Ks}$ versus period space for the same sample of data used in the XGBoost models. The projection does perform well for low periods but for high extinction, the O-rich stars enter into the low-extinction C-rich region leading to high false discovery rates/contamination in any C-rich sample (FD is around 95 per cent for $E(B-V) > 3$). However, this space is appropriate for removing C-rich stars from an O-rich sample. Restricting to $|b| \gtrsim 5$ deg largely removes the highly-reddened sources. Although we do not have access to the OGLE photometry for the majority of our sample, as the BP/RP spectra cover the entire range of the OGLE V and I bands we can use them to simulate what OGLE would see (De Angeli et al. 2022). We use the filters from the SVO filter service (Rodrigo, Solano & Bayo 2012; Rodrigo & Solano 2020) and sum the BP/RP spectra on the wavelength grid reported for the filter. A cross-check for those stars with measured OGLE photometry demonstrates this procedure performs reasonably well but there are large uncertainties, particularly for faint V. We display the resulting distribution in the right panel of Fig. 7. It largely resembles the $W_{\mathrm{RP,\,BP\text{-}RP}} - W_{Ks,\,J-Ks}$ versus period projection but slightly rotated. For this reason, a linear support vector classifier performs similarly and suffers the same issue with highly-reddened stars.

3 VALIDATION OF THE CLASSIFICATION SCHEME

In Fig. 8 we display the on-sky distributions of the O-rich and C-rich long-period variable stars based on our unsupervised and supervised schemes. We see that in agreement with previous works the Galaxy is dominated by O-rich variables and the C-rich variables are biased more towards the outer Galactic disc. We also note the comparative excess of C-rich variables in the Small Magellanic Cloud and the Sagittarius dwarf spheroidal galaxy. We also notice the Galactic bulge contains some C-rich stars – we will return to this later. Our classification seems to agree with previous work indicating the bulk of long-period variables in the Milky Way are O-rich and the relative fraction of C-rich variables increases in the outer disc (Blanco et al. 1984; Ishihara et al. 2011).

We further validate our classification procedure by comparison to previously employed schemes based on colour–colour infrared photometry diagrams. We perform cross-matches of the sample with BP/RP spectra classifications to GLIMPSE, MSX, AKARI, WISE, and 2MASS (using a 1 arcsec crossmatch radius except for AKARI where we use 3). We apply the four-band zero-point corrections to the All-WISE data listed at https://wise2.ipac.caltech.edu/docs/release/neowise/expsup/sec2_1civa.html. In Fig. 9 we show five commonly employed colour–colour diagrams and display our classified objects. We further overlay the set of dusty AGB models from Sanders et al. (2022b). The AKARI projection has been advocated by Ishihara et al. (2011) and Matsunaga et al. (2017), the WISE diagram by Lian et al. (2014), the WISE/2MASS diagram Suh & Hong (2017), the GLIMPSE diagram by Groenewegen & Sloan (2018), and the MSX/2MASS diagram by Lewis et al. (2020a) and Lewis et al. (2020b). Clearly in all but the GLIMPSE diagram, our classification produces two distinct clusters of points. In the GLIMPSE diagram, there is a low number of C-rich sources and also the extinction acts to make stars bluer in ([5.8] − [8.0]). We run a linear support vector machine classifier in each colour–colour space, balancing each class using weights inversely proportional to their number in the data set. Above each panel of Fig. 9 we give the false positive, false

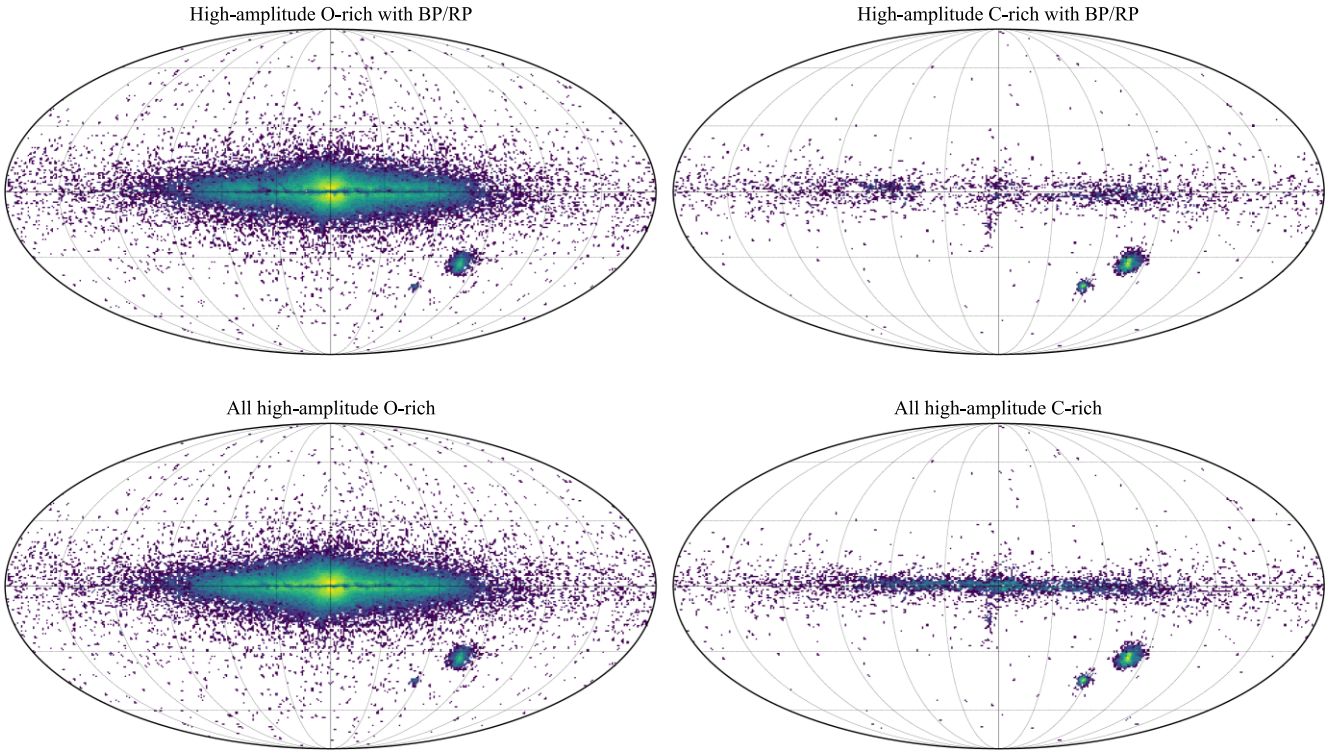

Figure 8. On-sky distribution in Galactic coordinates of the unsupervised classifications based on BP/RP spectra (top row) and the supervised classifications using photometry. The left panels are O-rich classification whilst the right are C-rich.

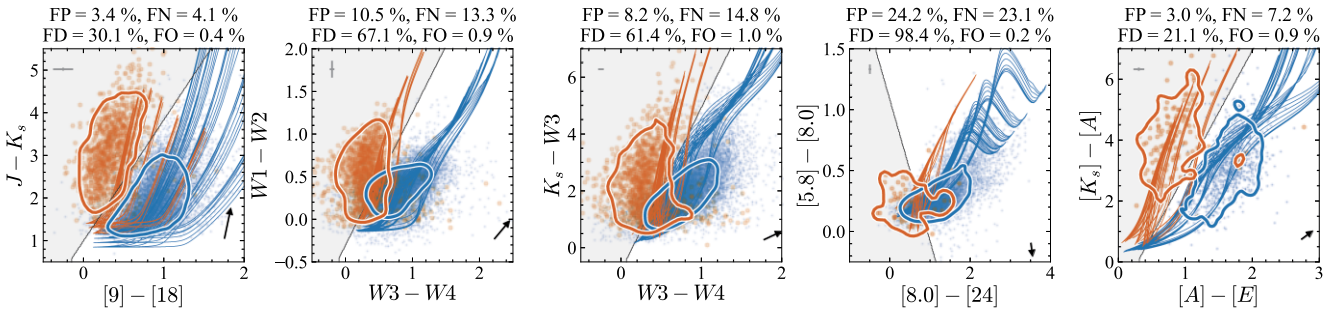

Figure 9. Infrared colour–colour diagrams with the classified O-rich (blue) and C-rich (orange) high-amplitude LPVs. The contour contains 80 per cent of each set. The lines are sequences of dusty O-rich and C-rich models. The grey region shows the best linear support vector classifier for separating O-rich and C-rich. We report the false positive (FP, percentage of O-rich stars classified as C-rich), false negative (FN, percentage of C-rich stars classified as O-rich), false discovery (FD, percentage of stars classified as C-rich that are O-rich), and false omission (FO, percentage of stars classified as O-rich that are C-rich) above each plot. The small arrow is the extinction vector direction using the results from Fritz et al. (2011, the photometry is not dereddened) and the grey errorbar is the median uncertainty in the photometry.

negative, false discovery, and false omission percentages of the C-rich classifications i.e. the fraction of 'true' O-rich stars classified as C-rich, the fraction of 'true' C-rich stars classified as O-rich, the fraction of C-rich classifications that are 'truly' O-rich and the fraction of O-rich classifications that are 'truly' C-rich. Both the AKARI/2MASS ($[9] - [18]$) versus ($J - K_s$) space (advocated by Matsunaga et al. 2017) and the MSX/2MASS ($[A] - [E]$) versus ($[K_s] - [A]$) space (advocated by Lewis et al. 2020a, b) produce good separations of the populations with false positive rates for O-rich and C-rich classification of around 5 per cent. The other diagrams are typically poorer due to the overlap in O-rich and C-rich stars in the bluer parts of the diagrams suggesting the differing circumstellar dust is the primary driver for the separation in these diagrams and when it is absent, there is limited photometric difference between the populations. The reported statistics for each colour–colour diagram do not reveal the true efficacy of each colour–colour diagram as they are biased towards those Mira variables that are optically detected in *Gaia*. This naturally misses very red sources possibly highly embedded in circumstellar dust. Such sources are preferentially C-rich, so we would typically expect more C-rich sources from an infrared catalogue. This suggests e.g. the false-positive rate for the O-rich classification that we report is an optimistic (under-) estimate. However, from the models, it is evident that the redder sources are more easily distinguishable suggesting that even with redder C-rich sources in our sample, the false-positive rate for the O-rich classification will not change significantly.

4 POTENTIAL C-RICH BAR-BULGE MEMBERS

We close this work by addressing some of the questions raised in the introduction, namely, how many C-rich stars are there in the Galactic bulge and what is their nature. Fig. 8 has already demonstrated that there is a low number of stars classified as C-rich from the BP/RP spectra. We first restrict to those stars with semi-amplitudes between 0.6 and 2 to remove any semiregular variables and any spurious high amplitudes due to aliasing (see Appendix B). We identify reliable spectroscopic C-rich stars as those lying on the C-rich spur from Fig. 2 and that have supervised cross-validated classification probabilities of being C-rich of >0.9. We remove potential young stellar object contaminants ensuring no stars have $G - 5\log_{10}(100/(\varpi - 3\sigma_\varpi)) > 2.5(G_{\rm BP} - G_{\rm RP}) - 5$ and also restrict to stars with *Gaia* DR3 classification probabilities >0.5 or those classified as 'SYST'. For this subset, we have generated 100 samples from the BP/RP coefficient covariance matrix and run them through the unsupervised classifier. If any of the per-star samples is classified as O-rich we remove the star from the sample. For the photometric C-rich candidates, we ensure similar criteria but also ensure any star is not spectroscopically classified as O-rich. In this way, we end up with 2018 and 2687 spectroscopically and photometrically classified C-rich Mira variables respectively across the entire sky.

We display these samples in the top left panels of Fig. 10 as viewed from the Galactic North Pole. We have assigned approximate distances to the stars using the Wesenheit $W_{Ks,\,J-Ks} = K_s - 0.48(J - K_s)$ versus period relation for those stars within 10 deg of the LMC as shown in the top right panel of Fig. 10. The extinction coefficient is appropriate for the Galactic bulge (Nishiyama et al. 2009; Fritz et al. 2011; Alonso-García et al. 2017; Sanders et al. 2022a). Note the relatively tight relation followed by the LMC stars giving confidence in our amplitude cut for isolating only those stars on the Mira variable sequence (Wood 2000). We have fitted the linear relation $W_{Ks,\,J-Ks} = -4.5(\log_{10}(P/{\rm d}) - 2.3) + 10.7$ by-eye to these stars and use the distance modulus of 18.477 for the LMC (Pietrzyński et al. 2019). From the Galactic distributions, it is evident that there are both spectroscopically and photometrically identified C-rich Mira variables within the Galactic disc and the Galactic bar-bulge. We also see a clump of stars associated with the Sgr dwarf spheroidal galaxy. In both samples, it appears there is a truncation in the radial distribution inside a radius of ~ 5 kpc which may correspond to the corotation of the bar. Inside this radius, we observe an approximate barred structure aligned at approximately 20 deg with respect to the line-of-sight (in agreement with other observations, e.g. Wegg & Gerhard 2013; Simion et al. 2017).

We isolate stars consistent in projection with bar-bulge membership as $|\ell| < 20$ deg and $|b| < 20$ deg. These are shown as black points in the top right panel of Fig. 10. We see three peaks in $W_{Ks,\,J-Ks}$ – a foreground disc population, a bulge population and the Sgr stars. The separation between the bulge and Sgr peak is not particularly clean possibly due to background disc stars. Interestingly, the mean period is smallest in Sgr and largest in the foreground population. We will discuss this further below. We also display the C-rich Mira variables stars from Matsunaga et al. (2017) using their infrared photometry and the symbiotic star from Miszalski et al. (2013). We isolate the Galactic bulge population as $\Delta s = |s - s_0| < 3$ kpc where s_0 is the distance to the Galactic centre. This region is shown in the left panels of Fig. 10. There are 56 spectroscopically and 269 photometrically-identified C-rich stars in this region. In the lower right panel of Fig. 10 we show the median spectrum of these stars – it is evident that they are predominantly C-rich.

In the lower row of Fig. 10, the velocities of the C-rich stars in the on-sky bar-bulge region are shown where the members of Sgr are visible. We also display the transverse velocity distributions of the two samples. Fitting a Gaussian to the photometric sample we find dispersions of 132.9 and 95.1 km s^{-1} in the longitudinal and latitudinal directions, respectively (the proper motion uncertainties are of order 10 km s^{-1} so unimportant here). These dispersions are very similar to what is observed for the red clump giant stars (Sanders et al. 2019). Also, the distributions have limited evidence of substructure. This then suggests that the C-rich stars are drawn from approximately the same population as the red clump giant stars and more generally the bulk bar-bulge population.

4.1 C-rich bar-bulge member scenarios

In the introduction we highlighted three possible reasons for C-rich bar-bulge stars: (i) there is recent (metal-poor) star formation in the Galactic bar-bulge, (i) they are accreted metal-poor stars, or (iii) the C-rich stars are formed primarily through binary interaction. These three scenarios are shown schematically in Fig. 11. We discuss the evidence for each scenario in turn.

4.1.1 In situ bar-bulge star formation

The formation of C-rich stars through dredge-up is easier at lower metallicities as less carbon is required to counteract the already present oxygen. Lower mass stars have weaker dredge-up episodes meaning C-rich star production is a function of both mass and metallicity. Fig. 8 from Boyer et al. (2013) shows that the upper limit in age at a given metallicity for the formation of C-rich stars is given approximately by

$$\log_{10}\frac{\tau_{\rm C}}{\rm Gyr} \approx 0.95 - \exp(1.3([{\rm Fe/H}]-0.35)+0.8([{\rm Fe/H}]+0.6)^3), \quad (3)$$

as depicted in Fig. 11. This means at metallicities of ~ -2 dex, C-rich stars can be as old as ~ 8 Gyr suggesting that C-rich bar-bulge stars could be remnants from the very earliest metal-poor phase of the bar-bulge region. However, the oldest C-rich Mira variables will also have the shortest periods. From C-rich Mira variables in the solar neighbourhood, Feast, Whitelock & Menzies (2006) concludes stars with $\log_{10}P \approx 2.62$ have ages of ~ 2.5 Gyr which would correspond to masses of $\sim 1.6 M_\odot$. A compilation of literature results (Wyatt & Cahn 1983; Feast & Whitelock 1987; Eggen 1998; Feast & Whitelock 2000; Feast et al. 2006; Feast & Whitelock 2014; Catchpole et al. 2016; López-Corredoira 2017; Grady et al. 2020; Nikzat et al. 2022; Sanders et al. 2022b) suggests a simple approximation for the Mira variable period–age relation of $\tau \approx 6.5(1 + \tanh((330 - P/{\rm d})/250))$ although recent theoretical relations (Trabucchi & Mowlavi 2022) predict younger ages at fixed period. It is likely there is some metallicity dependence to the period–age relation for the Mira variables (Trabucchi & Mowlavi 2022) but this is unlikely to make the ages at fixed periods significantly older than this. Utilizing this relation, our sample of C-rich stars with $2.45 < \log_{10}P/{\rm d} < 2.75$ have ages between 1.7 and 7.7 Gyr as shown by the band and the full distribution in Fig. 11.

Early investigations of the bar-bulge star formation history concluded that it was predominantly an early ~ 10 Gyr old burst (Zoccali et al. 2003). However, evidence from the spectroscopic study of main-sequence turn-off stars (Bensby et al. 2013) has pointed towards a small fraction of younger (~ 3 Gyr) stars. This younger minority population is supported by proper-motion-cleaned colour-magnitude

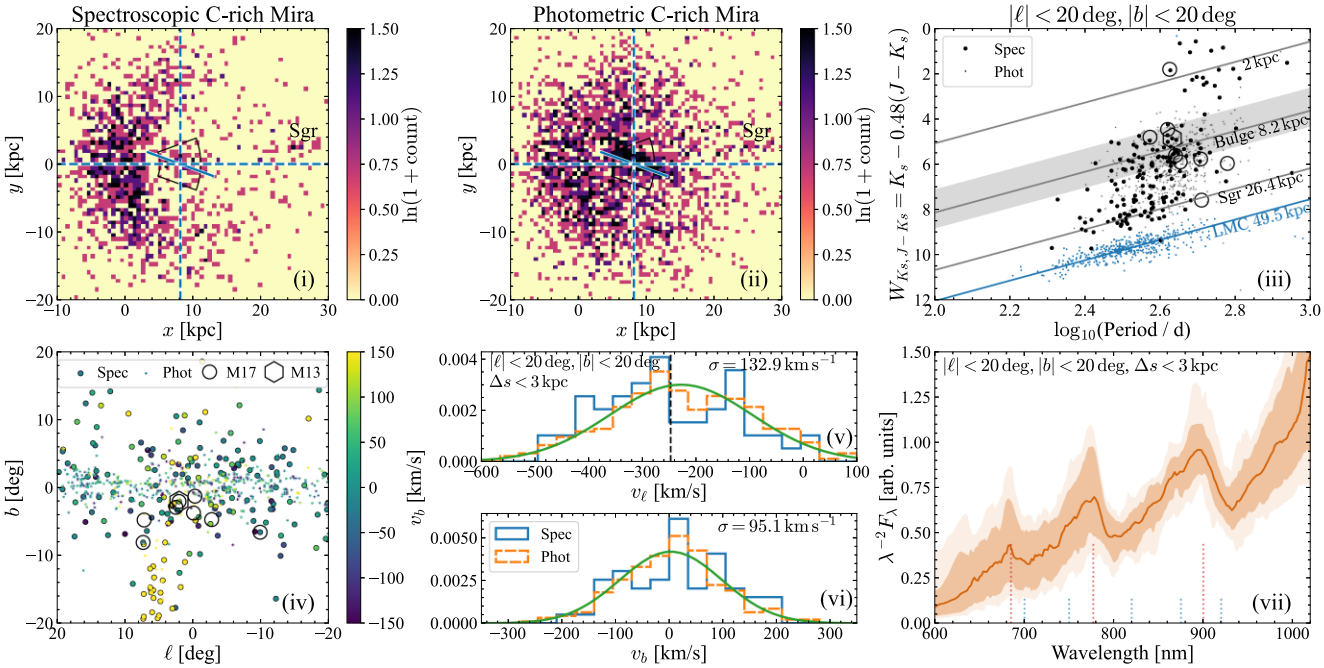

Figure 10. C-rich Mira variables within the Galactic bulge – panels (i) and (ii) show the view from the Galactic North Pole of those high-amplitude (> 0.6 mag) stars (i) spectroscopically and (ii) photometrically classified as C-rich. The vertical line is at the Galactic centre distance and the small tilted line is at an angle of 20 deg. Panel (iii) shows the Wesenheit magnitude $W_{Ks,J-Ks} = K_s - 0.48(J - K_s)$ against period for spectroscopically identified C-rich stars within 10 deg of the LMC (blue dots) and then those within $|\ell| < 20$ deg and $|b| < 20$ deg both spectroscopically (large black) and photometrically (small black) identified. The hexagon is the symbiotic C-rich star from Miszalski et al. (2013) and the circles are the C-rich Mira variables identified by Matsunaga et al. (2017, using their mean photometry). Panel (iv) shows stars in this region coloured by Galactic latitude proper motion with the Miszalski et al. (2013) and Matsunaga et al. (2017) stars also plotted. The histograms in panels (v) and (vi) show the transverse velocity distributions of stars in this region within 3 kpc of the Galactic centre (dashed is photometric identifications) with the best-fitting Gaussians to the photometric identifications in green. Panel (vii) gives the median BP/RP spectrum (with $\pm 1, 2\sigma$ brackets) for the spectroscopic C-rich bulge identifications.

diagrams (Bernard et al. 2018) and corroborated by further age indicators (as discussed by Nataf 2016). In Fig. 11 we show the age-metallicity relation derived by Bernard et al. (2018) along with the star formation history they infer. We see that although there is a weak tail of star formation in the bar-bulge that extends to more recent times, the bar-bulge has enriched to on average supersolar metallicities by this time. The typical metallicity dispersion at each age is not well characterized, but from the results of Bernard et al. (2018) 0.2 dex is a reasonable value. We can estimate the fraction of C-rich stars within the Mira variable star population using

$$\frac{N_\mathrm{C}}{N_\mathrm{total}} = \int_{t_\mathrm{min}}^{t_\mathrm{max}} \int_{-\infty}^{\infty} \mathrm{d}Z\mathrm{d}t\, \Gamma(t)\mathcal{N}(Z|Z_b(t), 0.2)\Theta(\tau_\mathrm{C}(Z) - t), \qquad (4)$$

where $\Gamma(t)$ is the star-formation history and $\mathcal{N}(Z|Z_b(t), 0.2)$ is a Gaussian with mean $Z_b(t)$ tracing the age-metallicity relation of the bar-bulge and dispersion 0.2 dex (both shown in Fig. 11). $t_\mathrm{min} = 1.07$ Gyr and $t_\mathrm{max} = 8.48$ Gyr are the minimum and maximum ages corresponding to the observed period spread ($2.4 < \log_{10}P < 2.8$). $\Theta(x)$ is a Heaviside step function (evaluating to 1 for $x > 0$ and 0 otherwise). This equation assumes that all stars are Mira variables for a similar time irrespective of the period. Although the TP-AGB phase is shorter for more massive stars, a higher fraction of this time is spent pulsating in the fundamental mode, so the relationship between Mira variable lifetime and mass is not simple (Trabucchi et al. 2019). Furthermore, we are assuming the two fields inspected by Bernard et al. (2018) are representative of the entire bar-bulge region. Using this relation, we find $N_\mathrm{C}/N_\mathrm{total} \approx 1 \times 10^{-4}$. In the spectroscopically-classified sample, we have $N_\mathrm{C}/N_\mathrm{total} \approx 3 \times 10^{-3}$. This theoretical calculation is slightly sensitive to the poorly-constrained low star-formation rate tail. Reasonable variations consistent with the star formation history from Bernard et al. 2018 typically produce a factor of two variation in $N_\mathrm{C}/N_\mathrm{total}$ and to produce $N_\mathrm{C}/N_\mathrm{total} \approx 3 \times 10^{-3}$ requires star formation histories strongly inconsistent with Bernard et al. (2018). This then suggests that the star-formation history of the bar-bulge cannot explain the C-rich stars we observe. Furthermore, in the single-star model (blue line) the C-rich stars are predominantly skewed to younger ages/longer periods such that $N_\mathrm{C}/N_\mathrm{total} \approx 5 \times 10^{-9}$ for $2.4 < \log_{10}P/\mathrm{d} < 2.6$ and the model would also predict significant numbers of C-rich stars with $\log_{10}P/\mathrm{d} > 2.8$. This means the full age distribution of the C-rich Mira variables is a poor match to the data (see lower panel of Fig. 11). However, this calculation also shows that the lack of C-rich stars for $\log_{10}P > 2.75$ is putting strong limits on the star formation in the bar-bulge in the last ~Gyr, while from Fig. 10 we see the local disc stars have younger C-rich members.

The spatial and kinematic distributions suggest the C-rich population follows the bulk population in the bar-bulge despite being younger in the *in situ* formation scenario. However, the sample of microlensed dwarfs from Bensby et al. (2017) shows a similar extended distribution even for the young to intermediate-age stars. Debattista et al. (2017) argue that the spatial (and kinematic) distributions of different age/metallicity populations are a reflection of the different velocity dispersions of the populations prior to bar formation meaning different age populations should be distinguishable spatially and kinematically but younger populations still are anticipated to be present at higher latitude. This line of evidence alone does not completely rule against the *in situ* scenario. However,

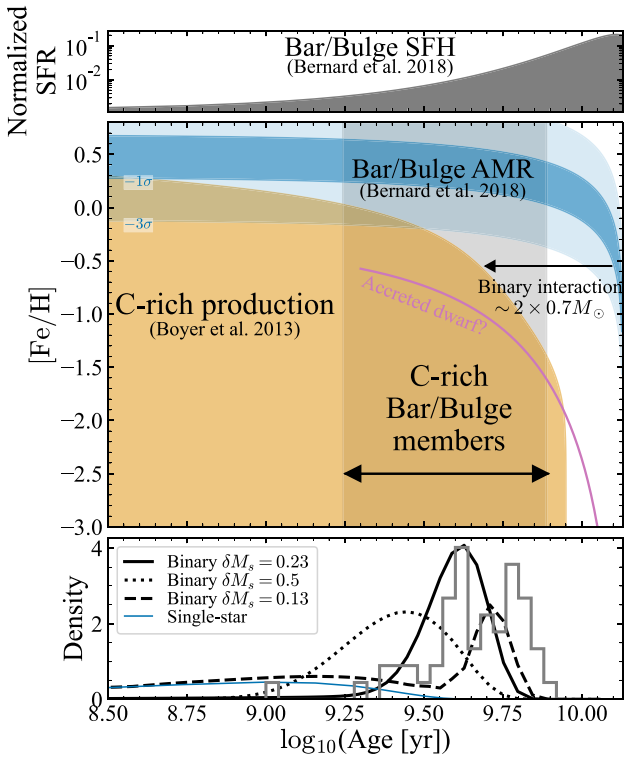

Figure 11. Likely scenarios for C-rich formation in the bar-bulge. The grey band shows the approximate age range of the spectroscopic C-rich candidates with the full distribution shown in the bottom panel. The orange-shaded region shows the range of age–metallicity combinations that give rise to C-rich TP-AGB production (Boyer et al. 2013). The blue band shows the ± 1 and 3σ of the age–metallicity relation for the Galactic bar-bulge corresponding to the bar-bulge star formation history shown in the top panel (Bernard et al. 2018). The pink line is an example age–metallicity relation for a dwarf galaxy that may have hypothetically merged into the bar-bulge. The horizontal leftwards arrow shows the approximate shift in remaining stellar age for the merger of two 0.7 $M_\odot$ stars – in this way, binary interaction can produce old C-rich stars. The bottom panel shows the predictions from the single-star scenario in blue, and three binary scenario variants in black labelled by the width of the blue straggler mass distribution, δM_s.

when combined with the predictions from the star formation history, the *in situ* star formation scenario is difficult to reconcile.

4.1.2 Accreted metal-poor population

We have observed that the bar-bulge population reaches too high metallicity at too early an epoch to explain the C-rich population observed. The next natural explanation is to invoke a more metal-poor star formation environment which subsequently merged into the bar-bulge region. For example, we have drawn a purely hypothetical age-metallicity track for a dwarf galaxy on Fig. 11. A very ancient merger such as the suggested Kraken/Heracles (Kruijssen et al. 2020; Horta et al. 2021a) would lack C-rich stars, and the dwarf must have been accreted in the last ~ 5 Gyr. The Sgr dwarf galaxy is a likely candidate here. The lower right panel of Fig. 10 shows all stars in the bar-bulge region of the sky coloured by proper motion. The Sgr dwarf is visible in proper motions with the suggestion there are other C-rich stars along the stream north of the Galactic plane. However, these stars are all at much further distances. The C-rich bar-bulge members are morphologically neither similar to the Sgr distribution, nor do their kinematics suggest any association with Sgr. Another merger event that is perhaps more radial and more phase-mixed than Sgr is required. Inspecting Fig. 10 we see that on average the C-rich bar-bulge stars are longer period, or younger, than their counterparts in both Sgr and the LMC (the bar-bulge population has a mean of $\log_{10}P = 2.6$, while the LMC and Sgr have 2.51 and 2.55, respectively). This then requires us to invoke a slightly peculiar star formation history for this suggested dwarf galaxy where there is only significant star formation recently. The problem is exacerbated if the dwarf galaxy is more metal-poor than Sgr and LMC. Furthermore, the minimum period of the bar-bulge C-rich Mira variables is also longer than that of the Sgr and LMC Mira variables. If we assume each group represents an approximately mono-metallicity population, Fig. 11 shows that a longer period minimum suggests a more metal-rich population. This suggests the progenitors of the bar-bulge C-rich Mira variables are more metal-rich than Sgr and LMC giving further evidence against an accreted population explanation. As discussed when considering Sgr, the spatial distribution (both on-sky and in three-dimensional) and the kinematic distributions do not give any suggestion of being distinct from the broader bar-bulge population. Therefore, while there likely exist some merger configurations and star formation histories that could reproduce all observational constraints, the merger scenario explaining a significant number of the observed bar-bulge C-rich Mira variables does seem improbable.

4.1.3 Binary channels

In addition to the single star channels, C-rich stars can form through a binary channel. Binary mass transfer increases the mass of the secondary potentially to the extent that it is of high enough mass to later become a C-rich star. In extreme cases, a stellar collision can approximately double the mass of a star. If the primary companion is itself a C-rich star, then it may require lower mass transfer to make the secondary C-rich. In Fig. 11 we show the shift in apparent age produced by the merger of two ~ 12 Gyr old $0.7M_\odot$ stars (assuming $\tau \propto M^{-2.5}$). At metallicity ~ -0.5 dex this change in mass is sufficient to bring the star into the C-rich formation region. These binary products would first appear as blue straggler stars before eventually evolving to C-rich stars through dredge-up.

The production of C-rich Mira variables in old environments is evidenced by the presence of a C-rich Mira variable in the globular cluster Lyngå 7 (Feast, Menzies & Whitelock 2013). Its radial velocity is consistent with membership of the globular cluster although the *Gaia* DR3 proper motion measurement is inconsistent possibly due to contamination in the cluster environment (there are two nearby sources with only two-parameter astrometric solutions). Feast et al. (2013) hypothesized that this star was formed through the collision/merger of two $\sim 0.8M_\odot$ stars producing a blue straggler star which subsequently evolved to be a C-rich Mira variable. There is the suggestion that ~ 27 per cent of bar-bulge stars were formed in globular clusters (Horta et al. 2021b) such that it is possible any blue stragglers in the bar-bulge are in fact the result of cluster evolution. However, the typical time-scale for blue stragglers to survive is ~ 1 Gyr meaning we will only be sensitive to cluster evolution products that formed in a cluster that very recently dissolved. None of the identified stars appears to be associated with globular clusters (the minimum separation relative to the Harris 2010 globular cluster list is 0.35 deg for the spectroscopic classifications and 0.17 deg for the photometric classifications). Globular clusters show no correlation between blue straggler fraction and density (Knigge, Leigh & Sills 2009) suggesting binary evolution rather than collisions form the

majority of blue stragglers in older systems. This is evidenced by the presence of blue stragglers in the field (e.g Carney et al. 2001). Complete mergers of close binary systems, rather than collisions in dense environments, are also a subdominant channel with old clusters producing $\lesssim 20$ per cent of blue stragglers via this channel (Geller, Hurley & Mathieu 2013; Leiner & Geller 2021). This suggests mass transfer is the dominant blue straggler production channel in old clusters and in particular in the field.

Although the binary fraction is lower in denser environments (Milone et al. 2012), the products of binary evolution have been observed in the bar-bulge. Clarkson et al. (2011) discovered ∼30 blue straggler bar-bulge members using proper-motion-cleaned colour-magnitude diagrams and photometric lightcurves in the *Hubble Space Telescope* SWEEPS field. They optimistically classify 29−37 stars as blue stragglers and more conservatively 18−22 depending on the assumption of a young bar-bulge population. There is also evidence of carbon-enhanced metal-poor stars and CH stars in the bar-bulge, although potentially at a lower fraction than the local disc fraction possibly due to the binary fraction variation with metallicity or density (Arentsen et al. 2021). Azzopardi et al. (1988) and Azzopardi et al. (1991) discovered a series of C-rich giant stars towards the bar-bulge that are too faint to be AGB stars so are likely products of binary evolution that could go on to be C-rich Mira variables.

Recently, Marigo et al. (2022) has studied the occurrence of C-rich TP-AGB in open clusters using the more reliable membership probabilities now possible using *Gaia*. They concluded that for the intermediate age (∼ 1.5 Gyr) clusters NGC 7789 and NGC 2660, the single star channel produces ∼10 − 1000 more C-rich TP-AGB stars than the binary channel (by anchoring to the observed number of blue stragglers in each cluster). Following the calculation in Marigo et al. (2022), we can relate the observed number of blue stragglers in the bar-bulge to the expected number of C-rich Mira variables as

$$\frac{N_{\rm C}}{N_{\rm BSS}} = \frac{\int {\rm d}M{\rm d}t{\rm d}Z\ \Gamma(t)p(Z)\tau_{\rm Mira}(M)p(M|t)\Theta(\tau_{\rm C}(Z)-\tau(M))}{\int_{1.41M_\odot}^{2.11M_\odot}{\rm d}M{\rm d}t\ \Gamma(t)\tau_{\rm MS}(M)p(M|t)} \tag{5}$$

where $p(Z) = \mathcal{N}(Z|Z_b(t), 0.2)$. The integration ranges in the numerator cover all valid Z and t (up to ∼ 13 Gyr) and from the main sequence turn-off mass up to infinity for M. In the denominator, we again consider all valid t but restrict to only considering blue-stragglers with masses $1.41 < M/M_\odot < 2.11$ as Clarkson et al. (2011) reports only being sensitive to these blue stragglers. We assume the relationship between main sequence age and mass is simply $\tau(M) \approx 10\,{\rm Gyr}(M/M_\odot)^{-2.5}$. $\tau_{\rm Mira}(M)$ is the approximate lifetime of the Mira phase which we assume is a constant 0.2 Myr based on the results from Trabucchi et al. (2019). $p(M|t)$ is the probability of producing a blue straggler of mass M in a population of age t. Both Leiner & Geller (2021) and Jadhav & Subramaniam (2021) provide estimates for this distribution in terms of the mass in excess of the main sequence turn-off mass, δM, based on results from *Gaia* for clusters. We fit an approximate half-Gaussian centred on zero to these distributions [the plotted 7 Gyr distribution from Leiner & Geller 2021 and Table 1 from Jadhav & Subramaniam 2021 for the 9.75 − 10 log(age) clusters] finding a standard deviation of $\delta M_s \approx 0.5 M_\odot$. Assuming a constant remaining blue straggler lifetime $\tau_{\rm MS}(M)$ with mass M, we find this calculation yields $N_{\rm C}/N_{\rm BSS} = 5.4 \times 10^{-5}/(\tau_{\rm MS}/\,{\rm Gyr})$. The largest uncertainty arises from the remaining blue straggler lifetime. Leiner & Geller (2021) consider several models for binary mass transfer finding the L2/L3 overflow model produces the best match to the cluster blue straggler distribution although not completely reproducing all features. For an old 7 Gyr population, the remaining main sequence lifetime from this model ranges from 400 Myr to 4 Gyr depending on the mass ratio. We adopt $\tau_{\rm MS} \approx 1$ Gyr giving $N_{\rm C}/N_{\rm BSS} = 5.4 \times 10^{-5}$ but note a factor ∼2 uncertainty in this number.

To compare with the number of blue stragglers found by Clarkson et al. (2011, $N_{\rm BSS,\,C11}$) we normalize by the respective stellar masses contained in the two areas considered:

$$\frac{A_{\rm SWEEPS}}{A_{\rm Bar-Bulge}} = \frac{\int_{\ell_{\rm SWEEPS}-\Delta\ell}^{\ell_{\rm SWEEPS}+\Delta\ell}{\rm d}\ell \int_{b_{\rm SWEEPS}-\Delta b}^{b_{\rm SWEEPS}+\Delta b}{\rm d}b\,{\rm d}\ell\cos b\rho(\ell,b)}{\int_{-\ell_{\rm max}}^{\ell_{\rm max}}{\rm d}\ell\int_{b_{\rm min}}^{\pi/2}{\rm d}b\,{\rm d}\ell\cos b\rho(\ell,b)} \tag{6}$$

where $\rho(\ell, b)$ is the bar-bulge density profile approximated as an exponential in ℓ and b with scalelengths of 3.5 deg and 1.3 deg, respectively (Wegg & Gerhard 2013). The density profile is integrated from $-\ell_{\rm max} = -20$ to $\ell_{\rm max} = 20$ deg in ℓ and for $b > b_{\rm min} = 1$ deg as extinction reduces the density of BP/RP C-rich detections from *Gaia* below this latitude. This calculation gives $A_{\rm SWEEPS}/A_{\rm Bar\text{-}Bulge} \approx 3.5 \times 10^{-5}$. Expanding to the full bar-bulge gives $A_{\rm SWEEPS}/A_{\rm Bar\text{-}Bulge} \approx 1.5 \times 10^{-5}$ i.e. 2.3 times more C-rich stars whereas we find ∼5 times more stars. This might reflect more contamination in the photometric samples or an inappropriate density law employed for the low-latitude regions. We estimate the expected number of bar-bulge spectroscopic C-rich Mira variables as $N_{\rm C} = N_{\rm BSS,\,C11}(N_{\rm C}/N_{\rm BSS})/(A_{\rm SWEEPS}/A_{\rm Bar\text{-}Bulge}) \approx 44$. This very nicely matches the 56 observed spectroscopic C-rich Mira variables but as discussed the uncertainty on the estimate is probably around a factor 2−3 as we can vary the blue straggler mass distribution (as discussed below), the remaining main sequence lifetime of blue stragglers, the lifetime of Mira variable stars and the specifics of the density modelling.

The advantage of this channel relative to the *in situ* formation channel described in Section 4.1.1 is that the peak of the predicted period distribution shifts to lower periods. We compute the expected period distribution by not integrating over M in the numerator of equation (5) to find the blue straggler mass distribution. We convert this mass distribution into effective age ($\tau \approx (10\,{\rm Gyr})(M/M_\odot)^{-2.5}$) and period distributions as shown in Fig. 11 using the relation given in Section 4.1.1 and the appropriate Jacobians. The mode of the distribution is $\log_{10}P \approx 2.70$ with width 0.05 dex. As shown in the lower panel of Fig. 11, this is not a particularly good match to the data which has median $\log_{10}P$ of ∼2.61 and width 0.07. The location of the peak period is a balance of increased mass to produce more C-rich stars at fixed metallicity while keeping the mass low enough to not overly bias towards longer period (younger) stars. It is the lower mass blue-stragglers that contribute to the lowering of the period distribution so to produce a lower mean period we must narrow the blue straggler mass distribution width, a not unreasonable suggestion considering the uncertainties and the use of a perhaps inappropriate $p(M)$ based on cluster stars. When we narrow the standard deviation of $p(M)$ to $\delta M_s = 0.13 M_\odot$ as shown in Fig. 11, we produce a high effective age (shorter period) peak from the blue straggler stars with masses greater than the turn-off mass, along with a broader low effective age (longer period) peak from stars with masses around the turn-off mass that approximately resembles the single-star distribution discussed in Section 4.1.1. Seeking a compromise we set the standard deviation of the $p(M)$ distribution as $\delta M_s = 0.23 M_\odot$ (solid black line in Fig. 11) and find a better match to the data with the mode of the distribution at $\log_{10}P \approx 2.64$ with width 0.09 dex. This choice reduces the expected number of C-rich Mira variables to around 20 but again the other uncertainties are large.

We have demonstrated that the binary channel can reproduce the observed number of C-rich Mira variables under reasonable assumptions and that it provides a better match to the period distribution than the single-star channel. Furthermore, of the three considered scenarios, the binary channel scenario is perhaps most consistent with the observation that both the spatial and kinematic distributions of the C-rich stars are very similar to that of the red clump giant stars in the Galactic bar-bulge. The red clump star distribution predominantly traces the properties of the bar-bulge for stars formed around the peak of star formation with a small bias towards red clump stars preferentially being found in younger populations. All these lines of evidence then indicate that the bulk of our observed bar-bulge C-rich Mira variable sample is likely formed through binary evolution.

5 CONCLUSIONS

The separation of O-rich and C-rich long-period variables is crucial for their precision use as distance tracers and indicators of the age/metallicity of stellar populations. Here, we have demonstrated the power of the *Gaia* BP/RP spectra for this task. Using a simple unsupervised approach based on the UMAP algorithm, we have naturally identified two broad groups of spectra that are associated with O-rich and C-rich objects. We have discussed how the unsupervised approach can be used to learn about the nature of the stars beyond their simple O/C separation, in particular how we can find some S-stars and also emission line objects that are possibly symbiotic. We have demonstrated how utilizing the information from the entire spectrum offers an improvement in the classification over simpler diagnostics. Our classification scheme has been further validated on the basis of infrared colour–colour diagrams and we have shown that a supervised scheme using *Gaia* and 2MASS photometry and the unsupervised classifications offers an improvement over simpler colour–colour cuts.

Using both the spectroscopic/unsupervised and photometric/supervised classifications we have identified a small population of C-rich stars in the Galactic bar-bulge region. Their spatial and kinematic distributions are in agreement with other bar-bulge tracers such as red clump giants suggesting they are an *in situ* population associated with the bulk of the bar-bulge. Their production via single-star evolution typically produces a factor of ten too few stars than observed because of the bar-bulge's predominantly early episodes of star formation that quickly enriches the inner Galaxy to high metallicity. Old high metallicity populations do not form C-rich stars. If we instead consider these stars as the products of binary evolution, we expect them to be the evolved versions of blue straggler stars. A rather simple model of the blue straggler production in the bar-bulge approximately reproduces the period distribution of our sample and the observed number of C-rich Mira variables across the entire bar-bulge when referencing against the observed number of blue-straggler stars in the SWEEPS field. This demonstrates that the entire population of C-rich Mira variables can be attributed to binary evolution and there is limited evidence for a significant young in-situ or accreted population.

Note that we have restricted our analysis to the long-period variables as the Mira variables are of particular interest for Galactic and cosmological studies. However, our analysis is simply extended to all stars in *Gaia*. Indeed it would be interesting to perform a dimensionality-reduction analysis on the entire BP/RP data set to identify and separate the gross stellar types but to also identify unusual outlier groups of stars, galaxies, or quasars.

ACKNOWLEDGEMENTS

JLS thanks the support of the Royal Society (URF\R1\191555), the hospitality of the Flatiron Institute, and the organizers of the Gaia Fête where this project was started. The authors thank Jo Ciucă for useful conversations in the preparation of this work. This paper made use of the Whole Sky Database (wsdb) created by Sergey Koposov and maintained at the Institute of Astronomy, Cambridge by Sergey Koposov, Vasily Belokurov, and Wyn Evans with financial support from the Science & Technology Facilities Council (STFC) and the European Research Council (ERC). This software made use of the Q3C software (Koposov & Bartunov 2006). This research has made use of the SVO Filter Profile Service (http://svo2.cab.inta-csic.es/theory/fps/) supported from the Spanish MINECO through grant AYA2017-84089. This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. This research has made use of the International Variable Star Index (VSX) database, operated at AAVSO, Cambridge, Massachusetts, USA. This paper made use of NUMPY (van der Walt, Colbert & Varoquaux 2011), SCIPY (Virtanen et al. 2020), MATPLOTLIB (Hunter 2007), SEABORN (Waskom et al. 2017), PANDAS (McKinney 2010) ASTROPY (Astropy Collaboration et al. 2013; Price-Whelan et al. 2018), and VAEX (Breddels & Veljanoski 2018).

DATA AVAILABILITY

All data used in this work are in the public domain. The authors have made our classifications and the UMAP coordinates for the full data set available here: https://www.homepages.ucl.ac.uk/~ucapjls/data/gaia_dr3_lpv_classifications.fits.

REFERENCES

Aaronson M., Blanco V. M., Cook K. H., Olszewski E. W., Schechter P. L., 1990, ApJS, 73, 841
Alonso-García J. et al., 2017, ApJ, 849, L13
Anders F., Chiappini C., Santiago B. X., Matijevič G., Queiroz A. B., Steinmetz M., Guiglion G., 2018, A&A, 619, A125
Andrae R. et al., 2022, preprint (arXiv:2206.06138)
Arentsen A. et al., 2021, MNRAS, 505, 1239
Astropy Collaboration, 2013, A&A, 558, A33
Azzopardi M., Lequeux J., Rebeirot E., 1988, A&A, 202, L27
Azzopardi M., Lequeux J., Rebeirot E., Westerlund B. E., 1991, A&AS, 88, 265
Belokurov V., Erkal D., Deason A. J., Koposov S. E., De Angeli F., Evans D. W., Fraternali F., Mackey D., 2017, MNRAS, 466, 4711
Belokurov V., Vasiliev E., Deason A. J., Koposov S. E., Fattahi A., Dillamore A. M., Davies E. Y., Grand R. J. J., 2022, MNRAS, 518, 6200
Bensby T. et al., 2013, A&A, 549, A147
Bensby T. et al., 2017, A&A, 605, A89
Bernard E. J., Schultheis M., Di Matteo P., Hill V., Haywood M., Calamida A., 2018, MNRAS, 477, 3507
Bhardwaj A. et al., 2019, ApJ, 884, 20

Blanco V. M., McCarthy M. F., Blanco B. M., 1984, AJ, 89, 636
Bobrovnikoff N. T., 1933, ApJ, 78, 211
Bovy J., Leung H. W., Hunt J. A. S., Mackereth J. T., García-Hernández D. A., Roman-Lopes A., 2019, MNRAS, 490, 4740
Boyer M. L. et al., 2013, ApJ, 774, 83
Breddels M. A., Veljanoski J., 2018, A&A, 618, A13
Brewer J. P., Richer H. B., Crabtree D. R., 1995, AJ, 109, 2480
Carney B. W., Latham D. W., Laird J. B., Grant C. E., Morse J. A., 2001, AJ, 122, 3419
Carrasco J. M. et al., 2021, A&A, 652, A86
Catchpole R. M., Whitelock P. A., Feast M. W., Hughes S. M. G., Irwin M., Alard C., 2016, MNRAS, 455, 2216
Chen T., Guestrin C., 2016, in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining. KDD '16. ACM, New York, NY, USA, p. 785
Chen Y.-P., Trager S. C., Peletier R. F., Lançon A., Vazdekis A., Prugniel P., Silva D. R., Gonneau A., 2014, A&A, 565, A117
Clarkson W. I. et al., 2011, ApJ, 735, 37
Creevey O. L. et al., 2022, preprint (arXiv:2206.05864)
De Angeli F. et al., 2022, preprint (arXiv:2206.06143)
De Marco O., Izzard R. G., 2017, PASA, 34, e001
Deason A. J., Belokurov V., Erkal D., Koposov S. E., Mackey D., 2017, MNRAS, 467, 2636
Debattista V. P., Ness M., Gonzalez O. A., Freeman K., Zoccali M., Minniti D., 2017, MNRAS, 469, 1587
Eggen O. J., 1998, AJ, 115, 2435
Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proc. Second Int. Conf. Knowledge Discovery Data Mining. KDD'96. AAAI Press, Palo Alto, California, U.S., p. 226
Evans D. W. et al., 2018, A&A, 616, A4
Feast M. W., Glass I. S., Whitelock P. A., Catchpole R. M., 1989, MNRAS, 241, 375
Feast M. W., Menzies J. W., Whitelock P. A., 2013, MNRAS, 428, L36
Feast M. W., Whitelock P. A., 1987, in Kwok S., Pottasch S. R., eds, Late Stages of Stellar Evolution, p. 33
Feast M. W., Whitelock P. A., 2000, MNRAS, 317, 460
Feast M. W., Whitelock P. A., Menzies J. W., 2006, MNRAS, 369, 791
Feast M., Whitelock P. A., 2014, in Feltzing S., Zhao G., Walton N. A., Whitelock P., eds, IAU Symp. Vol. 298, Setting the scene for Gaia and LAMOST. Cambridge University Press, Cambridge, p. 40
Fouesneau M. et al., 2022, preprint (arXiv:2206.05992)
Fraser O. J., Hawley S. L., Cook K. H., 2008, AJ, 136, 1242
Fritz T. K. et al., 2011, ApJ, 737, 73
Gaia Collaboration, 2016, A&A, 595, A1
Gaia Collaboration, 2022a, preprint (arXiv:2206.05870)
Gaia Collaboration, 2021, A&A, 649, A1
Gaia Collaboration, 2022b, preprint (arXiv:220800211)
Gavel A., Andrae R., Fouesneau M., Korn A. J., Sordo R., 2021, A&A, 656, A93
Geller A. M., Hurley J. R., Mathieu R. D., 2013, AJ, 145, 8
Glass I. S., Evans T. L., 1981, Nature, 291, 303
Gonneau A. et al., 2016, A&A, 589, A36
Gonneau A. et al., 2020, A&A, 634, A133
Grady J., Belokurov V., Evans N. W., 2019, MNRAS, 483, 3022
Grady J., Belokurov V., Evans N. W., 2020, MNRAS, 492, 3128
Groenewegen M. A. T., 2004, A&A, 425, 595
Groenewegen M. A. T., Sloan G. C., 2018, A&A, 609, A114
Grondin S. M., Webb J. J., Leigh N. W. C., Speagle J. S., Khalifeh R. J., 2023, MNRAS, 518, 4249
Harris W. E., 2010, preprint (arXiv:1012.3224)
Hasselquist S. et al., 2021, ApJ, 923, 172
Herwig F., 2005, ARA&A, 43, 435
Höfner S., Olofsson H., 2018, A&A Rev., 26, 1
Holl B. et al., 2018, A&A, 618, A30
Horta D. et al., 2021a, MNRAS, 500, 1385
Horta D. et al., 2021b, MNRAS, 500, 5462
Huang C. D. et al., 2018, ApJ, 857, 67
Huang C. D. et al., 2020, ApJ, 889, 5
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ishihara D., Kaneda H., Onaka T., Ita Y., Matsuura M., Matsunaga N., 2011, A&A, 534, A79
Ita Y. et al., 2004, MNRAS, 347, 720
Ita Y., Matsunaga N., 2011, MNRAS, 412, 2345
Iwanek P., Soszyński I., Kozłowski S., 2021, ApJ, 919, 99
Jadhav V. V., Subramaniam A., 2021, MNRAS, 507, 1699
Karakas A. I., 2014, MNRAS, 445, 347
Karakas A. I., Lattanzio J. C., 2014, PASA, 31, e030
Kastner J. H., Thorndike S. L., Romanczyk P. A., Buchanan C. L., Hrivnak B. J., Sahai R., Egan M., 2008, AJ, 136, 1221
Kim Y., Telea A. C., Trager S. C., Roerdink J. B., 2022, Inform. Visualization, 21, 197
Knigge C., Leigh N., Sills A., 2009, Nature, 457, 288
Kobayashi C., Karakas A. I., Lugaro M., 2020, ApJ, 900, 179
Koch A., McWilliam A., Preston G. W., Thompson I. B., 2016, A&A, 587, A124
Koposov S., Bartunov O., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, Astron. Soc. Pac. Conf. Ser. Vol. 351, Astronomical Data Analysis Software and Systems XV. Astronomical Society of the Pacific 390 Ashton Avenue, San Francisco, California 94112, p. 735
Kraemer K. E., Sloan G. C., Price S. D., Walker H. J., 2002, ApJS, 140, 389
Kruijssen J. M. D. et al., 2020, MNRAS, 498, 2472
Lançon A., Mouhcine M., 2002, A&A, 393, 167
Lançon A., Wood P. R., 2000, A&AS, 146, 217
Lebzelter T. et al., 2022, preprint (arXiv:2206.05745)
Lebzelter T., Mowlavi N., Marigo P., Pastorelli G., Trabucchi M., Wood P. R., Lecoeur-Taïbi I., 2018, A&A, 616, L13
Leiner E. M., Geller A., 2021, ApJ, 908, 229
Lewis M. O., Pihlström Y. M., Sjouwerman L. O., Quiroga-Nuñez L. H., 2020b, ApJ, 901, 98
Lewis M. O., Pihlström Y. M., Sjouwerman L. O., Stroh M. C., Morris M. R., BAaDE Collaboration, 2020a, ApJ, 892, 52
Lian J., Zhu Q., Kong X., He J., 2014, A&A, 564, A84
Liu C., Bailer-Jones C. A. L., Sordo R., Vallenari A., Borrachero R., Luri X., Sartoretti P., 2012, MNRAS, 426, 2463
Lloyd Evans T., 2010, JA&A, 31, 177
López-Corredoira M., 2017, ApJ, 836, 218
Lucey M. et al., 2022, preprint (arXiv:2206.08299)
MacConnell D. J., 1988, AJ, 96, 354
Marigo P. et al., 2022, ApJS, 258, 43
Matsunaga N., Menzies J. W., Feast M. W., Whitelock P. A., Onozato H., Barway S., Aydi E., 2017, MNRAS, 469, 4949
McInnes L., Healy J., Melville J., 2018, preprint (arXiv:1802.03426)
McKinney W., 2010, in van der Walt S., Millman J., eds, Proceedings of the 9th Python in Science Conference. p. 56
Milone A. P. et al., 2012, A&A, 540, A16
Miszalski B., Mikołajewska J., Udalski A., 2013, MNRAS, 432, 3186
Montegriffo P. et al., 2022, preprint (arXiv:2206.06205)
Mowlavi N. et al., 2018, A&A, 618, A58
Mürset U., Schmid H. M., 1999, A&AS, 137, 473
Nassau J. J., Velghe A. G., 1964, ApJ, 139, 190
Nataf D. M., 2016, PASA, 33, e023
Ng Y. K., 1997, A&A, 328, 211
Ng Y. K., 1998, A&A, 338, 435
Nikutta R., Hunt-Walker N., Nenkova M., Ivezić Ž., Elitzur M., 2014, MNRAS, 442, 3361
Nikzat F. et al., 2022, A&A, 660, A35
Nishiyama S., Tamura M., Hatano H., Kato D., Tanabé T., Sugitani K., Nagata T., 2009, ApJ, 696, 1407
Olnon F. M. et al., 1986, A&AS, 65, 607
Pietrzyński G. et al., 2019, Nature, 567, 200
Poličar P. G., Stražar M., Zupan B., 2019, bioRxiv
Price-Whelan A. M. et al., 2018, AJ, 156, 123
Reid M. J., Goldston J. E., 2002, ApJ, 568, 931
Reis I., Rotman M., Poznanski D., Prochaska J. X., Wolf L., 2019, preprint (arXiv:1911.06823)

Riebel D., Meixner M., Fraser O., Srinivasan S., Cook K., Vijh U., 2010, ApJ, 723, 1195
Riello M. et al., 2021, A&A, 649, A3
Rimoldini L. et al., 2019, A&A, 625, A97
Rix H.-W. et al., 2022, ApJ, 941, 45
Rodrigo C., Solano E., 2020, in Contributions to the XIV.0 Scientific Meeting (virtual) of the Spanish Astron. Soc. Spanish Astronomical Society, p. 182
Rodrigo C., Solano E., Bayo A., 2012, SVO Filter Profile Service Version 1.0, IVOA Working Draft 15 October 2012
Sanders J. L., Matsunaga N., Kawata D., Smith L. C., Minniti D., Lucas P. W., 2022b, MNRAS, 517, 257
Sanders J. L., Smith L., Evans N. W., Lucas P., 2019, MNRAS, 487, 5188
Sanders J. L., Smith L., González-Fernández C., Lucas P., Minniti D., 2022a, MNRAS, 514, 2407
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Secchi A., 1868, MNRAS, 28, 196
Semczuk M., Dehnen W., Schönrich R., Athanassoula E., 2022, MNRAS, 517, 6060
Sharpless S., 1956, ApJ, 124, 342
Simion I. T., Belokurov V., Irwin M., Koposov S. E., Gonzalez-Fernandez C., Robin A. C., Shen J., Li Z. Y., 2017, MNRAS, 471, 4323
Soszyński I. et al., 2009, AcA, 59, 239
Suh K.-W., Hong J., 2017, J. Korean Astron. Soc., 50, 131
Trabucchi M., Mowlavi N., 2022, A&A, 658, L1
Trabucchi M., Wood P. R., Montalbán J., Marigo P., Pastorelli G., Girardi L., 2019, MNRAS, 482, 929
Traven G. et al., 2017, ApJS, 228, 24
van der Maaten L., Hinton G., 2008, J. Machine Learning Res., 9, 2579
van der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., 13, 22
Van Eck S. et al., 2017, A&A, 601, A10
Verro K. et al., 2022, A&A, 660, A34
Virtanen P. et al., 2020, Nature Methods, 17, 261
Wallerstein G., Knapp G. R., 1998, ARA&A, 36, 369
Waskom M. et al., 2017, seaborn: v0.8.1. Zenodo
Wegg C., Gerhard O., 2013, MNRAS, 435, 1874
Wenger M. et al., 2000, A&AS, 143, 9
Whitehouse L. J., Farihi J., Green P. J., Wilson T. G., Subasavage J. P., 2018, MNRAS, 479, 3873
Witten C. E. C. et al., 2022, MNRAS, 516, 3254
Wood P. R., 2000, PASA, 17, 18
Wyatt S. P., Cahn J. H., 1983, ApJ, 275, 225
Xylakis-Dornbusch T., Christlieb N., Lind K., Nordlander T., 2022, A&A, 666, 58
Yao Y., Liu C., Deng L., de Grijs R., Matsunaga N., 2017, ApJS, 232, 16
Yuan W., He S., Macri L. M., Long J., Huang J. Z., 2017a, AJ, 153, 170
Yuan W., Macri L. M., He S., Huang J. Z., Kanbur S. M., Ngeow C.-C., 2017b, AJ, 154, 149
Zoccali M. et al., 2003, A&A, 399, 931

APPENDIX A: CLASSIFICATION OF THE ENTIRE *GAIA* DR3 LONG-PERIOD VARIABLE CATALOGUE

In the main body of the paper, we have focused on the high-amplitude variable stars in the *Gaia* DR3 long-period variable candidates catalogue. These are likely highly reliable but also contain the interesting Mira variable subset useful as a Local Group and cosmological distance and age tracer. In this Appendix, we extend the analysis to the entire LPV catalogue of 1 205 121 stars with *Gaia* BP/RP spectra. We run the same UMAP computation described in Section 2 on the full data set and display the results in Fig. A1. As with the high-amplitude sample, we see two distinct regions – a crescent of O-rich sources and a spur of C-rich sources. Again, the C-rich spur forms a near one-dimensional sequence corresponding primarily to variations in extinction (as the C-rich features are only weakly sensitive to effective temperature). However, unlike Fig. 2 we observe the spur is almost two overlaid one-dimensional sequences which we identify as due to the LMC/SMC sources and the Galactic sources respectively. The O-rich crescent is more structured than the C-rich spur due to the combination of extinction and effective temperature variation. If we consider the right part of the crescent, there are three overdense features. The right feature is composed of low extinction cooler stars often in the LMC/SMC while the middle sequence tends to be hotter stars without significant spectral features and with on average higher extinction. The left feature appears to be due to blended/contaminated sources as evidenced by their large ratio of the $G_{\rm RP}$ standard deviation to the equivalent in G. As we move clockwise around the crescent, the sources are typically higher extinction but also have the tendency to be cooler with more pronounced spectral features. The high-amplitude Mira variables sit on the right edge of the left part of the crescent. The directions in the UMAP diagram are awkward to map to physical dimensions but we have found moving across the crescent approximately maps into temperature variation while moving around the crescent maps into extinction variations. However, as evidenced by the diagram coloured by extinction, this is not a perfect mapping.

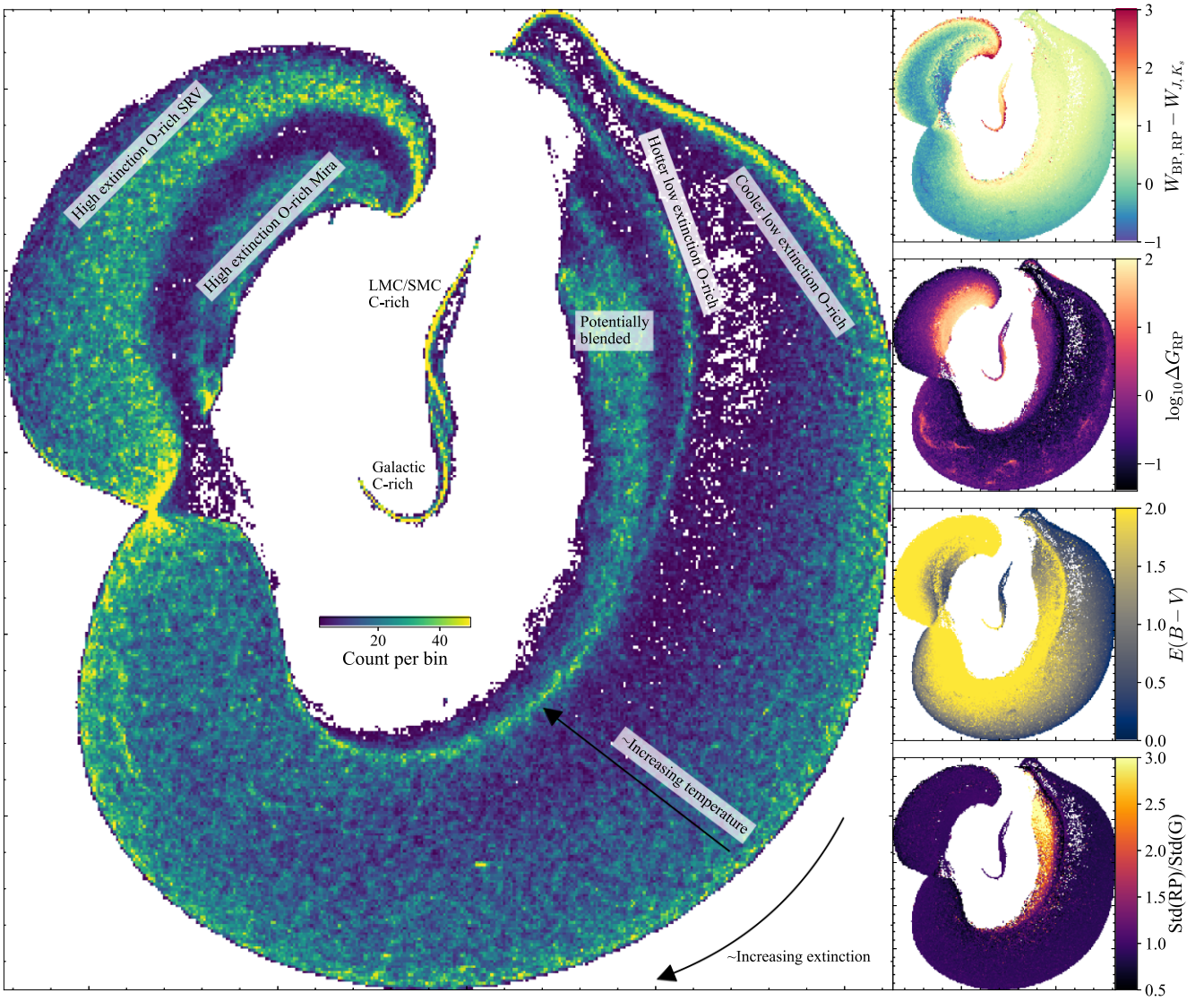

Figure A1. Two-dimensional UMAP projection of *all* Gaia DR3 long-period variable candidates. The main panel shows the counts per bin whilst each of the right subpanels shows the same diagram coloured by different properties.

APPENDIX B: COMPARISON OF AMPLITUDE MEASURES FOR LONG-PERIOD VARIABLES

Within the *Gaia* DR3 variable star catalogues, there are multiple measures of the variability amplitude. As highlighted by Belokurov et al. (2017), the *Gaia* photometric uncertainties contain variability information. As they are computed as errors in the mean of the epoch photometric measurements, the semi-amplitude can be estimated as

$$\Delta G_{\rm phot} = \frac{2.5\sqrt{2}}{\ln 10}\frac{\sqrt{\texttt{phot\_g\_n\_obs}}}{\texttt{phot\_g\_mean\_flux\_over\_error}}. \quad \text{(B1)}$$

For those *Gaia* sources classified as variable (Holl et al. 2018; Rimoldini et al. 2019), the standard deviation of the epoch photometry is reported as `std_dev_fov_g` from which the semi-amplitude can be estimated as

$$\Delta G_{\rm std} = \sqrt{2}\,\texttt{std\_dev\_fov\_g}. \quad \text{(B2)}$$

Furthermore, the 95th-5th percentile, `trimmed_range_mag_g_fov`, is reported for these sources from which we can find

$$\Delta G_{\rm range} = \frac{1}{2\cos(\pi/20)}\texttt{trimmed\_range\_mag\_g\_fov}. \quad \text{(B3)}$$

Finally, for those variables in the long-period variable catalogue (Lebzelter et al. 2022) the semi-amplitude, `amplitude`, has been estimated using a Fourier fit to the epoch photometry. We denote this $\Delta G_{\rm Fourier}$. This quantity is only reported if a period has been assigned to the source.

In Fig. B1 we show a comparison of the different amplitude measures for the high-amplitude long-period variable sample with $\Delta G_{\rm Fourier}$ and $80 < P/\,{\rm day} < 1000$. In general, there is a very good agreement between the different measures. We see that for the stars with periods within 20 d of 190 or 380 d (troublesome periods for *Gaia*) $\Delta G_{\rm Fourier}$ is larger than the other measures. This is due to the clustering of measurements around a small range of phases so any amplitude measurement based on the data is underestimated and any model fit is unconstrained over a wide range of phases and thus can be overestimated. The amplitude measure based on the photometric uncertainties is biased low relative to the Fourier amplitude when

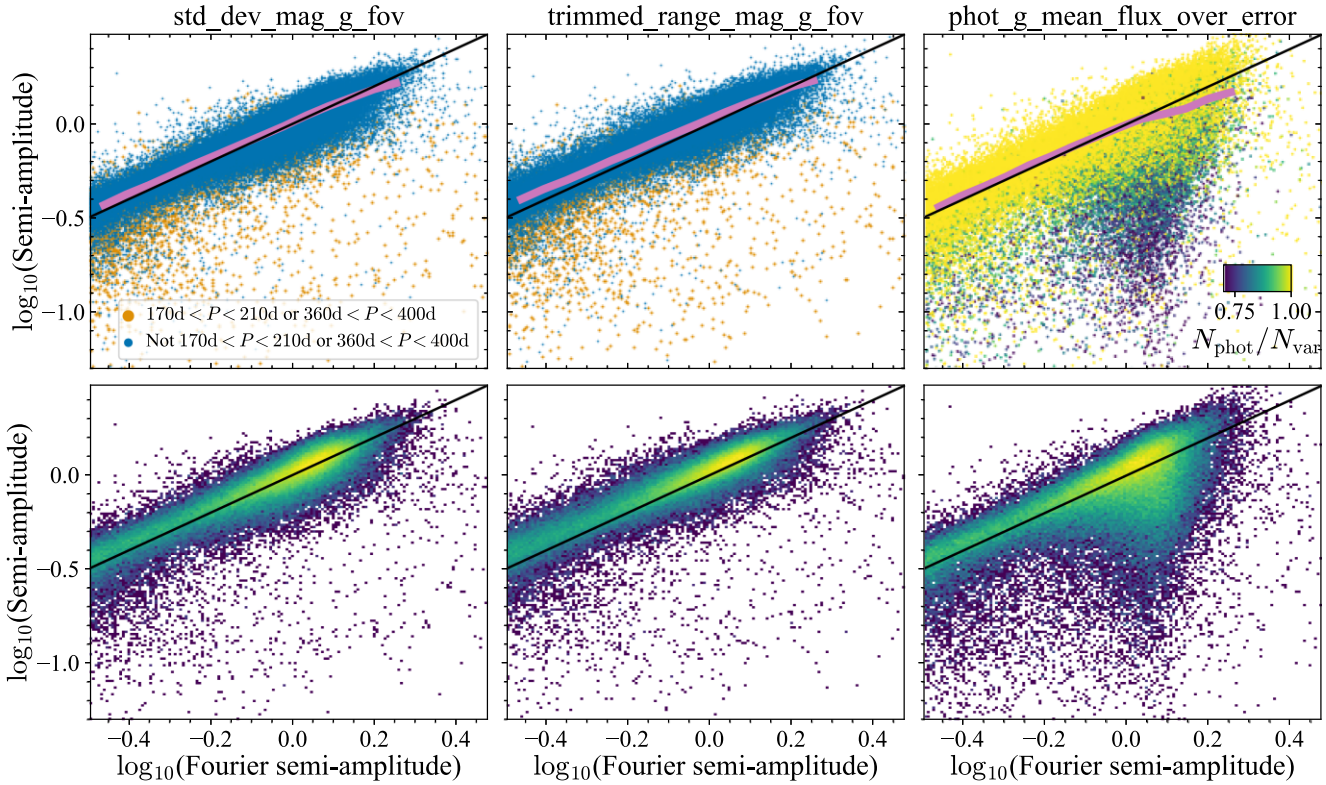

Figure B1. Comparison of amplitude measures for *Gaia* long-period variable stars. Every panel shows the semi-amplitude of a Fourier fit (denoted $\Delta G_{\rm Fourier}$ in the text) against one of the other amplitude measures: $\Delta G_{\rm std}$ from the standard deviation of the epoch photometry in the first column, $\Delta G_{\rm range}$ from the 95th-5th percentile range in the second column and $\Delta G_{\rm phot}$ from the photometric uncertainties in the third column. Both rows show the same data – the lower row is a logarithmically-coloured histogram. The top row is scatter plots coloured by whether the period is near an alias (orange) or not (blue) in the left two rows, and by the ratio of the number of measurements used in the mean photometry compared to the number used in the variable star epoch photometry processing. The solid pink line gives the median trend. The black line is a one-to-one relation. Note sources with periods near 190 and 380 day typically have overestimated Fourier amplitudes and/or underestimated amplitudes using the other methods as they only measure the scatter of the available data. If a low number of photometric points are used in the mean photometric pipeline, the amplitude estimated from the photometric uncertainties is typically smaller than the Fourier amplitude.

there are fewer measurements used in the photometric pipeline (Evans et al. 2018) than used in the variable star processing. This is possibly due to the variability of these stars leading to observations being sigma-clipped from the photometric pipeline.[4] There is also the suggestion that more outliers are removed for sources that fluctuate around the windowing configuration changes. However, in the main, the agreement between the different amplitude measures is very good. Removing sources with periods within 20 day of 190 or 380 d we find the median ratios $\Delta G_{\rm std}/\Delta G_{\rm Fourier} = 1.032$, $\Delta G_{\rm range}/\Delta G_{\rm Fourier} = 1.069$ and $\Delta G_{\rm phot}/\Delta G_{\rm Fourier} = 1.040$ where we have removed sources with the number of mean photometric measurements less than 95 per cent the number of measurements used in the variability pipeline.

[4] https://gea.esac.esa.int/archive/documentation/GDR3

This paper has been typeset from a TEX/LATEX file prepared by the author.