

Carbon-enhanced metal-poor star candidates from BP/RP spectra in *Gaia* DR3

Madeline Lucey¹,¹★ Nariman Al Kharusi,¹ Keith Hawkins¹, Yuan-Sen Ting,^{2,3} Nesar Ramachandra,^{4,5} Adrian M. Price-Whelan,⁶ Timothy C. Beers,^{7,8} Young Sun Lee⁹ and Jinmi Yoon^{10,8}

¹Department of Astronomy, The University of Texas at Austin, 2515 Speedway Boulevard, Austin, TX 78712, USA

²Research School of Astronomy & Astrophysics, Australian National University, Cotter Road, Weston Creek, ACT 2611, Canberra, Australia

³Research School of Computer Science, Australian National University, Acton ACT 2601, Australia

⁴Computational Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA

⁵High Energy Physics Division, Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA

⁶Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA

⁷Department of Physics and Astronomy, University of Notre Dame, Notre Dame, IN 46556, USA

⁸JINA Center for the Evolution of the Elements (JINA-CEE), East Lansing, MI 48824 USA

⁹Department of Astronomy and Space Science, Chungnam National University, Daejeon 34134, Republic of Korea

¹⁰Space Telescope Science Institute, 3700 San Martin Dr, Baltimore, MD 21218, USA

Accepted 2023 May 31. Received 2023 May 30; in original form 2022 June 16

ABSTRACT

Carbon-enhanced metal-poor (CEMP) stars comprise almost a third of stars with $[\text{Fe}/\text{H}] < -2$, although their origins are still poorly understood. It is highly likely that one sub-class (CEMP-*s* stars) is tied to mass-transfer events in binary stars, while another sub-class (CEMP-no stars) are enriched by the nucleosynthetic yields of the first generations of stars. Previous studies of CEMP stars have primarily concentrated on the Galactic halo, but more recently they have also been detected in the thick disc and bulge components of the Milky Way. *Gaia* DR3 has provided an unprecedented sample of over 200 million low-resolution ($R \approx 50$) spectra from the BP and RP photometers. Training on the CEMP catalogue from the SDSS/SEGUE database, we use XGBoost to identify the largest all-sky sample of CEMP candidate stars to date. In total, we find 58 872 CEMP star candidates, with an estimated contamination rate of 12 per cent. When comparing to literature high-resolution catalogues, we positively identify 60–68 per cent of the CEMP stars in the data, validating our results and indicating a high completeness rate. Our final catalogue of CEMP candidates spans from the inner to outer Milky Way, with distances as close as $r \sim 0.8$ kpc from the Galactic centre, and as far as $r > 30$ kpc. Future higher resolution spectroscopic follow-up of these candidates will provide validations of their classification and enable investigations of the frequency of CEMP-*s* and CEMP-no stars throughout the Galaxy, to further constrain the nature of their progenitors.

Key words: stars: abundances – stars: carbon – stars: Population II – Galaxy: abundances.

1 INTRODUCTION

Stellar chemical abundances act as a fossil record of the interstellar medium (ISM) from the time a star is formed, given that a star's atmospheric abundances are not expected to change over its lifetime, aside from mass-transfer events and evolutionary changes on the giant branch. Therefore, it is generally true that more metal-poor stars formed at earlier times when the Universe contained fewer metals. Furthermore, the detailed chemical composition of metal-poor stars can illuminate the early chemical evolution of the Universe that resulted from the lives and deaths of the first generations of stars.

Many studies of metal-poor stars have focused on the Galactic halo, where the metallicity distribution function is dominated by metal-poor stars (e.g. Beers & Christlieb 2005; Frebel & Norris 2015). These studies have found a significant fraction of metal-poor

stars in the halo that exhibit high levels of carbon enhancement ($[\text{C}/\text{Fe}] > +0.7$), and are referred to as carbon-enhanced metal-poor (CEMP) stars (e.g. Beers, Preston & Shectman 1992; Beers & Christlieb 2005; Christlieb et al. 2008). This fraction increases with decreasing metallicity in the Galactic halo, with CEMP stars making up 10–30 per cent of stars with $[\text{Fe}/\text{H}] < -2$ and ≈ 80 per cent of stars with $[\text{Fe}/\text{H}] < -4$ (Lucatello et al. 2006; Lee et al. 2013; Placco et al. 2014; Yoon et al. 2018). However, as cautioned by Arentsen et al. (2022), it can be difficult to compare these fractions across different samples of CEMP stars, given the various selection effects and differences in the abundance analysis from study to study.

Further analysis of CEMP stars have identified a number of sub-classes (Beers & Christlieb 2005). A significant fraction of CEMP stars exhibit enhancements of slow neutron-capture (*s*-process) elements (such as Ba), and are thus called CEMP-*s* stars. There also exist small numbers of CEMP-*r* stars, which show enhancements in rapid neutron-capture (*r*-process) elements (such as Eu), CEMP-*r/s*

* E-mail: m_lucey@utexas.edu

stars, which exhibit enhancements in both *r*- and *s*-process elements (Gull et al. 2018), and CEMP-*i* stars, which exhibit enhancements of intermediate neutron-capture (*i*-process) elements (Frebel 2018). The CEMP-no sub-class of stars does not exhibit over-abundances of neutron-capture elements. The CEMP-*r*, CEMP-*r/s*, and CEMP-*i* sub-classes are sparsely populated in extant samples, while the CEMP-*s* and CEMP-no stars are the most common (see e.g. Zepeda et al. 2023).

It is thought that CEMP-*s* stars, which are more common at $[\text{Fe}/\text{H}] > -3.0$, are the result of chemical enrichment by mass-transfer events from (post-)asymptotic giant branch (AGB) stars (Lugaro et al. 2012; Placco et al. 2013). This hypothesis is strongly supported by the high rate of binarity among CEMP-*s* stars (McClure & Woodsworth 1990; Preston & Sneden 2001; Lucatello et al. 2005; Bisterzo et al. 2010; Abate et al. 2015; Jorissen et al. 2016; Hansen et al. 2016b). In fact, binarity rates as high as 82 per cent have been reported for CEMP-*s* stars (Hansen et al. 2016b).

On the other hand, CEMP-no stars have a lower rate of binarity than CEMP-*s* stars, thus are less likely to have experienced a mass-transfer event (Starkenburg et al. 2014; Hansen et al. 2016a; Arentsen et al. 2019). Hence, CEMP-no stars likely formed from an ISM that was already carbon enhanced. Given their low metallicity and increasing frequency at lower metallicities, it is thought that these are truly ancient, and were primarily enriched by the first generations of stars. It has been suggested that massive first stars may have had high rotation rates, which would lead to large carbon production (Chiappini et al. 2006; Meynet, Ekström & Maeder 2006). Furthermore, it is possible that the first stars exploded as faint supernovae, which also overproduce carbon (Umeda & Nomoto 2003; Nomoto, Kobayashi & Tominaga 2013; Tominaga, Iwamoto & Nomoto 2014). Yoon et al. (2016) have associated CEMP-no stars with their Morphological Groups III and II, respectively, corresponding to these two primary carbon-production sources.

Initial studies indicate that the frequency of CEMP-*s* and CEMP-no stars varies throughout the Galaxy. Specifically, the number of CEMP stars appears to increase with increasing distance from the Sun, although we note that these studies are mostly focused on the Galactic halo (Frebel et al. 2006; Carollo et al. 2012; Lee et al. 2017; Yoon et al. 2018). Furthermore, the relative fraction of CEMP-no stars compared to CEMP-*s* stars also increases at larger distances (Carollo et al. 2014; Lee, Beers & Kim 2019). Ultrafaint dwarf galaxies have shown similar fractions of CEMP-no stars as the Milky Way, but dwarf spheroidal galaxies have a clear deficit of CEMP-no stars (Norris et al. 2010; Lai et al. 2011; Frebel, Simon & Kirby 2014; Salvadori, Skúladóttir & Tolstoy 2015; Skúladóttir et al. 2015). In a comparative study between Galactic halo and dwarf galaxy CEMP stars, Yoon et al. (2019) suggest that the majority of Galactic halo CEMP-no stars have been accreted from dwarf galaxies. Furthermore, CEMP-no and CEMP-*s* stars have been discovered in the metal-weak thick disc (MWTD; Beers et al. 2017). Dietz et al. (2021) tentatively associated the retrograde MWTD CEMP-no population with the Gaia-Enceladus system, while suggesting that the equivalent prograde population has both *in situ* and *ex situ* origins.

There are fewer studies of metal-poor stars towards the center of the Galaxy compared to the Galactic halo. This is partly due to the difficulty of targeting metal-poor stars in a region of the Galaxy that is dense with metal-rich stars. Furthermore, high levels of extinction demand long exposure times and large-aperture telescopes in order to achieve sufficient signal-to-noise spectroscopic observations for metallicity measurements. Fortunately, the advent of metallicity-sensitive photometric surveys (e.g. SkyMapper and Pristine; Bessell

et al. 2011; Starkenburg et al. 2017; Wolf et al. 2018) have led to studies of thousands of metal-poor inner Galaxy stars. Studies using SkyMapper photometry have found a much lower fraction of CEMP stars in the inner Galaxy compared to the Galactic halo (Howes et al. 2014, 2015, 2016; Lucey et al. 2022). However, metallicity estimates from SkyMapper photometry have proven to be biased against CEMP stars (Da Costa et al. 2019; Chiti et al. 2020). Targeting with Pristine photometry, Arentsen et al. (2021) found a CEMP frequency that is consistent with the Galactic halo for stars with $[\text{Fe}/\text{H}] < -3$, but also found that it is much lower than the halo at higher metallicities.

Measuring and understanding the frequency and relative rates of CEMP-no/CEMP-*s* stars throughout the Galaxy will be crucial for shedding new light on the origins and formation mechanisms of these stars, including whether or not CEMP-no stars are true inheritors of the elements created by the first stars. Given that the measured properties of CEMP samples have been shown to vary across different samples (Arentsen et al. 2022), creating a uniformly analysed sample with limited selection effects across the Milky Way will be essential for achieving this goal. The release of the *Gaia* BP/RP spectra in DR3 presents a unique opportunity to identify the largest all-sky sample of CEMP stars to date (Witten et al. 2022). However, the BP/RP spectra have quite low resolution ($R = \lambda/\Delta\lambda \approx 50$), and require unconventional methods for analysis.

In this work, we present a novel method for detecting CEMP stars in the *Gaia* BP/RP spectra with machine learning, specifically the XGBoost classification algorithm, and apply it to the spectra released in DR3. In Section 2, we describe the BP/RP spectra, along with other data used in our analysis. We introduce XGBoost, our chosen classification algorithm, in Section 3. We evaluate the accuracy and sensitivity (i.e. completeness) of our classification in Section 4. In Section 5, we interpret the XGBoost model. Finally, in Section 6 we present the sample of CEMP candidate stars, along with an investigation of their metallicity and Galactic distributions.

2 DATA

The *Gaia* mission has revolutionized Milky Way astronomy and beyond, primarily by providing astrometric data for billions of stars (Gaia Collaboration 2016, 2022). Simultaneously, the *Gaia* mission has also been collecting low-resolution spectra ($R \approx 50$), with the blue photometer (BP) and red photometer (RP; De Angeli et al. 2022). These spectra have provided effective temperature (T_{eff}), surface gravity ($\log g$), and metallicity ($[\text{M}/\text{H}]$) estimates (Liu et al. 2012; Andrae et al. 2022), but have too low of a resolution to provide further elemental abundances (Gavel et al. 2021). Molecules, however, absorb large bands of light, and therefore may be easier to detect in the BP/RP spectra. Given the wavelength coverage of 3300–10 500 Å (Carrasco et al. 2021), we expect to be able to detect carbon-enhanced stars from the plethora of carbon molecular bands in that range (e.g. C₂ Swan bands at ≈ 4500 –6000 Å, and CN bands at ≈ 7000 –10 500 Å). The ability to measure carbon abundances from mock BP/RP spectra has been explored by Witten et al. (2022), but they make use of a different method than this work and also do not model the impact of dust extinction.

To explore the impact of carbon on the BP/RP spectra, we create mock synthetic spectra. We employ the MARCS carbon-enhanced model atmosphere grids (Gustafsson et al. 2008) and the TURBOSPECTRUM radiative transfer code (Alvarez & Plez 1998; Plez 2012) to construct these spectra. We also use the fifth version of the Gaia-ESO atomic linelist (Heiter et al. 2021) with the addition of molecular lines for CH (Masseron et al. 2014), CN, NH, OH, MgH,

C₂ (Masseron, private communication), SiH (Kurucz linelists¹), and TiO, ZrO, FeH, CaH (Plez, private communication). To apply the instrumental profile of the BP/RP photometers, we use the DR3 calibrated passbands (Riello et al. 2021). The true spectral resolution is a function of wavelength for both BP and RP spectra, with the BP spectra ranging from $R \approx 100$ at the blue edge to $R \approx 30$ at the red edge, and the RP spectra ranging from $R \approx 100$ at the blue edge to $R \approx 70$ at the red edge. To simplify the calculation, we assume a resolution of 50 for the BP spectra and 70 for the RP spectra. We also model the impact of dust extinction using the DUST_EXTINCTION² package with the extinction curve from Fitzpatrick (2004), assuming $R_V = 3.1$.

Fig. 1 shows examples for a number of noiseless synthetic mock spectra, along with the impact of carbon on the calculated flux. The left-hand panels shows eight spectra, in sets of two with the same stellar parameters, except one spectrum has $[C/Fe] = +0.5$ (blue dotted line) and the other has $[C/Fe] = +1.0$ (green solid line). We start with the stellar parameters of a typical metal-poor giant star ($T_{\text{eff}} = 4500$ K, $\log g = 2.5$ and $[Fe/H] = -2.0$) in the top row. We then increase the T_{eff} to 5500 K in the second row and $T_{\text{eff}} = 6500$ K in the third row. The spectra in the bottom row have $T_{\text{eff}} = 4500$ K but $A_V = 9.0$. This is a worst-case example, since we expect very few spectra released in *Gaia* DR3 to have $A_V \geq 9.0$, given that most stars have $G < 17.6$. In the right-hand panels, we subtract the spectrum with $[C/Fe] = +1.0$ from the spectrum with the same stellar parameters, but with $[C/Fe] = +0.5$.

The impact of carbon enhancement on the BP/RP spectra (Fig. 1) is very dependent on the T_{eff} and reddening. Compared to the standard metal-poor giant ($T_{\text{eff}} = 4500$ K), the hottest star ($T_{\text{eff}} = 6500$ K) has a significantly weaker signal. This is likely a consequence of the dissociation of carbon molecules in the atmospheres of hotter stars ($T_{\text{eff}} > 6000$ K). This is consistent with results from Witten et al. (2022), which found that the T_{eff} must be < 6000 K in order to achieve precision on the carbon abundance of $< +0.5$ for stars with $[Fe/H] = -2.0$ with *Gaia* BP/RP spectra. Furthermore, reddening greatly reduces the flux, and therefore the strength of the carbon-enhancement signal in the bluest wavelengths. Fortunately, there are carbon molecules (e.g. CN bands at ≈ 7000 – $10\,500$ Å) that impact the RP spectra, so we are still able to detect reddened carbon-enhanced stars. These four combinations of T_{eff} and reddening values lead to significantly different results for the impact of carbon enhancement on the BP/RP spectra. As we hope to be able to detect carbon-enhanced stars across a wide range of stellar parameters and reddening, we require a complex model that can adapt to the different signals of carbon enhancement. To balance complexity with interpretability, we select XGBOOST as our algorithm for detecting carbon-enhanced stars.

2.1 Training and testing sample

In order to teach our model how to accurately detect carbon-enhanced stars, we require a sample of stars that are already classified. To acquire this, we employ the spectroscopic catalogues of parameters for stars from the SDSS survey, and its various extensions, including the Sloan Extension for Galactic Understanding and Exploration (SEGUE; Yanny et al. 2009; Lee et al. 2013; Rockosi et al. 2022), which obtained over 500 000 low-resolution ($R = 2000$) optical spectra. For simplicity, we refer to this collection of spectra as the SDSS/SEGUE sample.

The SEGUE Stellar Parameter Pipeline (SSPP; Allende Prieto et al. 2008; Lee et al. 2008a, b, 2011; Smolinski et al. 2011; Lee et al. 2013) has continued to be refined since its introduction. In the process, additional calibration stars with available high-resolution spectroscopic analyses have been used to improve estimates of the stellar parameters from the SDSS/SEGUE spectra. The most recent version of the SSPP has been run through the stellar samples we employ. Note that, at this stage, the spectra include examples of objects originally targeted as QSO or candidate galaxy candidates, but that turned out to have spectra that were stellar in appearance. This is important, since numerous late-type CEMP stars turned out to be originally targeted as QSOs, based on their photometry (strong carbon absorption features can lead to colors that mimic quasars).

After removal of duplicates (retaining the parameter estimates for the highest signal-to-noise (S/N) spectrum among repeated objects), the full set of spectra were then inspected visually (by Beers), for the identification of defective spectra that could perturb the stellar parameter estimates, the identification and rejection of white dwarfs, some of which were missed by the flags raised by the SSPP, as well as likely spectroscopic binaries, often comprising a white dwarf and a late-type star. For the recognized spectroscopic binaries that do not include a white dwarf, the estimated stellar parameters are not necessarily compromised, as the SSPP parameters primarily consider features in the bluer portion of the spectrum. Nevertheless, we conservatively dropped them from inclusion. Care is taken in order to evaluate the best available estimate of $[Fe/H]$, based on consideration of the various techniques available in the SSPP.

Note that, for the purpose of the present application, we use the measured carbon abundances, without corrections for evolutionary effects (e.g. the first and second dredge ups, see Placco et al. 2014). This means that there are stars that appear to have $[C/Fe]$ estimates below our adopted CEMP cutoff, but would appear above this cutoff once corrections are applied. This is appropriate since, at this stage, we are primarily interested in identifying candidate CEMP stars in the *Gaia* DR3 sample, and this conservative choice ensures that we do not miss-classify stars due to uncertainties in the corrections.

After the culling procedure described above, we are left with 569 874 stars, of which 29 399 have $[C/Fe] > +0.7$. We find that 233 604 of the original 569 874 stars have BP/RP spectra released in *Gaia* DR3, of which 9094 (≈ 4 per cent) have $[C/Fe] > +0.7$. In preliminary tests, we found that our algorithm is incapable of detecting carbon enhancement in warmer stars, due to the weakness of the molecular carbon bands, so we apply a cut in the $G_{BP} - G_{RP}$ colour. Specifically, we only include stars with $G_{BP} - G_{RP} > 0.8$, which roughly corresponds to $T_{\text{eff}} \approx 6000$ K (Andrae et al. 2018). This cut is consistent with results from Witten et al. (2022), who found they could achieve a carbon-abundance precision of ≈ 0.5 dex for stars with $[Fe/H] < -2$ only if they restrict their analysis to stars with $T_{\text{eff}} < 6000$ K. This final trim leaves samples of 1514 carbon-enhanced stars and 141 108 carbon-normal stars.

We randomly select ≈ 30 per cent of these data as our testing sample, while the remaining ≈ 70 per cent is used for training. Fig. 2 shows colour–magnitude diagrams of our training sample. The left-hand panel shows the *Gaia* DR3 $G_{BP} - G_{RP}$ colour on the x -axis and the absolute G magnitude calculated using the *Gaia* parallax and apparent G magnitude on the y -axis for the entire training/testing sample. The right-hand panel shows the part of the sample that has $[C/Fe] > +0.7$. To avoid extrapolation, we classify only stars that fall within this colour–magnitude distribution (dotted black lines). However, note that our training sample is not uniformly distributed in this space, which may introduce a bias in our classification. We investigate this by evaluating the false positive and true positives rates

¹<http://kurucz.harvard.edu>

²https://github.com/karlark/dust_extinction

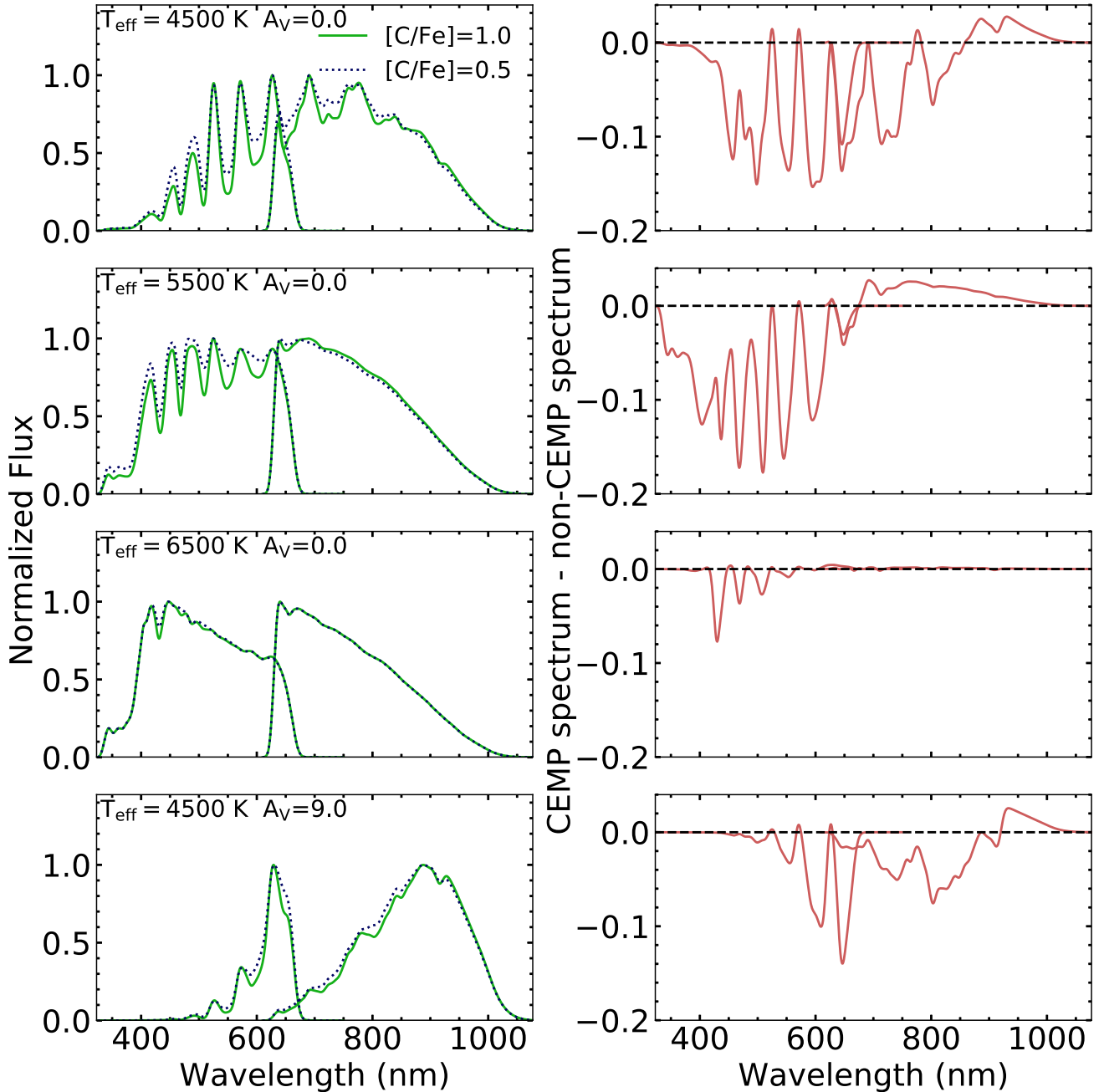


Figure 1. Examples of noiseless synthetic mock BP and RP spectra (left), and the impact of carbon enhancement on the spectra (right). Specifically, in the left-hand panels, we compare stars of the same stellar parameters and reddening, except with different carbon abundances. The BP and RP spectra are plotted separately, with the BP spectra at lower wavelengths. The BP and RP spectra overlap at ≈ 650 nm. The blue dotted lines have $[C/Fe] = +0.5$, while the green solid lines have $[C/Fe] = +1.0$. Starting from the top, which has typical stellar parameters for a metal-poor giant ($T_{\text{eff}} = 4500$ K, $\log g = 2.5$, and $[Fe/H] = -2.0$), we increase the T_{eff} to 5500 K in the second row and 6500 K in the third row. The fourth row has $T_{\text{eff}} = 4500$ K, but with increased extinction at $A_V = 9.0$ mag. In the right-hand panels, we have subtracted the dotted spectrum ($[C/Fe] = +0.5$) from the solid spectrum ($[C/Fe] = +1.0$) for each row. The impact of carbon on the spectra changes drastically with the stellar parameters with higher T_{eff} s having weaker signals and extinction erasing the signal in the bluest wavelengths. We therefore require a flexible classification model in order to achieve low contamination of our detected carbon-enhanced stars.

of the classification as a function of color and absolute G magnitude (see Section 4).

Fig. 3 shows the carbonicity as a function of metallicity (left-hand panel), as well as the apparent G magnitude as a function of extinction (right-hand panel) for the training/testing sample. In the left-hand panel, we also show our definition for carbon enhancement as a red dashed line at $[C/Fe] = +0.7$. It is interesting to note the

appearance of two sequences with different slopes in the carbonicity as a function of $[Fe/H]$ plane for stars with $[C/Fe] > +1$. Most of the 1514 carbon-enhanced stars are metal-poor ($[Fe/H] \leq -1$). However, there are 139 carbon-enhanced stars with $[Fe/H] > -1$, corresponding to 9 per cent of the stars. We find that the classification algorithm identifies CEMP stars more accurately when these are included as positive cases in the training. We expect the metallicity

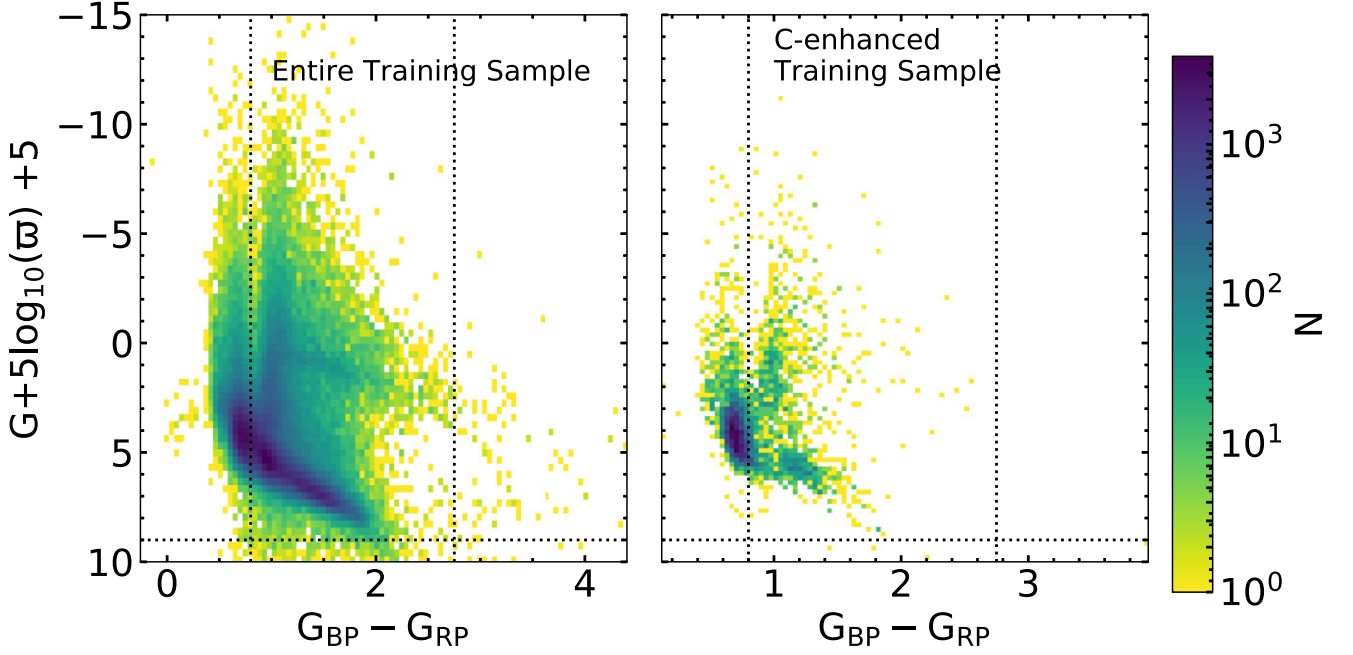


Figure 2. Colour–magnitude diagram of our training sample from SDSS/SEGUE. The x -axis is the *Gaia* $G_{BP} - G_{RP}$ colour; the y -axis is the absolute G magnitude, calculated from the distance modulus using the *Gaia* apparent G magnitude and parallax. The left-hand panel shows the entire training set, while the right-hand panel only shows stars with $[C/Fe] > +0.7$. The black dashed lines correspond to the colour and magnitudes cuts made on our training/testing sample. The logarithmic colour bar corresponds to the number of stars for each data point.

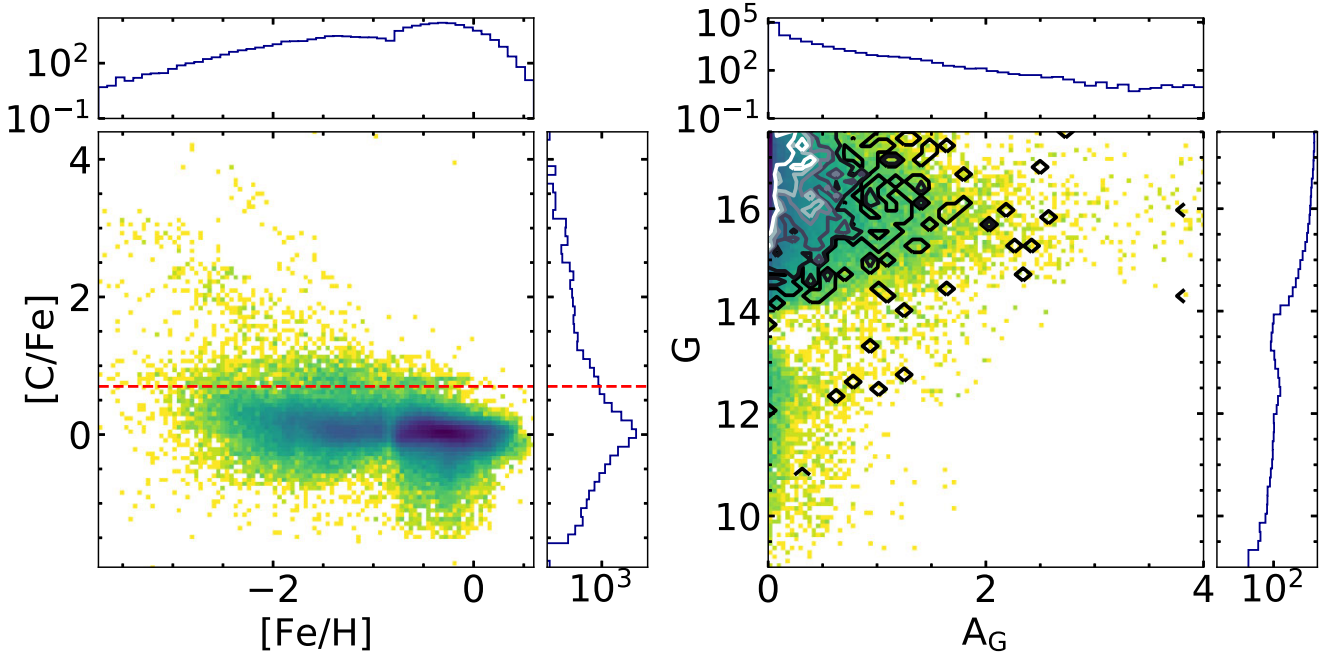


Figure 3. Relevant properties of the training/testing sample from the SDSS/SEGUE. The left-hand panel shows the carbonicity ($[C/Fe]$), as a function of metallicity ($[Fe/H]$), with our definition of carbon-enhanced ($[C/Fe] > +0.7$) marked as a red dashed line. We also show the marginal histogram of each parameter on the corresponding axis. The right-hand panel is a similar plot, but the axes are instead the extinction (A_G ; Andrae et al. 2022) and apparent *Gaia* G magnitude. From inspection, the training/testing samples span a large range of parameters, similar to what we expect for the data we classify. In both plots the colours of the data points correspond to the logarithmic colour bar shown in Fig. 2. In the right-hand panel, we also overlay the distribution of the CEMP stars in the training/testing sample as black/white contour lines.

distribution of carbon-enhanced stars in *Gaia* DR3 to be similar to our training/testing sample. Therefore, we call the stars positively classified by our algorithm as CEMP candidates, as we expect most of them to be metal-poor with only ≈ 9 per cent to be metal-rich.

In the right-hand panel of Fig. 3, the extinction values are from the *Gaia* DR3 GSP-Phot pipeline (Andrae et al. 2022). Our training/testing sample includes data with $A_V \approx 4$. Given that we constrain the data we wish to classify to $0.8 < G_{BP} - G_{RP} < 2.75$, it is unlikely we would include any data with $A_V > 4$. Therefore, our training/testing sample should sufficiently teach our algorithm to distinguish highly extincted stars from carbon-enhanced stars. In addition, we can use the apparent G magnitude distribution to investigate the signal-to-noise ratio (SNR) distribution of our testing/training sample. Given that the majority of the SDSS/SEGUE data is fainter than the apparent G magnitude cut for the BP/RP spectra released in DR3 ($G \leq 17.5$), we find the apparent G magnitude distribution of the SDSS/SEGUE data peaks at $G \approx 17.5$. Therefore, the G magnitude distribution of our training and testing data matches our expectations for the BP/RP spectra that we classify in that it peaks at the faintest magnitudes, but also includes stars as bright as $G \approx 10$.

2.2 *Gaia* BP/RP spectra

The *Gaia* BP/RP spectra are a unique data set, not only due to their wide wavelength coverage, very low-resolution ($R \approx 50$), and an unprecedented number of stars, but also because the spectra have been released as a linear combination of basis functions, specifically as Hermite function coefficients (Carrasco et al. 2021). This was done because of the complexity of the *Gaia* instrument, which has two wide fields of view and 14 detectors. To create the calibrated mean spectra, multiple epochs of observations with different instrumental conditions needed to be combined. In this work, we use the coefficients as the input data for our model rather than convert them to sampled spectra, which results in some information loss (Carrasco et al. 2021).

In total, the BP and RP spectra comprise 55 coefficients each, but also come with a recommended truncation. This is possible because the coefficients have been rotated to an optimized basis so that the bulk of the spectral information is contained in the first few coefficients. A truncation is then recommended based on the magnitude of the coefficients compared to their corresponding uncertainties. For more details see Carrasco et al. (2021). As XGBoost requires the input data to be vectors of the same length, we apply the largest recommended truncation to avoid losing potentially useful information. The largest recommended truncation is 55 for both BP and RP spectra (i.e. all coefficients are relevant). Therefore, we do not truncate the coefficients. Because we do not want to include apparent magnitude information, we normalize the coefficients by the first BP coefficient. Furthermore, since the coefficient values can span many orders of magnitude, we also divide the spectra by the mean normalized spectrum of the training sample. This decreases the orders of magnitudes spanned by the coefficients, which makes it easier for XGBoost to find the carbon-enhancement signal in the data.

The XGBoost algorithm cannot reliably extrapolate. Therefore, we ensure to only classify stars that are similar to our training sample. Specifically, we constrain our sample using the absolute G magnitude and $G_{BP} - G_{RP}$ colour. As shown in the right-hand panel of Fig. 2 by the black dotted lines, we only classify stars that satisfy the following criteria:

$$(i) \ 0.8 < G_{BP} - G_{RP} < 2.75$$

$$(ii) \ G + 5\log_{10}(\varpi) + 5 < 7.0$$

Although our training sample includes stars outside of this range, we chose to restrict to these values where most of the sample resides. Note that we do not place a lower limit on the absolute G magnitude, since stars with very low absolute G magnitudes likely have small, uncertain parallaxes that cause an over-estimation in the brightness, and therefore underestimate the absolute G magnitude. As we do not want to introduce a selection bias by removing stars with uncertain parallaxes, we chose to include them. In Section 4, we investigate how the false positive rate and completeness (i.e. the true positive rate) of our classification behaves at the edges of this region where the training sample is less dense. In total, we find 182 815 672 BP/RP spectra in *Gaia* DR3 that are within our colour and magnitude cuts.

3 XGBoost

In order to detect carbon-enhanced stars across a wide range of stellar parameters and reddenings, we require a flexible model. We chose to use XGBoost, which is powerful but still easy to interpret. Furthermore, XGBoost is optimized for efficiency, allowing fast training and inference. XGBoost is quickly becoming a popular machine learning algorithm in astronomy, with applications in a large variety of sub-fields (e.g. Hayden et al. 2022; Li et al. 2021; Machado Poletti Valle et al. 2021; He, Luo & Chen 2022; Pham & Kaltenegger 2022).

Fig. 4 shows the general architecture of XGBoost used for classification. For a detailed description of the algorithm see Chen & Guestrin (2016). In short, XGBoost sequentially builds decision trees to fit the residuals from the previous tree. XGBoost continues to train trees until it reaches the maximum number of trees set by the user or the residuals stop consistently shrinking. The results from each tree are then summed together, weighted by the learning rate, η . This value is then plugged into the sigmoid function, $\sigma(x) = 1/(1 + e^{-x})$, to calculate the probability that the star is carbon enhanced ($\hat{y}$). We provide these probability values so that the reader can choose their own sample depending on the completeness and contamination rate required for their science. In this work, we choose to classify a star as carbon enhanced if its probability is > 50 per cent.

XGBoost does not allow for the direct inclusion of uncertainties for each input, but it is able to learn how to distinguish noise from signal sufficiently if the noise distribution of the training sample is representative of the data to which the model will be applied. Given that the apparent G magnitude distribution of our training sample is similar to what we expect for the BP/RP spectra we classify (see Fig. 3), we conclude that the noise distribution of the training sample is representative and sufficient to train the XGBoost model.

To train the XGBoost algorithm, a number of hyperparameters need to be set. We can set the maximum number of trees, the learning rate (η), the percentage of the training sample and the percentage of input coefficients to use for each tree, as well as the maximum depth of each tree. We can also set limits on the purity of a sample for a given leaf to prevent overfitting. To explore the parameter space and find the optimal set of hyperparameters, we use RandomSearchCV from scikit-learn (Pedregosa et al. 2011).

4 CONTAMINATION AND COMPLETENESS

In order to estimate the contamination and completeness (i.e. true positive rate) of our sample of newly identified CEMP candidates, we use our testing sample (described in Section 2.1), where we already know the observed carbon abundances. Specifically, we

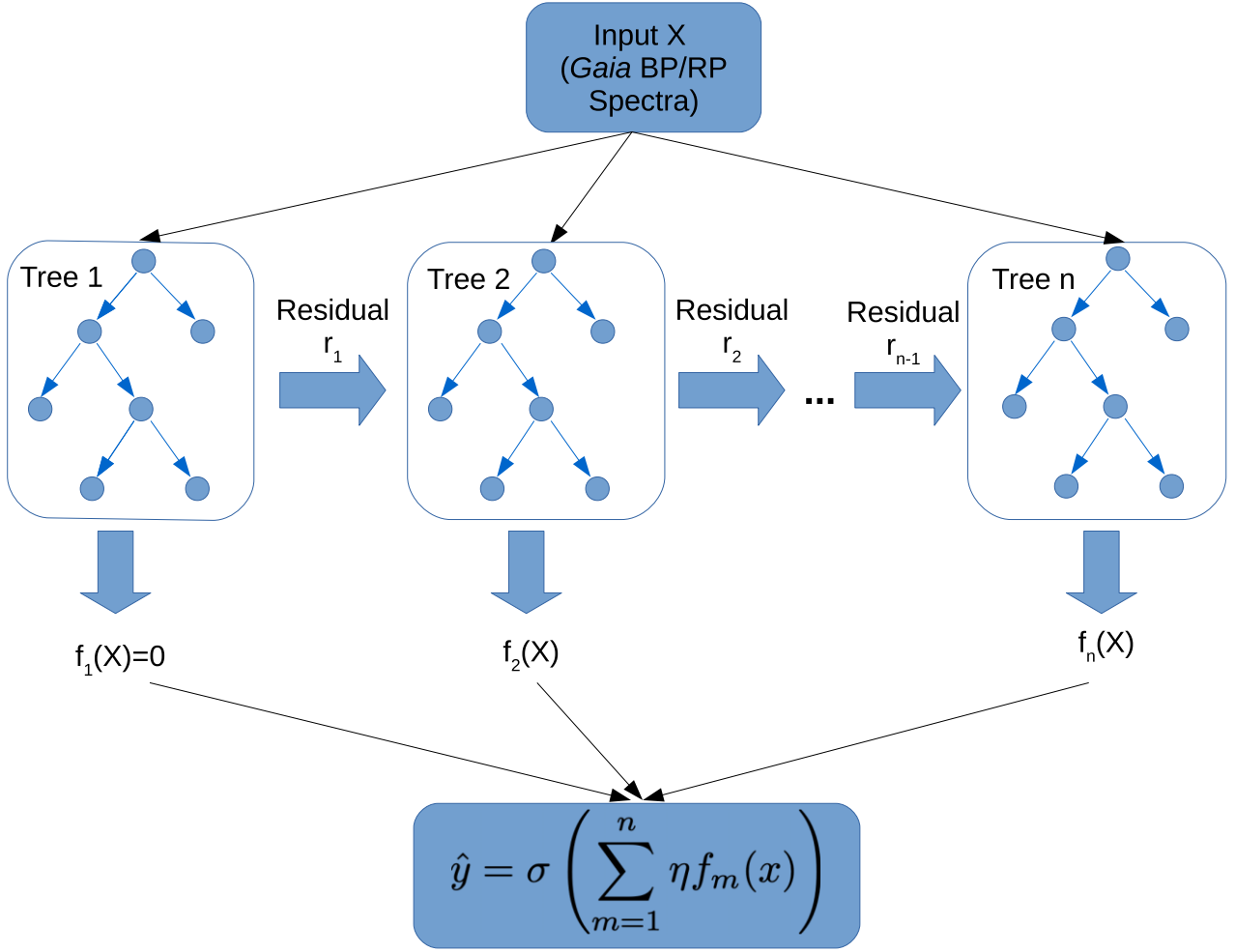


Figure 4. General architecture of XGBoost for classification. In short, XGBoost iteratively creates trees to fit the residuals from the prediction of the previous tree. The first tree provides a prediction of zero for each spectrum. For the final output, the predictions of each tree are summed after being multiplied by the learning rate, η . This value is then input into the sigmoid function, $\sigma(x) = 1/(1 + e^{-x})$, to calculate the final probability, $\hat{y}$. If $\hat{y} > 0.5$, the star is classified as carbon enhanced.

define contamination as the number of false positives divided by the sum of the true positives and false positives. In other words, the contamination estimates the rate of carbon-normal stars in the sample we classify as CEMP. On the other hand, we define completeness as the number of true positives divided by the sum of the true positives and false negatives. Therefore, the completeness estimates the fraction of true CEMP stars that we expect to detect. There may be non-stellar objects that contaminate our sample (e.g. quasars), but we mitigate these effects by constraining the full set of spectra that we classify by the absolute G magnitude and $G_{BP} - G_{RP}$ colour of the training/testing sample (see Section 2.2). Furthermore, stars hotter than our training/testing sample might mistakenly be included if high levels of extinction make them appear sufficiently red. However, we expect that this situation is rare, given that our training sample is representative of the full distribution in T_{eff} and extinction (see Figs 2 and 3). In addition, we have stellar parameters and carbon abundances for the testing sample, which allows us to study how the contamination and completeness behave as a function of various parameters, including observational effects.

Fig. 5 shows the receiver operating characteristic (ROC) curve, which describes the false positive rate (false positives divided by the sum of the true negatives and false positives) and true positive

rate (true positives divided by the sum of the false negatives and true positives) of the classification as a function of the probability cutoff assumed. As expected, the true positive percentage and false positive rate increase as the probability cut is decreased. The black dashed vertical and horizontal lines mark the true positive and false positive rates for a probability cut of > 50 percent, which we use to define the CEMP candidate sample in this work. Specifically, for a probability cut of > 50 percent we find a false positive rate of 0.04 percent and a true positive rate of 26 percent. The false positive rate is especially low because our classification is unbalanced, with only ≈ 1 percent of our sample being a positive case (i.e. carbon enhanced). On the other hand, the contamination rate (false positives divided by the sum of the false positives and the true positives) is ≈ 12 percent. In the final catalog, we provide the probability values for each star classified so that the reader can choose a probability cut best suited for their science case.

Fig. 6 shows the median carbon abundance of the testing sample, as a function of the assigned probability of being carbon enhanced, as a dark blue line, with the 1σ percentiles as the shaded region. We also show the completeness (green line) and contamination rate (red line) for $p(\text{CEMP}) > x$ with the scaling on the right y-axis. The vertical black dashed lines gives the $p(\text{CEMP})$ that above which

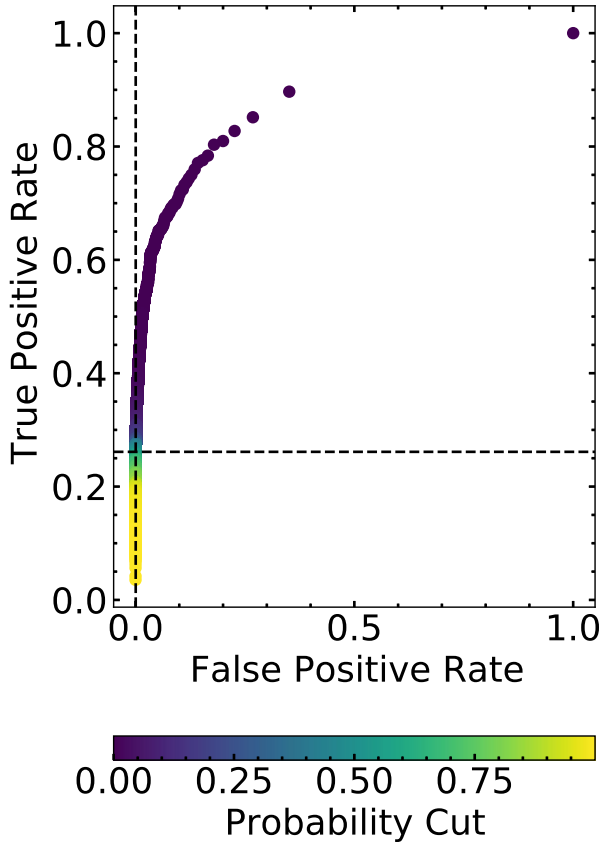


Figure 5. The ROC describing the false positive (false positives divided by the sum of the true negatives and false positives) and true positive rates (true positives divided by the sum of the false negatives and true positives) of our classification for the testing sample. We calculate this curve by assuming different probability cuts for our classification, which are shown by the colour of the points. We mark the point where the probability cut is >50 per cent with black dashed vertical and horizontal lines. Given that only ≈ 1 per cent of our training sample is carbon enhanced, the classification is very unbalanced. Therefore, the false positive rate is very small, even though the contamination rate (false positives divided by the sum of the true positives and false positives) is ≈ 12 per cent. The true positive percentage or completeness is 26 per cent.

a star is classified as CEMP ($p(\text{CEMP}) = 0.5$). Therefore, where this line intersects the completeness (26 per cent) and contamination rates (12 per cent) gives those properties for our final sample. The horizontal dashed line indicates the $[\text{C}/\text{Fe}]$ above which we define a star to be carbon-enhanced ($[\text{C}/\text{Fe}] = +0.7$). The median $[\text{C}/\text{Fe}]$ becomes larger than this at $\approx p(\text{CEMP}) = 0.5$, indicating that our algorithm learns to assign a $p(\text{CEMP}) > 0.5$ for stars with $[\text{C}/\text{Fe}] > +0.7$.

Fig. 7 shows 2D maps of the false positive (left) and true positive rates (right), as a function of absolute G magnitude, and $G_{BP} - G_{RP}$ colour (top), as well as for $[\text{C}/\text{Fe}]$ and $[\text{Fe}/\text{H}]$ (bottom). In these plots, we also show the underlying density distribution of stars as grey hexagonal bins. The left-hand panels show the false positive rate, which is calculated by taking the number of false positives divided by the number of true carbon-normal stars in that bin. The right-hand panels show the true positive rate, which is calculated by dividing the number of true positives (i.e. correctly identified CEMP stars) by the total number of true CEMP stars in that bin.

In the top left panel of Fig. 7, we do not see a trend in the false positive rate with the $G_{BP} - G_{RP}$ colour or absolute G magnitude. In the top right panel, we see that at the true positive rate is lowest

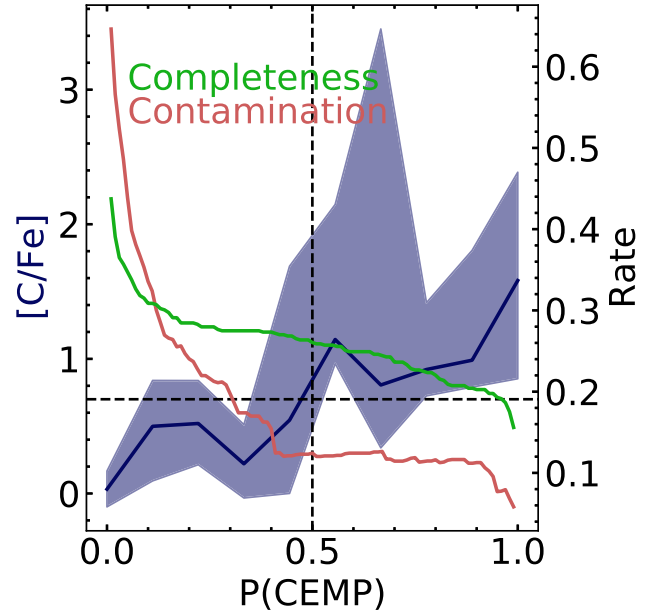


Figure 6. The median carbon abundance of the testing sample, as a function of the assigned probability of carbon enhancement, is shown by the dark blue line, with the 1σ percentiles shown in the blue shaded region. We also show the contamination rate as a red line, and the completeness as a green line, with the scaling shown on the right y-axis. The vertical dashed line corresponds to $p(\text{CEMP}) = 0.5$ and the horizontal dashed line corresponds to $[\text{C}/\text{Fe}] = +0.7$, above which is our definition of a carbon-enhanced star. It is clear that the assigned $p(\text{CEMP})$ is strongly correlated to carbon abundance. Furthermore, we find that the algorithm learns our definition of carbon-enhanced is $[\text{C}/\text{Fe}] > +0.7$, in that the median $[\text{C}/\text{Fe}]$ for $p(\text{CEMP}) \approx 0.5$ is $\approx +0.7$.

for dwarf stars and blue giant stars. We also find that the algorithm does not detect the most extincted, reddest giant stars in our testing sample.

In the bottom row of Fig. 7, we see a slight trend in the false positive and true positive rates with the $[\text{C}/\text{Fe}]$ abundance. Specifically, we see the false positive rate is higher for stars with $+0.5 < [\text{C}/\text{Fe}] < +0.7$, and the true positive rate is lowest for stars with $+0.7 < [\text{C}/\text{Fe}] < +1.0$. This indicates that our classification is most inaccurate for stars with $[\text{C}/\text{Fe}] \approx +0.7$, which is to be expected given that it is unlikely we could measure a $[\text{C}/\text{Fe}]$ abundance from these very low-resolution ($R \approx 50$) spectra that is more precise than ≈ 0.5 dex.

Fig. 8 shows the true positive percentage and contamination rate, as a function of extinction (top panel) and apparent G magnitude (bottom panel). The grey histogram shows the arbitrarily scaled underlying distribution of the testing sample for these parameters. The green line shows the contamination rate, with the scaling provided by the right y-axis, while the dark blue lines show the true positive percentage, with corresponding scaling on the left y-axis.

As expected, we find that the classification performs better at low extinction. Specifically, we see that the true positive percentage decreases from ≈ 30 per cent at $A_G = 0$ to ≈ 0 per cent at $A_G = 1$. We also see that the contamination rate increases from ≈ 10 to ≈ 30 per cent over the same range. Of our final sample of CEMP candidates, 77 per cent has $A_G < 0.5$. Furthermore, we find that the contamination rate also increases with fainter G magnitude, as expected, since the G magnitudes are directly related to the signal-to-noise for these spectra. We find that the false positive percentage increases from ≈ 5 per cent at G magnitude of ≈ 13 to ≈ 12 per cent at $G = 17.5$. The true positive percentage is lowest for bright stars. This is likely because there are few bright CEMP stars in our training

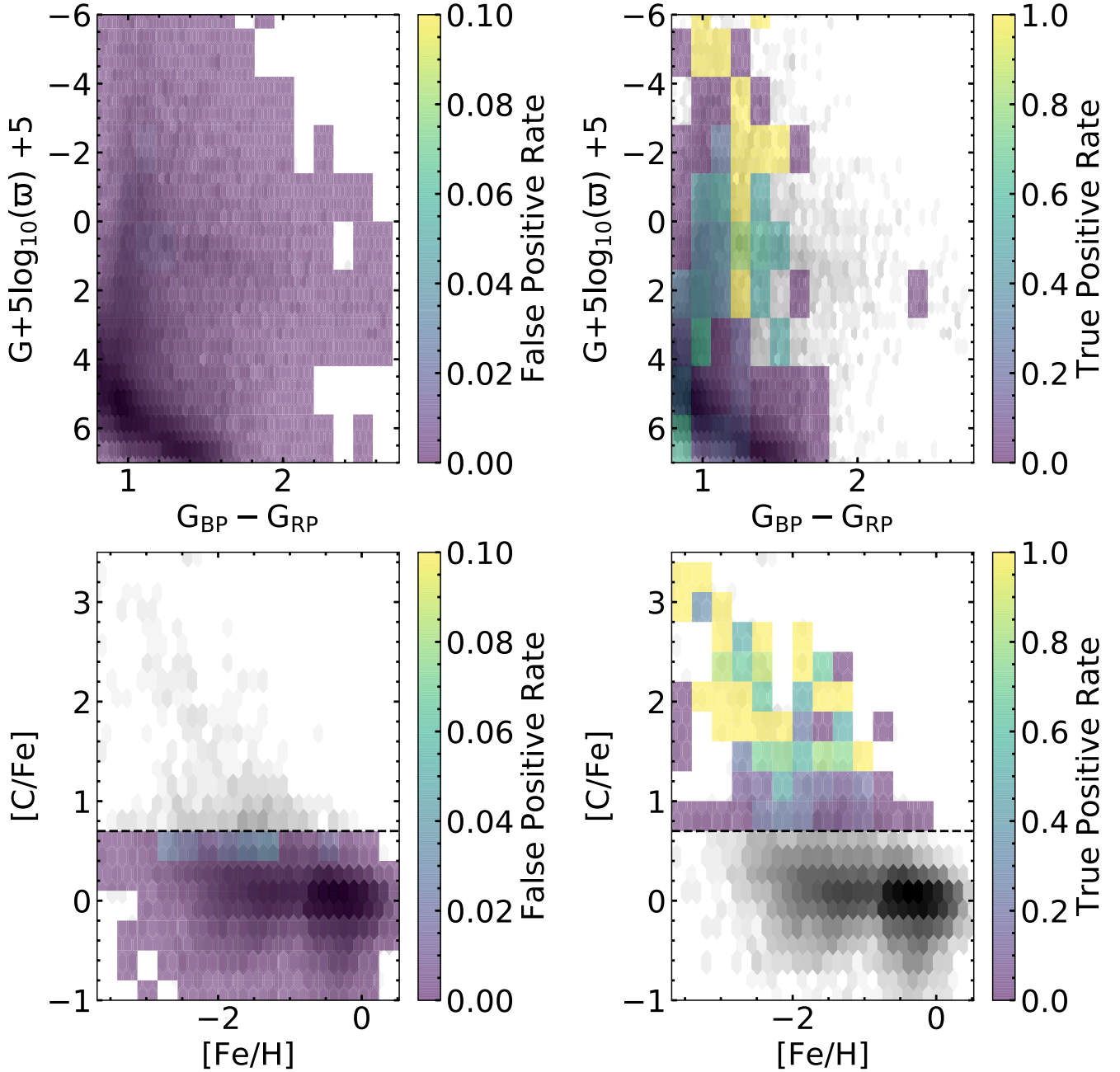


Figure 7. The false positive and true positive rates, as functions of absolute G magnitude, $G_{BP} - G_{RP}$ colour, $[C/Fe]$, and $[Fe/H]$. We show the distribution of stellar parameters for the testing sample in greyscale, with darker areas corresponding to more stars. In general, the model tends to struggle most with red dwarf stars and blue giant stars. Furthermore, the false positives tend to have $+0.5 \leq [C/Fe] \leq +0.7$, while the false negatives tend to have $+0.7 \leq [C/Fe] \leq +1.0$. Therefore, our model likely only can interpret the carbon abundance to ≈ 0.5 dex for some stars.

sample (see Fig. 3), which may introduce a bias against detecting bright CEMP stars in the *Gaia* data. We also see that the true positive percentage decreases at $G > 15$. This is likely due to lower signal-to-noise for these data.

To further validate our sample, we also compare to large spectroscopic surveys. However, as CEMP stars are (relatively) rare, large spectroscopic surveys generally do not correctly account for CEMP stars in their analysis pipelines. For example, APOGEE’s ASPCAP pipeline does not account for stars with $[C/Fe] > +1$ (Abdurro’uf et al. 2022). When cross-matched against our SDSS training/testing sample, we find 1765 stars in common with APOGEE DR17. Of

these, only 26 are identified as CEMP in the SDSS/SEGUE catalogue. However, the ASPCAP pipeline assigns only 1 of these stars an abundance of $[C/Fe] > +0.7$. This star has $[C/Fe] = +2.28$ from SDSS/SEGUE, while ASPCAP measures $[C/Fe] = +0.89$. From this comparison, we expect ASPCAP to report $[C/Fe] < +0.7$ for as much as 96 per cent of our final CEMP catalogue. Of the stars we classify as CEMP, we find that 442 stars have a match in APOGEE DR17 without flags on the ASPCAP derived $[C/Fe]$. Of these, 425 have $[C/Fe] < +0.7$ as measured by ASPCAP, corresponding to 95 per cent. As we found that ASPCAP measured $[C/Fe] < +0.7$ for 96 per cent of our CEMP training sample, it is not unexpected that

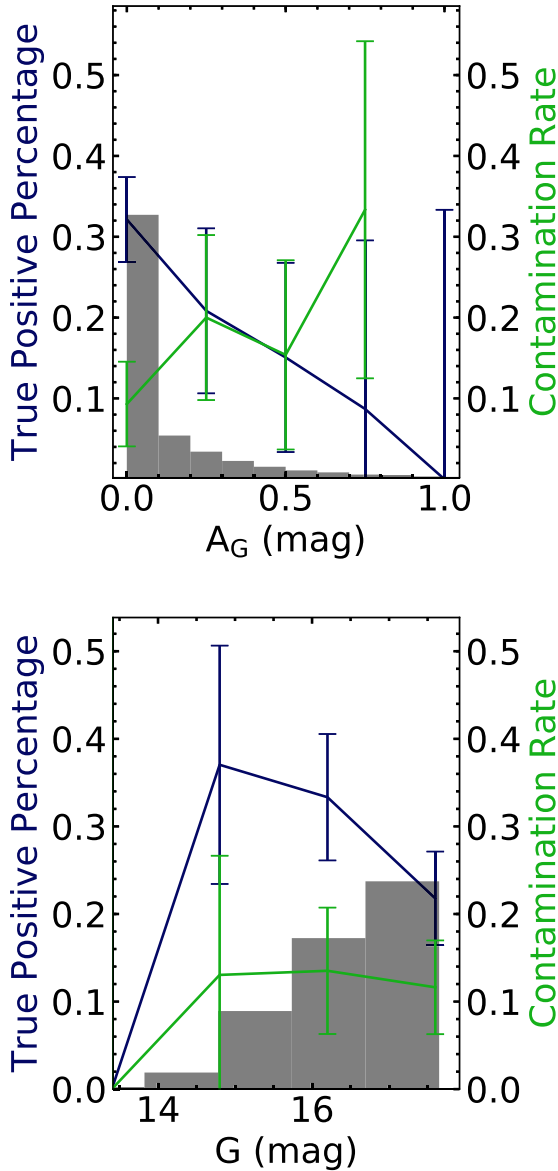


Figure 8. The contamination rate and true positive percentage, as a function of G magnitude and extinction (A_G). The grey bins show the arbitrarily scaled number density of stars in our training/testing sample. For each panel, the scale for the true positive percentage (dark blue) is on the left y-axis; the right y-axis shows the scale for the contamination rate (green). The error bars are $1/\sqrt{N}$, where N is the number of true positives. As expected, we find that our classification improves for bright stars, which likely have high signal-to-noise spectra, and for stars at low extinction.

we find a high percentage of our final sample has $[C/Fe] < +0.7$ from ASPCAP as well.

The metallicities measured by APOGEE of the stars that have $p(\text{CEMP}) > 0.5$ in our catalogue are generally metal poor, with mean $[Fe/H] = -0.95$ (compared to mean $[Fe/H] = -0.25$ for stars with $p(\text{CEMP}) < 0.5$) and minimum $[Fe/H] = -2.41$. However, we note that only 17 of these stars have $[C/Fe] > +0.7$ in APOGEE. Of these 17, APOGEE measures an average $[Fe/H] = -1.39$, with minimum $[Fe/H] = -2.23$.

To validate the completeness of our sample, we cross-match against a database of CEMP stars confirmed from high-resolution spectra (Zepeda et al. 2023). Of the 382 stars from Zepeda et al.

(2023) that have observed $[C/Fe] > +0.7$ (no evolutionary correction), 208 of them have BP/RP spectra within our color and absolute magnitude cuts. We correctly classify 143 of those as CEMP stars, which gives a completeness rate of 69 per cent. This is significantly higher than that estimated with our testing sample (26 per cent). Therefore, it is possible that our true completeness percentage is higher, and 26 per cent is a conservative estimate. In general, we find that the completeness relative to Zepeda et al. (2023) behaves as expected from Figs 7 and 8. Specifically, the completeness decreases with decreasing $[C/Fe]$, increasing T_{eff} , and fainter apparent G magnitudes. We also find that we tend to miss Group II stars in Zepeda et al. (2023) with lower $[Fe/H]$, because they have lower absolute carbon abundance, $A(C)$, and therefore weaker absorption features.

5 MODEL INTERPRETATION

As we have used a data-driven method to classify stars in this work, we want to ensure that the final model matches our physical intuition. It is also important to ensure that the model is not using any confounding variables (e.g. extinction or metallicity) that could be correlated with carbon enhancement but are not a direct measurement of the true carbon abundance. Typically, model interpretation is done by determining the importance of each input feature for the final inference and comparing it to expectations from physical models. Here, we investigate the importance of four key Hermite coefficients and how they relate to true carbon abundance. Furthermore, we compare BP/RP spectra of CEMP stars from the training sample to newly classified stars in order to ensure they match expectations.

Fig. 9 shows the average difference between CEMP spectra and carbon-normal stars for a subset of the training sample (green) and the newly classified data (black). We also show the difference between synthetic spectra (dark blue dashed line) with $[C/Fe] = +1$ and $[C/Fe] = +0.5$, assuming $T_{\text{eff}} = 4500$ K, $\log g = 2.5$, and $[Fe/H] = -2$. This is the same line as shown in the top right panel of Fig. 1. For the observed spectra, we chose to calculate the averages over a narrow colour and absolute G magnitude range in order to isolate the effect of carbon from the effects of T_{eff} , $\log g$, and extinction on the spectra. Specifically, we use stars with $1.20 < G_{BP} - G_{RP} < 1.25$ and $0 < G + 5\log_{10}(\varpi) + 5 < 3$, which is a narrow dense region of the red giant branch (RGB). For our training sample, this corresponds to a T_{eff} range of $4500 \text{ K} < T_{\text{eff}} < 5500 \text{ K}$. We calculate the average CEMP spectrum and carbon-normal spectrum in this region for both the training sample and the newly classified sample, in order to ensure that they look similar, which indicates that the classification algorithm successfully identified candidate CEMP stars. We also compare this to expectations from synthetic spectra. We find that the differences between CEMP and carbon-normal stars for the newly classified data matches the differences found in the training sample. Furthermore, the features seen in the subtracted spectra roughly match with expectations from mock spectra. The major discrepancies are likely due to imperfect modeling and different normalization procedures. Specifically, the observed spectra are normalized in the *Gaia* BP/RP coefficient space following the procedure described in Section 2.2, while the synthetic spectrum is normalized by the maximum flux. Therefore, the scaling of the features between the observed and model spectra should not be compared. In general, the model and observed spectra agree in the location of local minima and maxima, indicating agreement in the location of key absorption features. However, blue-ward of ≈ 480 nm, there is poor agreement between the model and observed spectra. This is likely due to a shortcoming in the model-spectra resolution. The resolution of the

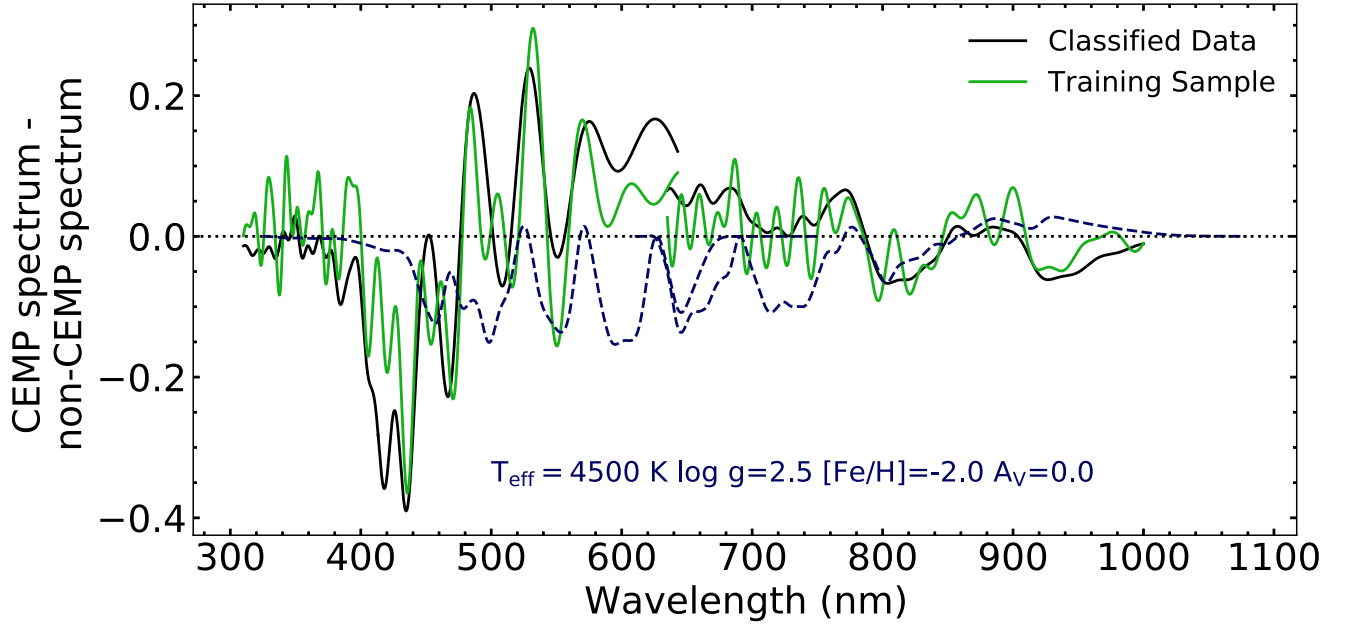


Figure 9. The average difference between CEMP spectra and carbon-normal spectra for a subset of the training sample (green) and newly classified data (black). We also include the difference between synthetic spectra (dark blue dashed line) with $[C/Fe] = +1$ and $[C/Fe] = +0.5$, assuming $T_{\text{eff}} = 4500$ K, $\log g = 2.5$, and $[Fe/H] = -2.0$, as described in Fig. 1. The green and black lines are calculated by subtracting the average spectrum of carbon-normal stars from CEMP stars for stars in a narrow range of color ($1.20 < G_{BP} - G_{RP} < 1.25$) and absolute G magnitudes ($0 < G + 5\log_{10}(\varpi) + 5 < 3$) which corresponds to a small dense region of the red giant branch. We do this both for the training sample and the newly classified data in order to ensure that the classification has worked, and that the newly classified CEMP stars have the expected carbon features. Given that the difference between the CEMP stars and the carbon-normal stars is quite similar for both the training and newly classified samples, and that the features generally match expectations from synthetic spectra, we conclude that our classification has correctly selected candidate CEMP stars.

BP spectra increases in the bluer regions to $R \approx 100$, but the mock model spectra assumes $R = 50$ for the entire wavelength region.

We measure the impact of each Hermite basis coefficient on the classification using SHAP (Shapley Additive exPlanations) values (Lundberg & Lee 2017). The SHAP values allow us to explore the importance of individual features, as a function of various parameters, since each star’s coefficients are assigned individual values. The SHAP values are defined so that their summation plus the average predicted value (ϕ_0) is equal to the predicted value ($\hat{y}$). Explicitly,

$$\hat{y} = \phi_0 + \sum_{i=1}^M \phi_i, \quad (1)$$

where ϕ_0 is the average value of $\hat{y}$ for all of the spectra, ϕ_i is the SHAP value for coefficient i , and M is the total number of coefficients per spectrum. Therefore, each SHAP value directly measures the impact of the selected coefficient on the inference of $\hat{y}$ for a given star. We refer the interested reader to Lundberg & Lee (2017) for further details on the calculation of SHAP values.

We calculate the SHAP values for all of the spectra in our testing sample. Fig. 10 shows how the SHAP values relate to the carbon abundance and coefficient value for four BP coefficients. These coefficients were selected because they have the largest SHAP values of all the BP and RP coefficients. Specifically, we have the $[C/Fe]$ on the y-axis, and the coefficient value on the x-axis. The points are colored by the SHAP values. The bottom panel shows $[C/Fe]$ as a function of the summation of the four SHAP values. As large positive SHAP values indicate that the given coefficient increased the probability that the star is carbon enhanced, it is expected that stars with $[C/Fe] > +0.7$ should have higher SHAP values than

carbon-normal stars. We find that the XGBoost model is able to pick up this sensitivity, in that the SHAP values for these coefficients are generally large and positive for carbon-enhanced stars. Although there are some coefficients where the SHAP value is negative for a carbon-enhanced star, these stars may still be classified as carbon enhanced based on the value of other coefficients. Similarly, there are many carbon-normal stars that have individual positive SHAP values, but the combined effect of all of the other coefficients effectively decreases the probability of the star being carbon enhanced (see the bottom panel), so that the false positive rate is not exceedingly high. Given the positive correlation between $[C/Fe]$ and the SHAP value, it is likely that XGBoost model is using the carbon information in the spectra to determine whether a given star is carbon enhanced rather than using another confounding variable.

6 PROPERTIES OF THE CEMP CANDIDATE SAMPLE

Out of the ≈ 180 million stars that we classify, we find 58 872 CEMP candidate stars. This is the largest, homogeneously identified sample of CEMP candidate stars to date. In this section, we briefly investigate a few properties of this sample, including their metallicity distribution and Galactic distribution.

Fig. 11 shows the colour–magnitude diagram of our CEMP candidate sample compared to a random sample of stars classified as carbon-normal stars of the same size. We show the CEMP candidate sample in a gradient from dark blue to beige, where dark blue shows the densest area of stars. The carbon-normal sample is shown as contour lines, where the highest density is shown with white contour lines; lower density areas have black contour lines. In general, the colour–magnitude distribution of the CEMP candidate and carbon-

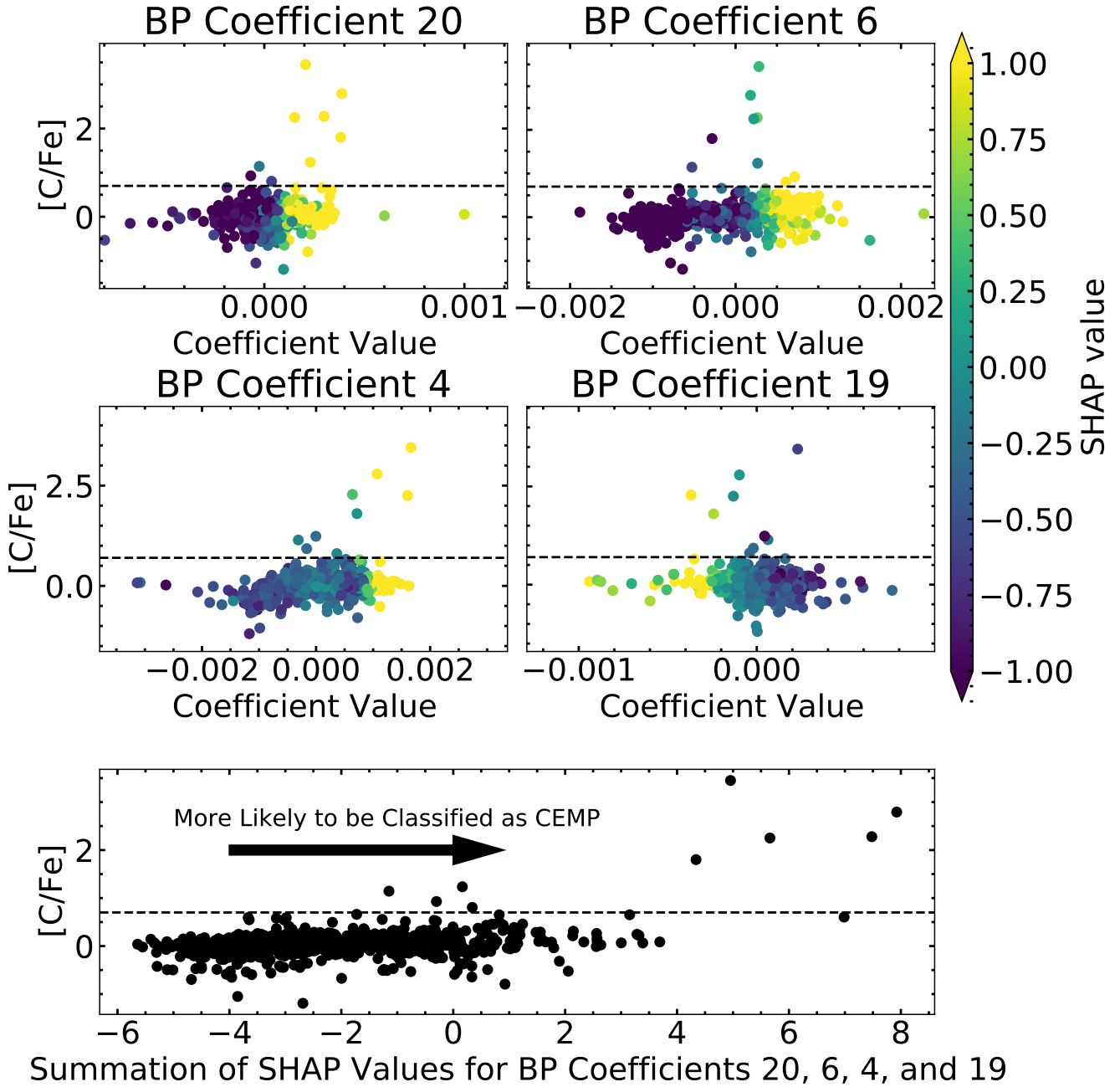


Figure 10. The relations between $[C/Fe]$, the spectral coefficient value, and the SHAP value for the four spectral coefficients with the largest SHAP values. We show a horizontal black dashed line at $[C/Fe] = +0.7$. The bottom panel shows $[C/Fe]$ as a function of the summation of the SHAP values of the four coefficients. Each point corresponds to an individual star in our testing sample. The SHAP values give the feature importance, in that large positive SHAP values indicate a large increase in the assigned probability of the star being carbon enhanced due to a given coefficient’s value. Therefore, if the model uses the carbon information in the coefficients, we expect that stars with $[C/Fe] > +0.7$ to have high positive SHAP values.

normal samples are similar, in that the highest density occurs where we expect bright dwarf and turn-off stars to reside ($G \approx 4\text{--}6$). However, one key difference occurs on the giant branch ($G < 4$). In general, the CEMP candidate giant stars are bluer than the carbon-normal giant stars. There is a clear red clump feature in the carbon-normal giant star distribution (black contour) at $G \approx 1$ and $G_{BP} - G_{RP} \approx 1.6$, but this does not appear in the CEMP distribution. This is consistent with expectations that the CEMP candidate sample is more metal poor than the carbon-normal sample, given that metal-poor stars will become blue-horizontal branch (BHB) after the RGB phase

rather than red-clump stars like their more metal-rich counterparts. Another interesting feature of the CEMP distribution is a distinct group of dwarf stars at $1.25 < G_{BP} - G_{RP} < 1.50$ and $G < 5$. This is similar to the distribution of dwarf stars in the training sample (see the left-hand panel of Fig. 2). This group appears to be brighter than the typical main-sequence stars in that range of $G_{BP} - G_{RP}$. It is possible that this may be due to an unresolved binary companion or a reddening effect from strong molecular absorption feature in the blue for the corresponding T_{eff} range. However, higher resolution follow-up is required to confirm these hypotheses.

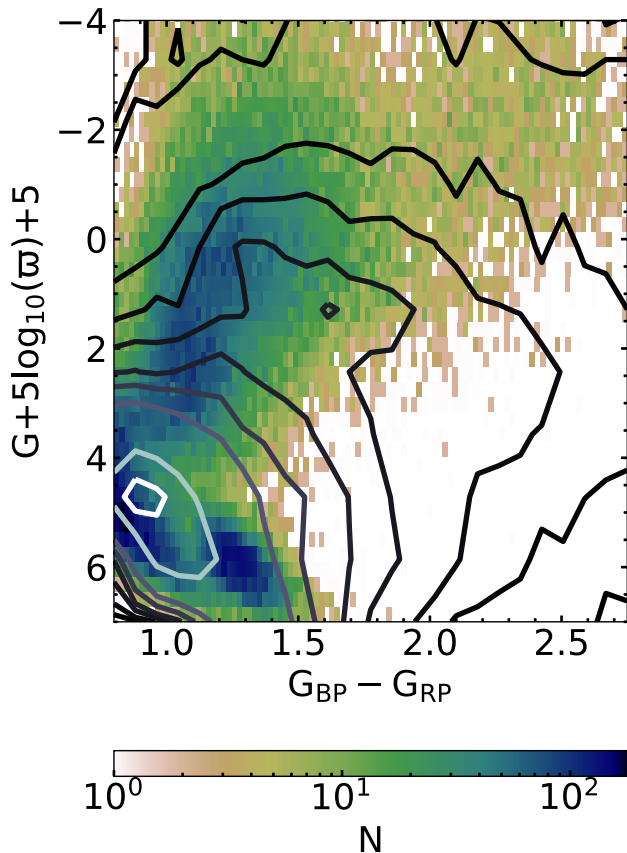


Figure 11. Colour-magnitude diagram of the final sample of candidate carbon-enhanced stars (dark blue/beige gradient), compared to a random sample of stars classified as not carbon-enhanced in white (high density) to black (low density) gradient contour lines. In general, the colour-magnitude distribution of our CEMP candidate sample is similar to the carbon-normal sample, in that the majority of stars are bright dwarf/turn-off stars. On the giant branch, however, the CEMP candidate sample tends to be bluer, while the carbon-normal sample shows a clear red clump at absolute $G \approx 1$. This indicates that the CEMP candidate sample is more metal-poor than the carbon-normal sample, as expected.

Fig. 12 shows the metallicity distribution of our sample of CEMP candidate stars in a grey histogram, using data-driven results from Andrae, Rix & Chandra (2023). In dark blue, we also show the ratio of the number of stars classified as CEMP to the total number of BP/RP spectra we analysed for $[M/H] < x$. In other words, the dark blue line gives the fraction of CEMP candidate stars for stars with $[M/H] < x$. It is important to note that this plot is only meant to explore the trends in our sample, and not meant as a measurement of the true occurrence rate of CEMP stars. With the uncertain metallicities, undefined selection function, and low completeness percentage, our sample currently cannot provide a robust estimate of the occurrence rate as a function of metallicity. However, the overall trend is consistent with previous work (Lucatello et al. 2006; Lee et al. 2013; Placco et al. 2014; Yoon et al. 2018; Arentsen et al. 2022) in that the occurrence rate increases with decreasing metallicity. Unexpectedly, we find that 57 per cent of our CEMP candidate sample has $[M/H]$ estimates from Andrae et al. (2023) higher than -1 , indicating that they are not metal poor. However, given that the $[M/H]$ estimates are data-driven and CEMP stars are outliers, it is unlikely that the $[M/H]$ values are robust for CEMP stars. It is likely that they are, in fact, over-estimated given the carbon enhancement, and the true $[Fe/H]$

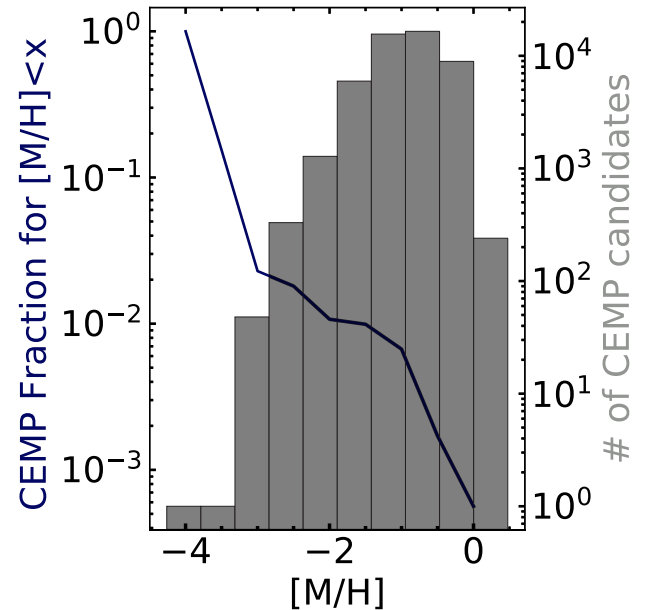


Figure 12. The behaviour of the occurrence rate of our classified CEMP candidate stars, as a function of metallicity ($[M/H]$) provided by Andrae et al. (2023). Specifically, we show the fraction of stars with $[M/H] < x$ that are classified as CEMP out of the total number of BP/RP spectra that we classify in a given metallicity bin. We also plot the metallicity distribution of our CEMP candidate sample in grey. We note this plot is only to explore the properties of our sample, and is not meant as a measure of the true occurrence rate of CEMP stars. The CEMP fraction increases with decreasing metallicity, consistent with previous results from high-resolution samples.

would be much lower. For example, using the 17 stars that have $p(\text{CEMP}) > 0.5$ in our catalogue and also $[C/Fe] > +0.7$ measured from APOGEE, we find an average $[Fe/H] = -1.39$, with minimum $[Fe/H] = -2.23$ and maximum $[Fe/H] = -0.82$. On the other hand, the average $[M/H]$ given by Andrae et al. (2023) for the same stars is $[M/H] = -0.03$, with minimum $[M/H] = -0.51$ and maximum $[M/H] = +0.49$. Comparing with the SEGUE training/testing sample, we find that the $[M/H]$ given by Andrae et al. (2023) is over-estimated compared to $[Fe/H]$ by an average of $+0.88$ dex for stars with $[C/Fe] > +0.7$. The over-estimation increases with increasing $[C/Fe]$ in that stars with $[C/Fe] > +4$ have $[M/H] - [Fe/H] > +3$ dex.

Fig. 13 shows the Galactic distribution of the CEMP candidate sample. Here, we use the geometric distances from Bailer-Jones et al. (2021). We choose not to use distances that are calculated using photometry, since it is likely that the carbon enhancement will bias these results. The top panel simply shows the sky-projected distribution in Galactic coordinates, l and b , which is independent of the parallax. Here, the Galactic centre is along the line-of-sight towards $(l, b) = (0, 0)^\circ$. We also show the distribution of our CEMP candidate sample in Galactic X and Y coordinates in the bottom right-hand panel, while the bottom left-hand panel shows the distribution in Galactic X and Z coordinates. The Galactic centre is located at $(0, 0, 0)$ kpc with the Sun located at $(8.3, 0, 0)$ kpc (Reid et al. 2014).

The CEMP candidate sample is spread throughout the Galaxy in a halo-like distribution. However, towards the inner Galaxy at low $|b|$ there is an under-density of stars, likely caused by extinction. The Large and Small Magellanic Clouds (LMC and SMC) stand out as clear features in the sky projected distribution, with peaks in the number density of CEMP candidate stars. This is consistent with previous work that has identified thousands of carbon-rich

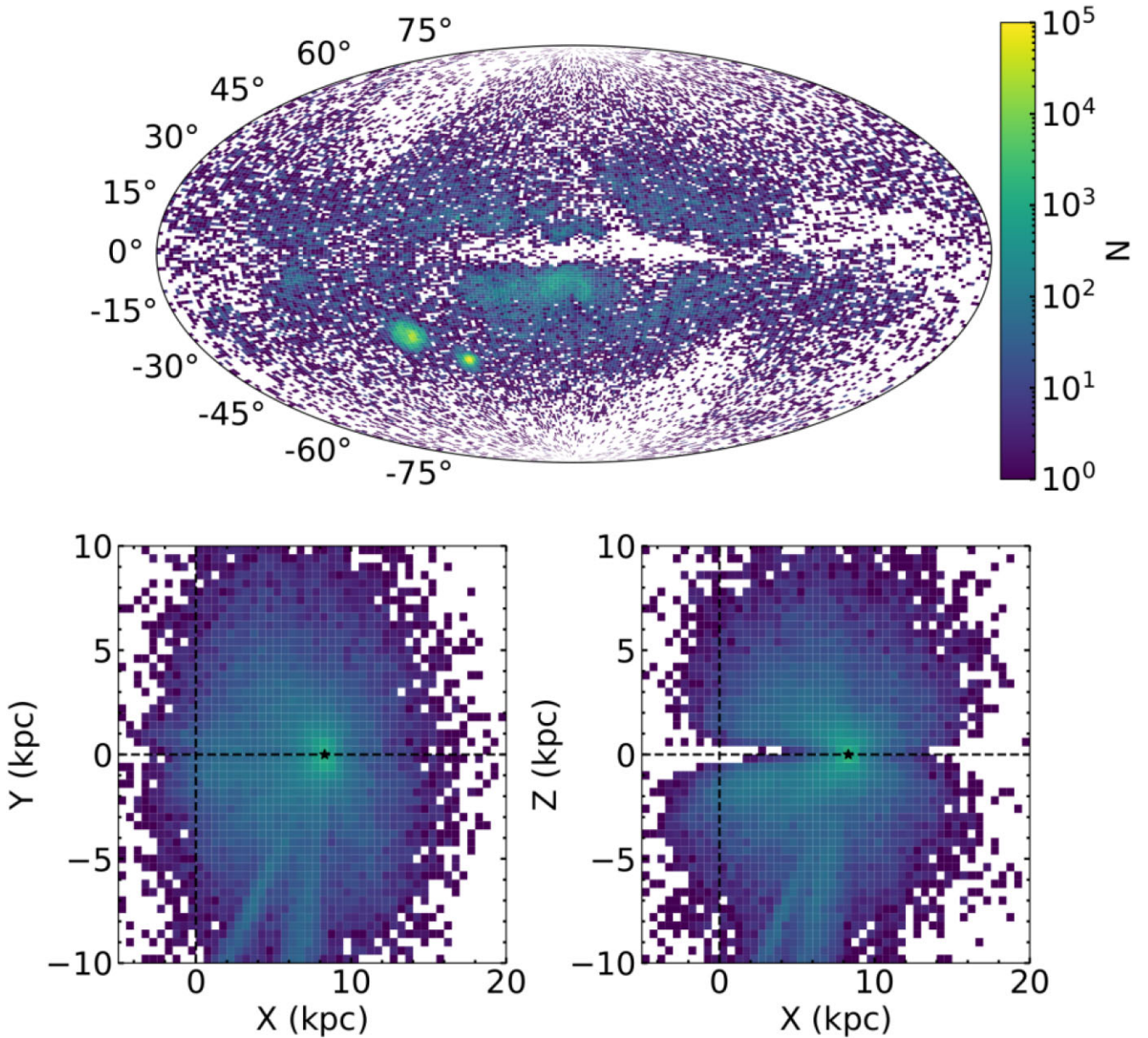


Figure 13. The Galactic distribution of our CEMP candidate sample. Specifically, in the top panel we show the sky projected distribution of the sample in Galactic coordinates, l and b , where $(l, b) = (0, 0)^\circ$ is the line-of-sight towards the Galactic centre. The bottom left-hand panel shows the distribution of stars in the Galactic coordinates X and Y . The bottom right-hand panel shows the distribution in the Galactic X and Z coordinates. The Galactic centre is located at $(0, 0, 0)$ kpc, while the Sun is at $(8.3, 0, 0)$ kpc. The LMC and SMC are clear features in the sky projected distribution.

(post-)AGB stars in the Magellanic clouds (Rebeiro, Azzopardi & Westerlund 1993; Kontizas et al. 2001). Similarly, the over-density of CEMP candidate stars towards the Galactic centre at negative b is likely associated with the Sagittarius dwarf galaxy. However, further work on the dynamics of these stars is required to tag them to a specific Galactic component. Additionally, spectroscopic observations are required to calculate evolutionary corrections, and determine which of these stars are potentially natal CEMP-no stars, rather than (post-)AGB carbon stars or CEMP- s stars. Impacts of the *Gaia* DR3 selection function can also be seen in the sky projected distribution at high $|b|$ with sweeping under-density features (see fig. 29 in De Angeli et al. 2022).

Consistent with a kinematically hot Galactic population, our CEMP candidate sample is extended to large distances from the

Galactic centre (see Fig. 13). It is important to note that some of the largest distances may be unreliable, given that the fractional parallax uncertainty for faint stars can be quite large. Given that the majority of BP/RP spectra released in DR3 have $G \approx 17.6$, and assuming an absolute G magnitude of -2.5 (roughly the tip of the RGB), we expect to have detected CEMP stars at distances of up to ≈ 30 kpc. We find that 8707 stars in our CEMP sample have parallaxes corresponding to distances from the Galactic centre > 30 kpc. However, the Galactic prior used in Bailer-Jones et al. (2021) brings most of these to within ≈ 10 kpc.

We also investigate the relative parallax distribution of CEMP candidate stars (red), compared to carbon-normal metal-rich stars (dark blue) and carbon-normal metal-poor stars (light blue) in Fig. 14. Specifically, we use the metallicity estimates from Andrae et al.

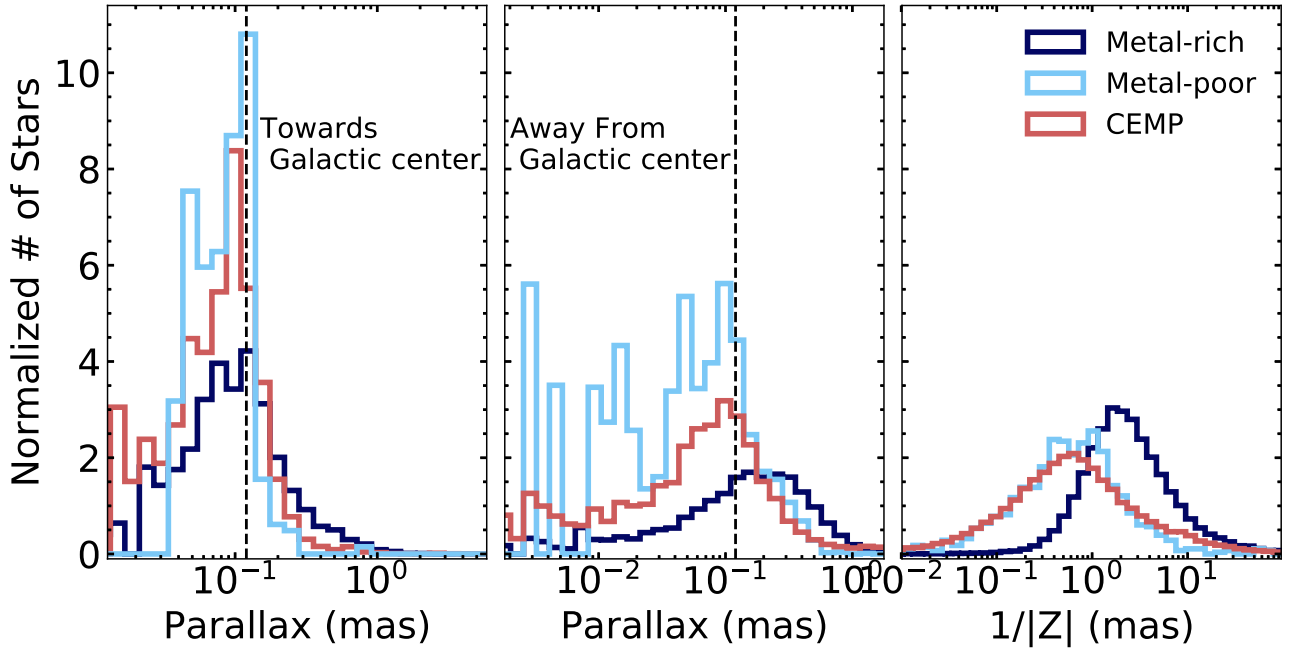


Figure 14. The parallax distribution of our CEMP candidate sample (red), compared to the metal-rich (dark blue) and metal-poor (light blue) carbon-normal sample. Specifically, in the left-hand panel, we show the parallax distributions for stars towards the Galactic centre with $|l| < 10^\circ$ (or $350^\circ < l < 10^\circ$) and $|b| < 10^\circ$. The middle panel shows the parallax distributions for stars towards the Galactic anticentre with $170^\circ < l < 190^\circ$ and $|b| < 10^\circ$. We also mark the parallax of the Galactic centre (0.12 mas) with a black vertical dashed line in both the left-hand and middle panels. The right-hand panel shows the inverse distance from the Galactic plane ($|Z|$) calculated by dividing the parallax by $\sin(b)$. The CEMP candidate stars generally follow the distribution of metal-poor stars.

(2023) with $[M/H] > -1$ defined as metal rich and $[M/H] < -1$ as metal poor. The left-hand panel shows the parallax distribution of CEMP candidate stars towards the Galactic centre by choosing stars with $|l| < 10^\circ$ and $|b| < 10^\circ$. We find that the distribution of CEMP candidate stars peaks at a parallax consistent with that of the Galactic centre (0.12 mas). The distribution of metal-rich and metal-poor carbon-normal stars also peaks at this parallax, but the metal-rich stars have a stronger tail towards large parallaxes, i.e. closer distances to the Sun. The middle panel shows the parallax distributions of CEMP candidate and carbon-normal stars towards the Galactic anticentre with $|l| < 170^\circ$ and $|b| < 10^\circ$. We find that the parallax distribution of CEMP candidate stars and carbon-normal metal-poor stars peaks at a parallax of ≈ 0.1 mas. Given that the average parallax precision for the faintest stars in our sample ($G \approx 17.5$) is ≈ 0.1 mas (Lindgren et al. 2021), this is consistent with CEMP stars peaking at distances ≥ 10 kpc. The carbon-normal metal-rich stars, on the other hand, peak at large parallaxes, indicating they are generally closer to the Sun than CEMP stars. This is consistent with previous work, which found the frequency of CEMP stars to increase with increasing distance from the Sun (Frebel et al. 2006; Carollo et al. 2012; Lee et al. 2013, 2017; Yoon et al. 2018). In the right-hand panel, we show the inverse of the distance from the Galactic plane ($1/|Z|$), which is calculated by dividing the parallax by $\sin(b)$. Again, we find that the CEMP candidate and metal-poor carbon-normal stars follow the same trends, and are generally farther from the Galactic plane than the carbon-normal stars. In general, we find that the CEMP candidate stars are more distant from the Sun than the carbon-normal stars, consistent with a Galactic halo population. Further work looking at the rate of carbon enhancement for metal-poor stars as a function of Galactic position is required to determine if CEMP stars have different origins than carbon-normal metal-poor stars.

7 SUMMARY

The origins of CEMP stars are poorly understood, even though they comprise ≈ 30 per cent of stars with $[\text{Fe}/\text{H}] < -2$ (Lucatello et al. 2006; Lee et al. 2013; Placco et al. 2014; Yoon et al. 2018; Arentsen et al. 2022). A significant fraction of CEMP stars have enhancements in s -process elements, and are called CEMP- s stars (Beers & Christlieb 2005). These stars are thought to receive their over-abundant carbon and s -process elements from a mass-transfer event with their binary companion, which has evolved to or past the AGB (Lugaro et al. 2012; Placco et al. 2013). On the other hand, CEMP stars without neutron-capture enhancements, the CEMP-no stars, are thought to have been primarily enriched by material from the first generations of stars (Umeda & Nomoto 2003; Chiappini et al. 2006; Meynet et al. 2006; Nomoto et al. 2013; Tominaga et al. 2014). However, there are many remaining questions about these unique stars, including why they seem to be less frequent in the central regions of our Galaxy, where we expect the highest concentration of ancient stars to reside (Howes et al. 2014, 2015, 2016; Arentsen et al. 2021; Lucey et al. 2022). As suggested by Yoon et al. (2019), this dearth of CEMP stars may be caused by the dilution of CEMP stars in more massive subsystems (e.g. dwarf galaxies) with prolonged star-formation histories, which could be the origin of metal-poor stars in the inner Galaxy. However, given that the discrepancy in the CEMP fraction of the inner Galaxy is highest for $[\text{Fe}/\text{H}] > -2.5$, Arentsen et al. (2021) argue that it is due to a lower rate of CEMP- s stars caused by a lower binary fraction in the inner Galaxy.

In this work, we leverage the data from the all-sky *Gaia* survey to identify 58 872 CEMP candidates. Specifically, we use the ~ 180 million BP/RP spectra made available in *Gaia* DR3. Using the XGBoost algorithm for classification, we achieve a completeness of 26 per cent and a contamination rate of 12 per cent. When comparing

to high-resolution catalogues of CEMP stars (Zepeda et al. 2023), we find that we positively identify 60–68 per cent of the previously known CEMP stars in the *Gaia* DR3 BP/RP data. We ensure that the XGBoost algorithm matches our physical intuition and primarily performs the classification using spectral features that are correlated with the carbon abundance.

We briefly investigate a few of the properties of our CEMP candidate sample, including the metallicity distribution and the Galactic spatial distribution. As expected, the CEMP fraction increases with decreasing metallicity. In general, we find that the CEMP candidate stars tend to follow the distribution of metal-poor carbon-normal stars, and that they are farther from the Sun than metal-rich carbon-normal stars.

In future work, we plan to look at the orbital properties of a subset of these stars, as well as the rate of *s*-process enhancement. We plan to perform medium- and high-resolution spectroscopic follow-ups of many of these targets in order to confirm our contamination rate, measure radial velocities where needed, derive their dynamical properties (see e.g. Dietz et al. 2020, 2021), and identify chemodynamical groups (see e.g. Zepeda et al. 2023), and determine their neutron-capture abundances. We plan to specifically follow-up targets towards the central region of the Galaxy where the number of known CEMP stars is much lower. Following future *Gaia* releases, which will include many millions more BP/RP spectra, we expect to continue this work and again increase the number of CEMP candidate stars.

ACKNOWLEDGEMENTS

This project was developed in part at the *Gaia* Fête, held at the Flatiron Institute Centre for Computational Astrophysics in June 2022. We would like to thank David Hogg and Jason Hunt for their helpful discussions and computational resources. We also thank an anonymous referee for a careful review of this paper, which has improved the clarity of its presentation.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 000392968. KH acknowledges support from the National Science Foundation grant AST-1907417 and AST-2108736 and from the Wootton Center for Astrophysical Plasma Properties funded under the United States Department of Energy collaborative agreement DE-NA0003843. This work was performed in part at the Simons Foundation Flatiron Institute’s Center for Computational Astrophysics during KH’s tenure as an IDEA Fellow. YST acknowledges financial support from the Australian Research Council through DECRA Fellowship DE220101520. Work at Argonne National Laboratory was supported by the U.S. Department of Energy, Office of High Energy Physics. Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated by UChicago Argonne LLC under contract no. DE-AC02-06CH11357. NR acknowledges the Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. TCB and YSL acknowledge partial support for this work from grant PHY 14–30152; Physics Frontier Center/JINA Center for the Evolution of the Elements (JINACEE), and OISE-1927130: The International Research Network for Nuclear Astrophysics (IRENA), awarded by the US National Science Foundation. YSL also acknowledges support from the National Research Foundation (NRF) of Korea grant funded by the Ministry of Science and ICT (NRF-2021R1A2C1008679).

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.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III website is <http://www.sdss3.org/>.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

This research made use of Astropy,⁴ a community-developed core Python package for ASTRONOMY (Astropy Collaboration 2013, 2018). Other software used includes IPYTHON (Pérez & Granger 2007), MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), SCIKIT-LEARN (Pedregosa et al. 2011), and SCIPY (Virtanen et al. 2020).

DATA AVAILABILITY

The data underlying this article are publicly available through the *Gaia* Archive.³ The generated catalogue of CEMP candidate stars will be available in the online supplementary material and CDS upon acceptance. The XGBoost model and all code will be made available on Github upon acceptance.

REFERENCES

- Abate C., Pols O. R., Karakas A. I., Izzard R. G., 2015, *A&A*, 576, A118
- Abdurro’uf et al., 2022, *ApJS*, 259, 35
- Allende Prieto C. et al., 2008, *AJ*, 136, 2070
- Alvarez R., Plez B., 1998, *A&A*, 330, 1109
- Andrae R. et al., 2018, *A&A*, 616, A8
- Andrae R. et al., 2022, preprint([arXiv:2206.06138](https://arxiv.org/abs/2206.06138))
- Andrae R., Rix H.-W., Chandra V., 2023, preprint([arXiv:2302.02611](https://arxiv.org/abs/2302.02611))
- Arentsen A., Starkenburg E., Shetrone M. D., Venn K. A., Depagne É., McConnachie A. W., 2019, *A&A*, 621, A108
- Arentsen A. et al., 2021, *MNRAS*, 505, 1239
- Arentsen A., Placco V. M., Lee Y. S., Aguado D. S., Martin N. F., Starkenburg E., Yoon J., 2022, *MNRAS*, 515, 4082
- Astropy Collaboration, 2013, *A&A*, 558, A33
- Astropy Collaboration, 2018, *AJ*, 156, 123
- Bailer-Jones C. A. L., Rybizki J., Foesneau M., Demleitner M., Andrae R., 2021, *AJ*, 161, 147
- Beers T. C., Christlieb N., 2005, *ARA&A*, 43, 531
- Beers T. C., Preston G. W., Shectman S. A., 1992, *AJ*, 103, 1987

⁴<http://www.astropy.org>

³<https://gea.esac.esa.int/archive/>

- Beers T. C. et al., 2017, *ApJ*, 835, 81
- Bessell M., Bloxham G., Schmidt B., Keller S., Tisserand P., Francis P., 2011, *PASP*, 123, 789
- Bisterzo S., Gallino R., Straniero O., Cristallo S., Käppeler F., 2010, *MNRAS*, 404, 1529
- Carollo D. et al., 2012, *ApJ*, 744, 195
- Carollo D., Freeman K., Beers T. C., Placco V. M., Tumlinson J., Martell S. L., 2014, *ApJ*, 788, 180
- Carrasco J. M. et al., 2021, *A&A*, 652, A86
- Chen T., Guestrin C., 2016, preprint(arXiv:1603.02754)
- Chiappini C., Hirschi R., Meynet G., Ekström S., Maeder A., Matteucci F., 2006, *A&A*, 449, L27
- Chiti A., Frebel A., Jerjen H., Kim D., Norris J. E., 2020, *ApJ*, 891, 8
- Christlieb N., Schörck T., Frebel A., Beers T. C., Wisotzki L., Reimers D., 2008, *A&A*, 484, 721
- Da Costa G. S. et al., 2019, *MNRAS*, 489, 5900
- De Angeli F. et al., 2022, preprint(arXiv:2206.06143)
- Dietz S. E., Yoon J., Beers T. C., Placco V. M., 2020, *ApJ*, 894, 34
- Dietz S. E., Yoon J., Beers T. C., Placco V. M., Lee Y. S., 2021, *ApJ*, 914, 100
- Fitzpatrick E. L., 2004, in Witt A. N., Clayton G. C., Draine B. T., eds, ASP Conf. Ser. Vol. 309, *Astrophysics of Dust*. Astron. Soc. Pac., San Francisco, p. 33
- Frebel A., 2018, *Ann. Rev. Nucl. Part. Sci.*, 68, 237
- Frebel A., Norris J. E., 2015, *ARA&A*, 53, 631
- Frebel A. et al., 2006, *ApJ*, 652, 1585
- Frebel A., Simon J. D., Kirby E. N., 2014, *ApJ*, 786, 74
- Gaia Collaboration, 2016, *A&A*, 595, A2
- Gaia Collaboration, 2022, *A&A*
- Gavel A., Andrae R., Fouesneau M., Korn A. J., Sordo R., 2021, *A&A*, 656, A93
- Gull M. et al., 2018, *ApJ*, 862, 174
- Gustafsson B., Edvardsson B., Eriksson K., Jørgensen U. G., Nordlund Å., Plez B., 2008, *A&A*, 486, 951
- Hansen T. T., Andersen J., Nordström B., Beers T. C., Placco V. M., Yoon J., Buchhave L. A., 2016a, *A&A*, 586, A160
- Hansen T. T., Andersen J., Nordström B., Beers T. C., Placco V. M., Yoon J., Buchhave L. A., 2016b, *A&A*, 588, A3
- Harris C. R. et al., 2020, *Nature*, 585, 357
- Hayden M. R. et al., 2022, *MNRAS*, 517, 5325
- He X.-J., Luo A. L., Chen Y.-Q., 2022, *MNRAS*, 512, 1710
- Heiter U. et al., 2021, *A&A*, 645, A106
- Howes L. M. et al., 2014, *MNRAS*, 445, 4241
- Howes L. M. et al., 2015, *Nature*, 527, 484
- Howes L. M. et al., 2016, *MNRAS*, 460, 884
- Hunter J. D., 2007, *Comput. Sci. Eng.*, 9, 90
- Jorissen A. et al., 2016, *A&A*, 586, A158
- Kontizas E., Dapergolas A., Morgan D. H., Kontizas M., 2001, *A&A*, 369, 932
- Lai D. K., Lee Y. S., Bolte M., Lucatello S., Beers T. C., Johnson J. A., Sivarani T., Rockosi C. M., 2011, *ApJ*, 738, 51
- Lee Y. S. et al., 2008a, *AJ*, 136, 2050
- Lee Y. S. et al., 2008b, *AJ*, 136, 2022
- Lee Y. S. et al., 2011, *AJ*, 141, 90
- Lee Y. S. et al., 2013, *AJ*, 146, 132
- Lee Y. S., Beers T. C., Kim Y. K., Placco V., Yoon J., Carollo D., Masseron T., Jung J., 2017, *ApJ*, 836, 91
- Lee Y. S., Beers T. C., Kim Y. K., 2019, *ApJ*, 885, 102
- Li C. et al., 2021, *MNRAS*, 506, 1651
- Lindgren L. et al., 2021, *A&A*, 649, A2
- Liu C., Bailer-Jones C. A. L., Sordo R., Vallenari A., Borraero R., Luri X., Sartoretti P., 2012, *MNRAS*, 426, 2463
- Lucatello S., Tsangarides S., Beers T. C., Carretta E., Gratton R. G., Ryan S. G., 2005, *ApJ*, 625, 825
- Lucatello S., Beers T. C., Christlieb N., Barklem P. S., Rossi S., Marsteller B., Sivarani T., Lee Y. S., 2006, *ApJ*, 652, L37
- Lucey M. et al., 2022, *MNRAS*, 509, 122
- Lugaro M., Karakas A. I., Stancliffe R. J., Rijs C., 2012, *ApJ*, 747, 2
- Lundberg S. M., Lee S.-I., 2017, *Adv. Neur. Inf. Proces. Syst.*, 30
- Machado Poletti Valle L. F., Avestruz C., Barnes D. J., Farahi A., Lau E. T., Nagai D., 2021, *MNRAS*, 507, 1468
- Masseron T. et al., 2014, *A&A*, 571, A47
- McClure R. D., Woodsworth A. W., 1990, *ApJ*, 352, 709
- Meynet G., Ekström S., Maeder A., 2006, *A&A*, 447, 623
- Nomoto K., Kobayashi C., Tominaga N., 2013, *ARA&A*, 51, 457
- Norris J. E., Wyse R. F. G., Gilmore G., Yong D., Frebel A., Wilkinson M. I., Belokurov V., Zucker D. B., 2010, *ApJ*, 723, 1632
- Pedregosa F. et al., 2011, *J. Mach. Learn. Res.*, 12, 2825
- Pérez F., Granger B. E., 2007, *Comput. Sci. Eng.*, 9, 21
- Pham D., Kaltenegger L., 2022, *MNRAS*, 513, L72
- Placco V. M., Frebel A., Beers T. C., Karakas A. I., Kennedy C. R., Rossi S., Christlieb N., Stancliffe R. J., 2013, *ApJ*, 770, 104
- Placco V. M., Frebel A., Beers T. C., Stancliffe R. J., 2014, *ApJ*, 797, 21
- Plez B., 2012, *Astrophysics Source Code Library*. record ascl:1205.004
- Preston G. W., Sneden C., 2001, *AJ*, 122, 1545
- Rebeiro E., Azzopardi M., Westerlund B. E., 1993, *A&AS*, 97, 603
- Reid M. J. et al., 2014, *ApJ*, 783, 130
- Riello M. et al., 2021, *A&A*, 649, A3
- Rockosi C. M. et al., 2022, *ApJS*, 259, 60
- Salvadori S., Skúladóttir Á., Tolstoy E., 2015, *MNRAS*, 454, 1320
- Skúladóttir Á., Tolstoy E., Salvadori S., Hill V., Pettini M., Shetrone M. D., Starkenburg E., 2015, *A&A*, 574, A129
- Smolinski J. P. et al., 2011, *AJ*, 141, 89
- Starkenburg E., Shetrone M. D., McConnachie A. W., Venn K. A., 2014, *MNRAS*, 441, 1217
- Starkenburg E. et al., 2017, *MNRAS*, 471, 2587
- Tominaga N., Iwamoto N., Nomoto K., 2014, *ApJ*, 785, 98
- Umeda H., Nomoto K., 2003, *Nature*, 422, 871
- Virtanen P. et al., 2020, *Nat. Methods*, 17, 261
- Witten C. E. C. et al., 2022, *MNRAS*, 516, 3254
- Wolf C. et al., 2018, *PASA*, 35, e010
- Yanny B. et al., 2009, *AJ*, 137, 4377
- Yoon J. et al., 2016, *ApJ*, 833, 20
- Yoon J. et al., 2018, *ApJ*, 861, 146
- Yoon J., Beers T. C., Tian D., Whitten D. D., 2019, *ApJ*, 878, 97
- Zepeda J. et al., 2023, *ApJ*, 947, 23

APPENDIX A: ONLINE TABLE OF CEMP CANDIDATES

We show a section of the available online table in Table A1. The online table includes the source IDs for all 182 815 672 *Gaia* DR3 objects with BP/RP spectra within our color cuts. We also provide the inferred probability of being a CEMP star for each star.

Table A1. CEMP probability catalogue.

Source ID	$p(\text{CEMP})$
909506570173184	0.88
1012963742461952	0.00
1200018158226816	0.91
1235889725071360	0.99
1452184278048512	0.00
7683563349186048	0.00
7746888346949120	0.00
8731294851221504	0.63
8779982600469632	0.99
8797952743613440	0.52
...	...

Note. A sample of the provided online catalog of CEMP probabilities. We provide the *Gaia* DR3 source ID for each star and the corresponding probability of being a CEMP star ($p(\text{CEMP})$).

This paper has been typeset from a $\text{\LaTeX}$ file prepared by the author.