Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **516,** 5521–5537 (2022) https://doi.org/10.1093/mnras/stac2550
Advance Access publication 2022 September 9

A machine learning approach to photometric metallicities of giant stars

Connor P. Fallows★ and Jason L. Sanders

"Department of Physics & Astronomy, University College London, Gower St., London WC1E 6BT, UK

Accepted 2022 September 6. Received 2022 August 11; in original form 2022 April 13

ABSTRACT

Despite the advances provided by large-scale photometric surveys, stellar features – such as metallicity – generally remain limited to spectroscopic observations often of bright, nearby low-extinction stars. To rectify this, we present a neural network approach for estimating the metallicities and distances of red giant stars with 8-band photometry and parallaxes from *Gaia* EDR3 and the 2MASS and WISE surveys. The algorithm accounts for uncertainties in the predictions arising from the range of possible outputs at each input and from the range of models compatible with the training set (through drop-out). A two-stage procedure is adopted where an initial network to estimate photoastrometric parallaxes is trained using a large sample of noisy parallax data from *Gaia* EDR3 and then a secondary network is trained using spectroscopic metallicities from the APOGEE and LAMOST surveys and an augmented feature space utilizing the first-stage parallax estimates. The algorithm produces metallicity predictions with an average uncertainty of ± 0.19 dex. The methodology is applied to stars within the Galactic bar/bulge with particular focus on a sample of 1.69 million objects with *Gaia* radial velocities. We demonstrate the use and validity of our approach by inspecting both spatial and kinematic gradients with metallicity in the Galactic bar/bulge recovering previous results on the vertical metallicity gradient (-0.528 ± 0.002 dex kpc^{-1}) and the vertex deviation of the bar (-21.29 ± 2.74 deg).

Key words: methods: statistical – stars: distances – Galaxy: abundances – Galaxy: bulge – Galaxy: stellar content.

1. INTRODUCTION

One overarching goal of studying the Milky Way is to reveal its detailed formation and evolution, and place our Galaxy in the context of galaxy formation across the Universe (Bland-Hawthorn & Gerhard 2016; Barbuy, Chiappini & Gerhard 2018). With the advent of large-scale spectroscopic surveys (RAVE, Steinmetz et al. 2020; APOGEE, Ahumada et al. 2020; LAMOST, Cui et al. 2012; *Gaia*-ESO, Gilmore et al. 2012; SEGUE, Yanny et al. 2009; and GALAH, Buder et al. 2021, and in future DESI, DESI Collaboration et al. 2016; WEAVE, Dalton et al. 2014; 4-MOST, de Jong et al. 2019; and Milky Way Mapper, Kollmeier et al. 2017), we have highly detailed observations of $>10^6$ stars allowing characterisation of their effective temperatures, surface gravities, radial velocities, chemical compositions, masses, ages, and more, from which we can make progress on this goal by elucidating and separating the series of events and processes that have shaped our Galaxy over cosmic time.

However, despite the utility of spectroscopic data, these surveys do have limitations of scope when applied to some problems. As noted by Ivezić et al. (2008) and Huang et al. (2022), taking spectroscopic data for very distant or faint objects can quickly become difficult. This causes many surveys to have complex selection criteria to ensure good spectroscopic data can be taken. These criteria typically limit observations to specific object classes within a limited sky region making the application of such data to large-scale populations or structures difficult, as only a small portion of these groupings may be included in the selection criteria. For example, when attempting to study the inner regions of the Milky Way's disc and bulge, the large distances and extreme extinction effects put many stars beyond the reaches of spectroscopic observations. This area of the sky therefore tends to have relatively few spectroscopic observations, which makes analysis of these interesting populations difficult (although the infrared APOGEE and in future MOONS, Cirasuolo et al. 2014, surveys are rapidly changing this state of affairs).

On the other hand, we have large-scale photometric surveys, which typically are not bound by the same criteria that tend to limit spectroscopic observations. This allows them to be far more expansive, generally observing many classes of object across the whole sky (or often at least half) to a significantly greater depth. For example, *Gaia* (Gaia Collaboration 2021), 2MASS (Skrutskie et al. 2006), and WISE (Wright et al. 2010) have all observed the entire sky across the optical to infrared, whilst SDSS (Aihara et al. 2011), Pan-STARRS (Chambers et al. 2016), DES (Abbott et al. 2021), Sky-Mapper (Wolf et al. 2018), and GALEX (Bianchi, Shiao & Thilker 2017) among others have surveyed large fractions of the sky. However, unless designed with filters with specific sensitivity to stellar metallicity or surface gravity like Sky-Mapper's *u* and *v* bands (Keller et al. 2007), broad-band photometric data tend to struggle to accurately determine stellar parameters without additional input.

Thus, we reach our aim with this research: to develop a method that can determine stellar properties with the utility of spectroscopic data, while retaining the scope and scale of photometric surveys. From this, we would then be able to analyse, on a much deeper level, the stellar populations and structures that stretch across the Milky Way.

Attempts to determine stellar metallicity from photometry have had some past successes, through leveraging the subtle sensitivity of broad-band colours to metallicity. One early approach was the 'ultraviolet (UV) excess' method (Wallerstein 1962) which can be

★ E-mail: connor.fallows.20@ucl.ac.uk

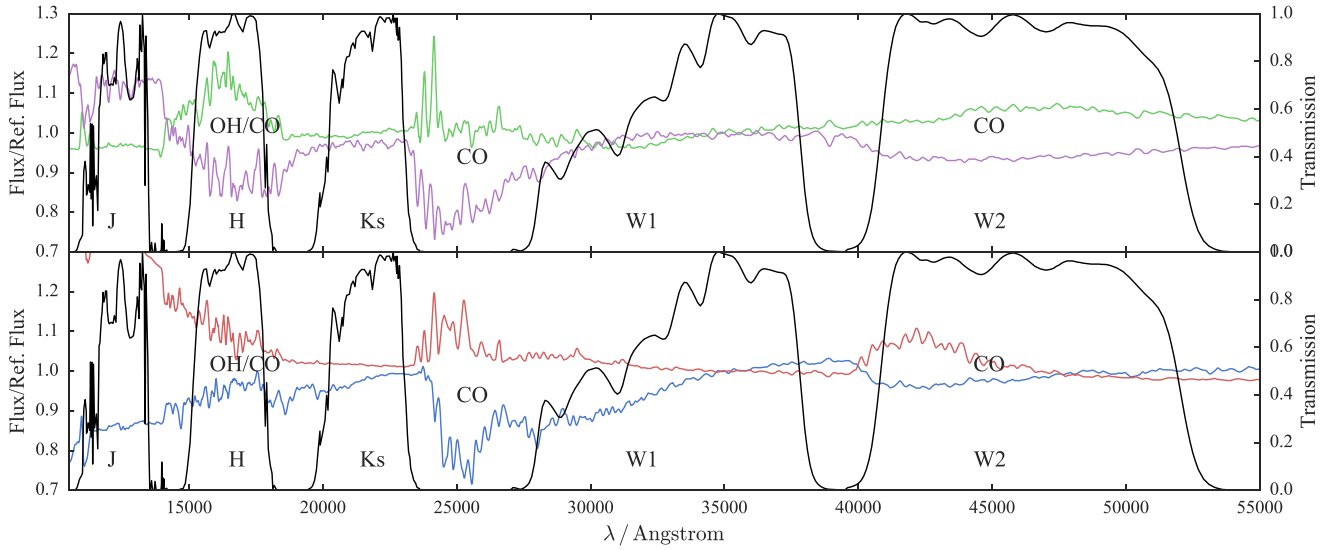

Figure 1. The sensitivity of WISE to stellar metallicity: Ratios of MARCS models with infrared bandpasses overplotted normalized at $\lambda = 35\,000$ Å. The reference model has $T_{\rm eff} = 3500$ K, $\log g = 0$ dex, and [M/H] = 0 dex. The green and purple lines show models with [M/H] = −1 dex and [M/H] = 1 dex, and blue and red lines $T_{\rm eff} = 3200$ K and $T_{\rm eff} = 4000$ K. Note the strong gradients in $W2$ due to the CO feature. The effects of temperature and metallicity variations in $W2$ can be distinguished using the bluer 2MASS bands.

calibrated to map both stellar temperatures and metallicities using the large number of metal lines in bluer and UV bandpasses. This method has been adapted for use with modern photometric surveys, using the SDSS (Ivezić et al. 2008) and Pan-STARRS (Thomas et al. 2019) ($g - r$) and ($u - g$) colours to estimate metallicities. In a similar vein, the metallicity sensitivity of the Ca H & K region at ∼3950 Å has been targeted using narrow-band filters in the PRISTINE (Starkenburg et al. 2017) and Sky-Mapper (Wolf et al. 2018) surveys (see Huang et al. 2022; Lin, Casagrande & Asplund 2022, for catalogues of stellar parameters derived from Sky-Mapper data). Due to the strong effects of extinction on UV/near-UV, these methods are less effective for studying faint or distant objects within the highly extincted inner Milky Way (although see Arentsen et al. 2020 for a study of metal-poor stars in the Galactic bulge using the PRISTINE survey).

For more highly extincted regions, infrared photometric surveys are more attractive. Schlaufman & Casey (2014), Koposov et al. (2015), Li et al. (2016), and Casey et al. (2018) have all demonstrated how the infrared WISE survey (Wright et al. 2010) can be used to both separate dwarf and giant stars and also estimate stellar metallicities for red stars. In particular, the WISE colour ($W1 - W2$) displays a strong correlation with stellar metallicity ($W1$ and $W2$ have effective wavelengths of 3.4 and 4.6 μm, respectively). This is primarily due to the presence of a CO feature in the spectrum of M giants. In Fig. 1, we show ratios of stellar spectra from the MARCS model grid (Gustafsson et al. 2008). Increasing the metallicity we observe the molecular features (particularly the CO band in the $W2$ bandpass) weaken whilst the flux in $K_{\rm s}$ and $W1$ are essentially unaffected leading to bluer ($K_{\rm s} - W2$) and ($W1 - W2$) for more metal-rich stars. These colours also vary with effective temperature (redder for hotter stars) again due to CO variation but this degeneracy can be removed by combining with bluer colours such as ($J - K_{\rm s}$). This metallicity sensitivity of the WISE bands was utilized most recently by Grady, Belokurov & Evans (2021), who used machine learning regression models with *Gaia*, 2MASS, and WISE bands to estimate metallicities of stars in the Magellanic Clouds. This improved on past works by allowing the subtler metallicity sensitivity of other photometric colours to be included. For example, they found that by including *Gaia* ($G_{\rm BP} - G_{\rm RP}$) and 2MASS ($J - H$) they were able to add additional metallicity information beyond that provided by ($W1 - W2$). This work provided metallicity estimations with high accuracy (±0.13 dex for $-1 \leq$ [Fe/H] ≤ -0.5) allowing for detailed mapping of the mean metallicity of the Magellanic Clouds. However, this method was not used on stars within the Milky Way.

Past research has therefore left a gap for broadly applying metallicity estimation to large-scale photometric surveys of the Milky Way. It should be noted that WISE information is often utilized in stellar characterization pipelines that provide metallicity estimates (e.g. Anders et al. 2022; Lin et al. 2022); although, these methods rely on theoretical stellar models, or isochrones, which can be uncertain for cool stars with significant molecular contributions to their atmospheres. Here, we provide a complementary data-driven approach to instead learn the correlations between photometric colours and metallicities obtained from large spectroscopic surveys. We thus bypass complexities in detailed stellar modelling. In doing this, we supplement *Gaia* EDR3 (Gaia Collaboration 2021) astrometry with metallicity information. Such a combination allows us to study the spatial, kinematic, and abundance trends within the Galaxy, and, thus, we are able use this new methodology to probe the evolution and origins of various Milky Way structures.

This paper is split into four main components: neural network (NN) set-up, distance estimation, metallicity estimation, and a brief analysis of the properties of a bar-bulge sample. Section 2 describes the general set-up of the NN algorithm we will use in the subsequent methods. Section 3 describes our machine learning-enhanced approach to refine the distances we use in our analysis, allowing us to improve object positional information and refine absolute magnitude calculations. Section 4 covers the estimation of metallicities through the use of our NN algorithms, and the creation of our final output catalogue. Section 5 describes an investigation of the spatial and kinematic gradients of a bar-bulge sample separated by our photometric metallicities, before we close with our conclusions in Section 6.

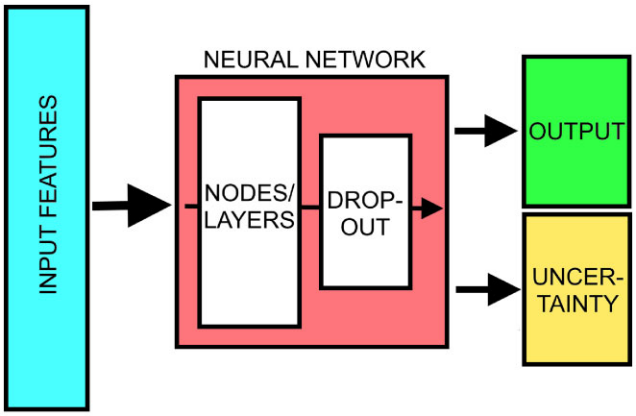

Figure 2. Diagram of the adopted NN architecture. Input features are fed into the NN, which trains the nodes/layers with the drop-out modifier active. Then, for predictions, the layers predict a value with a combined uncertainty from the drop-out stochasticity and the secondary 'uncertainty' output node.

2. NN SET-UP

For the most accurate predictions of photometric metallicity, we opt for a NN machine learning algorithm. Typically, NN architectures are trained with a set of input features, which are fed through a non-linear layered network to return an output value. The network layers are constructed from a set of inter-connected nodes, with the strength of the connections (or weights) tuned through training to allow the model to learn patterns in the input data. Training is guided by the network's 'loss function', which guides the penalty the model receives for returning poor predictions of the outputs compared to the training set, and which the network aims to minimize. The most common loss function is the mean squared error between the NN's output and desired target values – although this can be customized and tuned for the desired set-up.

Using the Python implementation of the Torch machine learning library, Pytorch (Paszke et al. 2019), we work with a NN with the architecture shown in Fig. 2 and described in-detail in Appendix A. For a set of input features, x, each layer in the network, h, follows

$$h = af(x) + b, \tag{1}$$

where $f(x)$ is the non-linear response function of the layer, and the matrix of weights a and vector of biases b are constants refined by the training process. Thus, for a network of n layers, we return an output value, y, from outputs of one layer being sequentially input to the next. This gives us:

$$y = a_n f(a_{n-1} f(\dots f(a_2 f(a_1 f(x) + b_1)) + b_2 \dots) + b_{n-1}) + b_n. \tag{2}$$

The network is trained iteratively with the network tuning a_i and b_i to improve the loss function. By improving the average loss function over the full training set, the network is able to learn the correlation between x and y and predict outputs for new sets of input features. This allows for accurate and robust fitting, while also allowing $f(x)$ to be customized and modified to best suit a chosen problem.

However, NNs do not tend have a measure of 'confidence' in their estimations and instead usually return a single value for a set of input features. For us to include a measure of the network's predictive confidence, we add two small modifications adapted from Leung & Bovy (2019a): an uncertainty output node and node drop-out.

2.1 Uncertainty node

We include a secondary output node into the NN architecture, as marked in Fig. 2. This node provides one uncertainty measure, $\sigma_{\rm pred}$, known as the model's 'predictive uncertainty'. This is the variance in the training data that is not accounted for by the uncertainties noted in the training set's output targets. Even with perfect data, there are 'hidden variables' that impact the outputs. This manifests as identical training inputs into the NN returning a range of outputs. During the training process, the output uncertainty was fed into the NN's customized loss function, and allows us to return an output value along with an uncertainty measure.

We adopt the loss function from Leung & Bovy (2019a). With y_i as the target value from the training set with uncertainty $\sigma_{{\rm data},\,i}$, and $\hat{y}_i$ the value returned by the NN with uncertainty $\sigma_{{\rm pred},\,i}$ (from the 'uncertainty node'), the logarithm of the joint variance is determined as $s_i = \ln(\sigma^2_{{\rm data},i} + \sigma^2_{{\rm pred},i})$. The loss function, $J(y_i, \hat{y}_i)$, is then defined as

$$J(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - \hat{y}_i)^2 e^{-s_i} + \frac{1}{2}s_i. \tag{3}$$

The predictive uncertainty, $\sigma_{\rm pred}$ is refined by the training process, with each iteration of training incrementally refining the uncertainty output when calculating the loss function. The function in equation (3) is designed such that the network minimizes loss from poor predictions by maximizing the predictive uncertainty. However, this drive is countered by the final additive term which increases loss for high predictive uncertainty. In this way, the network optimizes to find the largest predictive uncertainty for the given data, but is penalized for selecting extremely large or small values.

2.2 Drop-out

Drop-out (Hinton et al. 2012) is a common NN operation used to dissuade overfitting during the training stage by randomly 'dropping' a fraction of the nodes in each layer. This modifies equation (1) to be

$$h = a\ g(x) + b, \tag{4}$$

where $g(x) = {\rm P}\ f(x)$. Here, P is a function that applies a Bernoulli distribution to each node within a layer (and thus modifies the response of $f(x)$). The Bernoulli function causes some chosen fraction of nodes within a layer to be temporarily 'zeroed' out, and thus have no effect on the current training or prediction pass. This limits the effect one node or branch can have to the overall output, as other nodes in the network must learn to 'cover' for those hidden by the drop-out process. With drop-out active, the network tends to learn the problem as a cohesive unit, and avoids the creation a small number of overinfluential nodes that can dictate the network's predictions.

However, in our case, drop-out can have a secondary function to add stochasticity to the model (Gal & Ghahramani 2016). As each run of the network has a random fraction of nodes missing, we can consider each run to be a slightly different network. So, if a set of input features are repeatedly passed through the NN, the variations due to drop-out will cause different outputs to be returned each time. While predicting from our network, we return an ensemble of networks with slight variations due to the randomness of drop-out, all of which are consistent with the training data. When we input new features, we return a distribution of output values from the ensemble of networks – a distribution we can consider as a probability. We therefore consider our model to be a Bayesian NN, which returns

a probabilistic distribution rather than a single value. From such a distribution, we draw a prediction (mean) and an implicit uncertainty (standard deviation).

We return our uncertainty from the drop-out stochasticity as $\sigma_{\rm drop}$. Then, with both the drop-out uncertainty, $\sigma_{\rm drop}$, and the predictive uncertainty, $\sigma_{\rm pred}$, calculated, we can determine the final uncertainty of each prediction, $\sigma_{\rm total}$, by

$$\sigma_{\rm total} = \sqrt{\sigma_{\rm drop}^2 + \sigma_{\rm pred}^2}. \quad (5)$$

3. DISTANCE ESTIMATION

Before we begin estimating metallicities, our method requires a robust measure of stellar distances. Distances allow us to calculate absolute magnitudes for our sample stars, which can provide essential information on intrinsic stellar properties for the NN's model.

While using *Gaia* parallaxes directly would be the ideal choice for data-driven analysis, there are a number of limitations to such an approach. As described by Bailer-Jones et al. (2021), transforming between parallax and distance can lead to issues if done naively. Objects with $\varpi \approx 0$, even with well-constrained uncertainties, will tend to have very large fractional errors. This equates to extremely large distance uncertainties for stars beyond a few kpc. Additionally, valid parallaxes in the *Gaia* catalogue can have negative values due to the random scatter from uncertainties at small parallaxes, which makes the naive $r = 1/\varpi$ relation impractical to apply. The approach developed by Bailer-Jones (2015) instead adds a statistical prior to distance prediction that works to guide estimates for objects with poorly informative *Gaia* parallaxes. From this, a distance estimation can be drawn allowing us to avoid the limitations of the raw *Gaia* data.

However, for this work, we aimed to focus on a large proportion of the Milky Way's stars. Therefore, many of the objects in our samples exist in the distance regime where prior information becomes dominant over *Gaia* parallax information. While this was expected behaviour for this approach, we found some estimates to be strongly dependent on the parameters of the prior rather than being guided by *Gaia* measurements – which reduce the utility of these distances for our methodology.

In an attempt to reduce the impact of the prior, Bailer-Jones et al. (2021) adjust their method to also include a star's photometric information (producing 'photogeometric' distances). Briefly, this secondary approach uses a colour–magnitude prior (derived from *Gaia* photometric bands) to restrict the range of absolute magnitudes an object of a given colour can have. Thus, they constrain the distance probability function their method returns. With this addition, they find an improvement in the precision of stars with poorly informative parallaxes.

Our approach follows on from this idea, expanding the addition of photometric information through the inclusion of a wide range of additional bands (see Hogg, Eilers & Rix 2019, for a similar approach also utilizing spectroscopic information). However, instead of using this data as a constraint on our distance estimates, we instead used our photometric information and NN algorithm to estimate an independent parallax value. This 'photometric parallax' was then combined with the parallaxes from *Gaia*, and allowed us to return values with much lower uncertainties. Thus, we reduced the regime where parallax information is uninformative, and thus limited the number of objects where the Bailer-Jones et al. (2021) prior has a significant impact.

3.1 Data collection

In order to augment existing distance information with photometric data, we required accurate astrometry and a wide range of photometric colours.

We followed the lead of Grady et al. (2021), and selected our data from three photometric surveys: *Gaia* EDR3 (Gaia Collaboration 2016, 2021; Riello et al. 2021; Seabroke et al. 2021), 2MASS (Skrutskie et al. 2006), and the unWISE catalogue (Schlafly, Meisner & Green 2019). The *Gaia* survey is an optical photometric survey, with three bands (G, $G_{\rm RP}$, and $G_{\rm BP}$) between 330 and 1050nm, and focuses on observing accurate sky positions, proper motions, parallaxes, and radial velocity information. The 2MASS survey instead observes in near-infrared, with three bands, J, H, and Ks, with peak sensitivity at 1235, 1662, and 2159 nm, respectively, which grants information to separate giant and main-sequence stars (Majewski et al. 2003) as well as bolster extinction measurements (as will be discussed later). Finally, the unWISE survey is built upon the results of the WISE catalogue described previously (Wright et al. 2010), but with altered image processing to retain observation resolution in star-dense regions. This increases the available number of objects with WISE bands ($W1$, $W2$, $W3$, and $W4$ at 3.4, 4.6, 12, and 22μm, respectively), and thus greater coverage at large distances and within high-density sky regions. For our sample, we avoided the $W3$ and $W4$ bands due to the small number of objects with accurate observations, which would have limited the maximum potential size of our sample.

With access to the H and $W2$ bands, we made use of the Rayleigh–Jeans Colour Excess (RJCE) Method (Majewski, Zasowski & Nidever 2011) to determine accurate extinction corrections for objects in our sample. This approach relies on the fact that, for most stellar types, intrinsic (H-$W2$) colour is nearly constant. Therefore, significant reddening in this colour can provide a good measure of the extinction effects on a star-by-star basis. To transform the extinction to the other photometric bands, we used the extinction coefficients from Wang & Chen (2019).

Objects were chosen to ensure good photometry by filtering for high-quality observations. We limit the *Gaia* BP/RP flux excess to $\leq$3.0, limit the astrometric renormalized unit weight error to values $\leq$1.4, and select only for objects with 'good' $W1$ and $W2$ photometry from the UnWISE quality flags. We further ensured our entire sample had velocity information (proper motion, radial velocity) from *Gaia*, which provided kinematic information for stars within our sample. This kinematic information, when combined with the distances from our method, could then allow us to calculate three-dimensional (3D) velocities for each star, and thus analyse the kinematic distributions of our sample objects. Due to the limitations of *Gaia*'s radial velocity measurements, requiring radial velocities remained the largest limit on our sample size, with only 0.4 per cent of the full survey catalogue having radial velocity information. Corrections to *Gaia* parallaxes were also made at this stage, accounting for zero-point errors in the measurements described in Lindegren et al. (2021).

3.2 Methodology

Our method leveraged the NN architecture described in Section 2, and a 'pseudo-absolute magnitude' measure described by Arenou & Luri (1999). We anticipated that, if we chose to have the NN predict a value of parallax directly from photometric data, the algorithm would struggle to learn a direct correlation between a star's photometric colour and its distance. However, as the absolute magnitude of a star

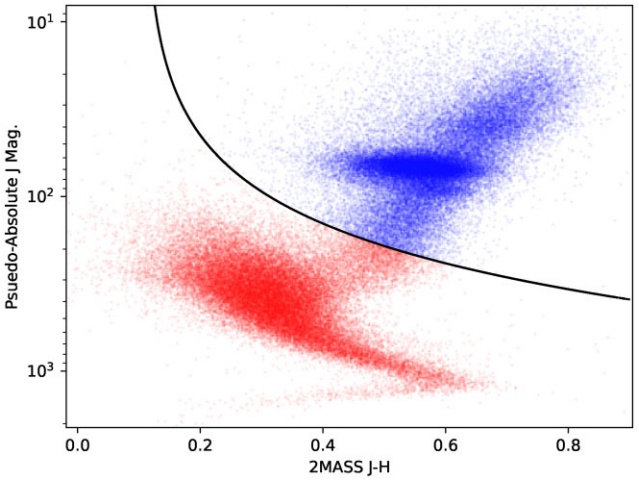

Figure 3. Plot of object pseudo-absolute J magnitudes against J-H colour. Both axes have been corrected for extinction using the RJCE method (Majewski et al. 2011). The giant-dwarf cut is shown with the main sequence in red, and the red clump/giant branch in blue.

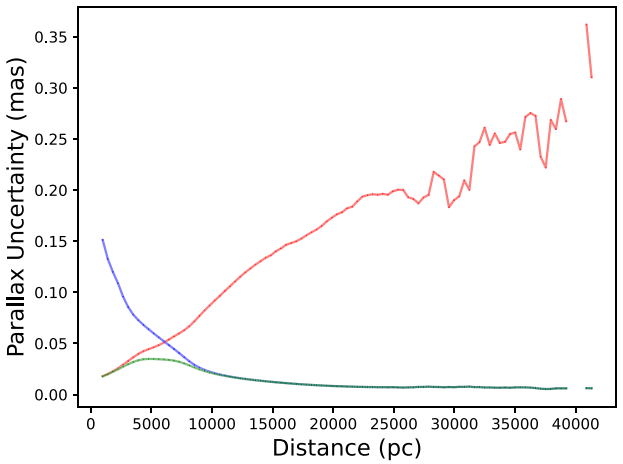

Figure 4. Plot of parallax uncertainty against object distance for *Gaia* parallaxes (red), our NN parallaxes (blue), and the combined unified parallax (green). Note that this sample is not limited to only objects with *Gaia* radial velocities.

is intrinsic to the star, it is therefore independent of object distance. We were thus able to use this as a target for the NN to predict, rather than attempting to estimate parallax directly.

We used the pseudo-absolute magnitude defined in the 2MASS J-band, $M_{J,\,\mathrm{pseudo}}$, as the basis for our analysis, where ϖ was the *Gaia* parallax and J_c was the extinction-corrected apparent J-band magnitude. $M_{J,\,\mathrm{pseudo}}$ was therefore defined as

$$M_{J,\mathrm{pseudo}} = \varpi\, 10^{0.2 J_\mathrm{c}}. \tag{6}$$

This value acted as a good proxy for absolute magnitude by combining parallax and magnitude information. The NN therefore made its predictions within the pseudo-absolute magnitude parameter space, rather than parallax space, and was therefore generalizable beyond the scope of the training data. Had we estimated parallax alone, the NN would struggle to predict reliably towards (and beyond) the edges of the parameter space – especially towards distant object parallaxes at the smallest end of our range. Furthermore, this formulation for pseudo-absolute magnitude allowed the Gaussian uncertainties in parallax to be translated into Gaussian uncertainties in pseudo-absolute magnitude space.

Initially, we used this value to filter our sample for giant stars. Due to their intrinsic brightness, giant stars are ideal targets for long distance analysis of the Milky Way's population. Therefore, to avoid the NN's attention being split between dwarf and giant stars while training – and thus lowering the model's overall performance – we removed non-giant stars from our sample. The giant and dwarf populations were clearly visible in colour–pseudo-absolute magnitude space, and so we were able to apply a simple cut in these parameters. With M_J and M_H being 2MASS J and H extinction-corrected apparent magnitudes, respectively, and the J-band pseudo-absolute magnitude being $M_{J,\,\mathrm{pseudo}}$, we selected only objects where

$$M_{J,\mathrm{pseudo}} < 492.101(M_J - M_H) - 53.827. \tag{7}$$

This cut is shown clearly in Fig. 3, separating the red clump and giant branch from the main sequence.

From this, we set the NN to accept 16 photometric colours as our input array, x, (described in Appendix A), and to predict the pseudo-absolute magnitude, y. As mentioned previously, using $M_{J,\,\mathrm{pseudo}}$ had the notable advantage of inheriting the Gaussian uncertainties of the *Gaia* parallaxes. We then returned a NN-refined parallax value, μ_NN, as

$$\mu_\mathrm{NN} = M_{J,\mathrm{pseudo}} 10^{-0.2 J_\mathrm{c}}, \tag{8}$$

and similar for the uncertainty σ_NN from the uncertainty node.

To train the network, we followed a method of 'cross-training' which functioned similarly to common cross-validation methods.

As every object in our data sample has a *Gaia* parallax, we chose to train the NN on our sample rather than some external source. In order to train our sample, we split our sample into eight equal 'chunks' which we iterated through. For each chunk, the remaining ~88 per cent was used to train our network, and returned new parallaxes for objects within the chosen chunk. After each iteration, we reset the NN's training for the new chunk which avoids biases arising from objects appearing in both the training and prediction data sets.

With the network's predictions applied to our entire sample, we had derived a set of parallaxes from stellar photometry alone. We therefore considered these results as independent measurements to the parallaxes reported by *Gaia*. Thus, we combine the two values to improve the overall parallax uncertainty. For an object with a *Gaia* parallax, μ_Gaia, and associated uncertainty, σ_Gaia, and with a NN-predicted parallax, μ_NN, and associated uncertainty, σ_NN, we calculate our combined parallax, μ_new, and combined uncertainty, σ_new as

$$\mu_\mathrm{new} = \left(\frac{\mu_\mathrm{Gaia}}{\sigma^2_\mathrm{Gaia}} + \frac{\mu_\mathrm{NN}}{\sigma^2_\mathrm{NN}}\right)\left(\frac{1}{\sigma^2_\mathrm{Gaia}} + \frac{1}{\sigma^2_\mathrm{NN}}\right)^{-1}, \tag{9}$$

and

$$\frac{1}{\sigma_\mathrm{new}} = \sqrt{\frac{1}{\sigma^2_\mathrm{Gaia}} + \frac{1}{\sigma^2_\mathrm{NN}}}. \tag{10}$$

Therefore, we produced a unified parallax value with much narrower error than the initial *Gaia* parallaxes, reducing the number of stars with poorly-informative parallaxes – and so reduced the proportion of objects for which the Bailer–Jones statistical prior was dominant for distance estimation.

We note this improvement in Fig. 4, where we show how parallax uncertainties vary with distance for our *Gaia* data, our NN's outputs, and for the unified parallax value. For this comparison, we have removed the limit of only selecting objects with *Gaia* radial velocity

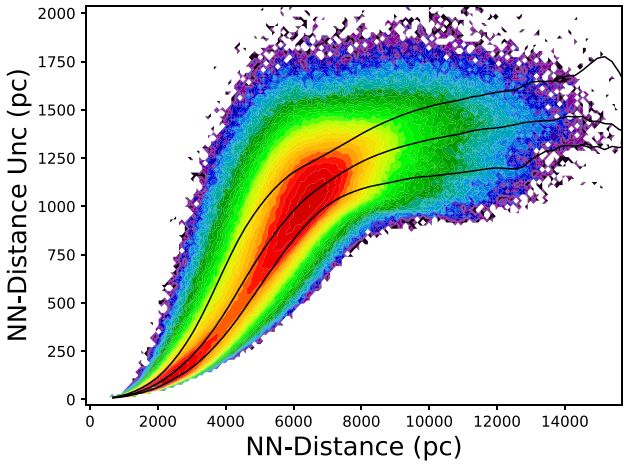

Figure 5. Plot of NN-estimated distance uncertainties against the corresponding absolute distances. Note that these contour plots are logarithmic, and the upper and lower curves are the 84th and 16th percentiles, respectively. Further, note that this sample is not limited to only objects with *Gaia* radial velocities.

information. As objects with radial velocities will tend to be brighter, *Gaia* parallaxes tend to be good, and our NN-based approach has limited impact. Removing this limit shows a more general comparison between the *Gaia* and NN parallax performance, and highlights clearly where our method provides improvement. It is clear that, at around 6.1 kpc, our NN parallax uncertainties become smaller than those from *Gaia*. At distances beyond this, our parallaxes are therefore more informative than those from *Gaia*, and our unified value retains a low uncertainty out to larger distances.

With each object given a NN-enhanced parallax, we calculated new distance estimations. We apply the method of Bailer-Jones et al. (2021), which uses the parallax information and simulation-backed prior distributions to return an estimated parallax. Our NN-enhanced parallaxes form a notable reduction in the number of uninformative parallaxes, decreasing the parallax uncertainty for around 58 per cent of our overall sample. When we focus only on objects with *Gaia* parallax/uncertainty < 2.0, we find around 89 per cent of objects see an improvement from our method. These distances were taken forward to calculate objects positions and absolute magnitudes.

3.3 Validation

To validate the accuracy of the distance predictions, we had three measures: the uncertainty output calculated by the NN, and two samples with comparison distance estimates. These comparison distance samples were those calculated by Bailer-Jones et al. (2021), and those calculated from the AstroNN algorithm (Leung & Bovy 2019b).

3.3.1 *Network uncertainty*

From the NN, we obtained a predicted value of distance (and an associated uncertainty) for each object. We found that this value is low for the majority of our sample, with the mean uncertainty of our whole sample being ± 159.7 pc. We plot these uncertainties versus estimated distances in Fig. 5, with uncertainties binned by absolute distance shown in Table 1. We note that, as with Fig. 4, this plot is not limited to only objects with *Gaia* radial velocity information. As discussed in Section 3.2, this gives us a better sense of our NN's performance than if we only focus on the brighter sample with radial velocity data.

Table 1. Table of mean distance uncertainties for binned absolute distance ranges. All distances are reported in pc. Percentage uncertainties are taken with respect to the mid-point of the bin.

Distance bounds	Distance unc.	Unc. (per cent)
$0 < d < 2000$	$\pm$ 35.018	3.5
$2000 < d < 4000$	$\pm$ 110.750	3.7
$4000 < d < 6000$	$\pm$ 369.182	7.38
$6000 < d < 8000$	$\pm$ 748.284	10.67
$8000 < d < 10\,000$	$\pm$ 1142.063	12.69
$10\,000 < d < 12\,000$	$\pm$ 1571.712	14.28
$12\,000 < d < 14\,000$	$\pm$ 2072.115	15.94
$14\,000 < d < 16\,000$	$\pm$ 2672.450	17.81

As expected, distance uncertainties remained small for closer objects, and become larger for distant objects. The distance uncertainties remained below 10 per cent for objects closer than approximately 6 kpc, with the furthest objects in our sample having distance uncertainties less than 20 per cent.

3.3.2 *Bailer-Jones et al. (2021) distance comparison*

We compared our network's performance in comparison to the photogeometric values calculated by Bailer-Jones et al. (2021). These reference distance values were the values we initially hoped to improve upon with our method. We used much the same method, but applied our NN to reduce the impact of prior terms on the distance estimates. It was therefore expected for there to be good agreement between the two data sets where parallaxes are highly-informative, and significant divergence in the regime where fractional *Gaia* parallax uncertainties were large (i.e. very high uncertainty, or very small parallaxes) and the NN had a stronger influence. If our approach returned accurate distances, we expected to see the majority of stars match between the two samples, with divergences in parallax space remaining symmetric and becoming more prominent for objects further away. We show this comparison for our sample in the left-hand panel of Fig. 6, where we see a clear correlation between the two methods (for good parallaxes) with a large scatter due to the impact of the NN.

We further highlight the right-hand panel of Fig. 6, where we selected a larger comparison sample without the restriction of requiring radial velocity data for all stars. This allowed us to observe additional objects at large distances (≥ 10 kpc) as well as fainter objects at closer ranges. The minor overestimation bias for Bailer-Jones et al. (2021) distances between 4 and 8 kpc appeared to be due to a divergence in the underlying methods. As our distance estimates used the same prior choices as Bailer-Jones et al. (2021), the primary differences between the results arose from our NN providing an improvement over the base *Gaia* parallaxes. Thus, this bias maps the regime where our distances were more weakly constrained by the statistical prior distribution than in Bailer-Jones et al. (2021) work. Beyond this region, where parallax measurements became too noisy for our NN-based approach to improve upon, the two methods again converge as the prior distribution comes to dominate the distance estimates.

3.3.3 *AstroNN distance comparison*

Finally, we compared our distance estimations to those calculated by Leung & Bovy (2019b) with the AstroNN machine learning

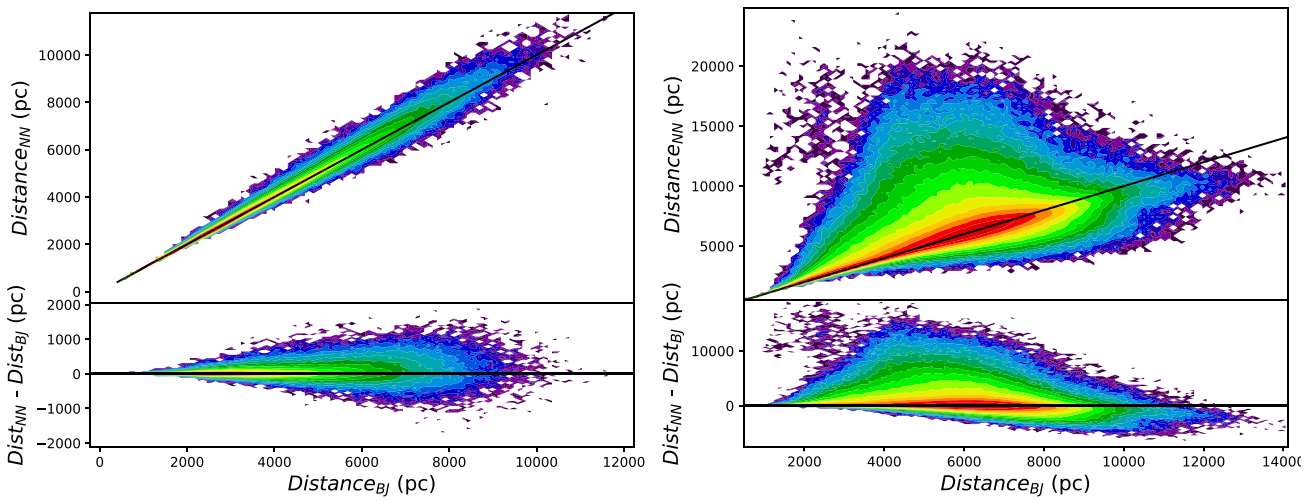

Figure 6. Logarithmic contour plot of NN distances from our work versus distances from Bailer-Jones et al. (2021) for our main sample (left-hand panel). We also perform the same comparison for a sample without the prerequisite of radial velocity data allowing comparisons out to greater distance (right-hand panel).

package. The AstroNN package is based on similar NN algorithms to our own, and uses APOGEE DR17 spectral data to estimate astrophysical parameters such as stellar abundances, ages, and distances. Therefore, we used these distances as an independent sample from which we could draw comparisons to our own results.

Using a sample of 11 318 common stars (not limited to only those with *Gaia* radial velocities), we plotted the comparison in Fig. 7. It was clear there was a strong correlation between the two methods with narrow deviations. The differences were also symmetric, suggesting no significant systematic errors in our method that had caused notable biasing. However, as the majority of our sample overlap existed at distances less than $\sim$4 kpc, a large proportion of sample objects had informative parallaxes. Therefore, we expected to see this strong agreement when comparing these two approaches. Overall, we concluded that our method has very good agreement with the AstroNN distances, and further confirms the reliability of our distance estimates.

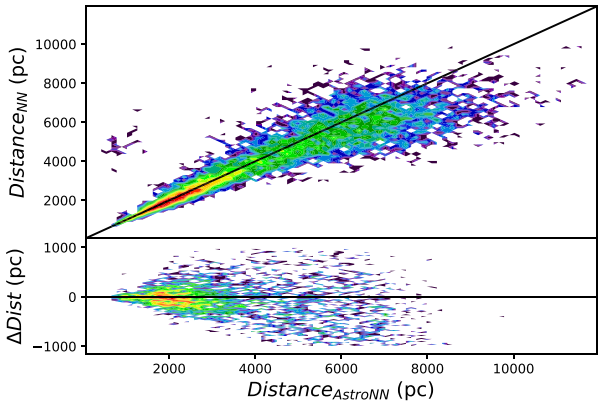

Figure 7. Plot of NN distances from our sample versus distances from AstroNN (above), with comparison residuals (below). Note, this contour plot is logarithmic, and is not limited to only stars with *Gaia* radial velocities.

4. METALLICITY ESTIMATION

With an accurate measure of distance determined for each object, we applied our method to predict stellar metallicities.

4.1 Data collection

We built two samples from which we can estimate metallicities: a training (TG) sample, and a photometric-only (PO) test sample.

The PO sample followed the approach detailed in Section 3.1, drawing astrometric and photometric data from *Gaia* EDR3 and the 2MASS and UnWISE surveys. We also included the distance estimations determined in Section 3, and applied the same filtering to ensure *Gaia* astrometry includes radial velocities for kinematic analysis. This sample acts as our 'output' sample, upon which we will be applying our method for predicting metallicities.

Our TG sample contained the data we will use to train our NN algorithms. This was built from matching objects from our PO sample with iron abundance measurements ([Fe/H]) derived from two spectroscopic surveys, APOGEE-2 (SDSS DR16) (Majewski et al. 2017; Ahumada et al. 2020) and LAMOST DR6 (Cui et al. 2012). This sample covers the magnitude range of $9 \leq G \leq 15.6$ in the *Gaia* G-band. This spectroscopic information can then be used as the data set our NN is trained to estimate from photometric data. We acknowledge that while broad-band photometry will be sensitive to overall stellar metallicity, we use spectroscopic iron abundance as an accurate proxy for this value.

We removed objects with poor spectroscopic data by excluding sources with $\sigma_{T_{\rm eff}}/T_{\rm eff} > 1$ and $\sigma_{\log g}/\log g > 1$. As the range of metallicities in the training data crosses [Fe/H] = 0, using fractional uncertainties causes us to filter valid objects with small absolute metallicities. Thus, we do not apply this filter to metallicity. We instead incorporate training data uncertainties as part of the NN's training process (as described in Section 2) which accounts for metallicity uncertainties in the training data set.

Together, these two spectroscopic surveys provided a large sample of objects, mainly due to the large sky region and depths observed by the LAMOST survey. We were therefore confident our training sample had high-quality metallicities with minimal bias from spatially unbalanced data sets. We note that, thanks to calibration between giant stars in LAMOST and APOGEE data sets, the two spectroscopic surveys shared a good agreement with their metallicity observations (Anguiano et al. 2018). Thus, while small discrepancies

may occur, we felt confident using the two surveys concurrently. Additionally, in situations where objects appear in both APOGEE and LAMOST, we preferred the higher resolution APOGEE data and included only this value in our sample.

4.2 Methodology

We built our network with architecture described in Section 2, and selected input features constructed from 16 photometric colours and eight absolute magnitudes (as described in Appendix A). For training, we used the input features alongside our TG sample's spectroscopic metallicities to optimize the network to predict metallicities from photometry.

A major deviation from the method used in Section 3 was the inclusion of an extra weighting term to the network's loss function which worked to down-weight objects with [Fe/H] ≈ 0. This aimed to oppose the significant overabundance of near-solar metallicities in our TG sample. Without mitigation, the network would learn this imbalance as a trend in the data, and return values which follow this bias. Thus, we would have expected the algorithm preferentially return metallicities close to zero, as (when averaged over the entire data set) these predictions would be generally accurate.

Our weighting term took the form of a linear multiplier on the network's loss function. Modifying equation (3), with this weighting term as $W = |[\mathrm{Fe/H}]| + \mathrm{C}$ (where C is a constant), the weighted loss function is

$$J(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{W}{2}(y_i - \hat{y}_i)^2 e^{-s_i} + \frac{1}{2}s_i. \quad (11)$$

This weighting acted to decrease the 'loss' penalty when training on objects with [Fe/H] ≈ 0, and increased the penalty linearly for objects with much larger or smaller metallicities. Thus, the network put less effort into accurately predicting objects with solar-like metallicities, as penalties were significantly smaller for poor estimations. The weighting was also tuned with the constant, C, which changed the minimum (and maximum) weight an object can be allocated. For our training, we selected C = 0.5, such that the penalty multiplier for an object with [Fe/H] = 0 was $\times$ 0.5 and an object with [Fe/H] = -2 was $\times$ 2.5. We chose this value to increase the network's sensitivity to very low- and high-metallicity objects, while reducing the priority of metallicities between $-0.5 <$ [Fe/H] < 0.5 (the metallicity region of the majority of our TG sample). This ensured that objects with near-solar metallicities still retained a small impact on the network training, while maximizing the relative weighting between the high and low ends of the metallicity range.

One small side-effect of this weighting procedure was a reduction in accuracy for objects with [Fe/H] ≈ 0, due to the network considering them as lower priority. However, this had a negligible effect on the overall prediction accuracy: The larger population of objects with [Fe/H] ≈ 0 somewhat offset this effect, while the improvements to high-/low-metallicity predictions provided much more significant enhancement.

We also note that the inclusion of the weighting term in equation (11) may have also caused a small increase in the uncertainties output by the network. As s_i incorporates the predictive uncertainty of the NN, the network may have returned slightly larger uncertainties to account for the weighting term. For objects at high- and low-metallicities, which would be most affected by the weighting term, this uncertainty increase would be the most severe. In this case, we would expect the potency of the weighting term's bias-reduction would have been reduced.

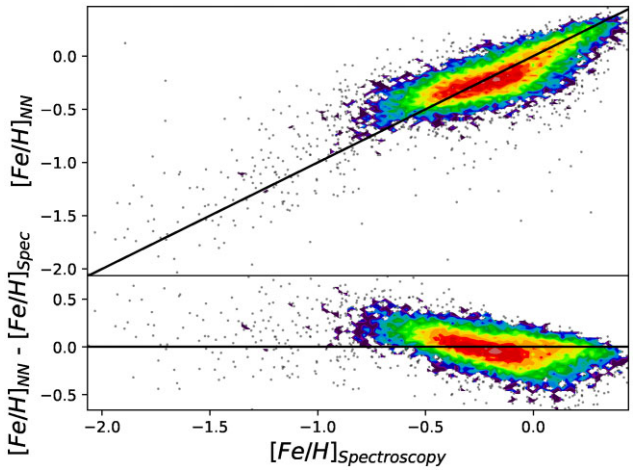

Figure 8. Plot of NN-estimated photometric metallicities from our cross-validation versus metallicities from spectroscopic observations (above), with comparison residuals (below). Note, this contour plot is logarithmic.

The success of this approach was not perfect, as we found that uncertainties still vary with respect to predicted metallicity. As shown in Fig. 8, even with the weighting term included, the prediction uncertainty was far larger for the highest and lowest metallicity objects. However, for the majority of our sample, the uncertainties remained small enough to be sufficient for our purposes.

There are two potential approaches to mitigate this in future work: more complex weighting criteria, to better reduce the impact of unbalanced data; or observing a greater number of objects with extremely high/low metallicities. While weighting may work to successfully mitigate this issue in some instances, removal of the imbalance altogether would be preferred, which can only be achieved through the latter of these two solutions.

The use of narrower bands, especially those bluer than in our data, may form a notable improvement over using broad-band photometry alone. The benefits of these bands for measuring stellar parameters have been shown by Keller et al. (2007) and Arentsen et al. (2020). However, the extreme extinction effects in these bands within regions such as the mid-plane or central bulge add additional complexities to their inclusion into our data set.

4.3 Validation

To validate the prediction accuracy, we had two measures: the uncertainty output calculated by the NN, and its performance compared against spectroscopically determined metallicities.

From the network's uncertainty measure, we found a very high confidence in the metallicity predictions being made. We returned a mean uncertainty output of ± 0.185 dex over our entire sample. We show our metallicity uncertainties binned by predicted metallicity in Table 2. This reiterates the correlation shown in Fig. 9. It is clear that within the range $-0.5 <$ [Fe/H] < 0.5 our predictions perform the best with an uncertainty of ± 0.15, and we have worse performance at low metallicities ([Fe/H] < -1.5). We also found that, while there is a large tail of high uncertainty predictions, these outputs only make up a small fraction of our entire sample: 97.49 per cent of our PO sample has uncertainties below ± 0.5 dex. In comparison, the spectroscopic metallicities in the range $-0.5 <$ [Fe/H] < 0.5 have a mean uncertainty of ± 0.046 dex, meaning our best-case metallicities have uncertainties about three times that of the spectroscopic data. We find these uncertainties are comparable to the results of other

Table 2. Table of mean metallicity uncertainties for binned metallicity ranges. Note that the top-most row shows the mean metallicity uncertainty over the entire sample.

[Fe/H] bounds	[Fe/H] unc.	Obj. counts
$-3.5 <$ [Fe/H] < 1.5	$\pm$ 0.185	1 697 077
$-3.5 <$ [Fe/H] < -2.5	$\pm$ 4.042	713
$-2.5 <$ [Fe/H] < -1.5	$\pm$ 1.713	12 392
$-1.5 <$ [Fe/H] < -0.5	$\pm$ 0.404	145 177
$-0.5 <$ [Fe/H] < 0.5	$\pm$ 0.150	1 538 047
$0.5 <$ [Fe/H] < 1.5	$\pm$ 0.537	736
$1.5 <$ [Fe/H] < 2.5	$\pm$ 1.009	12

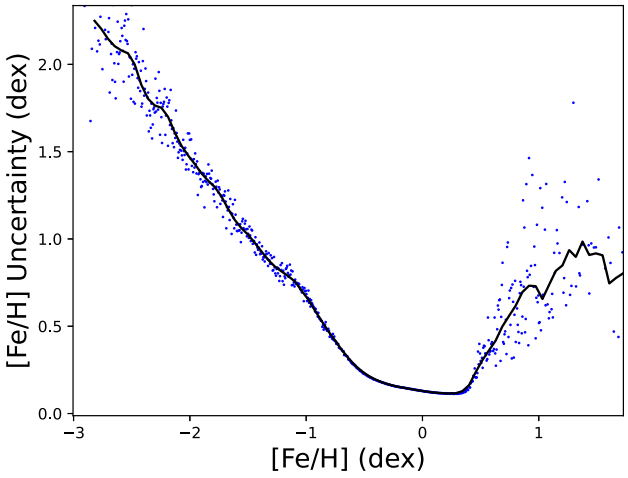

Figure 9. Plot of metallicity uncertainty against metallicity for our NN-estimated photometric metallicity values. The lowest uncertainty predictions are those with absolute metallicity close to 0.0 dex (the highest population region), with uncertainties becoming more significant at the edges of our distribution.

photometric metallicity methods, with Grady et al. (2021) finding an uncertainty of $\pm$0.21 ([Fe/H] > -0.5), Huang et al. (2022) finding an uncertainty of approximately $\pm$0.12 ($-0.5 <$ [Fe/H] < 0.5), and Lin et al. (2022) finding an uncertainty of $\pm$0.2 dex. Furthermore, we directly compare our metallicities to those from Huang et al. (2022) and Lin et al. (2022) in Appendix B, and find reasonable agreement to these methods.

We note that the range of Fig. 9 extends beyond the metallicity range of our TG sample. The minimum metallicity from the spectroscopically measured giant stars was -2.49 dex, and the maximum metallicity being 0.74 dex. Outside of this range, the NN must extrapolate beyond the training data – and thus causes returned uncertainties to be very large. This is most apparent above 0.74 dex, where uncertainties become extremely large beyond the extent of the training data. We therefore recognize that metallicities at the extreme edges of our metallicity distribution should be ignored in further analyses (either by specific cuts to metallicity, or by filtering for extreme metallicity uncertainties).

We further compared the predictions made by our NN to metallicities from APOGEE and LAMOST, providing a measure of the 'recovery accuracy' of the network. This worked to cross-check the uncertainty values outputted by the NN, ensuring that the network retains its high accuracy when compared to 'true' data values. This validation is achieved through a method of out-of-bag cross-validation. We selected a fraction of our TG sample to be removed from the network's training process, which we then used

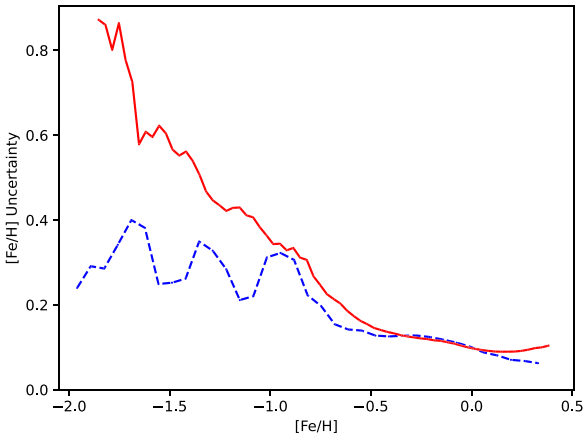

Figure 10. Plot of NN metallicity uncertainties (solid red) and the residuals between the NN and spectroscopic metallicities (blue dashed) from Fig. 7's lower panel, plotted against absolute predicted metallicity. These uncertainties diverge notably from the trend shown in Fig. 9, as these are objects from our spectroscopically-matched TG sample – and thus, tend to be closer and brighter than many objects in our output sample.

to validate the model's predictions. We chose a validation sample split of 15 per cent of our TG sample, leaving 85 per cent to train the network. The network's predictions on the validation sample were then compared to the spectroscopic measurement, with the comparison shown in Fig. 8. We find there is a good correlation between our method and the spectroscopic data, suggesting our approach is successful in accurately reproducing metallicity values. However, we do confirm the minor bias apparent in the residuals at high and low metallicities, with an overestimation of metallicities below [Fe/H] < -0.5 and a smaller biasing of underestimated metallicities for high [Fe/H] objects. This 'regression to the mean' effect is a common issue for NN algorithms using unbalanced data sets, and so suggests our weighting term has not fully removed these effects. Analyses using lower metallicity objects must take this into account.

We finally analyse the effect the weighting term may be having on the predicted metallicity uncertainties, as noted in the previous section. We compare the NN's output uncertainties to the residual scatter in the lower panel of Fig. 8. If the network is predicting larger uncertainties due to inclusion of the weighting term, we would expect the output uncertainties to be much larger than the scatter in the residuals. We plot this in Fig. 10. Note that we have significantly fewer objects at [Fe/H] < -1 (137 objects) than for [Fe/H] > -1 (15 430 objects), and so our trends are poor beyond this threshold. This figure shows clearly that, for the metallicity range where we have large numbers of objects, we see a good agreement between NN uncertainties and residual scatter. Thus, we conclude that the weighting term does not appear to be causing the NN uncertainties to be output significantly larger than expected. Furthermore, we note that the uncertainties shown in Fig. 9 and Table 2 may be overestimated at the low-[Fe/H] regime, as they are significantly larger than we would expect from the residual scatter trend.

5. RESULTS AND ANALYSIS

With the completion of the metallicity estimation, we returned our PO sample of 1 689 885 objects with: *Gaia* astrometry; eight photometric colours from *Gaia*, 2MASS, and WISE; kinematic information from *Gaia* proper motions and radial velocities; and

photometric metallicity estimations. Furthermore, we calculated 3D Galactocentric coordinates and velocities based on our distance estimates. Using a right-handed coordinate system, we converted *Gaia* astrometry (sky positions and velocities) into Galactic positions and velocities. In this system, the *X*-axis is along the Sun–Galactic Centre (GC) direction with positive towards the GC. The longitudinal axis, *Y*, sits perpendicular to *X* along the Galactic plane, with positive *Y* in the direction of positive Galactic longitude. The vertical axis, *Z*, is directed out of the Galactic plane with positive towards Galactic north. All axes have their origin at the GC.

We applied our catalogue to determine the out-of-plane metallicity gradient of the Galactic bulge, and to identify the vertex angle of the Milky Way's bar from stellar kinematics and metallicity.

5.1 Vertical metallicity gradient in the Galactic bulge

The presence of a metallicity gradient, vertically out of the Galactic plane, in the region of the bulge has been identified in many previous studies. This gradient is suggested by some to be the effect of overlapping populations within the bulge region (Barbuy et al. 2018). These intersecting structures include bulge and bar populations, as well as the surrounding disc and halo structures. As we observe away from the Galactic plane, we see the changing influence on each of these independent components, which creates a gradient in the observed metallicity distribution. Alternatively, other work proposes that this gradient instead forms from the kinematic separation of different populations during the formation of the bulge and bar (Debattista et al. 2017). Due to bursts of star formation during bulge formation, populations of metal-poor and metal-rich stars can become separated kinematically into hotter and colder velocity distributions. This causes a metallicity gradient to be observed, without the need for distinct, overlapping populations. As summarized by Ness & Freeman (2016), gradients have been observed in past literature of around -0.45 dex kpc^{-1} (Ness et al. 2013), with some methods observing as low as -0.6 dex kpc^{-1} (Zoccali et al. 2008) and as high as -0.35 dex kpc^{-1} (Minniti et al. 1995).

To draw this trend from our data, we first defined our selection region of the 'bulge'. Using our 3D Galactocentric Cartesian coordinates, we defined our bulge region to be within 2.5-kpc radius (along the plane) of the GC – selecting a cylindrical volume centred on the GC. We also applied filtering on selected objects, removing stars with metallicity uncertainties greater than ±0.5 dex, positional uncertainties greater than ±1 kpc, and velocity uncertainties greater than ±250 km s^{-1}. We note that the filter on metallicity uncertainty will ensure we are only selecting objects with 'good' metallicity estimations, but will also introduce a bias into the gradient observed. Filtering out objects with metallicity uncertainty greater than ±0.5 dex will predominantly remove objects with [Fe/H] < -1 dex and [Fe/H] > 0.5 dex. Thus, the trends we observe in metallicity will potentially ignore populations of high- or low-metallicity stars that would otherwise shift the mean metallicity at a chosen position in the Galaxy, and cause our recovered gradient to be under or overestimated.

Initially, we plot stellar metallicity against object height above/below the Galactic plane, *Z*, for the PO sample in the left-hand panel of Fig. 11. It is clear the trends visible are very noisy, with a large uncertainty across the range of *Z*-values shown. We find this is a limitation due to the small number of objects with *Gaia* radial velocity information within the volume, limiting us to only 22 280 objects. This small subsample leads to a large scatter in median metallicity with *Z*-height, and reduces the strength of trends we can draw. Fortunately, we do not need velocity information to draw a positional metallicity gradient, and thus we could remove this requirement when collecting our data set and expect to see more objects within the sample volume.

Including objects without radial velocities, we applied our NN to estimate metallicities for the larger sample. This is shown in the middle panel of Fig. 11, which mirrors the left-hand panel while showing a much stronger trend across the *Z*-height range. As we expect the gradient to be symmetric above and below the plane, we further plot the median metallicity against the absolute *Z*-height in the right-hand panel of Fig. 11, which increases the strength of observed trends and further constrains the level of uncertainty in a measured gradient.

We also note the clear gradient inversion visible within 500 pc of the plane. This appears to retain the well-constrained uncertainties between approximately 250 and 500 pc, before the trend becomes extremely scattered towards the mid-plane. This is unexpected, as Gonzalez & Gadotti (2016) note that many past works with spectroscopic data have recovered a smooth metallicity–height relation, from lower metallicity objects far from the plane, and higher metallicity objects towards the mid-plane.

This metallicity gradient change towards lower latitudes has been noted by Rich, Origlia & Valenti (2012), who observed the vertical gradient flattens below a vertical height of 550 pc. Furthermore, Babusiaux et al. (2014) found hints that the gradient indeed inverts close to the plane. This is proposed to have been due to early-forming stars becoming trapped in the inner regions of the Galaxy as it formed, and remaining bound in the mid-plane during bar buckling. Alternatively, this low-metallicity core population may be the result of metal-poor gas being funnelled into the bulge by the bar, forming this metallicity inversion towards the mid-plane. We therefore find our results agree with these past findings, and confirm the presence of a low-metallicity population towards the mid-plane.

We compute a metallicity gradient between $700 \leq Z \leq 1600$ pc, and return a value of -0.5278 ± 0.0022 dex kpc^{-1} (outwards from the galactic plane). It is useful to note that we are assuming a linear relationship between metallicity and *Z*-height within the quoted range only, and so does not account for the gradient flattening at values of *Z* outside of our selection.

We find a vertical metallicity gradient that is well within the literature range of values, although towards the steeper end. This suggests our observed metallicity distribution diverges significantly from that found by Minniti et al. (1995) and slightly from that of Ness et al. (2013). We suggest that such a discrepancy is expected between our method and those that use spectroscopic data. Due to the selection criteria used by spectroscopic surveys, we expect to find our photometric-based data to be sampling a slightly different stellar distribution, and return a slightly different metallicity gradient.

Furthermore, we note that the metallicity bias described in Section 4.3 may have also biased our recovered gradient. As we expect low-metallicity populations to be more common at higher latitudes, we expect the metallicity overestimation bias to have a stronger impact further from the plane. This would cause our gradient to be measured as shallower, as the mean metallicity at higher latitudes would be increased, while the metallicity of the mid-plane would remain mostly unchanged.

Overall, our main conclusion from this analysis remains that our data has successfully returned positions and metallicity estimations, which accurately trace known abundance trends within the Milky Way.

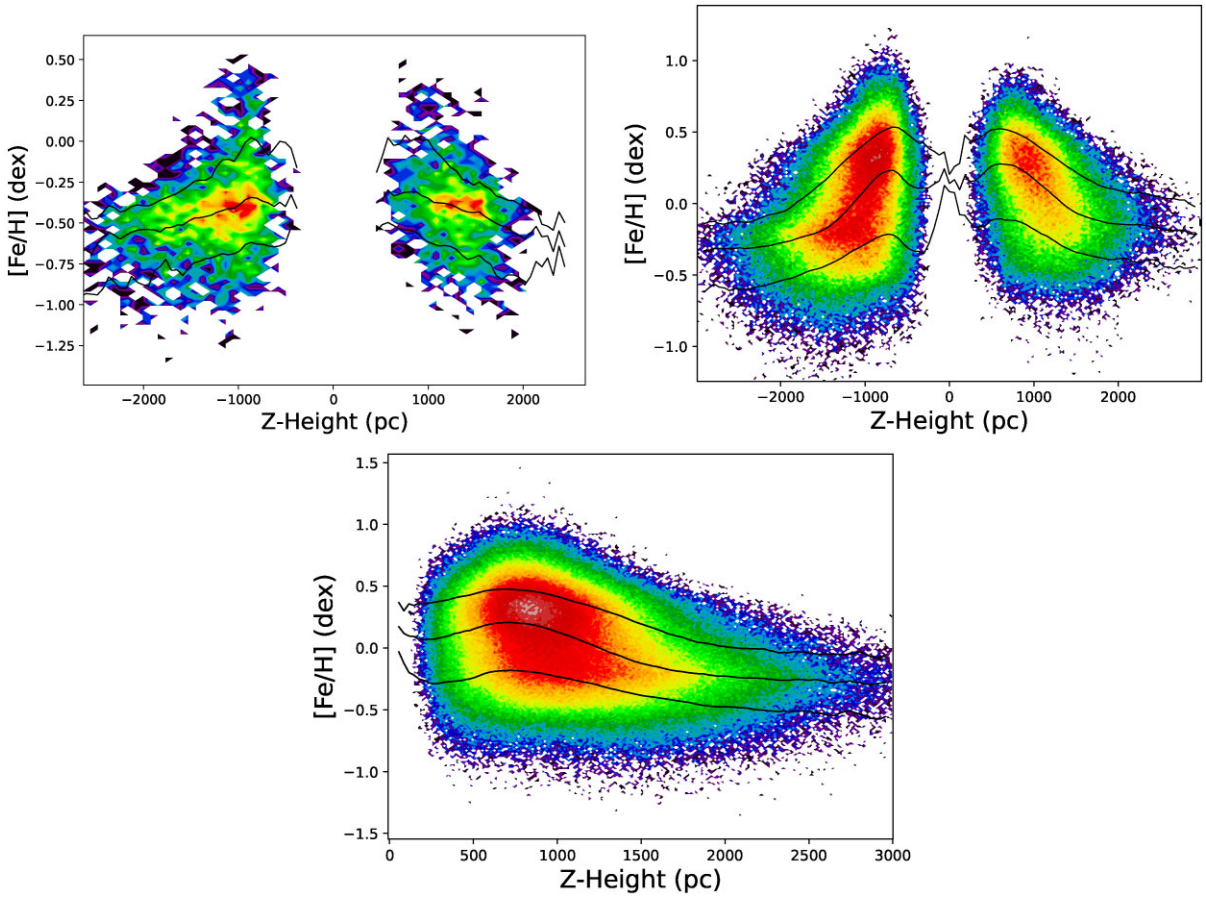

Figure 11. Logarithmic contour plot of median metallicity against an object's height, Z, for objects within the bulge selection volume. We plot objects with radial velocity information (left-hand panel), objects without radial velocities (right-hand panel), and objects without radial velocities plotted against *absolute* Z-height (bottom panel). For the bottom panel, we find the median curve to peak at approximately 560 pc with a metallicity of 0.173 dex. Note that the upper and lower curves are the 84th and 16th percentiles, respectively.

5.2 The vertex deviation of the bar

As we have radial velocities from *Gaia* DR2, we have also been able to use our catalogue to analyse the kinematics of the Galactic bar-bulge. One quantity useful for probing the kinematic properties of the bar is the vertex angle or vertex deviation, that is the angle of the major axis of the velocity ellipsoid relative to the GC direction giving an indication of the orientation of the bar (Zhao, Spergel & Rich 1994).

Vertex angles, l_v, are defined as

$$l_v = \frac{1}{2}\arctan\left(\frac{2\sigma_{\rm XY}^2}{\sigma_{\rm X}^2 - \sigma_{\rm Y}^2}\right), \tag{12}$$

where $\sigma_{\rm X}^2$ is the velocity dispersion on the Galactocentric X-axis, $\sigma_{\rm Y}^2$ is the dispersion in the Y-axis, and $\sigma_{\rm XY}^2$ is the correlation term. The angle is calculated from the Sun–GC line such that the value is within $|l_v| \leq 45^\circ$. The angle is positive in the direction of positive Galactic longitude (anticlockwise rotated bar), and negative in the direction of negative longitude (clockwise rotated bar). For an axisymmetric velocity distribution, the vertex angle is ill-defined (as the major and minor axes are equal) and we would expect any measurement to be unconstrained.

Existing literature has measured this value, and found a bar-like signal for metal-rich bulge objects. Zhao et al. (1994) note a vertex angle of $-65 \pm 9^\circ$ (for [Fe/H] $\geq$ 0.0), while Babusiaux et al. (2010) measures an angle of $-32 \pm 9^\circ$ (for [Fe/H] $\geq$ 0.3). Both studies also found that low-metallicity objects show a high-scatter, near-zero vertex angle — and thus not a bar-like signal. This suggests spherical or disc-like rotation in these metal-poor populations.

We note that, for our analysis, we use our Galactocentric coordinate system (X/Y/Z) to determine the angle of the velocity ellipsoid, rather than the usual Galactic coordinates (r/l/b). For the sky region of interest, these are broadly equivalent, but using this definition does alter the returned 'vertex angle' in comparison to past literature. We selected these coordinates as it ensures all object velocity vectors have parallel axes, which is not the case when using Galactic coordinates across large sky regions. Note that the orientations of our axes are described at the end of Section 3.2.

For this method, we developed a Bayesian inference process using a STAN implementation in Python (CmdPyStan, Stan Dev Team 2021). We constructed a Markov Chain Monte Carlo method (MCMC), which accepted 3D velocity vectors, $\boldsymbol{v}_i$, (in the XYZ coordinate system) and corresponding velocity uncertainty covariance matrices, $\Sigma_{\rm unc}$ and evaluated the log-likelihood as a two-component Gaussian mixture model with distribution means, $\boldsymbol{\mu}_n$, and covariance matrices $\Sigma_{{\rm v},n}$. Such a mixture model allows us to isolate a minor 'anomalous' component from the data, and return a stronger signal of interest. For this method, we evaluated the log-likelihood of each

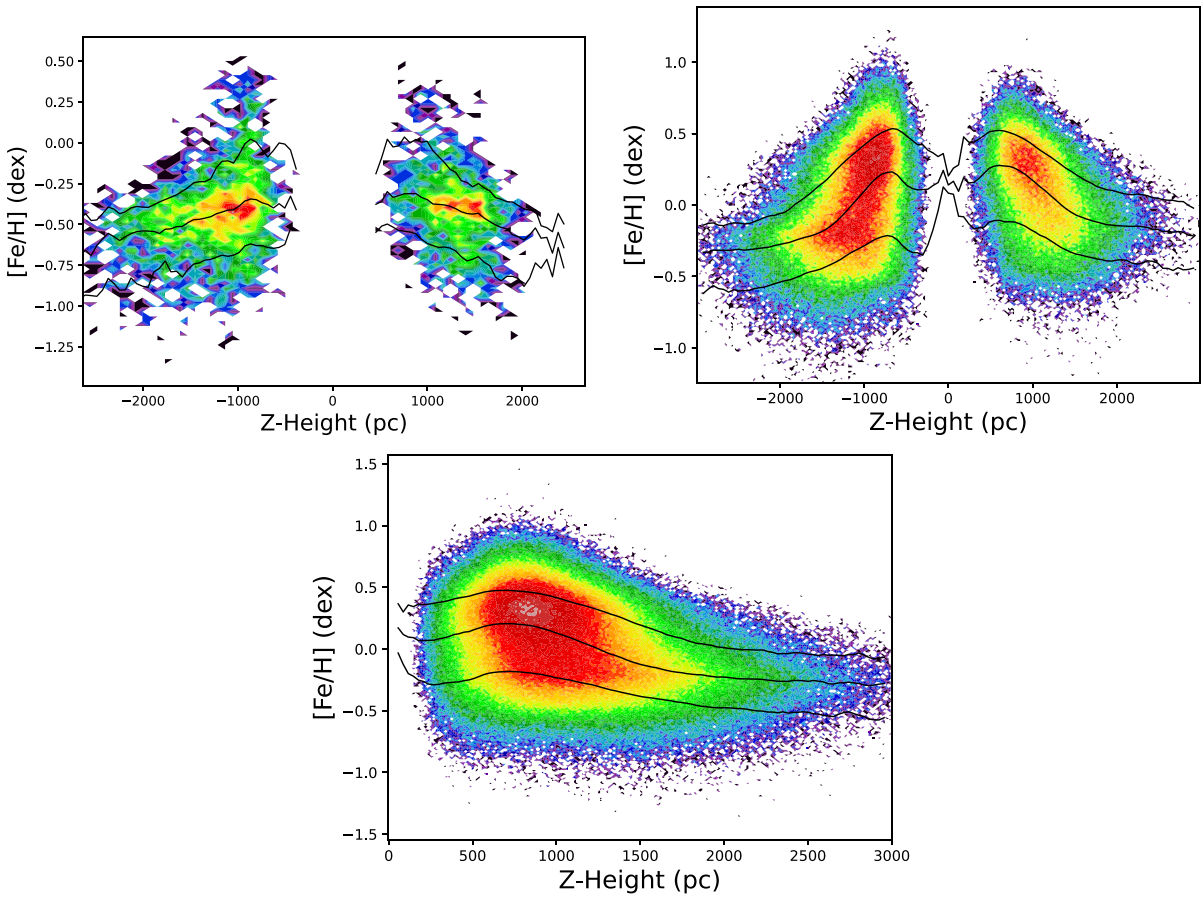

Figure 12. Plot of vertex angle calculated for seven metallicity bins, drawn from ∼4200 objects within the selected bulge region. The angles plotted in *red* are measured with the velocity ellipsoid centred at the mean velocity of the data, while angles plotted in *green* are measured with the velocity ellipsoid centred at the velocity of the GC. The bin object counts, from low- to high-metallicity, are: 4, 73, 600, 1882, 981, 418, and 56.

Gaussian component as

$$\ln \mathcal{L}_n = -\frac{1}{2}\sum_i \left((\boldsymbol{v}_i - \boldsymbol{\mu}_n)^{\mathrm{T}} (\Sigma_{\mathrm{v},n} + \Sigma_{\mathrm{unc},i})^{-1} (\boldsymbol{v}_i - \boldsymbol{\mu}_n) + \ln |\Sigma_{\mathrm{v},n} + \Sigma_{\mathrm{unc},i}| \right) \quad (13)$$

The two components are thus evaluated to assign member stars, with the two-component distribution given by

$$\mathcal{L}_{\mathrm{total}} = \lambda \mathcal{L}_a + (1-\lambda)\mathcal{L}_b, \quad (14)$$

where a and b denote the two components, and λ is a ratio of the two components where $0 \leq \lambda \leq 1$, and $\mathcal{L}_{\mathrm{total}}$ is the overall log-likelihood.

From the major component in the mixture model, we infer the mean velocity, $\boldsymbol{\mu}$ and the velocity ellipsoid, Σ_{v}, which has components

$$\Sigma_{\mathrm{v}} = \begin{pmatrix} \sigma_{\mathrm{X}}^2 & \sigma_{\mathrm{XY}}^2 & \sigma_{\mathrm{XZ}}^2 \\ \sigma_{\mathrm{XY}}^2 & \sigma_{\mathrm{Y}}^2 & \sigma_{\mathrm{YZ}}^2 \\ \sigma_{\mathrm{XZ}}^2 & \sigma_{\mathrm{YZ}}^2 & \sigma_{\mathrm{Z}}^2 \end{pmatrix}, \quad (15)$$

from which the vertex angle can be calculated using equation (12). As the MCMC method returns a distribution of covariance matrices, we output a distribution of vertex angles. From this, we calculated a median value and percentile uncertainties for our vertex angle.

We selected our sample as an on-sky region, with $|l| \leq 5^\circ$ and $|b| \leq 10^\circ$, and limited to distances between 6 and 10 kpc. This forms a volume approximately 700-pc wide on the Y-axis, 1.5 kpc on the Z-axis, and 2-kpc deep in the X-axis. We note here that this volume-based sample selection is strongly affected by the distances, and thus the distance uncertainties, reported for each star. We therefore filtered our objects for only those with distance uncertainties smaller than ±1 kpc, to limit the effect of non-bulge/bar objects with large uncertainties being included in the selection. We also applied filters on extreme metallicity uncertainty ($[\mathrm{Fe/H}]_{\mathrm{unc}} < 1.0$), extreme Galactocentric velocity uncertainty (velocity unc. $<$ 250 km s^{-1}), and perpendicular axis (Y and Z) positional uncertainty (position unc. $<$ 1 kpc). This selection region is strongly limited by the maximum depth of objects with radial velocity information, which causes our sample to be predominantly objects on the near-side of the bulge, with far fewer objects at greater distances.

To identify the cutoff between high- and low-metallicity samples, we split our sample by metallicity into bins of width 0.35 dex between the range of −1.65 and +0.8 dex. Due to the associated uncertainties at extremely high- and low-metallicities, we observe very few objects with $[\mathrm{Fe/H}]_{\mathrm{unc}} < 1.0$ outside of this range. We note that for objects beyond the range $-0.5 < [\mathrm{Fe/H}] < 0.5$, the metallicity uncertainty is larger than the 0.35-dex bins we use in Fig. 12. We therefore expect the bins within this 'good' range of metallicities to be accurately binned with predominantly objects within the bin range and little contamination. However, for metallicity ranges with higher mean metallicity uncertainties, we expect contamination to be higher between bins. This would cause us to return angles with large

uncertainties in these bins, as the contaminant objects will bring a larger distribution of object kinematics.

The vertex angle was calculated within each bin. This binned calculation is plotted in Fig. 12. For low-metallicity bins, the vertex deviation is approximately zero, whilst for higher metallicities a large negative angle is found.

We further compare these values against a slightly modified vertex angle calculation, where instead of centring the velocity ellipsoid on the fitted mean velocity of the data, we assume the mean velocity is zero. As our sample appears to be biased towards objects on the near-side of the bulge, we find that high-metallicity objects have a mean velocity dominated by Galactic rotation. Centring the ellipsoid on this mean removes this net motion from our vertex deviation calculations, and gives a better fit to the kinematic data.

However, this is not the only approach to determine the vertex angle from objects kinematics. If our data was more evenly distributed across the bulge region, rather than predominantly on the near-side, we would expect to find the mean of the velocity distribution close to zero in all axes (rather than dominated by galactic rotation). We therefore estimate the vertex angle with the ellipsoid means 'zeroed', to emulate the angle we would return from an unbiased sample. This 'zeroed' approach will likely be a poorer fit to our biased data set, but we find the comparison useful to understand the angle we expect to observe with a kinematically unbiased sample.

We note that our initial method with the means estimated by the algorithm is noted as the 'fitted' ellipsoid, while the method with ellipsoid means constrained at zero is the 'zeroed' ellipsoid.

From this binned selection we selected two main samples: a high-metallicity bin, and a low-metallicity bin. This maximized the number of objects used to calculate the vertex angle, and limited the potential uncertainties from small sample sizes. We selected our low-metallicity bin where [Fe/H] ≤ -0.7 and our high-metallicity bin where [Fe/H] ≥ -0.4. Using these we retained large samples, and recovered vertex angles with minimized scatter.

In the region between these two bins, where $-0.7 \leq$ [Fe/H] ≤ -0.4, we found a mean angle of -18.449 ± 8.644 deg, which fell between that of the high- and low-metallicity bins and retained a large uncertainty value. This suggested we were seeing an overlap of the two regimes, where scatter in metallicity predictions make it difficult to differentiate the distinct kinematic profiles. We therefore excluded this region from our analysis, and focused on the selected high- and low-metallicity samples.

We note that contamination between these two bins will be less severe than for the smaller bins used in Fig. 12. As we separate our two bins with the intermediate $-0.7 \leq$ [Fe/H] ≤ -0.4 region, there will be few objects with uncertainties extreme enough to contaminate the other bin. The most significant issues will arise for objects with [Fe/H] < -1.5 dex, as their uncertainties become large enough to potentially contaminate the binning. However, as noted previously, we have a very small sample of objects with these very low metallicities. Thus while there may be contamination, we expect this to have a minor influence on the angles calculated.

With the metallicity ranges set, we re-applied the Bayesian model to draw a vertex angle for each of these two samples. These results are shown in Table 3 and show that there was a clear difference between the velocity distribution of the low- and high-metallicity samples. The low-metallicity objects appear to have a small vertex angle with a high uncertainty, suggesting minimal bar-like signal in the data. Conversely, we show a much more negative, low uncertainty vertex angle present in the high-metallicity sample, with an angle of -21.29 ± 2.74 deg.

Table 3. Results of vertex deviation calculations of the Milky Way's bulge. The vertex angle is the value returned by the analysis, the uncertainty is the difference between the 16th and 84th percentiles of the angle distribution. We also include the range (1st to 99th percentile) of the distribution, to illustrate the edges – and thus the broadness – of the two prediction distributions.

[Fe/H] range (dex)	[Fe/H] ≤ -0.6	[Fe/H] ≥ -0.4
Vertex angle (deg)	6.8526	−21.2896
Angle uncertainty (deg)	± 16.4271	± 2.7367
Angle range (deg)	± 108.9450	± 13.2134

Our results therefore confirm the kinematic split of bulge populations by metallicity. We observe that lower metallicity bulge objects show more axisymmetric kinematics around the GC, suggesting they are populations found in the spheroidal-shaped bulge or thick disc. On the other hand, higher metallicity objects show a large, low-uncertainty vertex angle, suggesting these objects instead have a bar-like kinematic structure, and so will be members of the Milky Way's bar population.

However, the high-metallicity vertex angles are much lower than those found by past works, with our measured angle being around ∼10 deg smaller although still within the uncertainties of the measurement from Babusiaux et al. (2014). This suggests that we either measure a bar that is rotated to a smaller angle than prior works, or a weaker bar-like signal from a more axisymmetric velocity distribution.

We apply our 'zeroed' approach to these high- and low-metallicity samples, with both the 'zeroed' and fitted mean distributions shown in Fig. 13. It is clear that, for the low-metallicity sample, both the fitted means and the 'zeroed' means trace a similar spherically symmetric distribution, centred on the origin. On the other hand, for the high-metallicity sample, the two distributions diverge significantly, with a vertex angle of -47.32 ± 3.93 deg, compared to the fitted ellipsoid's vertex angle of -21.29 ± 2.73 deg. While the fitted ellipsoid is the better fit for the data set, it is noticeably more spheroidal than the much more extended 'zeroed' ellipsoid.

While the 'zeroed' approach is a slightly poorer fit to our data set, this larger angle is much closer to the vertex angle calculated by past works (from −32 deg to −65 deg). This suggests that while our ellipsoid with fitted means is a better fit to our data set, the velocity bias present means we return an angle that is smaller than expected. Our 'zeroed' ellipsoid being closer to the expected vertex angle suggests that mitigation of this bias is necessary to fully recover the bar vertex angle. This can be done either through centring the ellipsoids on zero in all axes, or by building a data set with greater depth to ensure a more balanced distribution of objects across the bulge. In this case, we would expect the means of the velocity distribution to tend towards zero, and so we would see a distribution closer to that of the 'zeroed' ellipsoid.

We do however conclude that we detect a clear difference in vertex angles measured for our high- and low-metallicity samples. The observation that galaxy bar-populations are metal-rich (in comparison to other bulge components) has been discussed by Wegg et al. (2019), who suggest this describes a formation process where the bar is formed from higher metallicity, kinematically cool stars which orbit outside of the central bulge, and thus form this separate population within the GC. They also note that a metal-rich bar has also been identified in other nearby galaxies (Gadotti et al. 2019), notably including M31 (Saglia et al. 2018).

Overall, we can confirm the success of our method in recovering this known bar-like signal from objects from photometrically esti-

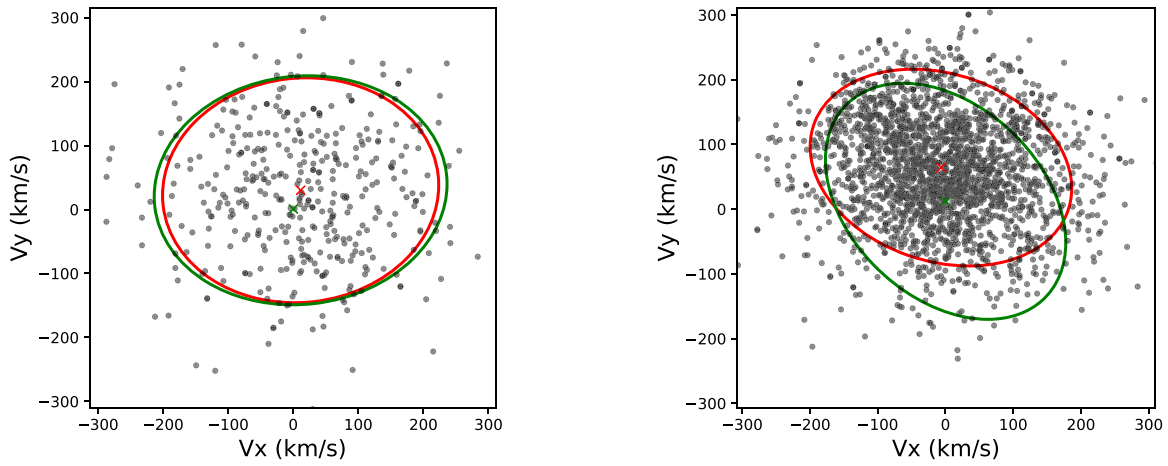

Figure 13. Plot of the velocity distribution on the x (Vx) and y (Vy) axes. The fitted velocity ellipsoids are shown, where the ellipsoids means are fitted to the sample mean velocity (*red*) and the origin of the plot (*green*).

mated distances and metallicities. We are also able to highlight the utility of our approach to be applied to structures like the Milky Way's bar, where debates on the metallicity and kinematic distributions are ongoing.

5.3 Limitations

We note there is a limitation in our approach to selecting a bulge sample for our two analyses. Our approach in both cases was to select target volumes using cuts in either on-sky Galactic coordinates or Galactocentric positions. These approaches predominantly selected bulge objects within the chosen region of the bulge, and so forms our stellar population of interest. However, we did not make any attempt to isolate any specific population or Galactic component. We therefore note that these selections contain non-bulge populations which overlap the chosen spatial region, such as from the Milky Way's disc or halo. In future work, we hope to include a more robust selection approach, which would account for additional parameters like stellar types or kinematics and allow us to target specific populations with specific analyses.

Our distance cuts also must account for the potential bias between metallicity and distance. This bias occurs due to low-metallicity stars being brighter than higher metallicity stars of the same effective temperature (Ahumada et al. 2020; Chiti et al. 2021). Low-metallicity stars are then overselected at greater distances, especially beyond 5 kpc from the Sun (as noted by Chiti et al. 2021). However, as our volume-based sample selections collect only a small range of possible distances, the near- and far-sides of out samples will have had approximately similar numbers of overselected low-metallicity stars. We therefore expect this bias to have only had minor effects on our analyses.

Furthermore, we also note a limitation in how we filtered our samples by metallicity uncertainty. As we wished to focus on how metallicity correlates with object positions and kinematics, we attempted to focus only on objects with 'good' metallicity measurements. However, as was noted in Section 4.3, our metallicity uncertainties vary with absolute metallicity. Therefore any filtering by metallicity uncertainty introduces a bias in our sample, due to removing very high- or low-metallicity objects. This bias was unlikely to cause a major deviation in the trends we observed, as 97.5 per cent of our output sample has metallicity uncertainties smaller than ± 0.5 dex. However, we note that the trends we observe are most strongly applicable to objects with solar-like metallicity, and may not fully account for very high- or low-metallicity populations.

6. CONCLUSION & DISCUSSION

Our method to determine metallicity information from photometric information was built on a three-step process: We first built a NN algorithm which enhanced *Gaia* parallax values with photometric information. This allowed us to determine distance estimations with greater accuracy, which we could then bring forwards to predicting metallicities. With accurate distances, we were then able to train our NN model to predict stellar metallicities from APOGEE and LAMOST spectra, allowing the NN to estimate metallicity from photometric colours and absolute magnitudes alone. From this, we could build a sample of objects with *Gaia* astrometry and metallicity information, allowing for analysis of the positional, kinematic, and metallicity trends in Milky Way populations. Finally, to test our method, we compared against known trends in the Milky Way. Firstly, we measured a vertical metallicity gradient within the Galactic bulge from our data, and compared this to known values in the literature. Then, we used a statistical model to estimate the vertex deviation of different metallicity populations in the Galactic bar.

6.1 Method improvements

Despite our confidence in our results, we still acknowledge there are some outstanding limitations in our method. The primary of these is the amount of data we had available to train the network. Due to the limited depth of *Gaia*'s (DR2) radial velocity data, there was a significant decrease in the number of stars available at large distances. This manifested quite clearly with our analysis in Section 5.2, where the limited object numbers in our bulge sample led to a lack of objects at the edges of our metallicity range. Therefore, analysis of the bar's vertex angle required us to choose large bins in our high- and low-metallicity regimes to maintain higher object counts. However, this approach increased the risk of contamination from overlap of the two regimes, as we attempted to maximize the available sample sizes.

Furthermore, our method can be applied to analyses that do not require kinematic information (such as gradients or trends in stellar positions), lacking kinematic information severely reduces

the trends and phenomena we can target in future. For example, without velocity information, our methodology would be unable to investigate the properties of bound groupings of stars (i.e. Milky Way structures, accreted substructures) where metallicity *and* kinematics are essential to identification and analysis.

With the *Gaia* DR3 release, the magnitude limit of the radial velocity is fainter (Gaia Collaboration 2021), and so we expect the number of objects we can return with full kinematic data will increase drastically. This would allow our primary output sample to be significantly larger, and permit us to expand the range of populations we can determine metallicities for. Furthermore, the addition of BP/RP spectra in the DR3 release will provide additional data from which our method can estimate metallicities.

We find another limitation in the lack of a good comparison sample for our metallicity estimations. While we are confident in our abundance estimates thanks to comparisons with APOGEE and LAMOST validation sets, the most robust comparison would be to use an independent survey sample. As we use these two spectroscopic samples as part of our training process, we cannot discern whether our metallicity outputs incorporate the biases or errors from these spectroscopic surveys. Thus, a comparison with an independent survey would ensure these biases could be accounted for.

In future work, we would be able to build a comparison sample from a selection of current and future surveys. In the immediate future, we could utilize the cross-match between our sample and the GALAH (Buder et al. 2019), RAVE (Steinmetz et al. 2020), or SEGUE (Yanny et al. 2009) surveys – each of which is on the scale of 10^4–10^5 objects, and would have notable overlap with our *Gaia*-based data. Furthermore, within the next couple of years, the large-scale WEAVE (Dalton et al. 2012) and 4-MOST (de Jong et al. 2019) survey releases would allow us to further compare our method against a wide selection of objects with high-resolution spectroscopic data. The *Gaia* DR3 release will also have the capability to estimate metallicities directly from BP/RP spectra, which could provide us with an additional sample of metallicities to compare our results against.

Additionally, we further hope to resolve the metallicity imbalance we see in our spectroscopic training data. As the majority of our training sample has near-solar metallicities, this creates a bias in our NN's predictions. While we add heavy weighting to our method to mitigate this, we still find some biasing in our estimations. To more robustly resolve this issue, we can instead augment our existing data with artificial objects, creating a more balanced data set from unbalanced samples. On one hand, this would require use of generative algorithms, such as variational autoencoders (Kingma & Welling 2019) or synthetic oversampling methods (Chawla et al. 2002), which would allow us to generate additional data which would be similar to our input samples. These algorithms could then be used to build samples with significantly more objects with extremely high or low metallicities, and thus work to mitigate the near-solar bias we see currently.

6.2 Final thoughts

Overall, we can conclude that our NN-based methodology has successfully estimated stellar properties that had previously been difficult to determine without spectroscopic data. Our approach retains high accuracy, with mean uncertainties (for $-0.5 < [\mathrm{Fe/H}] < 0.5$) of ± 0.15 dex. We return a catalogue (as described in Section 5) of 1.7 million *Gaia* objects with NN-enhanced distances, 3D kinematic information, and accurate metallicity information.

Future works will be able to leverage these results to draw conclusions which require very large samples with stellar abundance information. For example, the identification of substructure within the Milky Way would be an ideal target for our approach, as we have accurate distance, velocity, and metallicity measurements for our large sample. The detection of substructures such the '*Gaia*-Enceladus Sausage' (Belokurov et al. 2018) and 'Sequoia' (Myeong et al. 2019) merger remnants require the detection of a large number of bound low-metallicity objects to accurately define the structure's origin. While our method does not have the accuracy of spectroscopic approaches, the greater depth and objects counts we return can allow us to find deeper insights into these (and similar) substructures.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referee for their critiques, and their concise and pragmatic guidance.

JLS acknowledges support from the Royal Society (URF\R1\191555). This paper made use of the Whole Sky Database (wsdb) created by Sergey Koposov and maintained at the Institute of Astronomy, Cambridge by Sergey Koposov, Vasily Belokurov, and Wyn Evans with financial support from the Science & Technology Facilities Council (STFC) and the European Research Council (ERC). This research was supported in part at KITP by the Heising-Simons Foundation and the National Science Foundation under Grant No. NSF PHY-1748958. This project was developed in part at the 2019 Santa Barbara Gaia Sprint, hosted by the Kavli Institute for Theoretical Physics at the University of California, Santa Barbara. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical

Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University. Guoshoujing Telescope (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.

DATA AVAILABILITY

The data sets used in this article were derived from sources in the public domain: *Gaia* EDR3, https://gea.esac.esa.int/archive/; UnWISE, https://catalog.unwise.me/catalogs.html; 2MASS, https://irsa.ipac.caltech.edu/Missions/2mass.html; LAMOST DR6, http://dr6.lamost.org/catalogue; and SDSS DR17/APOGEE, https://skyserver.sdss.org/dr17.

REFERENCES

Abbott T. M. C. et al., 2021, ApJS, 255, 20
Ahumada R. et al., 2020, ApJS, 249, 3
Aihara H. et al., 2011, ApJS, 193, 29
Anders F. et al., 2022, A&A, 658, A91
Anguiano B. et al., 2018, A&A, 620, A76
Arenou F., Luri X., 1999, in Egret D., Heck A., eds, ASP Conf. Ser. Vol. 167, Harmonizing Cosmic Distance Scales in a Post-HIPPARCOS Era. Astron. Soc. Pac., San Francisco, p. 13
Arentsen A. et al., 2020, MNRAS, 491, L11
Babusiaux C. et al., 2010, A&A, 519, A77
Babusiaux C. et al., 2014, A&A, 563, A15
Bailer-Jones C. A. L., 2015, PASP, 127, 994
Bailer-Jones C. A. L., Rybizki J., Fouesneau M., Demleitner M., Andrae R., 2021, AJ, 161, 147
Barbuy B., Chiappini C., Gerhard O., 2018, ARA&A, 56, 223
Belokurov V., Erkal D., Evans N. W., Koposov S. E., Deason A. J., 2018, MNRAS, 478, 611
Bianchi L., Shiao B., Thilker D., 2017, ApJS, 230, 24
Bland-Hawthorn J., Gerhard O., 2016, ARA&A, 54, 529
Buder S. et al., 2019, A&A, 624, A19
Buder S. et al., 2021, MNRAS, 506, 150
Casey A. R., Kennedy G. M., Hartle T. R., Schlaufman K. C., 2018, MNRAS, 478, 2812
Chambers K. C. et al., 2016, preprint (arXiv:1612.05560)
Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2002, JAIR, 16, 321
Chiti A., Mardini M. K., Frebel A., Daniel T., 2021, ApJ, 911, L23
Cirasuolo M. et al., 2014, in Ramsay S. K., McLean I. S., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V. SPIE, p. 91470N
Cui X.-Q. et al., 2012, Res. Astron. Astrophys., 12, 1197
Dalton G. et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. SPIE, Bellingham, p. 84460P
Dalton G. et al., 2014, in Ramsay S. K., McLean I. S., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V. SPIE, p. 91470L
de Jong R. S. et al., 2019, The Messenger, 175, 3
Debattista V. P., Ness M., Gonzalez O. A., Freeman K., Zoccali M., Minniti D., 2017, MNRAS, 469, 1587
DESI Collaboration et al., 2016, preprint (arXiv:1611.00036)
Gadotti D. A. et al., 2019, MNRAS, 482, 506
Gaia Collaboration, 2016, A&A, 595, A1
Gaia Collaboration, 2021, A&A, 649, A1
Gal Y., Ghahramani Z., 2016, Proc. 33rd Int. Conf. on Machine Learning, PMLR Vol. 48. p. 1050
Gilmore G. et al., 2012, The Messenger, 147, 25
Gonzalez O. A., Gadotti D., 2016, in Laurikainen E., Peletier R., Gadotti D., eds, Astrophysics and Space Science Library, Vol. 418, Galactic Bulges. Springer, Berlin, p. 199
Grady J., Belokurov V., Evans N. W., 2021, ApJ, 909, 150
Gustafsson B., Edvardsson B., Eriksson K., Jørgensen U. G., Nordlund Å., Plez B., 2008, A&A, 486, 951
Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. R., 2012, preprint (arXiv:1207.0580)
Hogg D. W., Eilers A.-C., Rix H.-W., 2019, AJ, 158, 147
Huang Y. et al., 2022, ApJ, 925, 164
Ivezić Ž. et al., 2008, ApJ, 684, 287
Keller S. C. et al., 2007, Publ. Astron. Soc. Aust., 24, 1
Kingma D. P., Welling M., 2019, Foundations and Trends in Machine Learning, Vol 12, An Introduction to Variational Autoencoders. p. 307
Kollmeier J. A. et al., 2017, preprint (arXiv:1711.03234)
Koposov S. E., Belokurov V., Zucker D. B., Lewis G. F., Ibata R. A., Olszewski E. W., López-Sánchez A. R., Hyde E. A., 2015, MNRAS, 446, 3110
Leung H. W., Bovy J., 2019a, MNRAS, 483, 3255
Leung H. W., Bovy J., 2019b, MNRAS, 489, 2079
Li J. et al., 2016, ApJ, 823, 59
Lin J., Casagrande L., Asplund M., 2022, MNRAS, 510, 433
Lindegren L. et al., 2021, A&A, 649, A4
Majewski S. R., Skrutskie M. F., Weinberg M. D., Ostheimer J. C., 2003, ApJ, 599, 1082
Majewski S. R., Zasowski G., Nidever D. L., 2011, ApJ, 739, 25
Majewski S. R. et al., 2017, AJ, 154, 94
Minniti D., Olszewski E. W., Liebert J., White S. D. M., Hill J. M., Irwin M. J., 1995, MNRAS, 277, 1293
Myeong G. C., Vasiliev E., Iorio G., Evans N. W., Belokurov V., 2019, MNRAS, 488, 1235
Ness M. et al., 2013, MNRAS, 430, 836
Ness M., Freeman K., 2016, Publ. Astron. Soc. Aust., 33, e022
Paszke A. et al., 2019, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019. Curran Associates, Inc., Vancouver, Canada, p. 8024
Rich R. M., Origlia L., Valenti E., 2012, ApJ, 746, 59
Riello M. et al., 2021, A&A, 649, A3
Saglia R. P., Opitsch M., Fabricius M. H., Bender R., Blaña M., Gerhard O., 2018, A&A, 618, A156
Schlafly E. F., Meisner A. M., Green G. M., 2019, ApJS, 240, 30
Schlaufman K. C., Casey A. R., 2014, ApJ, 797, 13
Seabroke G. et al., 2021, A&A, 653, A160
Skrutskie M. F. et al., 2006, AJ, 131, 1163
Stan Dev Team., 2021, cmdstanpy, available at: https://pypi.org/project/cmdstanpy/
Starkenburg E. et al., 2017, MNRAS, 471, 2587
Steinmetz M. et al., 2020, AJ, 160, 82
Thomas G. F. et al., 2019, ApJ, 886, 10
Wallerstein G., 1962, ApJS, 6, 407
Wang S., Chen X., 2019, ApJ, 877, 116
Wegg C., Rojas-Arriagada A., Schultheis M., Gerhard O., 2019, A&A, 632, A121
Wolf C. et al., 2018, Publ. Astron. Soc. Aust., 35, e010
Wright E. L. et al., 2010, AJ, 140, 1868
Yanny B. et al., 2009, AJ, 137, 4377
Zhao H., Spergel D. N., Rich R. M., 1994, AJ, 108, 2154
Zoccali M., Hill V., Lecureur A., Barbuy B., Renzini A., Minniti D., Gomez A., Ortolani S., 2008, A&A, 486, 177

APPENDIX A: NN SET-UP

Here we describe the specific inputs and architecture we use to design our NNs.

A1 Input features

We select 24 features to use as input for our NN: 16 colours, and eight absolute magnitudes. Constructed from *Gaia G*, G_{RP}, and G_{BP}, 2MASS *J*, *H*, and K_s, and WISE *W*1 & *W*2 photometric bands, and extinction corrected following the RJCE method (Majewski et al. 2011), we select the following colours:

(J-K_s), (J-H), (H-K_s), ($W1$-$W2$), (G_{BP}-J), (G_{BP}-H), (G_{BP}-K_s), (G_{BP}-$W1$), (G_{BP}-$W2$), (G_{RP}-K_s), (G_{BP}-G_{RP}), (G_{BP}-G_G), (G_G-G_{RP}), (J-$W1$), (J-$W2$), (H-$W2$).

We also include the following absolute magnitudes when noted, calculated with distances from our NN-enhanced approach:

BP, RP, G, *W*1, *W*2, *J*, *H*, and K_s.

A2 Network architecture

We build our network out of four main layers: an input layer, two hidden layers, and an output layer.

Out input layer accepts the 24 input features, and assigns each of them to a node in the network. We have two hidden layers, both with 80 nodes, which are fully interconnected between each other, the input layer, and the output layer. These hidden layers also have drop-out applied, with a weighting of 20 per cent for each pass (i.e. one fifth of each hidden layer is 'dropped' each run) of the network during training or prediction. Two layers of 80 hidden nodes were chosen as the result of manual tuning, where we found a large network with drop-out gave the best recovery of initial data while maintaining the network's confidence in its predictions.

Our output layer contains only two nodes: the output node, where we return the 'final' output; and an uncertainty node which recovers the network's certainty in its prediction. This is explained fully in Section 2.

APPENDIX B: PHOTOMETRIC METALLICITY COMPARISONS

We compare the metallicities returned by our method to similar photometric-based techniques from Huang et al. (2022) and Lin et al. (2022), as shown in Fig. B1. Both papers use SkyMapper *u* and *v* photometry with Lin et al. (2022) comparing to theoretical isochrones and Huang et al. (2022) using a data-driven approach deriving polynomial colour relations for the metallicities fitted to SDSS (APOGEE DR14 and DR16) and LAMOST (DR7) data. We find a good correlation with these studies, especially at low metallicities ([Fe/H] < -1.5). There is, however, a notable overestimation in our metallicities visible in the Lin et al. (2022) comparison at $-1.5 <$ [Fe/H] < -0.5, where our method appears to predict a large portion of the sample with [Fe/H] ≈ -0.5. We note that the objects that make up this bias do tend to be stars with low $\log g$ values, suggesting this is may be a regime where the NN underperforms, possibly due to lack of training data. Alternatively, this bias could be due to discrepancies in the isochrones utilized by Lin et al. (2022) for cool stars.

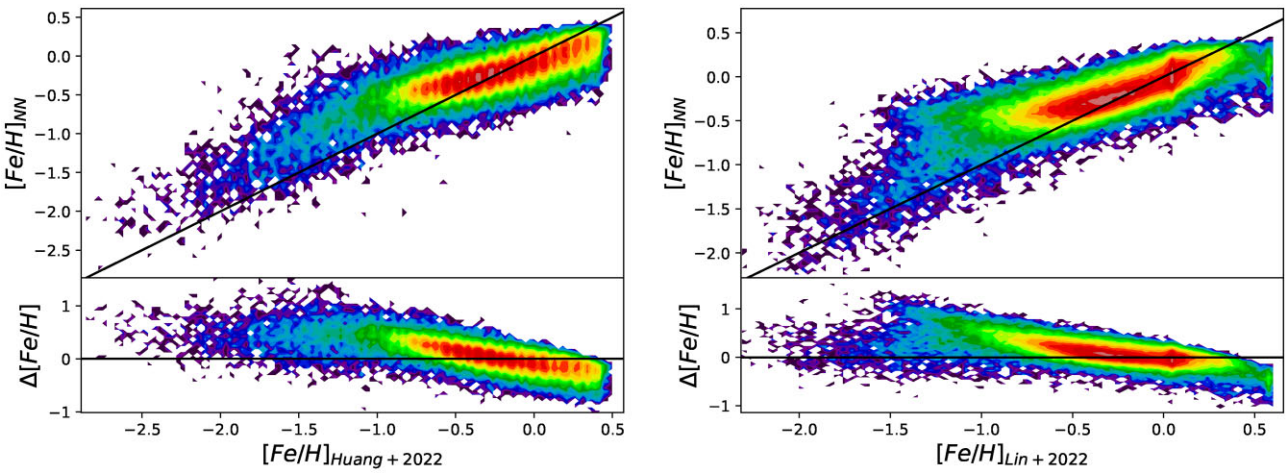

Figure B1. Plot of comparisons between our NN-estimated metallicities with those from Huang et al. (2022) (left-hand panel) and Lin et al. (2022) (right-hand panel).

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