Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **527,** 10937–10954 (2024) https://doi.org/10.1093/mnras/stad3775
Advance Access publication 2023 December 12

200 000 candidate very metal-poor stars in *Gaia* DR3 XP spectra

Yupeng Yao (姚宇鹏),[1]★ Alexander P. Ji [1,2] Sergey E. Koposov [3,4,5] and Guilherme Limberg [1,2,6]

[1]*Department of Astronomy and Astrophysics, University of Chicago, 5640 S Ellis Avenue, Chicago, IL 60637, USA*
[2]*Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[3]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[4]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[5]*Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[6]*Universidade de São Paulo, IAG, Departamento de Astronomia, Rua do Matão 1226, Cidade Universitária, SP 05508-090 São Paulo, Brazil*

Accepted 2023 November 20. Received 2023 November 15; in original form 2023 April 1

ABSTRACT

Very metal-poor stars ([Fe/H] < -2) in the Milky Way are fossil records of early chemical evolution and the assembly and structure of the Galaxy. However, they are rare and hard to find. *Gaia* DR3 has provided over 200 million low-resolution ($R \approx 50$) XP spectra, which provides an opportunity to greatly increase the number of candidate metal-poor stars. In this work, we utilize the XGBoost classification algorithm to identify ∼200 000 very metal-poor star candidates. Compared to past work, we increase the candidate metal-poor sample by about an order of magnitude, with comparable or better purity than past studies. First, we develop three classifiers for bright stars ($BP < 16$). They are Classifier-T (for Turn-off stars), Classifier-GC (for Giant stars with high completeness), and Classifier-GP (for Giant stars with high purity) with expected purity of 52 per cent/45 per cent/76 per cent and completeness of 32 per cent/93 per cent/66 per cent, respectively. These three classifiers obtained a total of 11 000/111 000/44 000 bright metal-poor candidates. We apply model-T and model-GP on faint stars ($BP > 16$) and obtain 38 000/41 000 additional metal-poor candidates with purity 29 per cent/52 per cent, respectively. We make our metal-poor star catalogues publicly available, for further exploration of the metal-poor Milky Way.

Key words: methods: statistical – techniques: photometric – techniques: spectroscopic – stars: Population II.

1 INTRODUCTION

Very metal-poor stars (VMP, [Fe/H]<-2[1]; Beers & Christlieb 2005) are fossil records of early chemical enrichment history of the Universe. The most metal-poor stars are likely to be some of the oldest stars that exist today, and their atmospheres contain information about the abundance pattern of gas in the early Universe (e.g. Frebel & Norris 2015). Chemical abundances of a large sample of metal-poor stars can advance our understanding of early nucleosynthesis and thus constrain the early stellar masses, rotation rates, mixing processes, explosion energies, compact remnant masses (neutron stars or black holes), thermohaline convection, and other stellar properties (e.g. Heger & Woosley 2010; Limongi & Chieffi 2012; Wanajo 2018; Jones et al. 2019; Ishigaki et al. 2021). Moreover, chemical abundances for these stars, together with kinematic data, can be utilized to understand the accretion history, and early formation of the Milky-Way (e.g. Hawkins et al. 2015; Das, Hawkins & Jofré 2020; Horta et al. 2021; Belokurov & Kravtsov 2022; Conroy et al. 2022; Rix et al. 2022, see Helmi 2020 for a review).

However, metal-poor stars are rare and difficult to find. Metal-poor stars only make up ∼0.1 per cent of Milky Way stars (e.g. Starkenburg et al. 2017; El-Badry et al. 2018), and only few thousands of metal-poor stars have been spectroscopically confirmed in past surveys (e.g. Li, Tan & Zhao 2018; Placco et al. 2018; Chiti et al. 2021a). The typical method to search for metal-poor stars is first finding metal-poor candidates and then following up these stars with medium/high-resolution spectra to get more detailed information (e.g. Beers & Christlieb 2005). Objective-prism surveys, photometric surveys, and some wide area spectroscopic surveys are the major ways to search for metal-poor stars. Objective-prism surveys (Bond 1970; Bidelman & MacConnell 1973; Bond 1980) were once the most effective method to search for candidate metal-poor stars, which utilized low-resolution spectra ($R \approx 400$) to estimate the strength of the Ca II K line at 393.36 nm. The HK-I, HK-II, and Hamburg/ESO surveys (Beers, Preston & Shectman 1985, 1992; Frebel et al. 2006; Christlieb et al. 2008; Beers et al. 2017) found a total of ∼4500 VMP stars (Limberg et al. 2021a). More recently, photometric surveys are utilized to identify candidate metal-poor stars. SkyMapper Southern Sky Survey (SMSS) utilizes SkyMapper v filter that reflect Ca II H&K absorption features, together with SkyMapper u, g, i photometry to derive metallicities (Onken et al. 2019; Chiti et al. 2021a). Analogously, Pristine utilizes a narrow-band filter that is centred on the Ca II H&K absorption lines, combined with SDSS broad-band g and i photometry to derive metallicities (Starkenburg et al. 2017; Aguado et al. 2019). Javalambre Photometric Local Universe Survey (J-PLUS) (Cenarro et al. 2019) and the Southern Photometric Local Universe Survey (S-PLUS) (Mendes de Oliveira et al. 2019) are also photometric surveys which utilize four SDSS-

★ E-mail: yyaoastro@gmail.com

[1]Standard nomenclature would be Very Metal-Poor for [Fe/H] < -2. From here we will refer to very metal-poor as just metal-poor.

like (g, r, i, z) and one modified SDSS (u), and seven narrow-band filters to identify low-metallicity stars in the Galactic halo (Placco et al. 2021; Galarza et al. 2022; Placco et al. 2022). Another photometric selection method is Best & Brightest (Schlaufman & Casey 2014) which utilizes all-sky APASS optical, 2MASS near-infrared, and *WISE* mid-infrared photometry to identify bright metal-poor star candidates through their lack of molecular absorption near 4.6 microns (Placco et al. 2019; Reggiani et al. 2020; Limberg et al. 2021b). Besides the aforementioned dedicated efforts, there are some large surveys that directly observe samples of stars at intermediate resolution spectra and estimate their metallicity, e.g. SEGUE, LAMOST, and RAVE surveys. These surveys have found several thousand of metal-poor stars. The Sloan Digital Sky Survey (SDSS; Eisenstein et al. 2011), and its Sloan Extension for Galactic Understanding and Exploration (SEGUE; Yanny et al. 2009) survey ($R \approx 2000$), SEGUE-1 and SEGUE-2, which motivated several high-resolution follow-up campaigns (e.g. Aoki et al. 2012). The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) survey ($R \approx 1800$; Deng et al. 2012), which has also triggered some high resolution observations (e.g. Li et al. 2022). LAMOST-I(DR7) released more than seven million spectra of stars in the Milky Way. The RAdial Velocity Experiment (RAVE; $R \approx 7000$) (Kunder et al. 2017) delivered spectra for about 480 000 stars. However, the number of candidate metal-poor stars found from each survey is about a few dozens to at most a few thousand, which is too small for a statistical investigation on metal-poor stars, especially for extremely metal-poor ([Fe/H] < -3) or ultra metal-poor regime ([Fe/H] < -4). Thus we need a survey that can provide a much larger number of stellar spectra to enable us to find such objects.

The *Gaia* mission has brought a revolutionary change to Milky Way astronomy, because it provides astrometric data for billions of stars (Collaboration et al. 2016, 2022). In *Gaia* Data Release 3 (DR3), it released 200 million low-resolution XP spectra ($R \approx 50$; De Angeli et al. 2023). Because of its low-resolution, the XP spectra cannot provide detailed element abundances of stars. Additionally, *Gaia* GSP-Phot also does not provide accurate metallicity estimations for the most metal-poor stars (Andrae et al. 2023a). However, some works have demonstrated that these low resolution XP spectra can be utilized to estimate effective temperature, surface gravity, and metallicity (e.g. Xylakis-Dornbusch et al. 2022; Andrae, Rix & Chandra 2023b; Zhang, Green & Rix 2023). Thus, these 200 million low-resolution XP spectra give us an opportunity to greatly increase the number of candidate metal-poor stars, if we can make full use of them.

In this work, we identify metal-poor stars in the *Gaia* DR3 XP spectra using the XGBoost classification algorithm. In Section 2, we describe the XP spectra and other data we utilized in this work. In Section 3, we introduce XGBoost, discuss the training process, and evaluate the performance of the models. Then, we utilize XGBoost models to make a prediction on the XP spectra, shown and discussed in Section 4. Then, we compare our work with other surveys and projects and utilize existing high-resolution spectroscopic data to validate the performance of our models in Section 5. Finally, we summarize this work in Section 6.

2 DATA

2.1 Data sets

In this work, the data utilized include *Gaia* DR3 XP spectra (De Angeli et al. 2023), *Gaia* DR3 photometry (Vallenari et al. 2023), LAMOST DR7 (Cui et al. 2012) metallicity, and Apache Point Observatory Galactic Evolution Experiment (APOGEE; Majewski et al. 2017) DR17 (Abdurro'uf et al. 2022) metallicity.

***Gaia* XP spectra:** *Gaia* DR3 released low-resolution blue and red photometer spectra (*BP*/*RP* or XP spectra) for 210 million stars. Metallicities were derived from these spectra in the *Gaia* GSP-Phot, but they are not accurate at low metallicities (Andrae et al. 2023a). Thus, it is not efficient to directly utilize the GSP-Phot metallicity [M/H] in *Gaia* DR3 to search for metal-poor stars. The XP spectra have wide wavelength coverage (330 to 1050 nm) and low-resolution. Because of its wide wavelength coverage, strong lines valuable for metallicity estimation are covered in it, such as Ca II K and Ca II infrared triplet, as well as broad-band or narrow-band photometry. Thus, in theory, XP spectra can be utilized to detect metal-poor stars. The XP spectra are released as Hermite function coefficients rather than fluxes versus wavelength (Carrasco et al. 2021). In order to avoid information loss (Carrasco et al. 2021), the input for XGBoost model are XP spectra coefficients, rather than corresponding sampled XP spectra. XGBoost requires the input vectors to be of the same length, so we do not truncate the XP coefficients.

Before inputting the XP coefficients to the model, we first normalize and deredden them. We normalized XP coefficients by their first coefficient to remove apparent magnitude information. Additionally, to take into account reddening, we determined the extinction coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ to correct the normalized XP coefficient vectors $\mathbf{C}$ for extinction $\mathbf{C}_{\rm corrected} = \mathbf{C} - (\boldsymbol{\alpha} + \boldsymbol{\beta}\hat{\mathbf{C}})E_{B-V} - \boldsymbol{\gamma}E_{B-V}^2$. Here, the $\hat{\mathbf{C}}$ is a truncated XP coefficient vector with first 10 elements, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ are vectors, and $\boldsymbol{\beta}$ is a matrix. We fit for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ by taking high extinction stars in APOGEE and matching them with stars with similar $\log g$ (surface gravity), $T_{\rm eff}$ (effective temperature), and metallicity, but at low extinction. The extinction utilized in this analysis is from a 2D map by Schlegel, Finkbeiner & Davis (1998).

Gaia DR3 photometry: We also utilized *Gaia* DR3 photometry (Vallenari et al. 2023) in this work. Gaia's G band covers a wavelength range from near ultraviolet (∼330 nm) to the infrared (∼1050 nm). The other two bands, denoted *BP* and *RP*, cover smaller wavelength ranges, from approximately 330 to 680 nm, and 630 to 1050 nm, respectively.[2] We utilize the extinction law, as described in https://www.cosmos.esa.int/web/gaia/edr3-extinction-law, to get the intrinsic colour $(BP - RP)_0$.

LAMOST DR7 and APOGEE DR17 metallicity: In order to train our model to identify metal-poor stars, we need a sample of stars that already have reliable metallicity estimates to provide true labels. We utilized the spectroscopic metallicity from the LAMOST DR7[3] and APOGEE DR17[4] LAMOST spectra ($R \approx 1800$) cover the optical band from 370 to 900 nm. APOGEE spectra ($R \approx 22\,500$) are a good complement to LAMOST, because they cover the infrared band from 1.51 to 1.70 μm, which is more suited for dust extincted regions, i.e. the Galactic disc and bulge. In total, we have 4×10^6 LAMOST and 6.5×10^5 APOGEE stars.

Data queries and quality cuts: We utilized the Whole Sky Database (WSDB)[5] for all queries (see Appendix C for the ADQL queries), which ingested the entire catalogue for APOGEE DR17 and LAMOST DR7. We did not do any significant quality cuts, but we do not think this will significantly affect the results for a few reasons. First, classification models are less sensitive to quality cuts than regression models. Secondly, after comparing the overlapping

[2] https://www.cosmos.esa.int/web/gaia/edr3-passbands
[3] https://dr7.lamost.org/
[4] https://www.sdss4.org/dr17/
[5] https://www.ast.cam.ac.uk/ioa/wikis/WSDB/index.php/Main_Page

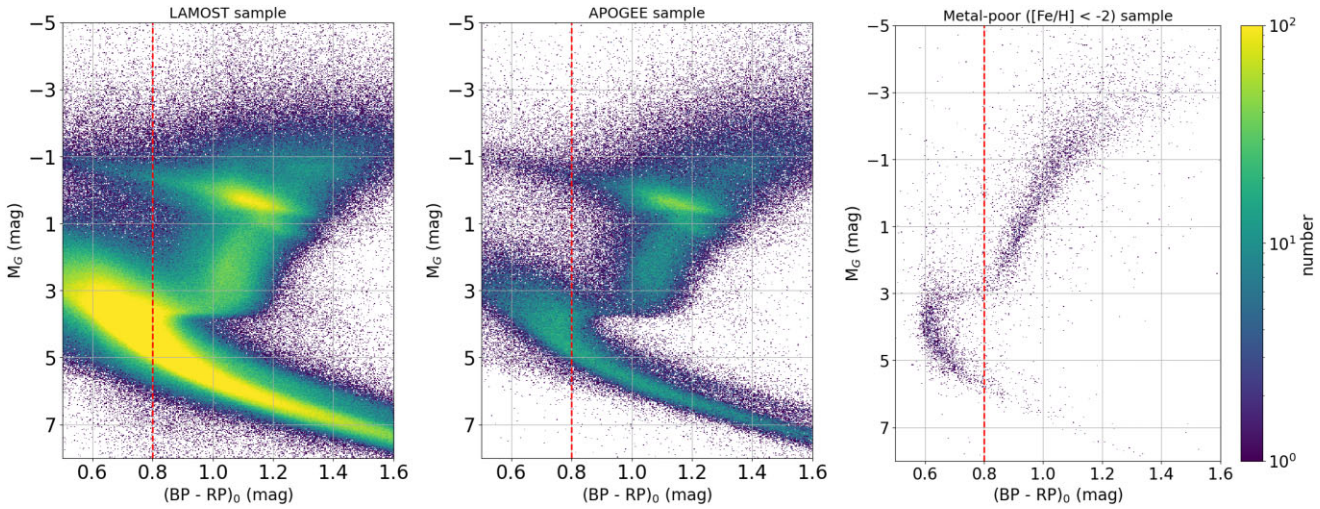

Figure 1. Colour–magnitude diagram of our training and testing sets. The horizontal axis is the *Gaia* intrinsic colour $(BP - RP)_0$, the vertical axis is the *Gaia* absolute G magnitude. The LAMOST and APOGEE samples, which primarily comprise main-sequence turn-off, giants, and dwarfs stars, are shown in the left and middle panels. The right panel shows the metal-poor stars from LAMOST and APOGEE. Metal-poor stars are primarily turn-off and giant stars. We divide the training and testing set into two parts, according to $(BP - RP)_0$, as shown in the red dashed line in the figure. On the left/right side of red dashed line are the samples utilized to train the model to identify the turn-off/giants metal-poor stars.

very metal-poor stars in LAMOST and APOGEE, we found that even if a star is flagged as bad spectral fitting solutions in either APOGEE or LAMOST, it often still carries sufficient information regarding being very metal-poor or not. For example, for LAMOST we adopted quality flags of SNR >20 and feh_err <0.5 (e.g. Zhang, Green & Rix 2023), which removed 22 per cent of our metal-poor training set. However, overlapping APOGEE spectra suggested that 84 per cent of these were actually still very metal-poor. For APOGEE, metal-poor stars run up against the edge of the spectral grid, so using quality flags (e.g. FE_H_FLAG = 0) removed all stars with [Fe/H] < -2.25 even though they are very metal-poor in LAMOST.

2.2 Training and testing sets

The *Gaia* XP spectra with the LAMOST or APOGEE metallicity form the training and testing set in this work. We directly put them together because the average difference between LAMOST and APOGEE[Fe/H]is 0.007 dex, which is well below the typical uncertainty in metallicity of LAMOST (>0.2 dex) or APOGEE (> 0.1 dex). Therefore, we conclude that these surveys are on similar metallicity scales within the range of parameters tested. Before the training process, we need to set some constraints on the training and testing set by intrinsic colour $(BP - RP)_0$, magnitude BP and extinction $E(B - V)$.

For the training set, we only consider stars with $(BP - RP)_0 > 0.5$, because, as shown in the right panel of Fig. 1, we do not have metal-poor samples with $(BP - RP)_0 < 0.5$. Note that the method utilized to calculate the $(BP - RP)_0$ excludes almost all of the $E(B - V) > 2$ stars, because the extinction coefficients should not be extrapolated outside the extinction range of this algorithm, as described in https://www.cosmos.esa.int/web/gaia/edr3-extinction-law. Additionally, we exclude fainter stars ($BP > 16$) in the training set, because the XP spectra with $BP > 16$ generally do not have high signal-to-noise ratio (S/N < 300). Thus, for the training set, we only consider stars that satisfy the following criteria:

(I) $0.5 < (BP - RP)_0 < 1.6$
(II) $E(B - V) < 2$
(III) $BP < 16$

However, for the testing set, we only constrain the data by $0.5 < (BP - RP)_0 < 1.6$ and $E(B - V) < 2$. We aim to see whether our classifiers that are trained on bright stars (BP <16) can be utilized to identify the faint metal-poor stars (BP >16). Thus, we include stars that satisfy the following criteria in the testing set:

(I) $0.5 < (BP - RP)_0 < 1.6$
(II) $E(B - V) < 2$

After applying cuts, we get 2.5×10^6 LAMOST stars and 4.5×10^5 APOGEE stars with XP spectra available, of which 4088 and 1295, respectively, are metal-poor stars with [Fe/H] < -2. In total, we utilize 2.9×10^6 spectra for training and testing, of which 0.2 per cent are metal-poor stars. We select 4×10^5 of them as testing set and 2.5×10^6 of them as training set.

Fig. 1 shows the colour–magnitude diagram of our training and testing sample. The horizontal axis is intrinsic colour $(BP - RP)_0$, the vertical axis is the absolute G magnitude (without any parallax cut here). The left and middle panels show the stars from the LAMOST and APOGEE surveys in the training and testing sets, which comprises main-sequence turn-off, dwarf, and giant stars. The right panel shows the distribution of metal-poor stars in the training and testing set. The majority of metal-poor stars are turn-off and giant stars. Fig. 1 suggests that our algorithm should only confidently identify metal-poor giants and turn-off stars, because metal-poor stars in other evolutionary stages would be extrapolation. Note that it is harder to find very metal-poor turn-off stars than giants. Because the resolution of XP spectra are low, the information we can get from them are close to what we can get from narrow band photometric surveys, but the photometric features of turn-off stars are less metallicity dependent, because they are hotter and absorption features are suppressed. Consequently, we utilize different models to find metal-poor turn-off and giant stars. We divide the training and testing sample into two parts, according to $(BP - RP)_0 < 0.8$ or > 0.8. The models trained on the former data set are responsible for finding turn-off metal-poor stars, and the other models trained on the latter data set are in charge of the giant metal-poor stars. As shown in the right panel of Fig. 1, our data set does not have many metal-poor dwarf stars, so we do not expect to find low-metallicity dwarf stars in this work.

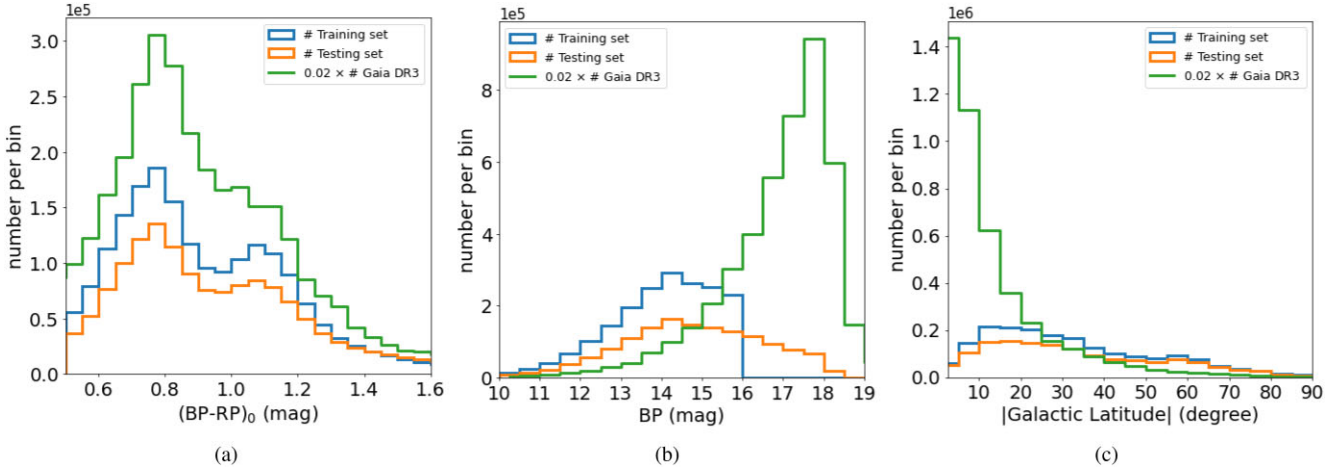

Figure 2. $(BP - RP)_0$, BP, and absolute Galactic latitude distribution of training, testing, and *Gaia* DR3 with $(BP - RP)_0 > 0.5$ and $E(B - V) < 2$ in this plot, which have XP spectra. We only randomly select 2 per cent of the *Gaia* DR3 data with XP spectra to display in this figure.

The $(BP - RP)_0$, BP, and $|b|$ distribution of *Gaia* DR3 data, training, and testing set are shown in Fig. 2. Note that, we only include *Gaia* DR3 data with $(BP - RP)_0 > 0.5$ and $E(B - V) < 2$ that have XP spectra in this plot. We see that the distributions of the *Gaia* data included in this plot are pretty different from our training and testing set, especially for the $(BP - RP)_0$ and Galactic latitude b distributions, which reminds us that the metal-poor candidates we find may only be a small fraction of the total.

3 MODEL TRAINING AND VALIDATION

We choose the XGBOOST algorithm to find metal-poor stars because it is a powerful and flexible algorithm that has been utilized in variety of sub-fields of astrophysics (e.g. Li et al. 2021; He, Luo & Chen 2022; Lucey et al. 2023; Pham & Kaltenegger 2022; Rix et al. 2022). The algorithmic principles for XGBOOST are not complex. In short, XGBOOST repeatedly builds decision trees to fit the residuals from the previous tree, until the residuals stop shrinking or it reaches the maximum number of trees, which is a free parameter. Then it sums the results from each tree, which are weighted by a learning rate (η), and plug this value into the Sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$, to calculate the probability of the input belonging to a certain category. For a detailed description of XGBOOST, see Chen & Guestrin (2016).

In this work, we utilize the coefficients of normalized and dereddened XP spectra together with their corresponding [Fe/H] from LAMOST or APOGEE to compose training and testing sets to train the XGBOOST model to identity metal-poor stars in *Gaia* DR3. We describe the training process and the performance of the well-trained models in this section.

3.1 Training process

In this work, we choose multiclassification algorithm to identify the metal-poor stars. The metallicity ([Fe/H]) of the training and testing samples ranges from -2.5 to $+1.0$. We utilize XGBOOST models to classify the stars into four metallicity intervals: [Fe/H]<-2.0, $-2.0 <$[Fe/H]< -1.5, $-1.5 <$[Fe/H]< -1.0, and $-1.0 <$[Fe/H]$< +1.0$, with probabilities P_0, P_1, P_2, P_3, respectively. For a star, when its P_0 is larger than the other probabilities, it will be classified as metal-poor star. The prediction uncertainty can be calculated from the probabilities of the multiclassification result, see Appendix A for more details. We choose the XGBOOST classification algorithm, rather than the regression algorithm, for following four reasons. (i) The minimum [Fe/H] of the training and testing set is -2.5, because of LAMOST and APOGEE analyses limitations, even though we do know there exist metal-poor stars with [Fe/H]< -2.5 in the data set. (ii) Regression would waste a lot of computational power on deciding the specific metallicity value for non-metal-poor stars ([Fe/H] > -2.0) which we do not care about. (iii) Unlike a regression algorithm, classification algorithm can more easily trade off completeness against purity. For samples that are difficult to identify, for example, turn-off stars and faint stars, we can sacrifice completeness for higher purity.

We utilize completeness and purity calculated on the test set to evaluate the performance of the models. Completeness refers to how completely our model can find all of the metal-poor stars. Purity refers to the fraction of true metal-poor stars for the set predicted to be metal-poor by our models. Completeness and purity are defined as:

$$\text{Completeness} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (1)$$

$$\text{Purity} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \quad (2)$$

Positive and negative samples here refer to the metal-poor ([Fe/H] <-2) and non-metal-poor ([Fe/H] >-2) stars, respectively. We divide the input samples into two training sets, according to their intrinsic colour: $0.5 < (BP - RP)_0 < 0.8$ and $0.8 < (BP - RP)_0$, as shown in Fig. 1, to find metal-poor turn-off and giant stars, respectively. Metal-poor giant stars make up 0.26 per cent of the training set with $0.8 < (BP - RP)_0$. However, metal-poor turn-off stars are much rarer, only make up 0.06 per cent of the training set with $0.5 < (BP - RP)_0 < 0.8$. Thus, it could be expected that metal-poor turn-off stars will be more difficult to find than metal-poor giant stars.

In preliminary tests, we found that the extreme imbalance between positive ([Fe/H] <-2) and negative ([Fe/H] >-2) samples badly hinders our training process. To solve this problem, we processed the training sets in the following two steps:

Step I: Utilize random undersampling to randomly remove over-represented metal-rich stars in the training set. The negative ([Fe/H] >-2) to positive ([Fe/H] <-2) ratio of the training set after

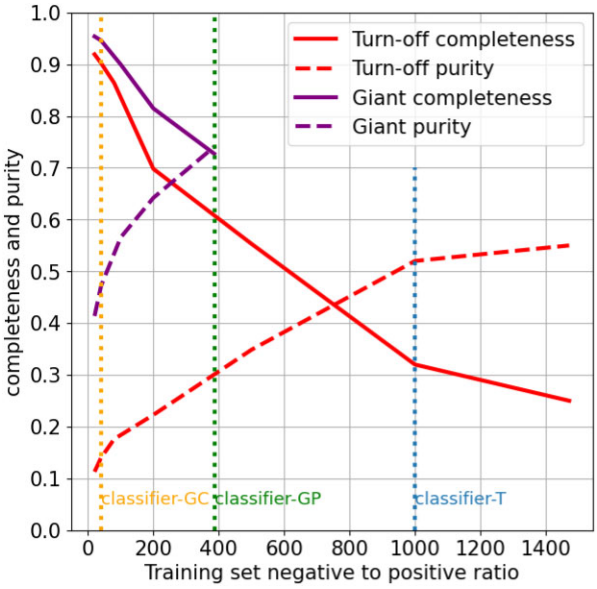

Figure 3. The completeness and purity of classifiers as a function of training NPR. The horizontal-axis is the negative to positive ratio of the training sample; the vertical-axis is completeness and purity of models. At each NPR, the classifiers were ran with the optimized set of hyper-parameters. The vertical lines with different colours refer to the NPR were chosen for Classifier-GC, Classifier-GP, and Classifier-T. The corresponding completeness and purity can be read from the vertical lines. The purple curves refer to the classifiers trained to find metal-poor Giants stars and the red curves refer to the classifiers aimed to find metal-poor turn-off stars.

undersampling is defined as NPR. We will change the NPR of the training set from 1 to the maximum value that the training set allowed.

Step II: Adopt oversampling algorithm Synthetic Minority Over-sampling Technique (SMOTE) to populate the metal-poor stars in the training set that has been under sampled. The SMOTE algorithm is an over-sampling method which synthesizes new examples from the minority class by selecting neighbouring examples in the feature space and then synthesizing a new sample at the point along the line connecting these two samples (Chawla et al. 2002).

We utilize RANDOMSEARCHCV from SCIKIT-LEARN (Pedregosa et al. 2011) to tune the XGBOOST hyperparameters. When training XGBOOST, a lot of hyperparameters can be adjusted, such as the learning rate (η), the maximum depth of a tree, and the minimum loss reduction required to make a further partition on a leaf node of the tree (γ). In order to find the optimal set of parameters, we utilize RANDOMSEARCHCV from SCIKIT-LEARN (Pedregosa et al. 2011). RANDOMSEARCHCV will go through points that are randomly selected from the predefined box in hyperparameter space, as shown below, to find the optimal set of parameters.

(i) n_estimators: from 100 to 1200 in steps of 50
(ii) max_depth: from 2 to 15 in steps of 1
(iii) learning_rate: from 0.05 to 1 in steps of 0.05
(iv) subsample: from 0.5 to 1 in steps of 0.05
(v) colsample_bytree: from 0.3 to 0.9 in steps of 0.05
(vi) min_child_weight: from 1 to 20 in steps of 1
(vii) gamma: from 0 to 0.7 in steps of 0.02

In this work, finding metal-poor stars trades off purity for completeness. For each NPR, we utilize RANDOMSEARCHCV to find the optimal set of parameters. Fig. 3 shows the completeness and purity of the well optimized model as a function of the training set NPR. The purple curves refer to the classifiers that are trained to find metal-poor giant stars, and the red curves refer to the classifier to find metal-poor turn-off stars. From Fig. 3 we see that increasing the NPR of the training set will increase the purity but decrease the completeness of the classifiers, and it is much easier to find metal-poor giant stars than metal-poor turn-off stars, just as we discussed before. The three vertical lines indicate the NPR that are chosen for Classifier-GP (Green, 386), Classifier-GC (Yellow, 40), and Classifier-T (Blue, 1000). Classifier-GC (Giant Complete) here denotes the model utilized to find metal-poor giants with high completeness, Classifier-GP (Giant Pure) denotes the model utilized to find metal-poor giants with high purity, and Classifier-T (Turn-off) denotes the model utilized to find turn-off metal-poor stars. The (completeness, purity) for our Classifier-T, Classifier-GC, Classifier-GP are (40.0 per cent, 47.2 per cent), (94.6 per cent, 47.2 per cent), (72.7 per cent, 74.1 per cent), respectively, which are derived by 3-fold cross-validation.

3.2 Models evaluation

After the training process, we utilize the testing sets to evaluate the performance of the classifiers on different [Fe/H], BP, $(BP - RP)_0$, and absolute Galactic latitude $|b|$. Typically, there are three factors that effect the performance of the classifiers: stellar species (turn-off or giants stars), brightness, and reddening. In this work, we utilize intrinsic colour $(BP - RP)_0$ to denote the type of stars, because we do not have metal-poor dwarf stars in the training and testing sets, as shown in Fig. 1. BP magnitude denotes the brightness of the stars. Additionally, the absolute $|b|$ can be utilized as an indicator of reddening, because stars in low $|b|$ regions, such as disc and bulge, often have severe extinction.

The metallicity distribution for stars in the testing set classified as metal-poor by different classifiers is shown in Fig. 4. The metallicity distribution for True Positive (TP), False Positive (FP), and False Negative (FN) samples in the testing set are shown in left and right panels, respectively. Comparing the distributions of Classifier-GC and Classifier-GP in the left panel, we see that the Classifier-GP can effectively remove the FP stars, although it loses some TP stars. On the other hand, the right panel shows that Classifier-GP loses some metal-poor stars with [Fe/H]< -2.8, which is the cost of high purity. This is why we provide Classifier-GC as supplement to Classifier-GP. Classifier-GC provides a high completeness data set and Classifier-GP provide a high purity data set. The good news for Classifier-GC is that most of the misclassified metal-poor still have rather low metallicity close to the [Fe/H] $= -2$ boundary.

The completeness and purity distributions of the classifiers on different $(BP - RP)_0$, BP, and $|b|$ intervals are shown in Fig. 5. We utilize different colours and symbols to denote different models, and dashed and solid lines to denote faint or bright stars. Let's discuss the performance of the classifiers on bright stars ($BP < 16$) first. Panel (a) and (d) show the performance of the classifiers as a function of $(BP - RP)_0$. We see that Classifier-T has a comparable purity at the blue end of the classifiers-GP and classifiers-GC, but its completeness is lower than these two models, because it is harder to find metal-poor turn-off stars, we have to sacrifice the completeness for high purity, just as we discussed before. Panels (b) and (e) show the performance of classifiers as a function of brightness. We can see that bright stars tend to have higher purity and completeness than faint stars, because bright stars typically have higher signal-to-noise ratio. Panels (c) and (f) show the performance as a function of $|b|$. The completeness and purity of our classifiers are lower in low-latitude region, because in this region extinction makes classification more difficult even with

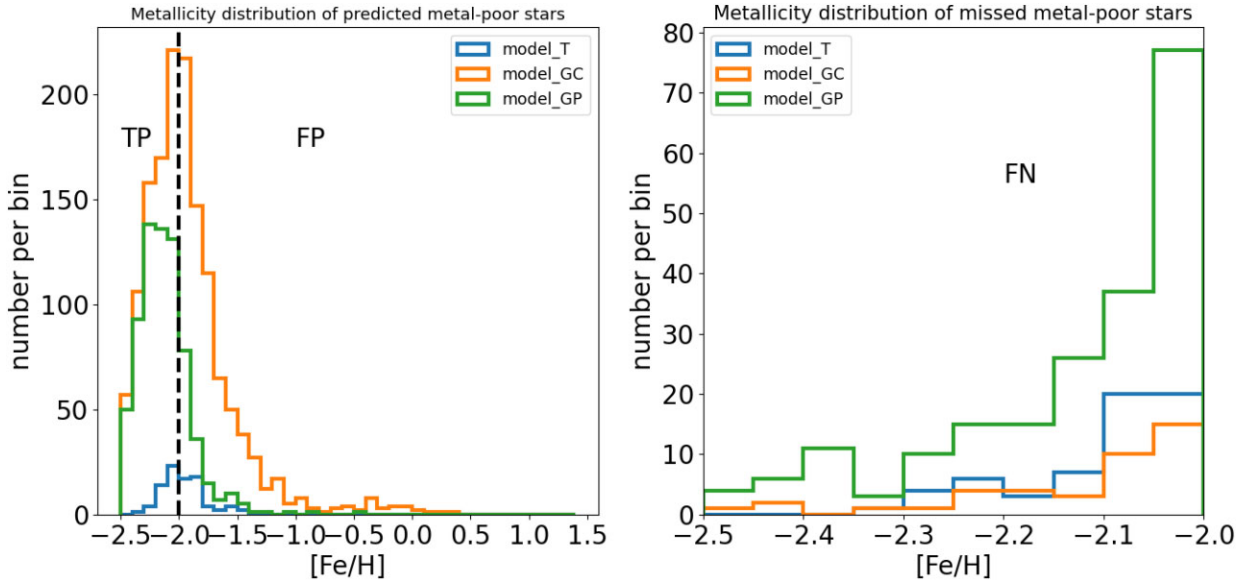

Figure 4. Left panel show the metallicity distribution of stars that are predicted to be metal-poor. The dashed line is the boundary of true positive samples and false positive samples. Right panel shows the metallicity distribution of false-negative stars (i.e. metal-poor stars missed by XGBOOST).

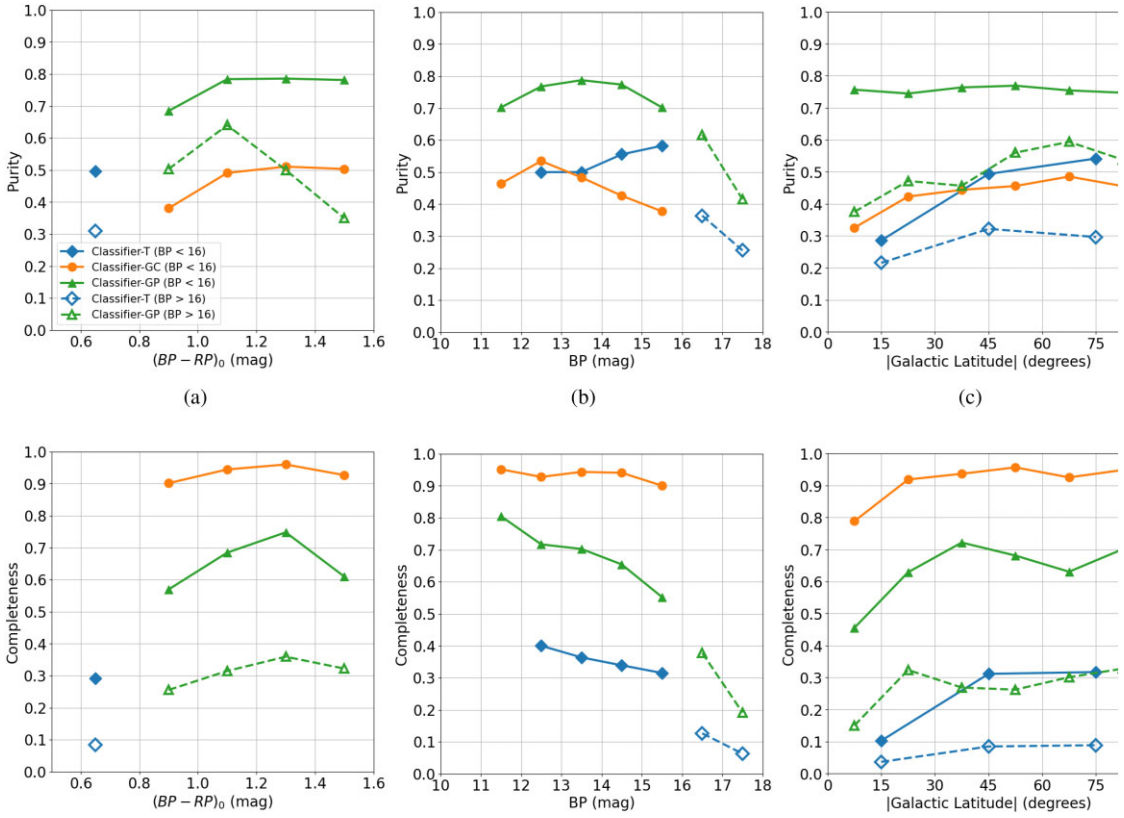

Figure 5. The completeness and purity of different classifiers as a function of intrinsic colour $(BP - RP)_0$, *BP* band magnitude *BP*, and absolute Galactic latitude $|b|$.

our coefficients extinction calibrations and higher contamination rate of metal-rich ([Fe/H] > -2) stars decrease the purity statistically. Note that, because there are few metal-poor turn-off stars at low or high Galactic latitude in our training and testing sets, we increased the bin size for turn-off stars in these two panels to avoid statistical fluctuations.

Most of the stars with XP spectra released by *Gaia* DR3 are faint ($BP > 16$), so it is worthwhile to evaluate the performance of the classifiers, which are trained on bright stars, on the faint stars. We utilize Classifier-T and Classifier-GP to make the prediction on faint stars. As shown in the dashed lines and open symbols of Fig. 5, the overall purity for Classifier-T is 29 per cent, for Classifier-GP is

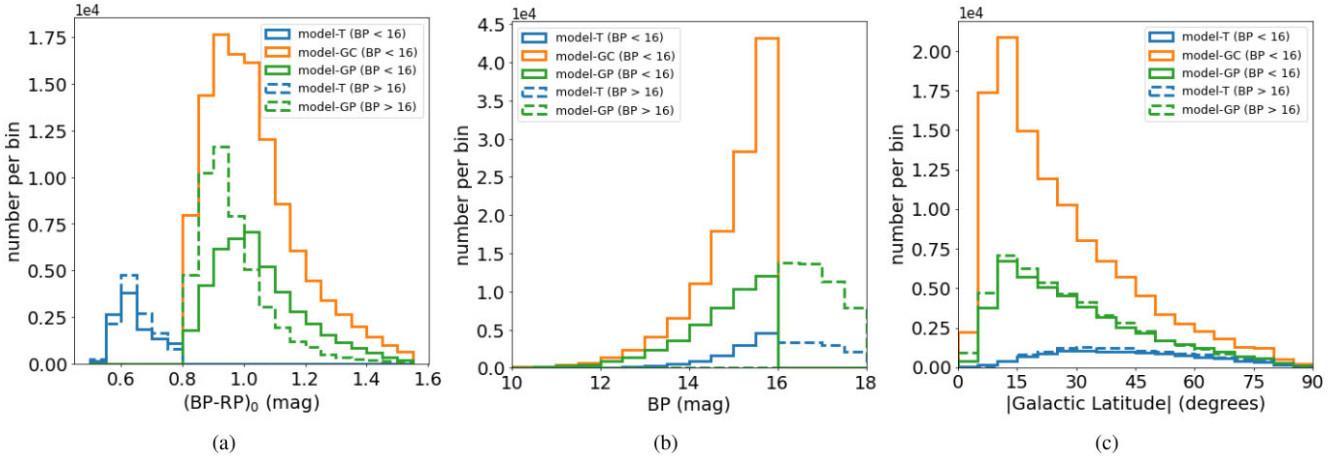

Figure 6. $(BP - RP)_0$, BP, and $|b|$ distribution of the metal-poor candidates we found in *Gaia* DR3 by different classifiers. Note that the $|b|$ for giants is skewed to very low $|b|$ is because those are mostly towards the inner Galaxy (bulge/inner halo), as seen in Fig. 8.

52 per cent. This purity is better than we expected, so we include the faint stars in our catalogue. However, as shown in panels (d), (e), and (f), the completeness for faint turn-off candidates is pretty low, less than 10 per cent, which means that the faint metal-poor turn-off stars we have in our final catalogues only make up a very small fraction of the total. Because of the low *S/N* ratio for faint stars, it is harder for us to find the genuine metal-poor ones. Thus, under this circumstance, purity has a higher priority than completeness. We can make a Shannon-Entropy cut on the final results to increase their purity. More details about the Shannon-Entropy cut are shown in Appendix A.

4 RESULTS

We have three reliable classifiers, Classifier-T, Classifier-GC, and Classifier-GP. We now classify the 200 million XP spectra released in *Gaia* DR3, and obtain three corresponding candidate metal-poor star catalogues, as shown in Tables 1, 2, and 3, which in total contain 200 000 metal-poor candidates. The distributions of the metal-poor candidates are shown in Fig. 6.

The colour–magnitude diagram for these candidate metal-poor stars, without any parallax quality cut, is shown in Fig. 7. The left/middle/right panel shows the colour-magnitude diagram for the candidate metal-poor stars identified by Classifier-T/Classifier-GC/Classifier-GP. From these panels we confirm that, in our catalogues, the candidate metal-poor stars are dominated by turn-off stars and giant stars. However, there are a small number of dwarf stars present in the cooler regions of the main sequence, as shown in the middle and right panels ($M_G > 4$, below the red dashed line). These red dwarf stars may be wrongly classified as metal-poor stars, because there are almost no red dwarf stars in the training sets for Classifier-GP and Classifier-GC. Table 4 shows that red dwarfs only make up a very small fraction of the metal-poor stars found by Classifier-GC and Classifier-GP, i.e. 1.7 per cent for Classifier-GC, 0.7 per cent for Classifier-GP ($BP < 16$) and 6.5 per cent for Classifier-GP ($BP > 16$). Since the risk of contamination is higher, we include the absolute G band magnitude M_G in our final catalogues if users would like to filter out any potential dwarf contamination.

The distance distributions and Galactic coordinate projections of the candidates are shown in Figs 8 and 9. Fig. 8 shows the distance distributions of the candidate metal-poor stars. The distances are calculated by inverting the *Gaia* DR3 parallax. The distance to the Galactic centre is marked by the red dashed line (∼8 kpc from the Sun Bland-Hawthorn & Gerhard 2016). The blue lines are the distribution of candidate turn-off metal-poor stars, and the orange and green lines are the distribution of candidate giant metal-poor stars. For the distance distribution, comparing to candidate metal-poor giant stars, the turn-off stars are located closer to the Sun, as expected given their lower luminosities. The giants are distributed around the Galactic centre. This result indicates that the Galactic centre contain a large amount of metal-poor stars, i.e. the Milky Way hosts an ancient, metal-poor, and centrally concentrated stellar population (e.g. Rix et al. 2022). Fig. 9 shows the skymap of the candidate metal-poor stars we found in *Gaia* DR3. Because the dereddening process excludes almost all of the high $E(B - V)$ stars ($E(B - V) > 2$), we do not obtain a lot of stars at low Galactic latitude, as shown in Fig. 9. Bulge stars and halo stars are the dominant stars for our sample.

The bright spots in the Galactic coordinate projections are globular clusters (Harris 2010). After testing, we found that comparing with Classifier-GC that includes many globular clusters with $-1.5 <$[Fe/H]< -1.0, Classifier-GP excludes all of the globular clusters with average metallicity larger than -1.5 and most of the globular clusters with average metallicity within -2 to -1.5, but keeps all of the globular clusters with metallicity less than -2, which is a demonstration that Classifier-GP has relatively higher purity than Classifier-GC. Note that the Galactic coordinate projections are also affected by the *Gaia* scanning law (see De Angeli et al. 2023) and crowding issues for XP spectra in globular clusters.

We created Table 5 to concisely summarize the main points from Section 3 and 4.

5 DISCUSSION

In this work, according to Table 4, we add up the numbers of metal-poor candidates found by Classifier-T, Classifier-GC ($BP < 16$ at all M_G), Classifier-GP ($BP > 16$ at all M_G) and obtained a total of 200 000 candidate metal-poor stars. Weighting each subsample by its purity in Table 4, we expect the catalogue contains 88 000 genuine metal-poor stars (overall purity of 44 per cent).

Though we only classify stars with [Fe/H] < -2, we can estimate how many stars are <-3 or <-4. We assume the slope of the metallicity distribution (Youakim et al. 2020; Chiti et al. 2021b), although there are also much more pessimistic slopes of the metal-

Table 1. Metal-poor turn-off candidates found by Classifier-T. P_0, P_1, P_2, P_3 refer to the probability of a stars with $-2.5 <$ [Fe/H] < -2, $-2 <$ [Fe/H] < -1.5, $-1.5 <$ [Fe/H] < -1, $-1 <$ [Fe/H] $< +1$. (This table is available in its entirety in the online supplementary material).

Candidates found by Classifier-T									
Gaia DR3 source id	$(BP-RP)_0$ (mag)	M_G (mag)	$E(B-V)$ (mag)	BP	P_0	P_1	P_2	P_3	Shannon entropy
6650038640545499264	0.59	3.66	0.07	15.65	0.97	0.03	0.0	0.0	0.17
6650111586271008256	0.58	2.96	0.06	16.34	0.84	0.15	0.0	0.0	0.68
6650144949575814016	0.61	− 0.24	0.07	17.65	0.83	0.15	0.0	0.02	0.77
6650193499886281088	0.62	3.76	0.07	16.65	0.84	0.14	0.01	0.0	0.7
6650230470965151360	0.78	3.14	0.08	16.9	0.85	0.1	0.0	0.05	0.76

Table 2. Metal-poor giant candidates found by Classifier-GC. (This table is available in its entirety in the online supplementary material).

Candidates found by Classifier-GC									
Gaia DR3 source id	$(BP-RP)_0$ (mag)	M_G (mag)	$E(B-V)$ (mag)	BP	P_0	P_1	P_2	P_3	Shannon entropy
4252405961205838208	0.81	− 1.55	0.68	15.77	0.6	0.01	0.01	0.38	1.12
4252433105401980800	1.5	− 8.02	0.61	15.65	0.39	0.21	0.39	0.01	1.57
4252454580242134912	0.84	− 2.07	0.78	15.1	0.93	0.0	0.02	0.04	0.42
6032351905927100928	0.98	− 0.61	0.42	15.69	0.56	0.02	0.09	0.34	1.4
6032356578851595392	1.13	nan	0.48	15.82	0.95	0.05	0.0	0.0	0.3

Table 3. Metal-poor giant candidates found by Classifier-GP. (This table is available in its entirety in the online supplementary material).

Candidates found by Classifier-GP									
Gaia DR3 source id	$(BP-RP)_0$ (mag)	M_G (mag)	$E(B-V)$ (mag)	BP	P_0	P_1	P_2	P_3	Shannon entropy
6032364236763645056	1.14	nan	0.51	18.31	0.45	0.28	0.02	0.25	1.63
6032371177430994048	1.24	− 0.40	0.44	16.68	0.4	0.15	0.39	0.07	1.73
6032371349229711616	0.96	0.40	0.45	17.26	0.42	0.01	0.16	0.42	1.55
6032372964137450240	1.19	1.64	0.47	17.9	0.51	0.32	0.12	0.04	1.59
6032408874375184896	1.06	nan	0.58	15.92	0.5	0.28	0.16	0.05	1.65

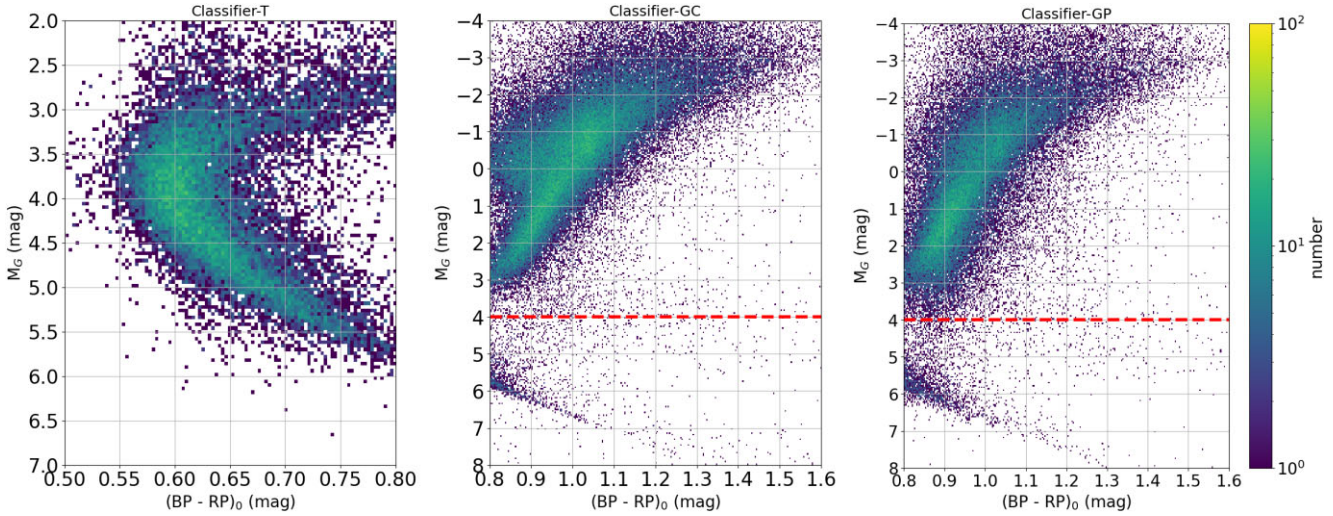

Figure 7. Colour–magnitude diagram of metal-poor stars we found in *Gaia* DR3 by different classifiers. The horizontal axis is dereddened colour $(BP-RP)_0$ and the vertical axis is the absolute G band magnitude (for stars without any parallax-quality cut).

Table 4. The number, purity, and completeness of metal-poor candidates we found by Classifier-T, Classifier-GC, and Classifier-GP in different M_G and BP ranges.

Classifier and brightness	Number of stars		Purity		Completeness	
Classifier-T (BP < 16)	10995		52 per cent		32 per cent	
Classifier-T (BP > 16)	37763		29 per cent		8 per cent	
$M_G < 4$ or $M_G > 4$	$M_G > 4$	$M_G < 4$	$M_G > 4$	$M_G < 4$	$M_G > 4$	$M_G < 4$
Classifier-GC (BP < 16)	1954	109493	27 per cent	45 per cent	56 per cent	94 per cent
Classifier-GP (BP < 16)	291	43514	50 per cent	76 per cent	10 per cent	66 per cent
Classifier-GP (BP > 16)	2542	38780	30 per cent	54 per cent	9 per cent	30 per cent

Table 5. A summary table for Sections 3 and 4.

Summary of the models					
Model name	Classifier-T	Classifier-GC	Classifier-GP	Classifier-T	Classifier-GP
$(BP - RP)_0$	< 0.8	> 0.8	> 0.8	< 0.8	> 0.8
BP	< 16	< 16	< 16	> 16	> 16
Shannon entropy cutoff	nan	nan	nan	< 0.8	nan
Percentage of MP-stars	0.06 per cent	0.28 per cent	0.28 per cent	0.06 per cent	0.28 per cent
NPR of training set	1000	40	386	1000	386
Test purity	47 per cent	47 per cent	74 per cent	40 per cent	53 per cent
Test completeness	40 per cent	94 per cent	65 per cent	8 per cent	7 per cent
Expected total purity	52 per cent	45 per cent	76 per cent	29 per cent	52 per cent
Expected total completeness	32 per cent	93 per cent	66 per cent	8 per cent	28 per cent
# Candidates	10 995	111 447	43 805	37 763	41 322

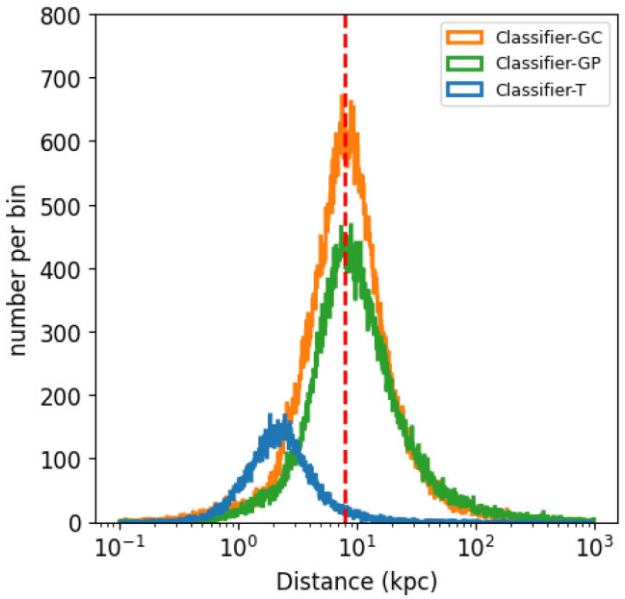

Figure 8. The distance distribution of the candidate metal-poor stars we found in *Gaia* DR3. The red dashed line in the left panel refers to the Galactic centre. Lines with different colour refer to the candidate metal-poor stars identified by different classifiers.

poor tail of the metallicity distribution function by Bonifacio et al. (2021):

$$\log \frac{\mathrm{d}N}{\mathrm{d[Fe/H]}} = \gamma \tag{3}$$

γ is 1 when −2.5 <[Fe/H]< −2.0 and 1.5 in −4 <[Fe/H]< −2.0. Based on this assumption, we can estimate the number of the actual metal-poor stars for these 188 000 candidate in each metallicity intervals. From −4 to −3.5, −3.5 to −3, −3 to −2.5, −2.5 to −2, the estimated number of actual metal-poor stars are 600, 2800, 17 000, and 64 000, respectively.

5.1 Comparing with other surveys

Table 6 shows our results compared to previous photometric selections. Huang et al. (2022) utilized SMSS DR2 and *Gaia* EDR3 photometry to estimate metallicity for 24 million stars. They obtained half a million very metal-poor ([Fe/H] < −2.0) stars, and over 25 000 extremely metal-poor ([Fe/H] < −3.0) stars. 48 270 very metal-poor candidates in Huang et al. (2022) are also predicted to be very metal-poor by our Classifiers. Chiti et al. (2021a) utilized SMSS DR2 photometry to derive photometric metallicities. They present more reliable metallicities of ∼280 000 stars with −3.75 ≤[Fe/H]≤−0.75 down to $g = 17$. 18 640 of them are candidate metal-poor stars ([Fe/H] < −2). After the validation by our training and testing set, we found their purity to be 49 per cent; and there are 9218 stars also predicted to be metal-poor by our Classifiers. Pristine survey does not publicly release their data, but according to Starkenburg et al. (2017) and Youakim et al. (2020), Pristine has covered a sky area of ∼2500 deg^2, at the time of those papers. In each ∼deg^2 field, they find ∼7 stars that have [Fe/H] < −2.5 down to magnitude of $V = 18$. The purity of Pristine to find stars with [Fe/H] < −2.5 is 49 per cent (Aguado et al. 2019). The Best & Brightest initiative selected over 11 000 candidate VMP ([Fe/H] < −2) and EMP stars ([Fe/H] < −3), with an overall purity of 30 per cent and 5 per cent, respectively (Schlaufman & Casey 2014; Placco et al. 2019; Limberg et al. 2021b). Comparing with other surveys, our work increases the number of candidate metal-poor stars by about an order of magnitude, but with similar or higher purity. The comparison results are shown in Table 6. Recently, Andrae, Rix & Chandra (2023b) utilized XGBoost and XP spectra, together with 38 narrowband colours derived from XP spectra and broad-band surveys (*Gaia*: G, BP, RP and CatWISE: W_1, W_2), to

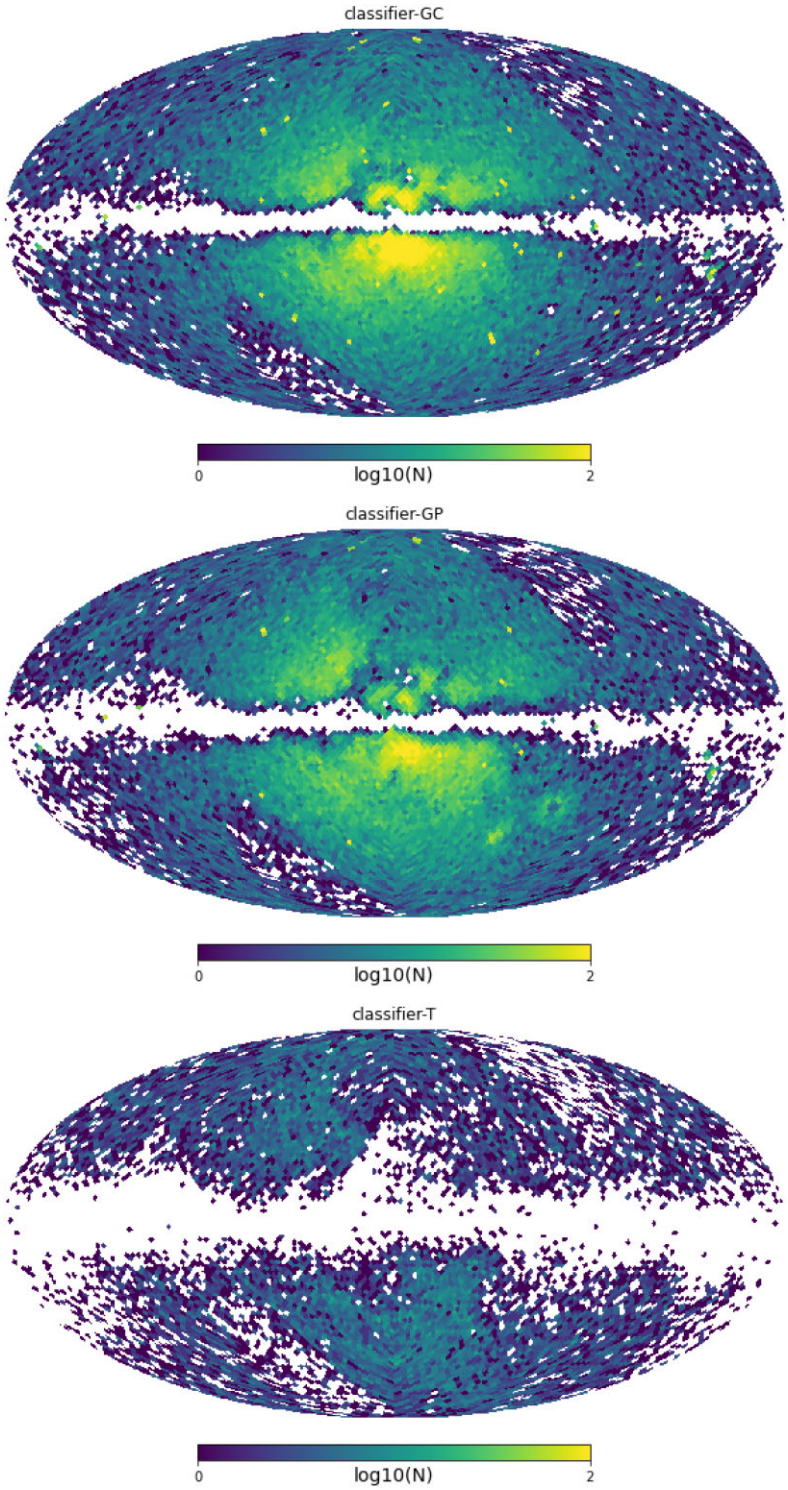

Figure 9. The Galactic coordinate projections of the candidate metal-poor stars we found through Classifier-GC, Classifier-GP, Classifier-T. The area of healpix pixel is 3.36 deg^2.

derive metallicity, $T_{\rm eff}$ and log g for 175 million stars. They reduced the temperature-extinction degeneracy by introducing CatWISE W_1 and W_2, which extend to the infrared regions, into the model. The metallicity were derived using the XGBoost regression model and the true labels came from APOGEE, and augmented by a set of very metal-poor stars (Li et al. 2022). Because we both utilize XGBoost algorithm and deal with the same data set, it is worth comparing our results with them. The comparisons are shown in Table 7. Tables 1 and 2 are two tables published by Andrae, Rix & Chandra (2023b). In short, table 2 is a high accuracy subset of bright ($BP < 16$) giant stars of table 1. Table 7 shows that, for giant candidates, Classifier-GP has higher purity and more candidates comparing with tables 1 and 2 (only including giant candidates with $BP < 16$). The purity for turn-off stars of table 1 is only 20 per cent, while our Classifier-T has

Table 6. Comparison to other photometric surveys. The purity mentioned above are obtained from the comparison with LAMOST DR7 and APOGEE DR17. Except for Pristine, for which it is from Aguado et al. (2019), and not for [Fe/H]< -2 but -2.5. Note that, for Classifier-T or Classifier-GP, 29 per cent and 52 per cent are the purity for faint ($BP > 16$) stars, 52 per cent and 76 per cent are the purity for bright ($BP < 16$) stars.

Comparison to photometric selections		
Photometic surveys	# [Fe/H] < -2	Purity
SMSS (Turn-off) (Huang et al. 2022)	548 518	10 per cent
SMSS (Giant) (Huang et al. 2022)	192 487	42 per cent
SMSS (Turn-off) (Chiti et al. 2021a)	522	46 per cent
SMSS (Giant) (Chiti et al. 2021a)	18 046	49 per cent
Best & Brightest (Placco et al. 2019; Limberg et al. 2021b)	11 000	30 per cent
Pristine (Starkenburg et al. 2017; Youakim et al. 2020)	18 000∗([Fe/H] < -2.5)	49 per cent
Classifier-T	48 758	29–52 per cent
Classifier-GC	111 447	45 per cent
Classifier-GP	85 127	52–76 per cent

Table 7. Comparison to Andrae, Rix & Chandra (2023b). The purity mentioned above are obtained from the comparison with LAMOST DR7 and APOGEE DR17.

Comparison to Andrae, Rix & Chandra (2023b)		
Table/Model	# [Fe/H] < -2	Purity
Andrae, Rix & Chandra (2023b) table 1 (Turn-off, $BP < 16$)	24 000	23 per cent
Classifier-T ($BP < 16$)	10 995	52 per cent
Andrae, Rix & Chandra (2023b) table 1 (Giants, $BP < 16$)	38 000	70 per cent
Classifier-GP ($BP < 16$)	43 805	76 per cent
Classifier-GC ($BP < 16$)	111 447	45 per cent
Andrae, Rix & Chandra (2023b) table 1 (Turn-off, $BP > 16$)	51 000	20 per cent
Classifier-T ($BP > 16$, Shannon entropy < 0.8)	37 763	29 per cent
Andrae, Rix & Chandra (2023b) table 1 (Giants, $BP > 16$)	35 000	53 per cent
Classifier-GP ($BP > 16$)	41 322	52 per cent
Andrae, Rix & Chandra (2023b) table 2	18 000	70 per cent

a higher purity of 29 per cent to 52 per cent. We suggest our models are better for finding metal-poor stars comparing with Andrae, Rix & Chandra (2023b) for the following reasons: (i) We have larger number of metal-poor stars, which provides the models a training set with greater diversity. (ii) Their model is a regression model which is trying to fit the metallicity for all stars, especially for metal-rich stars. As a result, their model may not do as well for metal-poor stars, which are only a very small part of the whole. In contrast, our models are more specialized, and only focus on finding metal-poor stars. (iii) Because we choose classification algorithm rather than regression, we can trade off completeness against purity. For stars that are difficult to classify, for example turn-off stars, we can sacrifice the completeness to the higher purity with NPR (see Fig. 3) and SMOTE. (iv) The *Gaia* XP spectra we utilized has been dereddened, which may make our predictions more accurate, even without *WISE* photometry. Out of 148 000 very metal-poor candidates in Andrae, Rix & Chandra (2023b), there are 65 949 stars are found to be very metal-poor with our Classifiers. Overall, we suggest that researchers and observers utilize this work together with Andrae, Rix & Chandra (2023b) to decide what metal-poor candidates to follow up.

Zhang, Green & Rix (2023) utilized a forward model to estimate stellar parameters ([Fe/H], $T_{\rm eff}$, and log$g$), revised distances and extinctions for 220 million stars with XP spectra. However, there is a trend that the metallicity derived by the forward model tend to be overestimated at very-metal-poor end, which is even more biased than the metallicity derived by Andrae, Rix & Chandra (2023b). We think this bias is caused by the imbalance of the numbers of the metal-poor and non-metal-poor stars in their training set.

Martin et al. (2023) used the spectroscopic and photometric information of 219 million stars from *Gaia* DR3 to calculate synthetic narrow-band *CaHK* magnitudes sensitive to metallicity. *CaHK* magnitudes mimic the observations of Pristine surveys. They derived the photometric metallicities for 30 million high signal-to-noise FGK stars. They identified 200 000 very metal-poor candidates and 8000 extremely metal-poor candidates ($[{\rm Fe/H}]_{\rm phot} < -2$ and $[{\rm Fe/H}]_{\rm phot} < -3$, respectively). Because their data were released while this paper was already in review, we do not consider their results for our comparisons.

5.2 Validation with existing high-resolution spectra

There are plenty of high-resolution follow-up observations to the candidate metal-poor stars that have been obtained by previous studies. We can utilize these confirmed metal-poor stars to evaluate the completeness of our XGBOOST models. The results are shown in Table 8. In this table, we utilize six metal-poor halo stars data sets, three metal-poor bulge data sets, one metal-poor disc star, and one carbon-enhanced metal-poor (CEMP; [C/Fe] $> +0.7$) data set to test our models. For each data set, we exclude stars of which dereddened colour $(BP - RP)_0 < 0.5$ and $E(B - V) > 2$. Then we divide each data set into turn-off metal-poor stars ($(BP - RP)_0 < 0.8$) and giant metal-poor stars ($(BP - RP)_0 > 0.8$). The total number of these stars are shown in third and fourth columns. Finally, we utilize the Classifier-T, Classifier-GC, and Classifier-GP to predict the metallicity of these turn-off and giant metal-poor stars, respectively, and get the corresponding completeness marked as completeness-T, completeness-GC, and completeness-GP. This table shows that the completeness from these data sets is very close to the results from our test set, especially for the halo stars. We also test our classifiers on CEMP stars as shown in the last row of Table 8. As might be expected, the completeness of the classifiers on CEMP stars is not as high as other metal-poor stars, potentially because the enhanced carbon makes the metal-poor star spectra look more metal-rich.

6 SUMMARY

Metal-poor stars ([Fe/H] < -2) record the chemical enrichment history, accretion events, and early stages of the Milky Way. However, they are rare and difficult to find. In this work, we train XGBOOST models to identify metal-poor stars in *Gaia* DR3. The input to the models are the coefficients of normalized and dereddened XP spectra. The classifiers split the stars into different [Fe/H] intervals of $-2.5 <$ [Fe/H] < -2, $-2 <$ [Fe/H] < -1.5, $-1.5 <$ [Fe/H] < -1, $-1 <$ [Fe/H] $< +1$. Because of the extreme imbalance between positive and negative samples, we randomly exclude some negative samples and utilize the SMOTE algorithm to oversample the training sets and, then, utilize them to train the models. Finally, we get three classifiers, Classifier-T, Classifier-GC, and Classifier-GP and utilize them to identify the metal-poor turn-off and giant stars in *Gaia* DR3 with XP spectra. We present the histogram of the testing

Table 8. Prediction results for metal-poor stars that are confirmed by high-resolution spectra.

Reference	Region or type	MP-Giants	MP-turn-off	Completeness-GC	Completeness-GP	Completeness-T
Abohalima & Frebel (2018)	Halo	266	115	98.9 per cent	82.3 per cent	54.8 per cent
Li et al. (2022)	Halo	152	79	92.8 per cent	71.1 per cent	55.7 per cent
Roederer et al. (2014)	Halo	68	111	100.0 per cent	77 per cent	56 per cent
Jacobson et al. (2015)	Halo	106	1	100.0 per cent	83 per cent	0 per cent
Cohen et al. (2013)	Halo	47	26	94 per cent	64 per cent	42 per cent
Cayrel et al. (2004)	Halo	25	0	100 per cent	48 per cent	NAN
Sestito et al. (2022)	Bulge	8	5	100.0 per cent	37.5 per cent	40.0 per cent
Howes et al. (2015)	Bulge	21	0	100.0 per cent	71 per cent	NAN
Howes et al. (2016)	Bulge	9	0	100.0 per cent	56 per cent	NAN
Schlaufman, Thompson & Casey (2018)	Disc	0	1	NAN	NAN	100 per cent
Yoon et al. (2016)	CEMP	106	34	84 per cent	55 per cent	41 per cent

result and the completeness/purity distributions for these models in Figs 4 and 5.

In total, we obtained 200 000 metal-poor candidates with overall purity 44 per cent. This number of metal-poor candidates is around an order of magnitude larger than previous work (e.g. Best & Brightest, SkyMapper, and Pristine), which has similar or even better purity.

We make the full catalogue available in the supplementary online material (Tables 1, 2, 3).

ACKNOWLEDGEMENTS

We thank Yang Huang, Xiaowei Ou, and Anirudh Chiti for helpful discussions. YY and APJ acknowledge support from the U.S. National Science Foundation (NSF) grant AST 2206264. GL acknowledges FAPESP (procs. 2021/10429-0 and 2022/07301-5). This research benefited from the 2022 *Gaia* DR3 Chicago Sprint hosted by the Kavli Institute for Cosmological Physics. We acknowledge the University of Chicago's Research Computing Center for their support of this work.

This work has made utilize of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa. int/gaia), the Large Sky Area Multi-Object Fiber Spectroscopic Telescope and Apache Point Observatory Galactic Evolution Experiment. Mission *Gaia* is 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. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences. Apache Point Observatory Galactic Evolution Experiment (APOGEE) is one of the programs in the Sloan Digital Sky Survey III. Funding for the creation and distribution of the SDSS Archive has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is http://www.sdss.org/. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Princeton University, the United States Naval Observatory, and the University of Washington.

This research makes utilize of public auxiliary data provided by ESA/Gaia/DPAC/CU5 and prepared by Carine Babusiaux. This paper made utilize of the Whole Sky Database (wsdb) created by Sergey Koposov and maintained at the Institute of Astronomy, Cambridge with financial support from the Science and Technology Facilities Council (STFC) and the European Research Council (ERC).

This research made utilize of ASTROPY[6] a community-developed core Python package for Astronomy (Robitaille et al. 2013). Other software utilized includes MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), SCIKIT-LEARN (Pedregosa et al. 2011), and SCIPY (Virtanen et al. 2020).

DATA AVAILABILITY

Our result, i.e. all contents in Tables 1, 2, 3, can be aquired at: https://doi.org/10.5281/zenodo.8360958.

REFERENCES

Abdurro'uf N. et al., 2022, ApJS, 259, 35
Abohalima A., Frebel A., 2018, ApJS, 238, 36
Aguado D. S. et al., 2019, MNRAS, 490, 2241
Andrae R. et al., 2023a, A&A, 674, A27
Andrae R., Rix H.-W., Chandra V., 2023b, ApJS, 267, 8
Aoki W. et al., 2012, AJ, 145, 13
Beers T. C., Christlieb N., 2005, Annu. Rev. Astron. Astrophys., 43, 531
Beers T. C., Preston G. W., Shectman S. A., 1985, AJ, 90, 2089
Beers T. C., Preston G. W., Shectman S. A., 1992, AJ, 103, 1987
Beers T. C. et al., 2017, ApJ, 835, 81
Belokurov V., Kravtsov A., 2022, MNRAS, 514, 689
Bidelman W. P., MacConnell D. J., 1973, AJ, 78, 687
Bland-Hawthorn J., Gerhard O., 2016, ARA&A, 54, 529
Bond H. E., 1970, ApJS, 22, 117
Bond H. E., 1980, ApJS, 44, 517
Bonifacio P. et al., 2021, A&A, 651, A79
Carrasco J. et al., 2021, A&A, 652, A86
Cayrel R. et al., 2004, A&A, 416, 1117
Cenarro A. J. et al., 2019, A&A, 622, A176
Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2002, J. Artif. Intell. Res., 16, 321
Chen T., Guestrin C., 2016, in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, p. 785

[6]http://www.astropy.org

Chiti A., Frebel A., Mardini M. K., Daniel T. W., Ou X., Uvarova A. V., 2021a, ApJS, 254, 31
Chiti A., Mardini M. K., Frebel A., Daniel T., 2021b, ApJ, 911, L23
Christlieb N., Schörck T., Frebel A., Beers T., Wisotzki L., Reimers D., 2008, A&A, 484, 721
Cohen J. G., Christlieb N., Thompson I., McWilliam A., Shectman S., Reimers D., Wisotzki L., Kirby E., 2013, ApJ, 778, 56
Conroy C. et al., 2022, preprint (arXiv:2204.02989)
Cui X.-Q. et al., 2012, Res. Astron. Astrophys., 12, 1197
Das P., Hawkins K., Jofré P., 2020, MNRAS, 493, 5195
De Angeli F. et al., 2023, A&A, 674, A2
Deng L.-C. et al., 2012, Res. Astron. Astrophys., 12, 735
Eisenstein D. J. et al., 2011, AJ, 142, 72
El-Badry K. et al., 2018, MNRAS, 480, 652
Frebel A., Norris J. E., 2015, ARA&A, 53, 631
Frebel A. et al., 2006, ApJ, 652, 1585
Gaia Collaboration, 2016, A&A, 595, A1
Gaia Collaboration, 2022, A&A, 657, A82
Galarza C. A. et al., 2022, A&A, 657, A35
Harris W. E., 2010, preprint (arXiv:1012.3224)
Harris C. R. et al., 2020, Nature, 585, 357
Hawkins K., Jofre P., Masseron T., Gilmore G., 2015, MNRAS, 453, 758
He X.-J., Luo A.-L., Chen Y.-Q., 2022, MNRAS, 512, 1710
Heger A., Woosley S. E., 2010, ApJ, 724, 341
Helmi A., 2020, ARA&A, 58, 205
Horta D. et al., 2021, MNRAS, 500, 1385
Howes L. M. et al., 2015, Nature, 527, 484
Howes L. M. et al., 2016, MNRAS, 460, 884
Huang Y. et al., 2022, ApJ, 925, 164
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ishigaki M. N. et al., 2021, MNRAS, 506, 5410
Jacobson H. R. et al., 2015, ApJ, 807, 171
Jones S., Côté B., Röpke F. K., Wanajo S., 2019, ApJ, 882, 170
Kunder A. et al., 2017, AJ, 153, 75
Li H., Tan K., Zhao G., 2018, ApJS, 238, 16
Li C. et al., 2021, MNRAS, 506, 1651
Li H. et al., 2022, ApJ, 931, 147
Limberg G. et al., 2021a, ApJ, 907, 10
Limberg G. et al., 2021b, ApJ, 913, 11
Limongi M., Chieffi A., 2012, ApJS, 199, 38
Lucey M. et al., 2023, MNRAS, 523, 4049
Majewski S. R. et al., 2017, AJ, 154, 94
Martin N. F. et al., 2023, preprint (arXiv:2308.01344)
Mendes de Oliveira C. et al., 2019, MNRAS, 489, 241
Onken C. A. et al., 2019, Publ. Astron. Soc. Aust., 36, e033
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pham D., Kaltenegger L., 2022, MNRAS, 513, L72
Placco V. M. et al., 2018, AJ, 155, 256
Placco V. M. et al., 2019, ApJ, 870, 122
Placco V. M. et al., 2021, ApJ, 912, L32
Placco V. M., Almeida-Fernandes F., Arentsen A., Lee Y. S., Schoenell W., Ribeiro T., Kanaan A., 2022, ApJS, 262, 8
Reggiani H., Schlaufman K. C., Casey A. R., Ji A. P., 2020, AJ, 160, 173
Rix H.-W. et al., 2022, ApJ, 941, 45
Robitaille T. P. et al., 2013, A&A, 558, A33
Roederer I. U., Preston G. W., Thompson I. B., Shectman S. A., Sneden C., Burley G. S., Kelson D. D., 2014, AJ, 147, 136
Schlaufman K. C., Casey A. R., 2014, ApJ, 797, 13
Schlaufman K. C., Thompson I. B., Casey A. R., 2018, ApJ, 867, 98
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Sestito F. et al., 2022, MNRAS, 518, 4557
Starkenburg E., Oman K. A., Navarro J. F., Crain R. A., Fattahi A., Frenk C. S., Sawala T., Schaye J., 2017, MNRAS, 465, 2212
Starkenburg E. et al., 2017, MNRAS, 471, 2587
Vallenari A. et al., 2023, A&A, 674, A1
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Wanajo S., 2018, ApJ, 868, 65
Xylakis-Dornbusch T., et al. 2022, A&A, 666, A58
Yanny B. et al., 2009, AJ, 137, 4377
Yoon J. et al., 2016, ApJ, 833, 20
Youakim K. et al., 2020, MNRAS, 492, 4986
Zhang X., Green G. M., Rix H.-W., 2023, MNRAS, 524, 1855

APPENDIX A: PREDICTION UNCERTAINTY

Shannon entropy is an indicator of the prediction uncertainty, which is defined as:

$$\text{Shannon entropy} = -\sum_{i=0}^{3} P_i \log_2(P_i) \tag{A1}$$

Shannon entropy is an indicator of prediction uncertainty, which can be utilized to filter the metal-poor candidates with high prediction uncertainty and increase the purity of catalogues. According to the definition A1, Shannon entropy increases as the probabilities become evenly distributed and decrease as they become skewed distributed. Thus, higher Shannon entropy typically indicates greater prediction uncertainty. In this project, since we utilize multiclassification algorithm, each star in our catalogues is assigned four probabilities P_0, P_1, P_2, and P_3 (summing to 1) that correspond to the probabilities of the star belonging to four metallicity intervals: [Fe/H] < -2, $-2 <$ [Fe/H] < -1.5, $-1.5 <$ [Fe/H] < -1, $-1 <$ [Fe/H] < -1. By comparing Figs A1 and A2, we see that most of our candidates have Shannon entropy smaller than 1.5, which indicates that most of the candidates have P_0 greater than 0.5, in other words, most of the candidates have low prediction uncertainty. Even so, we can still increase the purity of our catalogues by excluding the stars with high Shannon entropy (high prediction uncertainty). For example, as shown in Fig. A3, by excluding the candidates with Shannon entropy > 0.8, we can get a faint ($BP > 16$) metal-poor turn-off star catalogue with purity > 40 per cent.

Additionally, our catalogues are also useful for the science goals requiring stars with [Fe/H] < -1.5 or [Fe/H] < -1.0. As shown in Table A1, our Classifiers can also accurately and completely identify stars with [Fe/H] < -1.5 or [Fe/H] < -1.0. Comparing with finding stars with [Fe/H] < -2.0, finding stars with [Fe/H] < -1.5 or [Fe/H] < -1.0 is an easier task because there are many more positive samples in our training and testing sets for these tasks.

It is important to see probability-distribution situations for the stars with [Fe/H] close to -2 because, as shown in Fig. 4, there are a lot of stars with [Fe/H] close to -2 in our catalogues. Fig. A4 shows [Fe/H] versus P_0 of testing sets. The left and middle panels are for Classifier-GC and Classifier-GP on bright stars (BP < 16), and the right panel is for Classifier-GP on faint stars (BP > 16) and Classifier-T. In these panels, red points refer to the stars predicted to be metal-poor, and blue points refer to those predicted to be non-metal-poor. The left and middle panels of Fig. A4 show that, as the increase of [Fe/H] from -2.5 to -1.5, P_0 sharply decreases from 1 to nearly 0, which indicates that the P_0 of Classifier-GC and Classifier-GP are sensitive to the metallicity variance (for bright stars). Additionally, there are a lot of blue points (false-negative samples) with [Fe/H] < -2 in the right panel of Fig. A4, because we sacrificed the completeness of turn-off stars and faint giant stars to get higher purity, as discussed in Section 3. Fortunately, however, most of the red points in the right panels are still metal-poor, which is a sign of high purity.

Table A1. Completeness and purity of the classifiers for stars with [Fe/H]< −1.5 or −1.0.

	[Fe/H] < −1.5		[Fe/H] < −1.0	
	Completeness	Purity	Completeness	Purity
Classifier-T, BP < 16	36 per cent	79 per cent	61 per cent	90 per cent
Classifier-T, BP > 16	34 per cent	75 per cent	55 per cent	91 per cent
Classifier-GC, BP < 16, $M_G > 4$	67 per cent	86 per cent	77 per cent	93 per cent
Classifier-GC, BP < 16, $M_G < 4$	17 per cent	74 per cent	17 per cent	80 per cent
Classifier-GP, BP < 16, $M_G > 4$	71 per cent	87 per cent	83 per cent	93 per cent
Classifier-GP, BP < 16, $M_G < 4$	27 per cent	66 per cent	31 per cent	72 per cent
Classifier-GP, BP > 16, $M_G > 4$	48 per cent	75 per cent	58 per cent	88 per cent
Classifier-GP, BP > 16, $M_G < 4$	31 per cent	75 per cent	21 per cent	80 per cent

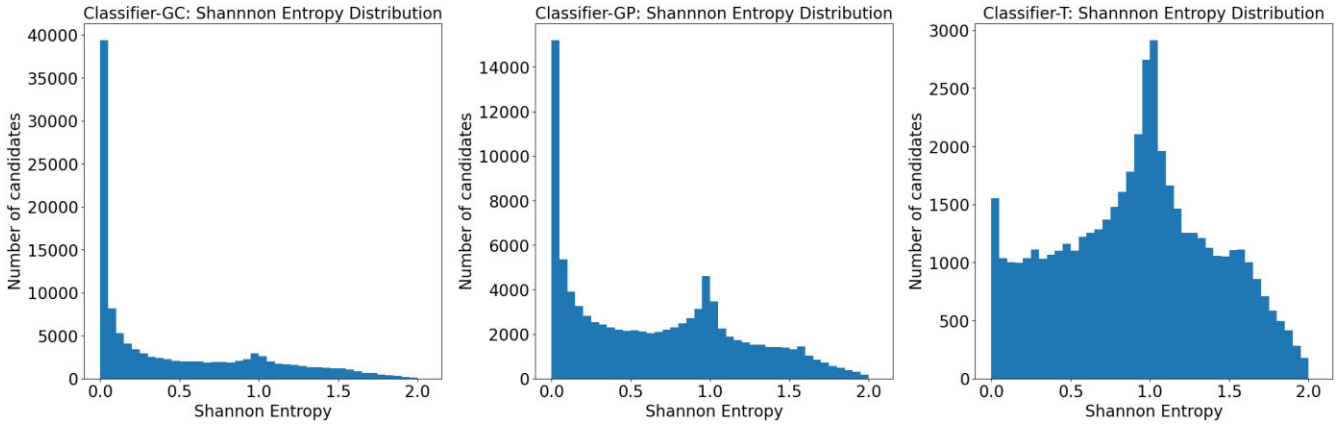

Figure A1. Shannon entropy distributions of the metal-poor candidates found by different Classifiers.

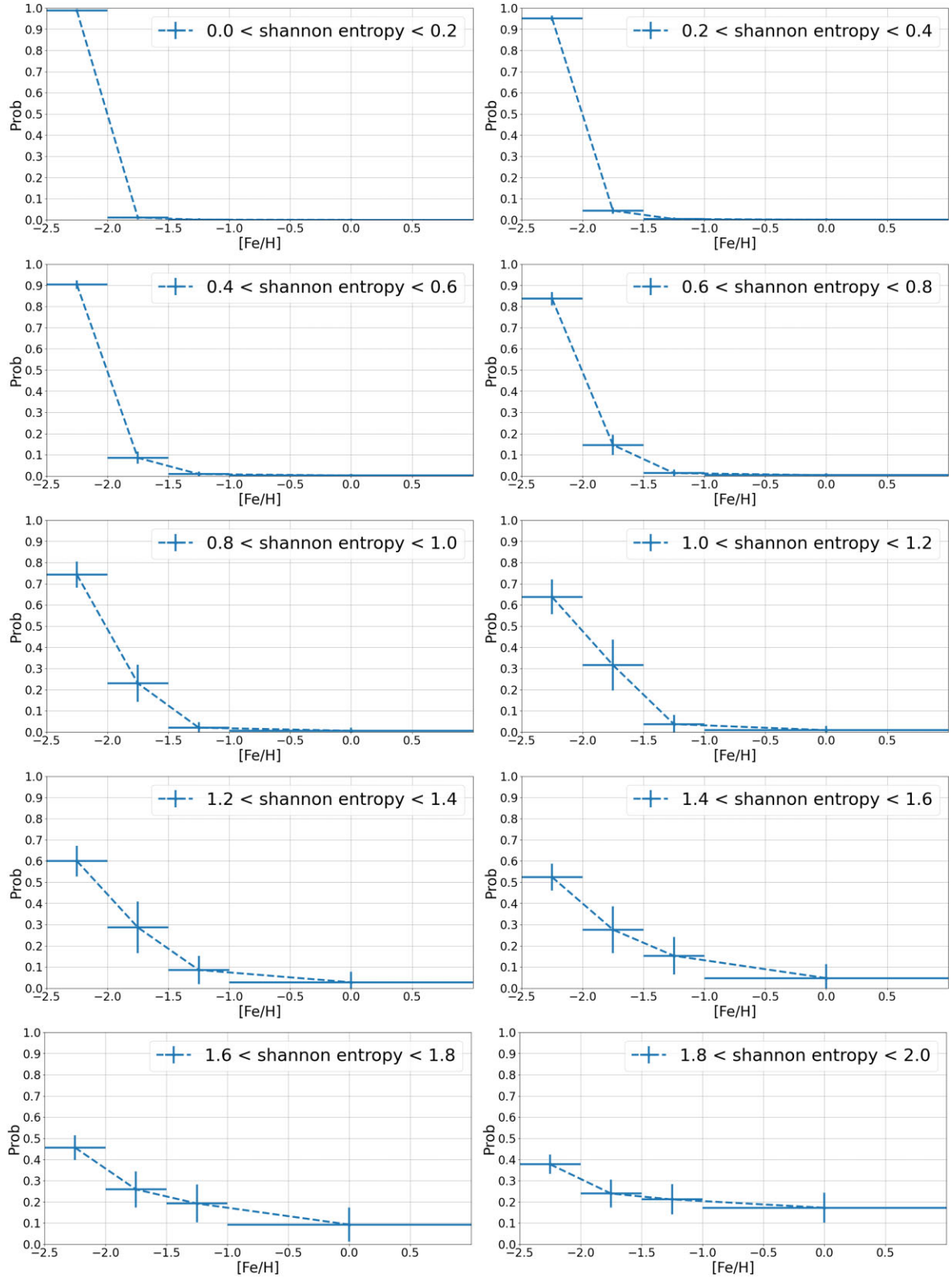

Figure A2. Distributions of mean P_0, P_1, P_2, P_3 with 1 σ error bars of our three catalogues in different Shannon entropy intervals.

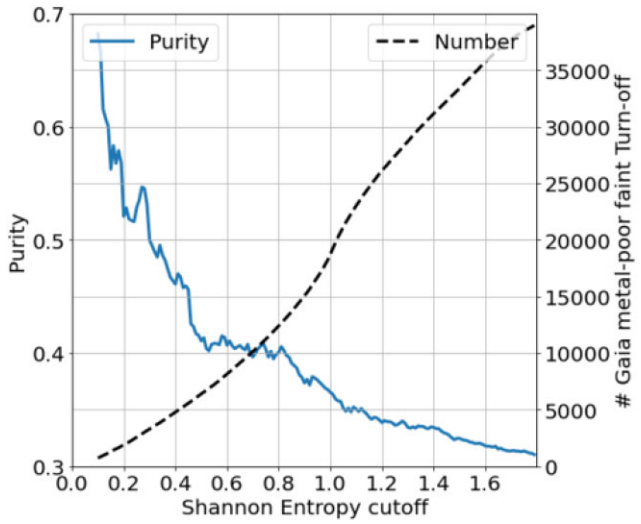

Figure A3. Number and purity of the remaining metal-poor faint ($BP > 16$) turn-off candidates as a function of Shannon entropy threshold.

APPENDIX B: EXTINCTION CORRECTION OF THE BP/RP COEFFICIENTS

Since the BP and RP coefficients of the star with extinction will differ from the coefficients of the same star without extinction, here we try to directly correct the BP/RP coefficients for extinction effects. The extinction model we assume is the following. If C is the BP/RP coefficient vector of a star without extinction (the coefficient vector is normalized by the first coefficient). We assume that the effects of extinction can be described as

$$C_{\rm extincted} - C = (\alpha + \beta C_0) E_{B-V} + \gamma E_{B-V}^2, \quad \text{(B1)}$$

where α, γ are vectors with the same number of elements as the length of the coefficient vector, and β is a matrix. The rationale behind this parametrization is that the first term in the right hand side of the equation is providing linear changes of coefficients with extinction and the extinction coefficients can differ for stars with different spectra (this is essentially a Taylor expansion). The final term allows some non-linearity of the coefficients with extinction (but without dependence on the coefficients themselves).

To fit for the coefficients we take the APOGEE DR17 catalogue with *Gaia* BP/RP coefficients. For each star with extinction E_{B-V}

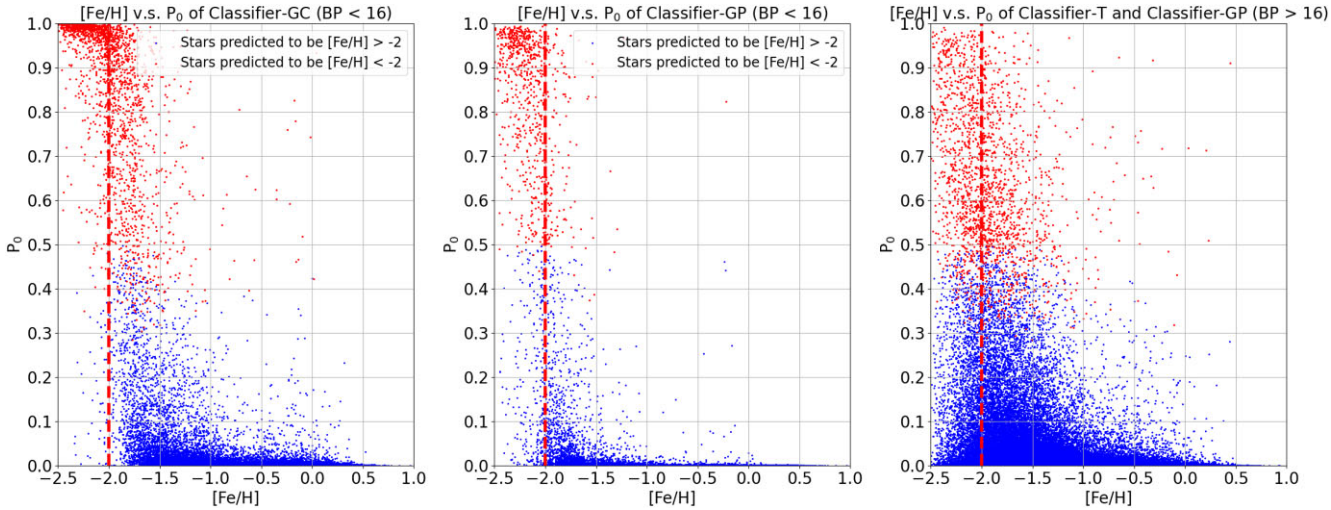

Figure A4. P_0 as a function of [Fe/H].

> 0.05 and BP/RP coefficients C we find a nearest neighbour in the space of T_{eff}, $\log g$, [Fe/H] but with $E_{B-V} < 0.05$. This provides us with the estimated unextincted BP/RP vector for that star. We then fit the relation (equation B1) between unextincted and extincted coefficients using regularized linear regression (as implemented in class `LassoCV` in `sklearn` package). We have found that the extinction coefficients mostly dependent on first few BP/RP coefficients (as those determine the broad spectral shape), thus we force the matrix β to only have first 10 non-zero rows. We provide the best fit α, β, γ for BP/RP in supplementary materials.

The Fig. B1 demonstrates the effect of the extinction corrections. The top rows shows the differences between the coefficients of extincted versus non-extincted stars versus extinction. We can clearly see that several coefficients show strong dependence on E_{B-V} as expected. The bottom panels show what happens after correcting the coefficients. We can see that the trends with extinction mostly disappeared.

In this extinction correction we rely on the Schlegel, Finkbeiner & Davis (1998) 2D maps thus we essentially make an assumption that all of the stars are behind the dust layer. When this assumption is broken we expect that our corrections will not be appropriate.

APPENDIX C: WSDB ARCHIVE QUERIES

Following ADQL queries were utilized to cross-match LAMOST and APOGEE to *Gaia* DR3 source id with XP spectra on WSDB:

Query utilized for APOGEE:

```
select
bp_chi_squared,
rp_chi_squared,
bp_degrees_of_freedom,
rp_degrees_of_freedom,
sfd_ebv,gaiaedr3_phot_g_mean_mag,
source_id,fe_h,alpha_m,
logg,teff,ra,dec,
gaiaedr3_parallax,
gaiaedr3_parallax_error, {COEFFS}

from apogee_dr17.allstar as a,
gaia_dr3.xp_continuous_mean_spectrum as s

where
s.source_id = a.gaiaedr3_source_id
```

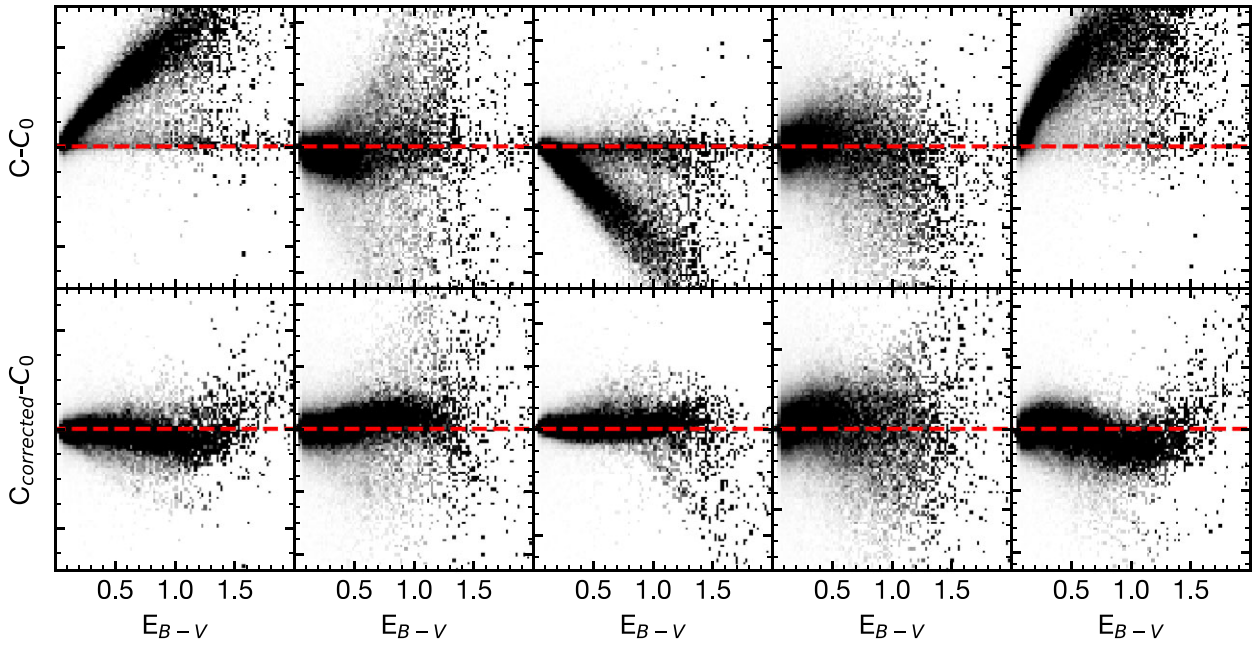

Figure B1. Effect of extinction correction. The top panel show the difference between BP coefficients between stars with zero extinction and stars with same stellar atmospheric parameters, but significant extinction versus value of extinction. The bottom panel shows the same but after applying the best-fitting extinction correction from Section B. The red line shows where zero is.

```
  and bp_chi_squared < 1.5*bp_degrees_of_freedom
  and rp_chi_squared < 1.5*rp_degrees_of_freedom
```

Query utilized for LAMOST:

```
  with
  x as(select gaia_source_id, feh, teff, logg,
  rank() over
  (partition by gaia_source_id or-
der by snrr desc)
  from lamost_dr7.lrs_stellar),


  y as (select gaia_source_id::bigint
  as sid,feh,teff,logg from x where rank = 1),


  z as (select feh,teff,logg,
  (select dr3_source_id
  from gaia_edr3.dr2_neighbourhood as g
  where g.dr2_source_id = sid
  order by angular_distance asc limit 1)
  as source_id from y where sid>0)


  select
  bp_chi_squared, rp_chi_squared,
  bp_degrees_of_freedom, rp_degrees_of_freedom,
  feh, ebv, phot_g_mean_mag,
  g.source_id, teff,
  logg, g.ra, g.dec, {COEFFS}


  from z as a,
  gaia_dr3.xp_continuous_mean_spectrum as s,
  gaia_dr3.gaia_source as g
  where g.source_id = a.source_id
  and s.source_id = g.source_id
```

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