Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **526,** 5735–5755 (2023) https://doi.org/10.1093/mnras/stad3165
Advance Access publication 2023 October 14

In situ or accreted? Using deep learning to infer the origin of extragalactic globular clusters from observables

Sebastian Trujillo-Gomez [1,2]★, J. M. Diederik Kruijssen [3,4], Joel Pfeffer [5], Marta Reina-Campos [6,7], Robert A. Crain [8], Nate Bastian[9,10] and Ivan Cabrera-Ziri [2]

[1]*Astroinformatics Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany*
[2]*Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Monchhofstraße 12-14, D-69120 Heidelberg, Germany*
[3]*Cosmic Origins Of Life (COOL) Research DAO, coolresearch.io*
[4]*Technical University of Munich, School of Engineering and Design, Department of Aerospace and Geodesy, Chair of Remote Sensing Technology, Arcisstr. 21, D-80333 Munich, Germany*
[5]*International Centre for Radio Astronomy Research (ICRAR), M468, University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia*
[6]*Department of Physics & Astronomy, McMaster University, 1280 Main Street West, Hamilton, L8S 4M1, Canada*
[7]*Canadian Institute for Theoretical Astrophysics (CITA), University of Toronto, 60 St George St, Toronto, M5S 3H8, Canada*
[8]*Astrophysics Research Institute, Liverpool John Moores University, 146 Brownlow Hill, Liverpool L3 5RF, UK*
[9]*Donostia International Physics Center (DIPC), Paseo Manuel de Lardizabal, 4, E-20018 Donostia-San Sebastián, Guipuzkoa, Spain*
[10]*IKERBASQUE, Basque Foundation for Science, E-48013 Bilbao, Spain*

Accepted 2023 September 25. Received 2023 September 15; in original form 2023 January 12

ABSTRACT

Globular clusters (GCs) are powerful tracers of the galaxy assembly process, and have already been used to obtain a detailed picture of the progenitors of the Milky Way (MW). Using the E-MOSAICS cosmological simulation of a $(34.4\ \mathrm{Mpc})^3$ volume that follows the formation and co-evolution of galaxies and their star cluster populations, we develop a method to link the origin of GCs to their observable properties. We capture this complex link using a supervised deep learning algorithm trained on the simulations, and predict the origin of individual GCs (whether they formed in the main progenitor or were accreted from satellites) based solely on *extragalactic* observables. An artificial neural network classifier trained on ∼50 000 GCs hosted by ∼700 simulated galaxies successfully predicts the origin of GCs in the test set with a mean accuracy of 89 per cent for the objects with $[\mathrm{Fe/H}] < -0.5$ that have unambiguous classifications. The network relies mostly on the alpha-element abundances, metallicities, projected positions, and projected angular momenta of the clusters to predict their origin. A real-world test using the known progenitor associations of the MW GCs achieves up to 90 per cent accuracy, and successfully identifies as accreted most of the GCs in the inner Galaxy associated to the *Kraken* progenitor, as well as all the *Gaia-Enceladus* GCs. We demonstrate that the model is robust to observational uncertainties, and develop a method to predict the classification accuracy across observed galaxies. The classifier can be optimized for available observables (e.g. to improve the accuracy by including GC ages), making it a valuable tool to reconstruct the assembly histories of galaxies in upcoming wide-field surveys.

Key words: galaxies: evolution – galaxies: formation – galaxies: haloes – galaxies: structure.

1 INTRODUCTION

One of the major goals of astrophysics is understanding the physical processes that gave rise to galaxies out of the tiny density perturbations that emerged from the epoch of recombination. The advent of modern cosmology has provided precise knowledge of these initial conditions in the context of the highly successful Λ cold dark matter (ΛCDM) cosmological paradigm. The ΛCDM model makes detailed predictions for the formation and evolution of dark matter (DM) haloes, which are the sites for baryonic material to condense into the galaxies we observe today (Blumenthal et al. 1984; Navarro, Frenk & White 1995; Springel et al. 2005). Hydrodynamical cosmological simulations in the ΛCDM framework predict that galaxies assemble through a combination of *in-situ* star formation in cold gas that is continuously accreted from cosmic filaments, and continuous infall of smaller satellite galaxies along with their DM, gas, and stars (Naab & Ostriker 2017; Crain & van de Voort 2023). Sophisticated dynamical models of external galaxies using integral field spectroscopic data are now able to recover the properties of kinematically and chemically distinct 'cold' and 'hot' populations that trace the *in-situ* and accreted stellar components, respectively (e.g. Zhu et al. 2020; Poci et al. 2021). However, reconstructing the detailed merger history of a galaxy from observations remains an extremely challenging task.

The first chemo-dynamical studies of galaxies date back to the 1960s, when observations of the stars and globular cluster (GC) populations of the Milky Way (MW) showed that the kinematics of stars contained important clues to the origin of the various components. This pioneering work by Eggen, Lynden-Bell & Sandage (1962)

★ E-mail: strujill@gmail.com

found that while young stars follow nearly circular orbits, older stars have eccentric radial orbits with lower angular momentum and higher vertical velocity dispersion, all indicative of their accreted origin. Later studies of the ages and metal abundances of GCs in the MW showed that while inner GCs follow a tight age–metallicity relation, the outer GCs have a broad range of ages at fixed metallicity (Searle & Zinn 1978). This simple observation confirmed the scenario where the MW disc stars and GCs formed early, while the halo formed slowly from material that continued to accrete long after the disc was in place. Several decades later, the Sloan Digital Sky Survey (SDSS; York et al. 2000) found the first evidence of a past accretion event in the MW, the Sagittarius stream, a remnant of the accretion of the *Sagittarius* dwarf galaxy (Ibata, Gilmore & Irwin 1994). Deep wide-field photometric surveys including SDSS, Pan-STARRS (Chambers et al. 2016), and DES (Abbott et al. 2018) have since found many spatial overdensities and streams in the MW stellar halo that correspond to recent smaller accretion events (for reviews, see Belokurov 2013; Grillmair & Carlin 2016). The first hints of kinematic halo substructures emerged from astrometric surveys combined with ground-based radial velocities (see Klement 2010; Smith 2016).

The *Gaia* survey (Gaia Collaboration 2018a) revolutionized the field of Galactic archaeology by precisely measuring the 3D positions and motions of millions of stars in the inner halo, enabling the search for the progenitor galaxies of the MW using the phase-space clustering of halo stars (for a review, see Helmi 2020). Over the last 5 yr these data, combined with other spectroscopic surveys, led to the identification of the stellar debris from one of the most massive galaxy ever accreted by the MW, *Gaia-Enceladus* (also known as the *Gaia Sausage*; Belokurov et al. 2018; Haywood et al. 2018; Helmi et al. 2018), and of at least six additional progenitors (e.g. Myeong et al. 2018a, b, c; Deason, Belokurov & Sanders 2019; Gallart et al. 2019; Iorio & Belokurov 2019; Mackereth et al. 2019; Myeong et al. 2019; Vasiliev 2019; Koppelman et al. 2019a, b; Necib et al. 2020a, b; Horta et al. 2021; Malhan et al. 2022). Their location in 6D phase-space together with their metallicities and alpha-element abundances, identifies these substructures as having formed in satellites with different masses and star formation histories (see Helmi 2020). The new data therefore allowed the global properties (such as mass and accretion redshift) to be determined for the most massive progenitors of the Galaxy. More recently, the H3 survey (Conroy et al. 2019) of high latitude stars in the MW found evidence of six chemo-dynamical substructures in the outer halo, beyond the reach of *Gaia* (Naidu et al. 2020). This brought the census of Galactic progenitors up to $\sim$10, accounting for $\sim$95 per cent of the mass of the stellar halo. Achieving a similarly detailed assembly reconstruction for large samples of galaxies would undoubtedly open an entirely new window into galaxy formation and cosmology.

Gaia also provided the precise orbits of nearly all of the Galactic GCs (Gaia Collaboration 2018b; Baumgardt et al. 2019; Vasiliev 2019). These data, along with the GC chemical abundances and ages, offered a novel and complementary way of reconstructing galaxy assembly. GCs are particularly powerful tracers of galaxy assembly because they can be studied at much larger distances than individual stars, up to $\sim$100 Mpc, have long phase-mixing time-scales, and their abundance relative to field stars increases in low-mass galaxies (Peng et al. 2008; Georgiev et al. 2010; Forbes et al. 2018). Using hydrodynamical cosmological simulations from the E-MOSAICS project, which include the formation and evolution of star clusters, Kruijssen et al. (2019a) demonstrated that GCs are excellent tracers of the properties of their progenitor galaxies. Kruijssen et al. (2019b) then used the age–metallicity relation of the MW GCs to obtain the most detailed reconstruction to date of the merger tree of the Galaxy. Trujillo-Gomez et al. (2021) found a surprising amount of galaxy assembly information encoded in the 3D GC system kinematics of simulated MW-mass galaxies, and applied a statistical method to the *Gaia* data to produce an independent and consistent reconstruction of the MW merger tree. Massari, Koppelman & Helmi (2019) used phase-space and age–metallicity information to associate most of the accreted GCs to each of the five most massive (likely) progenitors, *Gaia-Enceladus*, *Kraken*, *Sagittarius*, *Sequoia*, and the progenitor of the *Helmi streams*. Pfeffer et al. (2020) studied the relationship between the current phase-space distribution of GCs and the properties of their progenitors in cosmological simulations. Using machine learning to exploit this relation, along with the GC ages and metallicities in the simulations, Kruijssen et al. (2020) trained an artificial neural network (NN) to recover the masses and accretion redshifts of the five dominant MW progenitors. New studies continue to uncover further details of the MW assembly. For instance, Malhan et al. (2022) analysed the statistical 6D distribution of a large population of tracers in the Galactic halo (including GCs and stellar streams) to robustly search for phase-space substructures, and discovered a potential additional progenitor named *Pontus*.

In this study, we aim to provide the initial steps to extend the powerful methods that have been applied to the MW to recover the assembly histories of external galaxies based on their observed GC populations. First, we study the relation between the fraction of GCs accreted from satellites and fundamental galaxy properties in the simulations. We then investigate the link between GC origin (whether a GC was formed *in-situ* within the galaxy or was accreted), and its individual properties as determined by standard photometric and spectroscopic observations. The main result we highlight is that extragalactic GC observables contain a record of their progenitor properties, and that this information can be used to recover the origin of individual GCs using only a few key observables (their positions, radial velocities, and metallicities, and the stellar mass and effective radius of their host galaxy). With the goal of applying our classifier algorithm to upcoming deep, wide-field spectroscopic galaxy surveys, we provide the classification model code in a public repository.

The paper is organized as follows. Section 2 describes the simulations and the galaxy and GC sample selection. Section 3 shows how the accreted fraction depends on galaxy properties. Section 4 describes the deep learning model used to predict GC origin in external galaxies, and Section 5 provides a detailed analysis of its predictions for simulated galaxies, and the results of the a real-world test using the MW data. The results and discussed in Section 6, and summarized in Section 7.

2 SIMULATED GALAXY AND GC SAMPLE

In this work we use the simulated galaxies and star cluster populations from the E-MOSAICS simulations. Below we describe the simulations and sample selection criteria.

2.1 The E-MOSAICS simulations

E-MOSAICS (MOdelling Star cluster population Assembly In Cosmological Simulations within EAGLE) is a suite of hydrodynamical cosmological simulations that follow the formation and co-evolution of galaxies and their star cluster populations (Pfeffer et al. 2018; Kruijssen et al. 2019a). The physics of galaxy formation is implemented using the EAGLE model (Crain et al. 2015; Schaye et al. 2015), which uses a feedback prescription calibrated to reproduce the stellar mass

function and disc-galaxy sizes at $z = 0$. The EAGLE model also reproduces many additional key properties of the observed galaxy population, including their present-day luminosities and colours (Trayford et al. 2015), the evolution of the stellar mass function, star formation rates (Furlong et al. 2015), and galaxy sizes (Furlong et al. 2016), and the chemical abundances of stars in the MW (Mackereth et al. 2018).

To model the formation and evolution of star clusters, the simulations use an improved version of the MOSAICS subgrid model (Kruijssen et al. 2011; Pfeffer et al. 2018). Star clusters are treated as a subgrid population within each star particle, and form according to an environmentally-dependent prescription based on models for the fraction of stars formed in bound clusters (Kruijssen 2012), and for the upper truncation mass of the Schechter initial cluster mass function (Reina-Campos & Kruijssen 2017). Both of these quantities are calculated using the local gas conditions, and increase with the gas pressure. Clusters lose mass via stellar evolution, two-body relaxation, and tidal shocks, and may be completely disrupted by infall into the centres of galaxies via dynamical friction. Mass-loss due to tidal shocks and two-body relaxation is calculated self-consistently at each time step from the local tidal field.

The E-MOSAICS simulations have been shown to reproduce several key properties of GC populations. These include the massive end of the GC mass function (Pfeffer et al. 2018; Hughes et al. 2022), GC specific frequencies (Kruijssen et al. 2019a; Bastian et al. 2020), the colour–luminosity relation of metal-poor GCs (the 'blue tilt'; Usher et al. 2018), the GC radial distribution (Reina-Campos et al. 2021) and kinematics (Trujillo-Gomez et al. 2021), and the GC system mass–halo mass relation (Bastian et al. 2020). They also reproduce the age distribution of GCs in satellite streams (Hughes et al. 2019a), and the fraction of stars in the bulge of the Galaxy that were born in GCs (Hughes et al. 2019b). These simulations demonstrated that the properties of GC populations reflect the environment and assembly of their host galaxies (Kruijssen et al. 2019a), and this allowed the most detailed reconstruction so far of the merger tree of the MW (Kruijssen et al. 2019b), including the prediction of the masses and accretion times of its five most massive progenitors using the properties of its GCs (Kruijssen et al. 2020). We refer the reader to Pfeffer et al. (2018) for a complete description of the physical models in the simulations.

The E-MOSAICS simulations are unique in their ability to model star cluster populations in a cosmological volume to $z = 0$, and their success in reproducing galaxy and GC observables makes them an ideal tool to investigate how the intrinsic properties of GCs relate to their natal galaxies. In this work we use the galaxies and GCs from the E-MOSAICS $(34.4\ \mathrm{cMpc})^3$ periodic volume (Bastian et al. 2020). The gas particle mass is $2.26 \times 10^5 \mathrm{M}_\odot$, and the gravitational softening at $z = 0$ is $\epsilon = 0.35$ kpc. A Friends-of-Friends algorithm (FoF; Davis et al. 1985) is first used to identify DM groups with a linking length of 0.2-times the mean particle separation. Within each group, the SUBFIND algorithm (Springel, Yoshida & White 2001; Dolag et al. 2009) then identifies gravitationally bound structures, and identifies as the central subhalo/galaxy the one containing the particle with the lowest potential energy. All other subhaloes within the FOF group are then considered satellites of the central galaxy. Merger trees are constructed in the same way as for the EAGLE simulations, using the D-TREES algorithm (Jiang et al. 2014; Qu et al. 2017) to link between 10 and 100 of the most bound particles in each subhalo across the 28 snapshots (for further details on this procedure, see Qu et al. 2017). The simulation volume contains 2900 galaxies (resolved with at least 100 star particles), and [465, 69, 7] galaxies with $M_{\mathrm{halo}} > [10^{11}, 10^{12}, 10^{13}]\mathrm{M}_\odot$. We refer the reader to Bastian et al. (2020) for a detailed description of the simulation and its first results.

Table 1. Upper metallicity thresholds used to define the GC sample as a function of host galaxy stellar mass.

Galaxy $\log(M_*/\mathrm{M}_\odot)$	$[\mathrm{Fe/H}]_{\mathrm{thresh}}$	$[\mathrm{Fe/H}]_{\mathrm{split}}$
8.0–8.5	−1.0	−1.2
8.5–9.0	−1.1	−1.2
9.0–9.5	−0.8	−1.2
9.5–10.0	−0.5	−1.1
10.0–10.5	−0.5	−1.0
10.5–11.0	−0.5	−0.9
>11.0	−0.3	−0.8

The upper threshold is designed to remove artificially underdisrupted clusters (see Section 2.2). The last column shows the metallicity used to split the GC sample into metal-poor and metal-rich subpopulations.

2.2 Sample selection

For the analysis in this study we select all central galaxies (i.e. not satellites) in the $(34.4\ \mathrm{cMpc})^3$ periodic volume with stellar masses $M_* > 10^8\ \mathrm{M}_\odot$ and hosting at least 10 GCs. GCs are identified as star clusters with $M > M_{\mathrm{thresh}}$ and metallicities $-2.5 < [\mathrm{Fe/H}] < [\mathrm{Fe/H}]_{\mathrm{thresh}}$. The minimum mass threshold is chosen to increase with galaxy mass to follow the shift in the upper truncation of the GC mass function (Hughes et al. 2022). The metallicity range is designed to mitigate the effect of numerical cluster underdisruption due to the absence of cold and dense gas in the EAGLE model (Reina-Campos et al. 2021). It effectively removes most of the excess GCs (which are mostly metal-rich) that should have been effectively disrupted by tidal shocks (for a detailed discussion see Appendix D of Kruijssen et al. 2019a). To emulate the observationally motivated galaxy mass-dependent GC minimum mass used by Hughes et al. (2022), we use the smoothly varying function,

$$\log(M_{\mathrm{thresh}}) = 0.5 \log(M_*) - 0.5. \quad (1)$$

As a result, for the lowest mass galaxies in the sample we select as GCs star clusters with $M > 10^{3.5}\ \mathrm{M}_\odot$, while for the most massive ellipticals we use $M \gtrsim 10^5\ \mathrm{M}_\odot$. Table 1 shows the upper metallicity threshold values as a function of galaxy stellar mass.

These criteria result in a sample of 921 central galaxies hosting a total of 75 810 GCs. To classify the GCs into *in-situ* and accreted, we use the merger trees and examine the two snapshots that bracket the formation time of the host star particle. If the progenitor gas particle was assigned to the same branch of the merger tree as the resulting star particle, the GC has a clear origin, and it is labelled '*in-situ*' if it formed on the main branch, or 'accreted' (or 'ex-situ') if it formed in a different branch. GCs with a formation time that falls between two different branches do not have a determined origin (given the spacing between simulation snapshots), and they are labelled 'unclear'. After classification, the sample contains 39 158 *in-situ*, 36 652 accreted, and 6674 GCs with unclear origin. Since an unclear origin is simply an artefact of the merger trees, we remove those GCs from the final sample. The final sample contains 57 per cent *in-situ*, and 43 per cent accreted GCs.

We use this final sample of GCs and host galaxies for all the analysis in this work. The selection criteria are designed to avoid numerical artefacts in the simulations, and to prevent the classifier from learning this unphysical behaviour. As long as the classifier is applied to GCs and host galaxies within the scope of its training set, we expect that the predictions will be robust to the particular choice

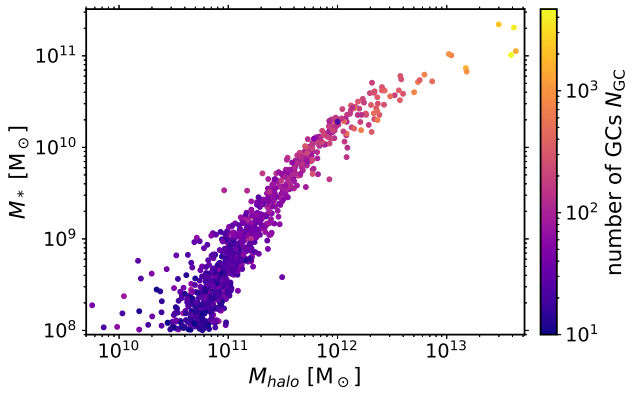

Figure 1. Number of GCs hosted by each simulated galaxy as a function of the galaxy stellar and halo mass. The number of hosted GCs increases steeply with both stellar and halo mass. At fixed halo mass, galaxies with larger stellar mass tend to host more GCs. See Section 2.2 for the sample selection criteria.

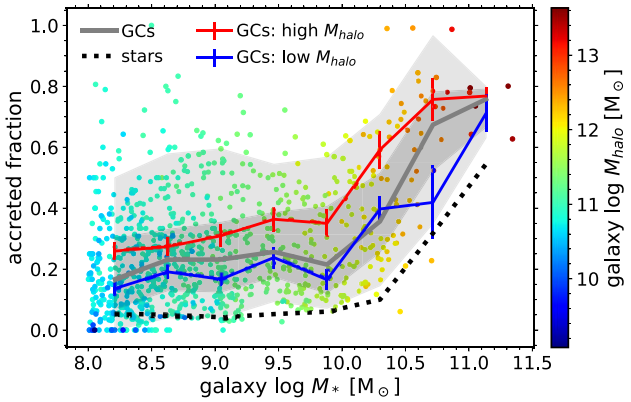

Figure 2. Fraction of stars and GCs accreted from satellites as a function of host galaxy stellar mass (points), coloured by halo mass. The grey line and shading show the median, [5,95], and [25,75] percentile range of the accreted fraction in bins of stellar mass (with error bars corresponding to the uncertainty in the median). The blue and red lines show the median accreted fraction in the bottom and top quartiles of halo mass in each bin, respectively. The dotted line shows the fraction of accreted stars. The median GC accreted fraction increases with stellar mass, and is always larger than for stars. At fixed stellar mass, galaxies hosted by more massive DM haloes have larger fractions of accreted GCs.

of parameter boundaries. Extension of the range of input observables to the most metal-rich observed GCs will require better treatment of ISM-driven cluster disruption (see Reina-Campos et al. 2022).

3 GC ORIGIN ACROSS THE GALAXY POPULATION

We begin by examining how GC origin is related to GC and host galaxy properties. Fig. 1 shows the stellar-to-halo mass relation of the simulated galaxies coloured by the number of GCs they host. The size of the GC population increases steeply with halo mass, reproducing the observed qualitative trend (e.g. Blakeslee, Tonry & Metzger 1997; Burkert & Forbes 2020). In a more detailed analysis, we found that the relation between halo mass and total mass in GCs in the simulations also matches observations (Bastian et al. 2020). At fixed galaxy stellar mass there is a weak secondary trend of increasing number of GCs with increasing halo mass.

Fig. 2 shows the fraction of accreted GCs and stars in each galaxy as a function of galaxy stellar mass. The fraction of accreted GCs in E-MOSAICS increases with galaxy stellar mass following the qualitative trend found for stars in semi-empirical and semi-analytical models, as well as in cosmological simulations including EAGLE (e.g. Rodriguez-Gomez et al. 2016; Qu et al. 2017; Clauwens et al. 2018; Tacchella et al. 2019; Davison et al. 2020; Moster, Naab & White 2020). This is not surprising, and is a direct consequence of the hierarchical nature of galaxy assembly combined with the shape of the fundamental stellar-to-halo mass relation (Fig. 1). The stellar-to-halo mass relation is very steep at low masses and becomes shallower for galaxies more massive than the MW. Massive galaxies are therefore partially assembled by hierarchical accretion of satellites with relatively high stellar masses, while dwarfs accrete only satellites with relatively low stellar masses. While the accreted fraction increases with galaxy mass for both stars and GCs, Fig. 2 shows that the fraction of accreted GCs is always larger. This is a result of the higher mean specific frequencies of satellites relative to centrals.

There is significant scatter in the GC accreted fraction at fixed galaxy stellar mass. To understand the physical drivers of the scatter, we search for secondary trends in the GC accreted fraction. Fig. 2 also shows the median GC accreted fraction of galaxies hosted by the least/most massive DM haloes in each stellar mass bin (in the lower/upper quartile of the distribution of $M_{\rm halo}$ in each bin). At fixed stellar mass, galaxies hosted by more massive DM haloes have larger GC accreted fractions, as expected from their larger fraction of accreted material from DM-rich satellites. The left panel of Fig. 3 shows how the accreted GC fraction varies with galaxy metallicity, with metal-poor galaxies typically hosting a larger fraction of accreted GCs (for $M_* \gtrsim 10^9$ M$_\odot$). This trend is driven by the decrease in the mean metallicity due to the accretion of a larger fraction of stars/GCs from satellites.[1] We also find that the mean metallicity of accreted GCs is higher in galaxies with higher accreted GC fractions due to the dominant contribution of the most massive satellite (which also contains the most metal-rich GCs).

In the right panel of Fig. 3 we show an even stronger trend found in the accreted fraction of metal-poor and metal-rich GC systems (i.e. considering the mean GC metallicity of each galaxy). At fixed stellar mass, galaxies with metal-poor GC systems have systematically higher accreted GC fractions compared to those with metal-rich systems. As in the case of the stellar component, this distinct metallicity dependence of the accreted fraction results from a combination of the overall effect of larger fractions of (metal-poor) GCs accreted from satellites on the mean GC metallicity, and the effect of DM halo formation times on the *in-situ* GCs.

Fig. 4 shows the accreted fraction of metal-poor and metal-rich GC subpopulations as a function of galaxy stellar mass. The subpopulations are defined using the stellar mass-dependent split shown in Table 1. Metallicity alone is not a direct proxy for GC origin. While metal-rich GCs tend to form *in-situ* in low-mass galaxies, the metal-poor population is typically a mix of *in-situ* and accreted objects. Fig. 5 shows the distribution of accreted and *in-situ* GCs as a function of GC metallicity and host galaxy mass. It confirms

[1]There is an additional weak trend where the *in-situ* GCs in massive galaxies with high GC accreted fractions tend to be less enriched than in those with low accreted fractions. This originates from the anticorrelation between halo mass (or formation time) and galaxy metallicity at fixed stellar mass discussed above.

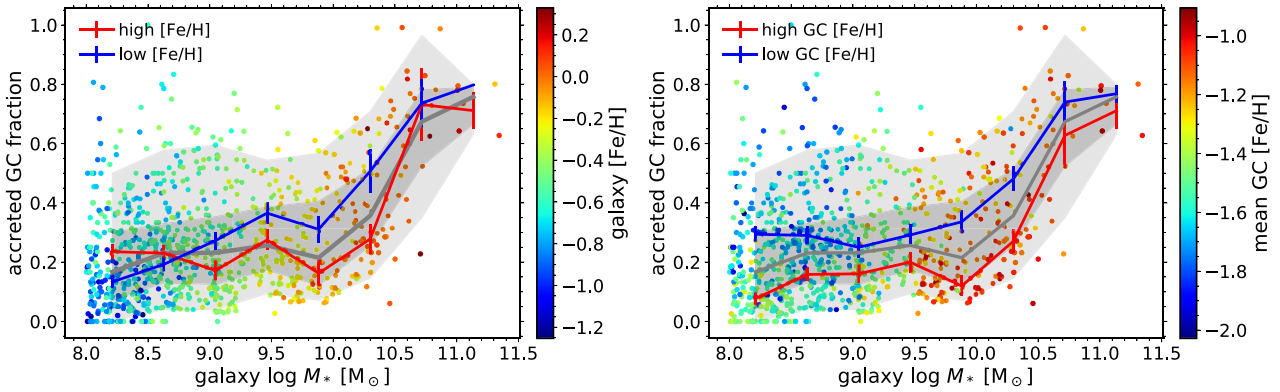

Figure 3. Accreted GC fraction as a function of galaxy stellar mass. Left: Coloured by mean galaxy metallicity [Fe/H]. The blue and red lines show the median accreted GC fractions for the galaxies with the 25 per cent lowest and highest metallicities, respectively. Right: Coloured by mean GC metallicity. The blue and red lines show the median accreted GC fractions for the galaxies in the bottom and top GC system metallicity quartiles, respectively. At fixed stellar mass, galaxies with higher metallicity stars and GCs have systematically lower accreted fractions than those at lower metallicities. This is driven by the fact that accreted stars and GCs from low-mass satellites are metal-poor.

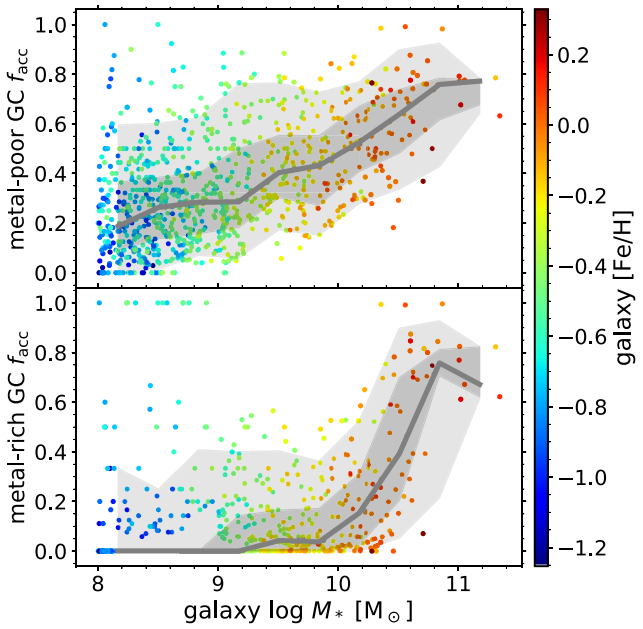

Figure 4. Accreted GC fraction as a function of galaxy stellar mass for metal-poor (top panel) and metal-rich (bottom panel) GC populations. The GC subpopulations are selected based on the stellar mass-dependent metallicity split in Table 1. Metal-poor GCs typically have a mixed origin. Metal-rich GCs in low-mass galaxies are almost exclusively formed *in-situ*, while in galaxies more massive than the MW they have a mixed origin.

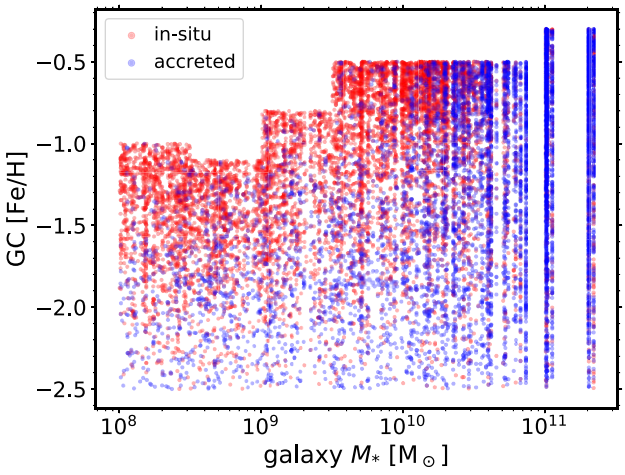

Figure 5. Origin of individual GCs in the simulation as a function of GC metallicity and host galaxy mass. The upper mass-dependent metallicity limit reflects the selection applied to reduce contamination by artificially underdisrupted GCs (see Section 2.2). Accreted GCs tend to have lower metallicities than *in-situ* GCs, but there is a significant overlap between the two populations. There is also significant variation in the metallicity distribution of the accreted and *in-situ* GCs across galaxies of similar mass due to differences in their assembly histories.

that while accreted GCs tend to be more metal-poor than *in-situ* GCs, there is significant overlap and galaxy-to-galaxy variation in the populations, and metallicity alone is generally not enough to determine GC origin.

4 PREDICTING GC ORIGIN USING MACHINE LEARNING

We now turn to the question of whether the origin of a particular GC can be predicted using its observable properties, and which observables are best suited for this purpose. We take advantage of the flexibility and predictive power of deep learning algorithms when applied to problems with highly non-linear relations between the input and output variables. In addition, we explore other supervised learning techniques to find possible alternatives with higher predictive power.

After exploring several classifier algorithms including k-nearest neighbours (Fix & Hodges 1989), Logistic Regression (Pearl & Reed 1920), Support Vector Machines (Boser, Guyon & Vapnik 1992), Decision Trees (Hunt, Marin & Stone 1966), and Random Forests (Breiman 2001), we find that their predictive accuracy is generally lower compared to deep learning, while most do not provide probabilistic outputs. The probabilistic output of NNs will be key for tuning and predicting the uncertainties in the model (see Sections 5.1 and 5.4).

Table 2. GC and host galaxy observables used as features in the fiducial NN classifier.

Feature	Object	Definition
$\log M_*^{\rm gal}$	Galaxy	Stellar mass
$\log R_{\rm e}^{\rm gal}$	Galaxy	Projected effective radius
$[{\rm Fe/H}]_{\rm gal}$	Galaxy	Mean metallicity
$[\alpha/{\rm Fe}]_{\rm gal}$	Galaxy	Mean oxygen abundance relative to iron [O/Fe]
$\sigma_{\rm gal}$	Galaxy	Stellar velocity dispersion
$\log N_{\rm GC}$	Galaxy	Total number of GCs
$\sigma_{\rm GCs}$	Galaxy	GC system velocity dispersion
[Fe/H]	GC	Metallicity
$[\alpha/{\rm Fe}]$	GC	Oxygen abundance relative to iron [O/Fe]
Δ[Fe/H]	GC/galaxy	Metallicity relative to the galaxy, $[{\rm Fe/H}] - [{\rm Fe/H}]_{\rm gal}$
$\Delta[\alpha/{\rm Fe}]$	GC/galaxy	Alpha-abundance relative to the galaxy, $[\alpha/{\rm Fe}] - [\alpha/{\rm Fe}]_{\rm gal}$
$\log R_{\rm p}/R_{\rm e}^{\rm gal}$	GC/galaxy	Projected distance from galaxy centre in units of the galaxy effective radius
$\sqrt{\lvert V_{\rm p}\rvert/\sigma_{\rm gal}}$	GC/galaxy	LOS velocity in units of the galaxy velocity dispersion
$\sqrt{\lvert V_{\rm p}\rvert/\sigma_{\rm GCs}}$	GC/galaxy	LOS velocity in units of the GC system velocity dispersion
$V_{\rm rot}/\sigma_{\rm gal}$	GC/galaxy	'Projected rotation velocity': dot product of LOS velocity and the unit vector pointing along the galaxy rotation velocity at the GC projected position in units of the galaxy velocity dispersion (see Section 4.2)
$\log R_{\rm p}\lvert V_{\rm p}\rvert$	GC	'Projected angular momentum': product of the projected galactocentric distance and magnitude of LOS velocity
$(R_{\rm p}V_{\rm rot})^{1/3}$	GC/galaxy	'Projected angular momentum vector': product of projected galactocentric distance and $V_{\rm rot}$ (see Section 4.2)

Projected positions and LOS velocities are calculated with respect to the position and velocity of the centre of the galaxy, assuming a single random orientation for each galaxy.

4.1 Algorithm description

For the fiducial model we employ a Multilayer Perceptron NN[2] architecture (MLP; Rumelhart, Hinton & Williams 1986) with dense, sequential layers. MLPs are powerful classifiers that are ideally suited for complex problems where the classes are not linearly separable, as we expect here. They are also advantageous compared to more traditional models because they automatically create useful new features from the provided inputs. We use the deep learning library KERAS (Chollet et al. 2015) implemented within the TENSORFLOW framework (Abadi et al. 2015).

The MLP architecture consists of several layers of artificial neurons that are connected in sequence, such that each neuron takes as input the combined outputs from all the neurons in the previous layer. To adapt the model to our specific classification task, we set the input layer to contain as many nodes (i.e. neurons) as the dimensions of the input data (i.e. the number of GC observables, $N_{\rm input}$), and the output layer to have two dimensions corresponding to the two possible classification labels: *in-situ* and *accreted*. The number of hidden layers $N_{\rm layers}$, and the number of nodes per layer ($N_{\rm nodes}$) are left as free parameters to be optimized using the validation data. The input and hidden layers use the standard 'Rectified Linear Unit' (ReLU) activation function, $h(x) = \max(0, x)$, and the output layer uses the sigmoid activation function to convert the output into a binary probability in the range [0,1], $P_{\textit{in-situ}} = 1 - P_{\rm accreted}$. The model is compiled using the 'Adam' optimizer (Kingma & Ba 2014), with the standard binary cross-entropy loss function used in binary classification tasks,

$$\mathcal{L} = -\frac{1}{N_{\rm GC}} \sum_{i=1}^{N_{\rm GC}} y_i \log(P_i) + (1 - y_i) \log(1 - P_i), \tag{2}$$

where y_i is the true label and P_i is the output probability for GC i (1 for *in-situ*, 0 for accreted). Below we describe the input features and training procedure.

[2]also known as feed-forward NN.

4.2 Training a NN classifier on simulated galaxies and their GCs

To select the set of input observables (i.e. the features) used by the model to predict GC origin, we first explore a large set of physically-motivated GC and host galaxy observables. We iteratively remove features that do not affect the accuracy of the predictions to reduce as much as possible the complexity of the model. This procedure yields a fiducial set of $N_{\rm input} = 17$ observables that we use in the final step to optimize the NN architecture and to train the fiducial model. Table 2 summarizes the features. These are all derived using physically-motivated combinations of GC observables (metallicity, alpha-element abundance, projected position on the sky, and line-of-sight velocity), and global galaxy properties (stellar mass, mean metallicity, effective radius, and stellar velocity dispersion). Since GC ages are notoriously difficult to measure precisely beyond the MW, we ignore them here and evaluate their contribution to the predictions in Section 5.7. In the E-MOSAICS simulations the GC alpha-element abundances follow the same trends as in the field stars, making them a good proxy for the alpha-enrichment history of galaxies (Hughes et al. 2019b).

We select a single random orientation for each galaxy corresponding to a projection onto the x–y plane of the simulation box, and calculate the positions and velocities in the reference frame of the centre of the galaxy obtained using SUBFIND. We define the GC 'rotation velocity' as the dot product of the GC LOS velocity $\mathbf{V}_{\rm p}$ and the unit vector pointing in the direction of net rotation of the galaxy at the projected GC position, $V_{\rm rot} \equiv \mathbf{V}_{\rm p} \cdot \mathbf{V}_{\rm rot}^{\rm gal}/|\mathbf{V}_{\rm rot}^{\rm gal}|$. We further test an augmented feature set by including an additional set of six features that describe the distribution of the projected distance and LOS velocity of the GC system (using the median, inter-quartile range, skewness, and kurtosis), and quantify the projected GC distance and velocity relative to the four nearest GC neighbours. We find that these additional features do not increase the model performance, and therefore keep only the original set of 17 input observables. Having chosen the final feature set, we follow the common practice of standardizing each feature by subtracting the mean and diving by the standard deviation to obtain distributions with a mean of 0 and standard deviation of 1.

To train the model we use the sample of 69 136 simulated GCs with clear origin hosted by 921 central galaxies (see Section 3). Normally, the model would be trained on a random subsample containing the majority of the simulated GCs (typically ∼70–80 per cent), and the remaining fraction would be used in model validation and testing. However, to avoid leakage of the information on host galaxy properties from the GCs in the training set to the GCs in the test set, we adopt a different approach. Instead, we split the host galaxies randomly into a training set containing all the GCs hosted by a subset comprised of 80 per cent of the galaxies (50 612 GCs from 736 galaxies), and a test set containing all the GCs in the remaining 20 per cent (18 524 GCs from 185 galaxies). This ensures that the model is not exposed to any of the test data during training, and increases its capacity to generalize to other data sets, including GCs in the real Universe.

After selecting the training and test sets, we perform a grid search to optimize the main hyperparameters of the network: the number of layers, and the number of nodes (neurons) per layer. To maximize the use of the simulation data, we choose to use the same data for both validation and testing. We have verified that using separate validation and test data has no effect on the ability of the model to generalize to new data.[3] We evaluate the validation accuracy of predictions using the test set for a model with parameters in the two-dimensional grid defined by the values $N_{\rm layers} \in [2, 3, 4, 5, 6, 7, 8, 9, 10]$, and $N_{\rm nodes} \in [10, 20, 50, 100, 200]$. In each iteration, the training is stopped after 30 epochs, or when the accuracy does not increase over 5 epochs. The architecture with $[N_{\rm layers}, N_{\rm nodes}] = [4, 20]$ results in the highest validation accuracy ≈80 per cent, and we select it for the fiducial model. Using these parameters we retrain the final model for 100 epochs, stopping early when the accuracy does not increase over the last 20 epochs. This model is saved and used to evaluate the predictions and performance in Section 5. We refer to it throughout the paper as the 'fiducial' model.

5 RESULTS

In this section we evaluate and tune the performance of the fiducial model on the simulated test data. We then analyse the detailed predictions and the relative importance of each observable, and evaluate the model confidence. We also perform the first real-world test of the algorithm by predicting the origin of the MW GCs. Lastly, we test the impact of observational uncertainties and evaluate the performance improvement when GC ages are included.

5.1 Model performance

We now evaluate the performance of the classifier on the test sample containing 185 galaxies (i.e. 20 per cent of the sample) drawn at random from the simulation, and the 18 524 GCs they host. Fig. 6 shows the distribution of predicted probabilities $P_{\rm in\text{-}situ}$ for the simulated test sample (top panel) along with the number of correct predictions. To calculate the accuracy (i.e. the fraction of correct predictions), we must map the predicted probabilities output by the classifier to binary class labels assuming a simple probability threshold $P_{\rm thresh} = 0.5$, such that a GC is labelled '*in-situ*' when $P_{\rm in\text{-}situ} > P_{\rm thresh}$, and 'accreted' otherwise. The overall accuracy of the model (measured across the entire test GC sample) is 80 per cent. We find that the classifier produces two distinct peaks in the distribution of probabilities, with each peak near the maximum probability for each class (i.e. $P_{\rm in\text{-}situ} \sim 0$ or $P_{\rm in\text{-}situ} \sim 1$). This shows that the NN reaches a high confidence when predicting the origin of the majority of the GCs in the test sample. The fraction of correct predictions in each bin is shown in the bottom panel of Fig. 6. The accuracy increases monotonically towards the most confident predictions (at $P_{\rm in\text{-}situ} \sim 0$, and $P_{\rm in\text{-}situ} \sim 1$). This is evidence that the classifier successfully predicts the correct labels with high confidence (i.e. high probabilities). A minority of the predictions lie in the ambiguous region with $P_{\rm in\text{-}situ} \sim 0.3$–$0.7$.

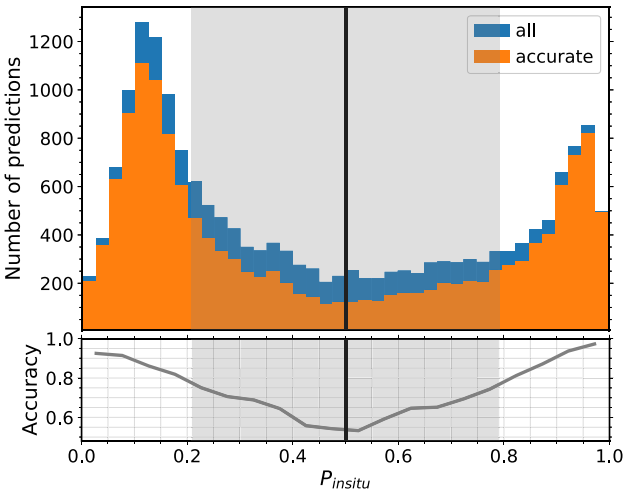

Figure 6. Accuracy distribution of the classifier on the simulated test sample as a function of the predicted probabilities. Top: Distribution of the predicted *in-situ* probability $P_{\rm in\text{-}situ}$ across the entire GC test sample compared to the distribution of only the correct predictions (assuming a decision threshold $P_{\rm thresh} = 0.5$). Bottom: Accuracy of the predictions in each probability bin. The vertical line marks the initial value of $P_{\rm thresh} = 0.5$ we adopt for labelling the predictions of the classifier, where $P_{\rm in\text{-}situ} > P_{\rm thresh}$ corresponds to *in-situ*, and $P_{\rm in\text{-}situ} \leq P_{\rm thresh}$ corresponds to accreted. The grey shaded region indicates ambiguous predictions as defined in Section 5.1. Both the probability distribution and the accurate predictions are peaked near the two extremes of $P_{\rm in\text{-}situ}$, showing that the classifier makes accurate predictions with high confidence.

To exploit the probabilistic nature of the model to improve the accuracy of the classifications, we introduce a new label, and define 'ambiguous' predictions as those with $P_{\rm in\text{-}situ} > P_{\rm thresh}$ and $P_{\rm accreted} \equiv 1 - P_{\rm in\text{-}situ} > P_{\rm thresh}$, where $P_{\rm thresh}$ is the decision threshold. Fig. 7 shows the effect of increasing the decision threshold on the fraction of unambiguous predictions, and on their accuracy. As expected, the accuracy increases with $P_{\rm thresh}$, while the completeness (i.e. the unambiguous fraction of predictions) decreases: 3/4 of the sample reaches an accuracy of 85 per cent, while only half of the sample reaches 90 per cent accuracy. To optimize both the accuracy and the sample completeness, we define the unambiguous predictions using a fiducial value of $P_{\rm thresh} = 0.79$. This results in an accuracy of ∼89 per cent (for a 60 per cent completeness). The ambiguous region is shown using grey shading in Fig. 6.

The results of the classification of the test set are shown in Fig. 8 using the standard confusion matrix. The columns represent the true labels, and the rows show the number of GCs in each column that are predicted to be *in-situ*, accreted, or ambiguous. After removing the ambiguous predictions, the model erroneously classifies 6 per cent

[3]We tested this by running an experiment where the model performance was tested using data that had not been used in the hyperparameter tuning. The accuracy was unaffected.

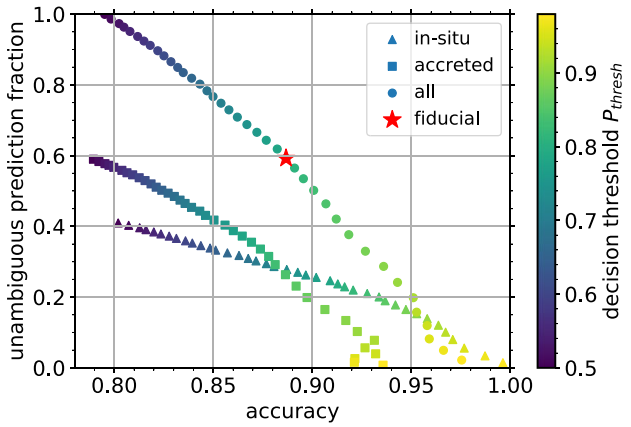

Figure 7. Effect of increasing the decision threshold of the NN classifier on the fraction of unambiguous predictions in the test sample and their accuracy. The colour bar shows the decision threshold $P_{\rm thresh}$ for each class as well as for the combined sample. A GC is labelled '*in-situ*' when $P_{\rm in\text{-}situ} > P_{\rm thresh}$, and 'accreted' when $P_{\rm accreted} \equiv 1 - P_{\rm in\text{-}situ} > P_{\rm thresh}$. The star symbol indicates the fiducial decision threshold adopted in this work. It corresponds to an accuracy of ∼89 per cent on 60 per cent of the GC sample.

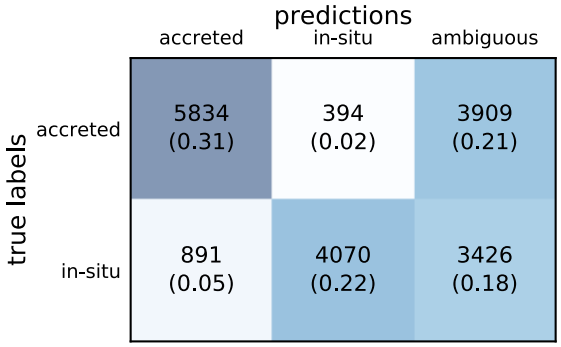

Figure 8. Confusion matrix showing the distribution of the predicted versus true labels of GCs in the test set. The 'ambiguous' label corresponds to predictions with low confidence, $P < P_{\rm thresh} = 0.79$. For each category the matrix shows the number of GCs, and the fraction relative to the total sample in parentheses. The background shading is darker for larger fractions. The model is excellent at classifying accreted GCs (with only 6 per cent falsely identified as *in-situ*), but misclassifies *in-situ* GCs in 18 per cent of the cases.

(394/5834) of the accreted GCs, and a much larger fraction of *in-situ* GCs (891/4070 = 18 per cent).

The fraction of predicted labels for each class is shown in Fig. 9. The ambiguous class represents a nearly constant fraction ∼25–45 per cent of the predicted labels across the entire range of host galaxy stellar masses. The figure also shows that the model correctly predicts the dominant GC origin as a function of host galaxy stellar mass. Indeed, we find that the predictive accuracy remains nearly constant as a function of galaxy mass. However, the fraction of the dominant class is slightly overpredicted in dwarfs and massive ellipticals (but still lies within the uncertainty defined by the grey band).

To evaluate the impact of global galaxy properties on the model performance, Fig. 10 shows the accuracy obtained across each galaxy as a function of galaxy stellar mass, metallicity, and GC accreted fraction. To properly account for the large class imbalance in some galaxies (i.e. where accreted or *in-situ* GCs dominate),

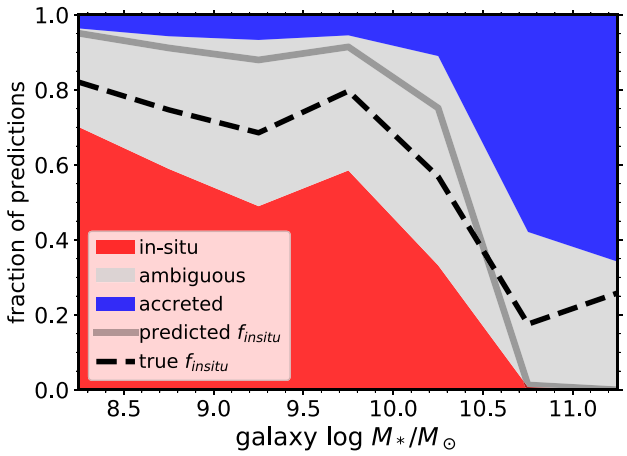

Figure 9. Fraction of test GCs in the predicted classes as a function of host galaxy stellar mass. The 'ambiguous' class corresponds to predictions below the confidence threshold, $P < P_{\rm thresh} = 0.79$. The grey line and grey shaded area show the predicted *in-situ* fraction and the uncertainty range (due to the ambiguous predictions). The dashed line indicates the true *in-situ* fraction. Even though ∼25–45 per cent of the sample is classified as ambiguous at a given mass, the model correctly predicts the majority class as a function of stellar mass.

we also show the balanced accuracy (defined as the average of the accuracies calculated separately for each class). The accuracy reaches >80 per cent for the majority of galaxies, while it drops below 60 per cent in only a few galaxies. The low values of balanced accuracy in the most massive galaxies and several dwarfs are due to poor performance in identifying GCs in the minority class (which corresponds to *in-situ* for massive galaxies, and accreted in some dwarfs). This is more common among the most massive galaxies due to the small number of these objects in the training set. In addition, accreted GCs in massive galaxies have properties that are very similar to *in-situ* GCs due to the high masses of their satellite progenitors (see Section 5.3). There is a weak trend of decreasing accuracy in metal-poor galaxies, which reflects the weak correlation between accreted fraction and metallicity (see Fig. 3).

5.2 Detailed predictions for simulated GC systems

To evaluate the performance of the classifier across individual GCs, we now look at a few specific examples of galaxies in the test set. Fig. 11 shows the projected distributions of GCs labeled by their predicted and true origin in four example galaxies selected randomly in each stellar mass bin, including a massive elliptical, a MW-mass galaxy, a massive dwarf, and a low-mass dwarf. In the massive elliptical galaxy, the model has difficulty identifying any of the *in-situ* GCs (as indicated by the low balanced accuracy), despite the galaxy containing 11 per cent of GCs in this class. Similarly, in the MW-mass galaxy that is currently undergoing a massive merger, the model achieves high accuracy overall but again has difficulty identifying the small fraction of *in-situ* GCs. Across the dwarf galaxies the model shows excellent performance on both classes (i.e. a high overall and balanced accuracy), despite the relatively low accreted fractions. The better performance in low-mass galaxies is consistent with their relative dominance across the training set. In general, the model seems to produce lower confidence predictions at intermediate galactocentric distances, where the *in-situ* and accreted GCs are co-spatial.

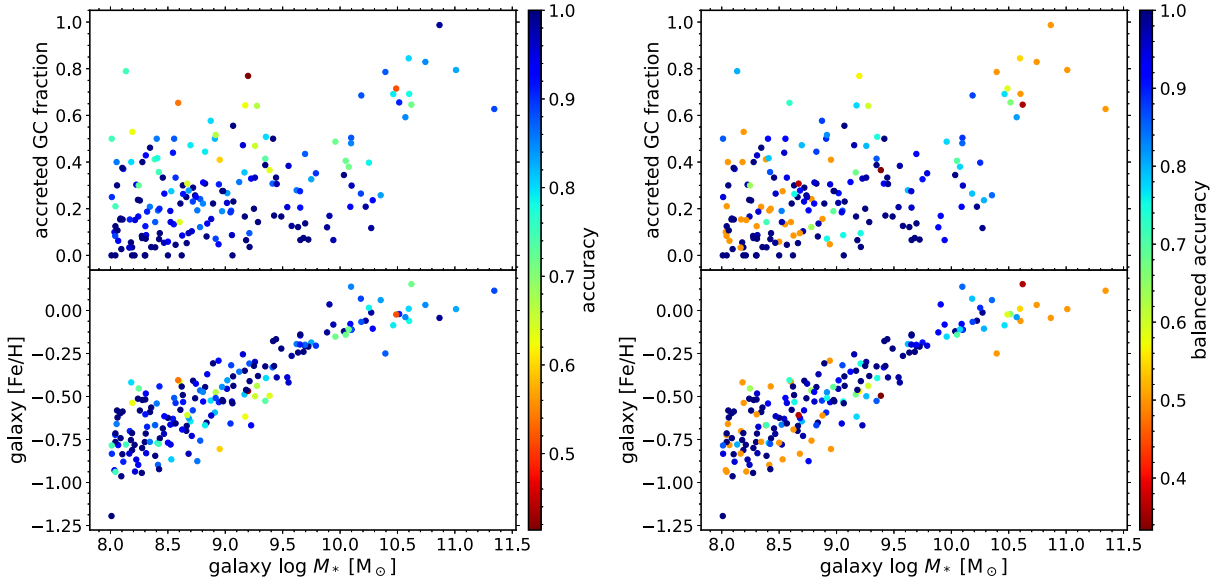

Figure 10. Predictive accuracy of the classifier on the simulated GC test set as a function of host galaxy properties. Left: Accuracy across each galaxy versus galaxy mass and accreted GC fraction (top) and galaxy [Fe/H] (bottom). Right: Same coloured by balanced accuracy. The accuracy tends to be lower in low-mass galaxies with high accreted fractions (the outliers in that mass range). The balanced accuracy in massive ellipticals is low because of the low performance of the model when predicting the origin of *in-situ* GCs (the minority class) as a result of the small number of training galaxies. The accuracy depends only weakly on galaxy metallicity.

To understand the role of the GC phase-space distribution in the model predictions we show the same galaxies in projected position-velocity space in Fig. 12. The importance of the 'projected angular momentum' $R_\mathrm{p}|V_\mathrm{p}|$ is evident in the massive elliptical, with the decision boundary of the algorithm describing a near circle in position-velocity space, equivalent to a nearly constant projected angular momentum. This separation is also clear in lower mass galaxies, but in those cases the model predicts a more complicated boundary based on additional GC properties including their chemical abundances (see Fig. 13). We investigate which GC properties are most important for predicting GC origin in the next section.

5.3 Importance of each GC and galaxy observable

To assess how important each GC and galaxy observable is for the predictions of the model, we calculate the 'permutation feature importances' (Breiman 2001). For a given feature, its importance is defined as the mean decrease in accuracy when the feature information in the test set is removed from the model input. For this, the feature vector of the desired feature is randomly shuffled while leaving the other features unchanged. The accuracy is then computed using the predictions over several random realizations of the shuffled data N_iter. The importance is then the difference in accuracy between the shuffled data and the fiducial model averaged over all realizations. The left panel of Fig. 13 shows the result using $N_\mathrm{iter} = 30$.

Surprisingly, the most important features are host galaxy properties: the 2D effective radius, velocity dispersion (which is very similar for GCs and stars), and alpha-element abundance. These are followed by the GC projected galactocentric radius, the projected angular momentum, galaxy metallicity, and GC alpha-abundance offset relative to the galaxy. The importance of the galaxy properties might seem counterintuitive at first glance. However, it can be explained in two ways. First, most of the galaxy properties we use here correlate strongly with stellar mass, such that the classifier can obtain galaxy mass or size information indirectly from any of them. As we show in Fig. 2, galaxy mass is the strongest predictor of GC accreted fraction, so it is natural for the algorithm to use it to estimate to first order the likelihood of a GC having formed *in-situ*. Secondly, highly covariant features can skew the results of the permutation technique, artificially reducing the importance of all features in a covariant cluster (Wei, Lu & Song 2015). This occurs because the model can always obtain the information on a permuted feature from one of its covariates.

To remove this possible bias, we perform a clustering analysis of all the features and split them into covariant groups based on a correlation threshold, and select only one feature from each group (see Appendix A for details). We train a new model with the fiducial architecture but using only the selected subset of features,[4] and obtain the new permutation importances (right panel of Fig. 13). The three remaining galaxy properties still have the highest importance, followed by the GC metallicity and alpha-abundance relative to the galaxy, and the projected angular momentum and projected galactocentric radius in units of R_e.

For galaxies in surveys with limited data (i.e. no alpha abundances), Fig. 13 also provides an estimate of the performance if the model when a specific observable is not included. However, the optimal solution in this case would be to retrain a new model with the reduced feature set (see Section 5.5 for a discussion of the performance of such a reduced model).

In Appendix B we show that the observables with the highest importances have distributions across the GC sample that lead to the most distinct separation of *in-situ* and accreted objects. To understand why the classifier performs poorly in massive elliptical galaxies, Fig. 14 shows the distribution of GC observables for galaxies with $M_* >$

[4]Since removing covariant features may lead to a slight loss of predictive power, we use this model variant only for the purpose of evaluating feature importance.

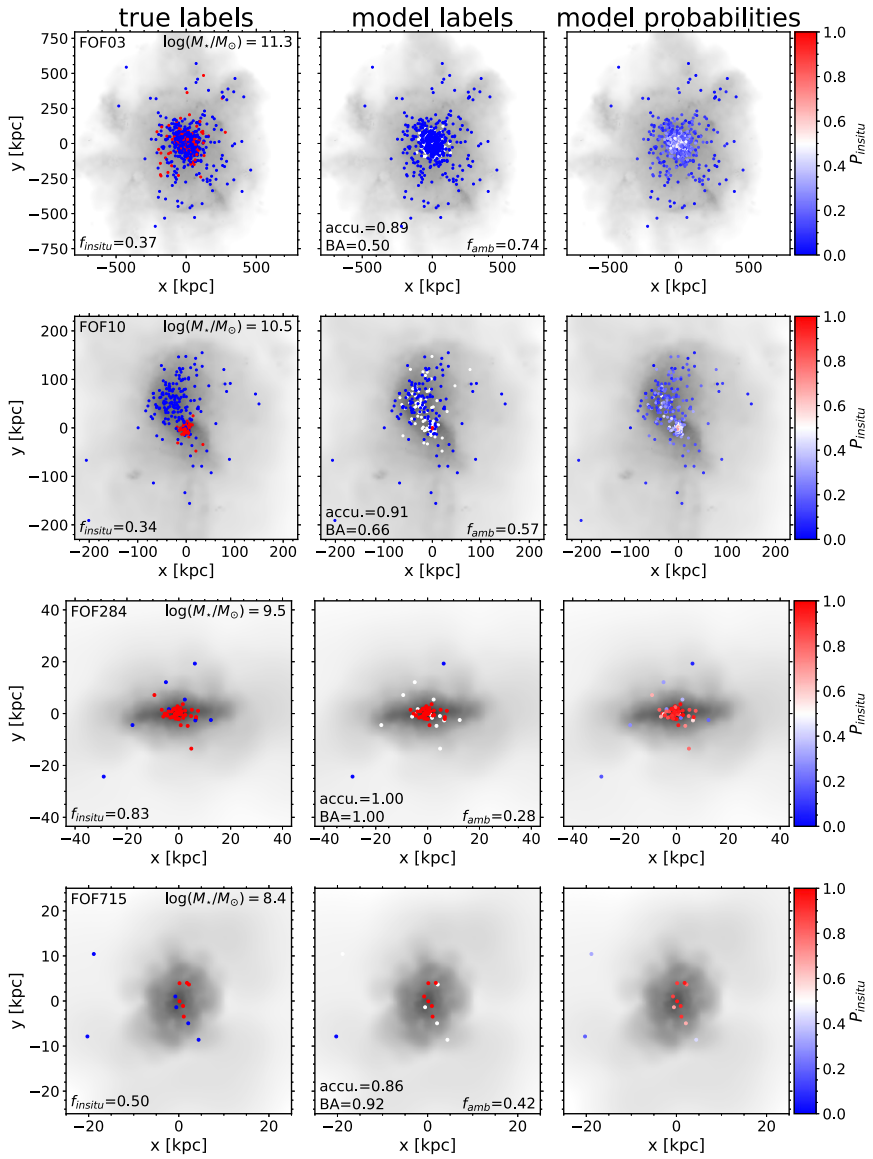

Figure 11. Projected distributions of GCs in the test sample and their predicted origin compared to their true origin. Each row shows the GCs hosted by selected galaxies (dots) in the random test set in each of four representative stellar mass bins, ranging from massive ellipticals (top row) to low-mass dwarfs (bottom row). The left column shows the true origin, while the middle and right columns show the predicted labels and probabilities of *in-situ* origin $P_{\text{in-situ}}$, respectively. GCs with ambiguous classifications are shown in white in the middle column. The stellar mass, GC *in-situ* fraction, accuracy, balanced accuracy, and fraction of ambiguous predictions are indicated in each row. The stellar surface density is shown in grey-scale. The model produces high accuracy predictions for low-mass galaxies but has difficulty identifying *in-situ* GCs in massive galaxies with high accreted GC fractions due to their rarity in the training set.

10^{11} $M_\odot$. The buildup of elliptical galaxies is dominated by massive satellites that contribute GCs with similar chemical abundances to the main progenitor GCs, and violent relaxation further mixes the two populations in phase space. The GC observables of *in-situ* and accreted populations entirely overlap, and this partly explains why the model cannot discriminate between the two classes from these data.

5.4 Estimating uncertainty in the model predictions

Ideally we would like to predict not only the GC origin of each GC in an external galaxy, but also to have an idea of the uncertainty in the prediction. To estimate this predictive uncertainty we formulate a new problem: can we predict the accuracy of the model across a galaxy using only the observed properties of the galaxy? This would provide an estimate of how much the predictions for a given observed galaxy can be trusted. We explored a variety of regression algorithms including a Multilayer Perceptron with a linear activation function for the output layer (Rumelhart, Hinton & Williams 1986), a Random Forest (Breiman 2001), and a Ridge Regressor (Hoerl & Kennard 1970). Each model was trained on all the galaxy features listed in Table 2, in addition to the features describing the distribution of GC galactocentric radii and LOS velocities in each galaxy (their mean, inter-quartile range, skewness, and kurtosis).

Perhaps unsurprisingly, we find that none of these algorithms can predict the accuracy of the fiducial classifier. To predict the galaxy-wide accuracy, the models would need to know the true GC origin labels, and this is precisely the information we lack for real galaxies.

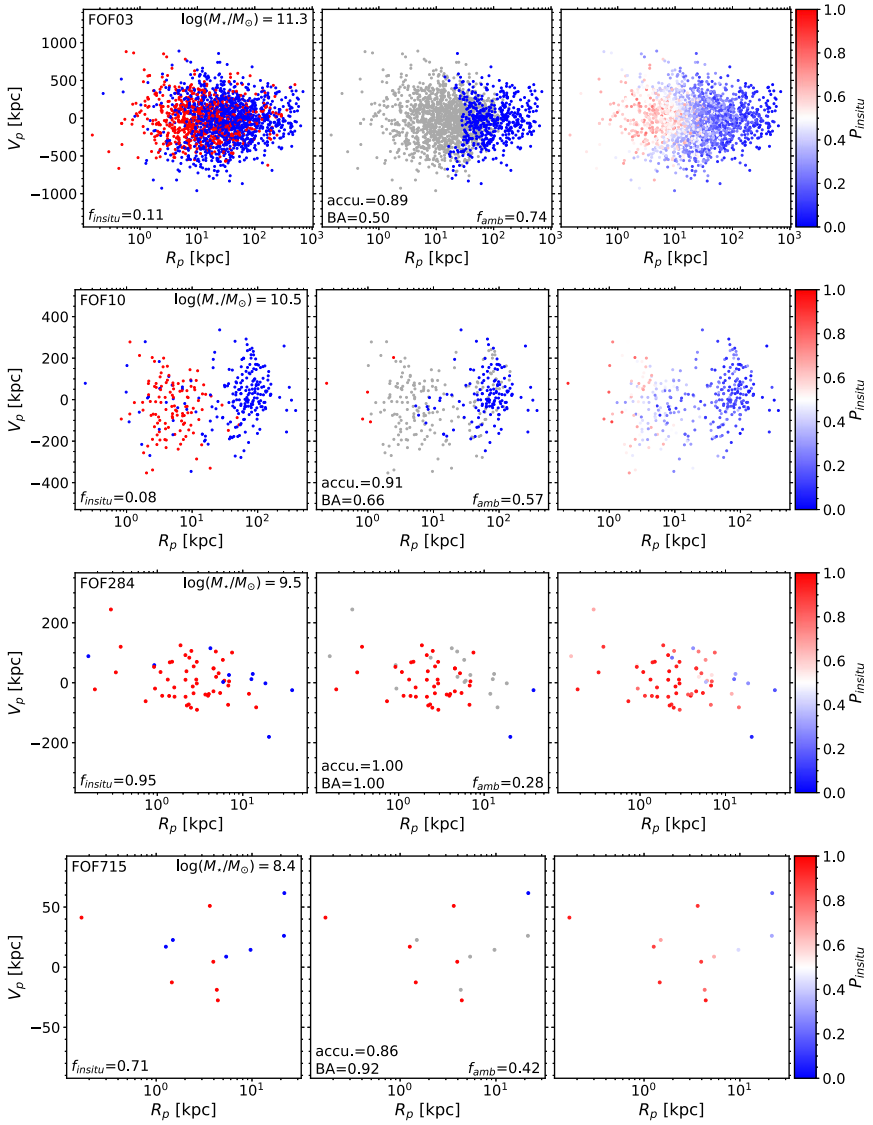

Figure 12. Projected position-velocity distributions of GCs in the test sample and their predicted origin compared to their true origin. The rows show the LOS velocity versus projected galactocentric radius for the randomly selected simulated galaxies in Fig. 11. The left column shows the true origin, while the middle and right columns show the predicted labels and probabilities of *in-situ* origin $P_{\text{in-situ}}$, respectively. GCs with ambiguous classifications are shown in grey in the middle column. The stellar mass, GC *in-situ* fraction, accuracy, balanced accuracy, and fraction of ambiguous predictions are indicated in each row. The decision boundary is clear in the right panels for massive galaxies, and highlights the predictive power of the 'projected angular momentum' $R_{\text{p}}|V_{\text{p}}|$.

Deep learning offers a possible solution: the output of MLP classifiers is a set of class membership probabilities. We may therefore exploit the correlation that was found between the label probabilities $P_{\text{in-situ}}$ and the full sample accuracy in Fig. 6 to predict the model uncertainty. We define the 'confidence' of the model predictions for a galaxy by how close on average the predicted probabilities get to complete certainty,

$$\text{mean confidence} = \frac{1}{N_{\text{GC}}} \sum_{i=1}^{N_{\text{GC}}} \max\left(P^i_{\text{in-situ}}, P^i_{\text{accreted}}\right). \quad (3)$$

We examine the relation between the galaxy-wide accuracy and mean prediction confidence using the simulation test set in Fig. 15. To calculate the mean confidence we use all the GC predictions, including those with $P < P_{\text{thresh}}$. Despite the large scatter, there is a highly significant correlation ($p = 3 \times 10^{-9}$) between mean prediction confidence and accuracy. The median accuracy increases from ∼0.8 to ∼1.0 as the mean confidence increases from ∼0.70 to ∼0.95. This shows that the NN successfully learned which regions of the high-dimensional feature space contain both *in-situ* and accreted GCs, and therefore lead to ambiguous predictions. We can then use the distribution of galaxy-wide accuracy in Fig. 15 to estimate the probability that the classifier will reach a given desired accuracy in a real galaxy. For instance, we expect that the classifier will be more than 90 per cent accurate in three out of four galaxies that reach a mean confidence ∼0.85. The dashed line in Fig. 15 shows a linear fit to the data with the parameters provided in the legend.

5.5 Testing the model on the MW GCs

Simulations are rough simplifications of the real Universe. As such, they may or may not capture the physical processes linking GC

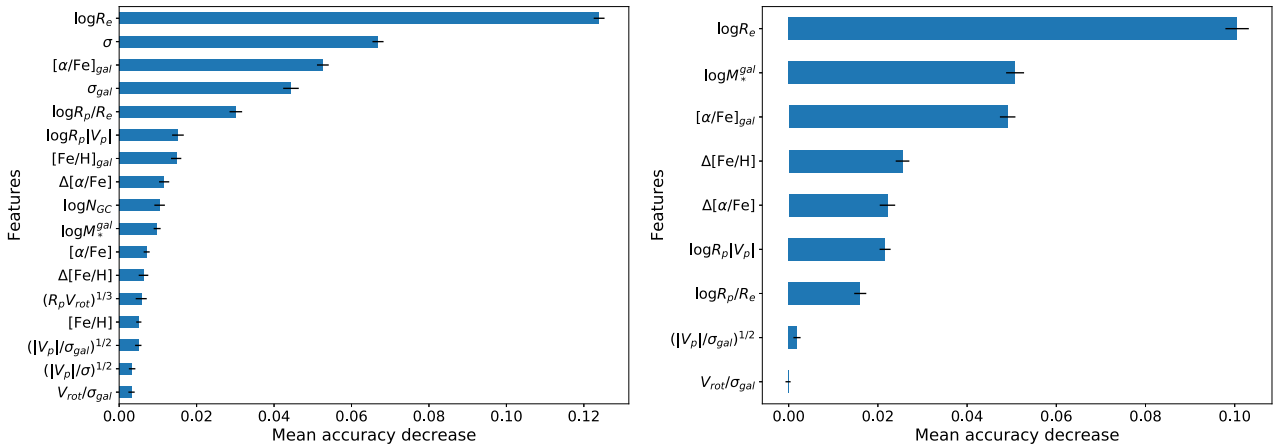

Figure 13. Permutation importance of each of the input features (i.e. observables) of the classifier. The value for each feature corresponds to the decrease in accuracy when the feature data in the test set is randomly shuffled before making predictions. Left: Using all features. Right: After removing highly covariant features and retraining the model with only independent ones (see Section 5.3 for details). The black lines show the standard deviation in the result over 30 random iterations. The projected galaxy effective radius, stellar mass, and alpha-element abundance are the most predictive host galaxy properties. The most predictive GC observables are GC metallicity and alpha-abundance relative to the host galaxy, and projected angular momentum $R_p|V_p|$ and relative projected radius R_p/R_e.

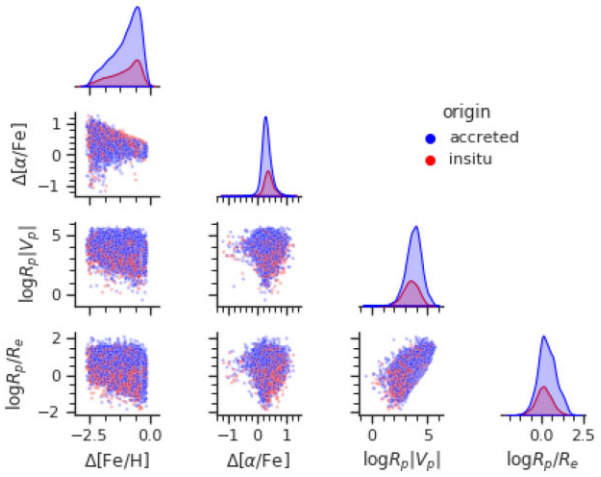

Figure 14. Joint and marginal distributions of GC origin across the observables with the most predictive power for GCs hosted by massive ellipticals. The panels show the distribution of *in-situ* and accreted GCs across simulated galaxies with $M_* > 10^{11}$ M$_\odot$. The overlap of the two classes across all the observables partly explains the underperformance of the classififer in the most massive galaxies.

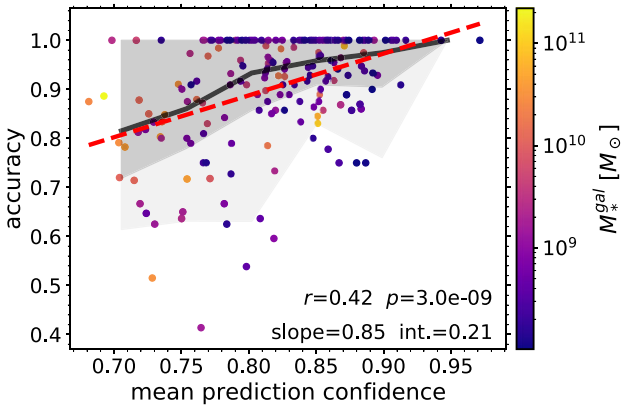

Figure 15. Accuracy of the NN classifier as a function of the mean confidence in the predictions across each galaxy in the test set. The black line shows the binned median, and the dark and light shading contain the top 75 and 95 per cent of the accuracy distribution in each bin. The dashed line shows a linear fit, with parameters given in the legend. Here we define confidence as the maximum of the predicted class probabilities for each GC, max ($P_{\rm in\text{-}situ}$, $P_{\rm accreted}$). There is a highly significant correlation between mean prediction confidence and accuracy. About 75 per cent of simulated galaxies with a mean prediction confidence ∼0.85 reach at least 90 per cent accuracy.

formation to their observable properties. With any supervised deep learning model trained on simulation data the question therefore arises: does the complex relationship between the features and target variables learned by the model resemble the actual relation in the real Universe? In other words, does the performance of the model using real data match the performance on the simulated test data? To answer this question we now perform a first, real-world test of the NN classifier using data for the MW GCs.

For this test we use the detailed data on the MW GC system that has been compiled over several decades, together with the progenitor associations determined recently using *Gaia* orbital information, chemical abundances, and ages (Massari, Koppelman & Helmi 2019; Kruijssen et al. 2019b, 2020). These associations may still contain substantial uncertainties, but we are only interested here in the binary *in-situ*/accreted labels, and these should generally be more robust. It is difficult to model the uncertainties in these labels, but given that they were obtained using much more information (i.e. 6D kinematics and detailed abundances), we assume them to be a good approximation to the 'ground truth' for the purpose of testing the extragalactic GC classifier.

To obtain the input observables for the classifier we use the compilation of GC metallicity data from Harris (1996, 2010 edition), and the 3D positions and velocities compiled by Baumgardt et al. (2019) from a combination of *HST* and *Gaia* data. To extend the applicability of the model to surveys that do not include the most difficult to obtain GC observables, we build a new 'minimal' ANN classifier using a reduced feature set (by removing the alpha-element abundances and velocity dispersions), and train it using the fiducial

Table 3. GC and host galaxy observables used as features in the 'minimal' classifier.

Feature	Object	Definition
$\log M_*^{\rm gal}$	Galaxy	Stellar mass
$\log R_{\rm e}^{\rm gal}$	Galaxy	Projected effective radius
$[{\rm Fe/H}]_{\rm gal}$	Galaxy	Mean metallicity
[Fe/H]	GC	Metallicity
Δ[Fe/H]	GC/galaxy	Metallicity relative to the galaxy, $[{\rm Fe/H}] - [{\rm Fe/H}]_{\rm gal}$
$\log R_{\rm p}/R_{\rm e}^{\rm gal}$	GC/galaxy	Projected distance from galaxy centre in units of the galaxy effective radius
$\log R_{\rm p}\lvert V_{\rm p}\rvert$	GC	'Projected angular momentum': product of projected galactocentric distance and LOS velocity

Projected positions and LOS velocities are calculated with respect to the position and velocity of the centre of the galaxy, assuming a single random orientation for each galaxy.

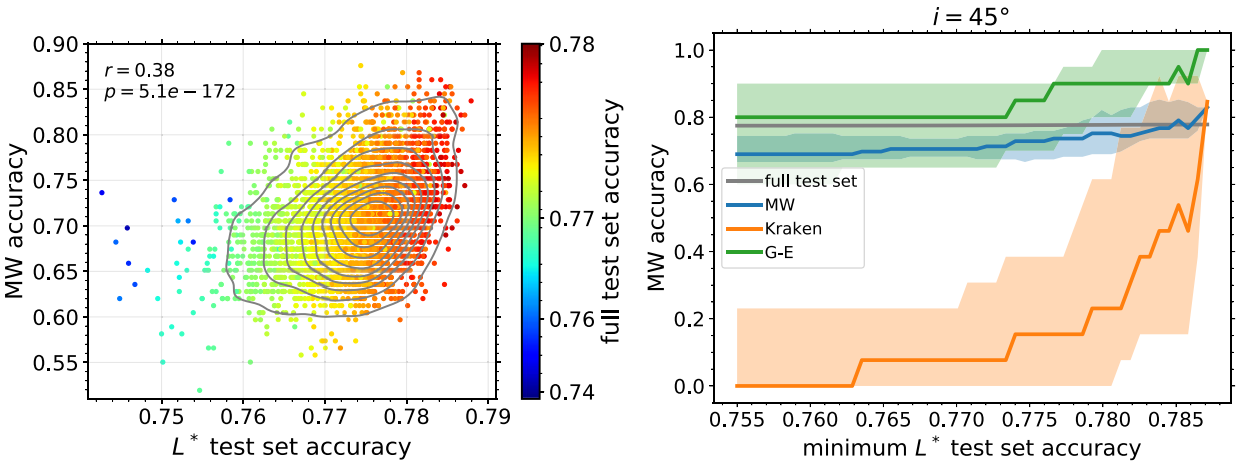

Figure 16. Performance of an ensemble of minimal classifiers on the simulated test set and on the MW GCs. Left: Correlation between the accuracy of each model (points) on the MW GCs and on the 24 simulated L^* galaxies in the test set (and its Spearman coefficient and p-value), with the colour indicating the accuracy on the full test set, and the contours showing a kernel density estimate of the underlying distribution. Right: Accuracy of a voting ensemble as a function of the minimum L^* galaxy test set accuracy used in the selection. The ensemble uses 5000 models trained on identical simulation data (and model architectures sampled from a grid of $[N_{\rm nodes}, N_{\rm layers}]$). To obtain the MW accuracy the models are tested on randomly inclined MW GC system observables. We exploit the strong correlation between test set accuracy and MW accuracy to select a model with the highest performance on both simulated and observed galaxies.

simulation training set. As we show below, this retraining procedure achieves a better perfomance than simply neglecting these features in the fiducial model (where the loss of accuracy would be >5 per cent; see Fig. 13). Table 3 summarizes the features of the minimal classifier.

For the global properties of the Galaxy we assume $M_*^{\rm gal} = 5 \times 10^{10}$ M$_\odot$, $R_{\rm e}^{\rm gal} = 3.8$ kpc (Cautun et al. 2020), and $[{\rm Fe/H}]_{\rm gal} = 0.0$ (Bland-Hawthorn & Gerhard 2016). We apply the same metallicity selection used for the simulation to the MW GCs (see Table 1), without imposing a GC mass cut (since this was only used to remove artefacts in the simulation). This results in a sample of 129 GCs with $-2.5 < [{\rm Fe/H}] < -0.5$. For the true origin labels we use the classification by Massari, Koppelman & Helmi [2019; as revised by Kruijssen et al. (2020) for Pal 1 and NGC 6441] based on the GC ages, metallicities, and orbits. To obtain the projected positions and LOS velocities we artificially incline the plane of the Galaxy by an angle i deg (around the x-axis) towards the observer.

As in the case of the fiducial model, we optimize the architecture using a grid search for the combination $[N_{\rm layers}, N_{\rm nodes}]$ that yields the highest accuracy on the simulation test set. For a decision threshold $P_{\rm thresh} = 0.5$, the resulting network achieves an accuracy of ∼78 per cent on the test data (using $N_{\rm layers} = 2$ and $N_{\rm nodes} = 50$). This corresponds to a decrease of ∼2 per cent compared to the fiducial model. During testing of the minimal model we found that the accuracy of the MW predictions varies significantly across identically trained models (with a dispersion of ≈3 per cent) as a result of the inherent stochasticity in the ANN training process.[5] This stochasticity is averaged out when considering the large simulated GC test sample, but becomes more important when evaluating the predictions for the small set of GCs in the MW system (see Appendix C).

To reduce the variance in the MW predictions we create an ensemble of 5000 models trained on identical simulation data, and vary the network complexity by sampling uniformly from the grid of $[N_{\rm nodes}, N_{\rm layers}]$ described in Section 4.2. We then tested each model on three different samples: the full simulation test set, the subset of 24 L^* galaxies (i.e. $10^{10} \leq M_*/{\rm M}_\odot \leq 10^{11}$) in the simulation test set, and the projected MW GC system (at random inclinations sampled uniformly from the range $0 \leq \cos i \leq 1$). The results are shown in Fig. 16 as a function of the performance on the L^* galaxy test set. We find an interesting statistically-significant correlation between the performance on the L^* simulations and on the real MW, and a weaker correlation with the fulll test set. This indicates that models with above-average performance on the simulations will in general also produce more accurate predictions on real galaxies. In other words, the best models tend to have the best generalization capacity,

[5]This is a well known trade-off of the computational efficiency necessary for estimating the gradient of the loss function in a high-dimensional feature space.

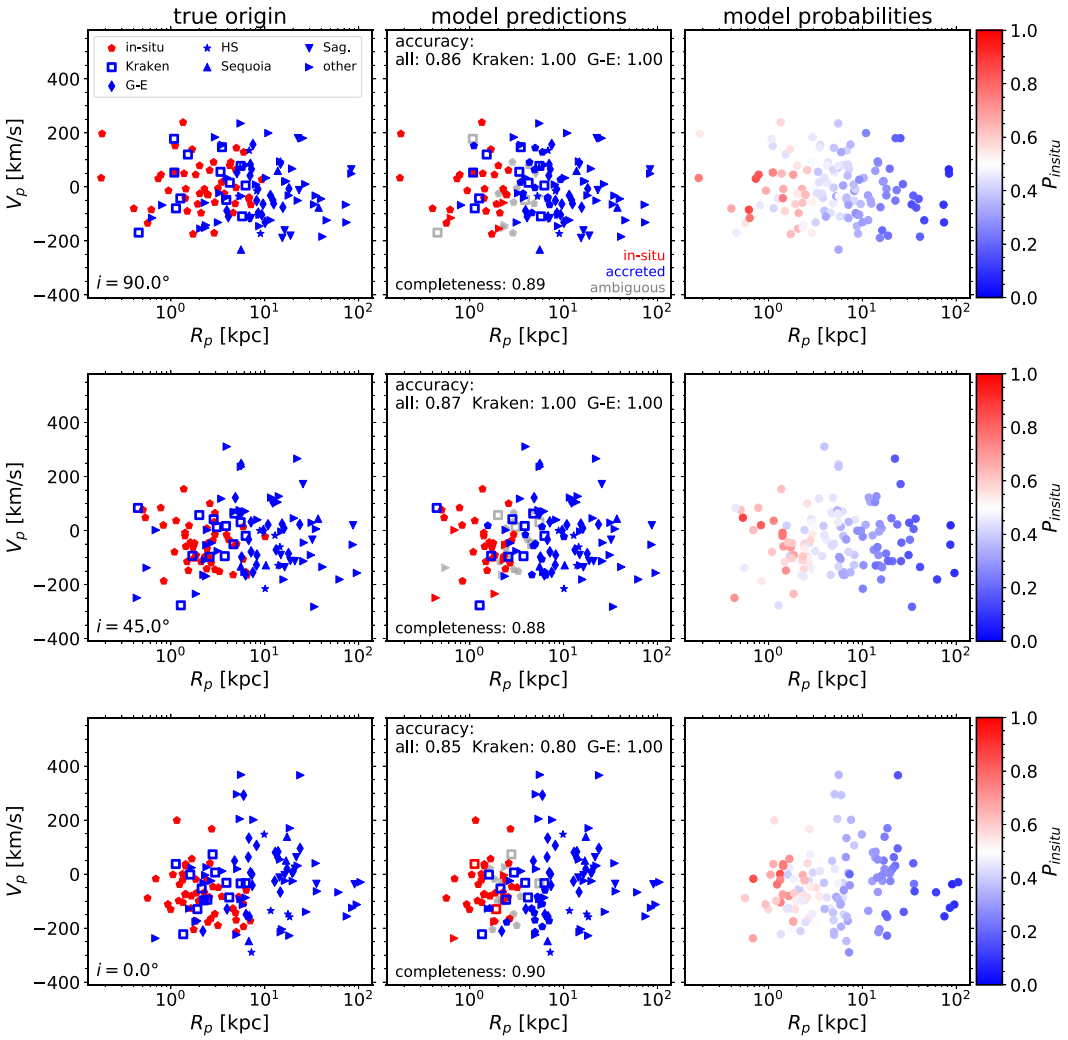

Figure 17. Predictions for the origin of the MW GCs as a function of line-of-sight velocity and projected galactocentric distance. Each row shows the results assuming that the MW is observed at a different inclination, as indicated in the legend. Left: True origin is indicated using different symbols for each progenitor galaxy, and colours indicating *in-situ* and accreted GCs. Middle: Predicted *in-situ* and accreted labels are indicated with using colour, with ambiguous classifications shown in grey (and symbols corresponding to each progenitor). Right: Predicted probability of *in-situ* origin $P_{\text{in-situ}}$. The accuracy for the entire GC system and for each of the two major progenitors is indicated in the middle panels. The performance of the minimal ANN classifier is robust to the assumed inclination of the Galaxy, and the model successfully identifies GCs in each of the five known progenitors, including at least 80 per cent of the GCs associated with *Kraken* (squares), the progenitor debris located closest to the centre of the MW, and all of the *Gaia-Enceladus* GCs.

and this would only be true if the simulation captures the physical processes responsible for the formation and evolution of the Galaxy.

The perfomance of the model ensemble on each sample as a function of the accuracy threshold is shown in the right panel of Fig. 16. In addition to the MW, we show the accuracy on the two main progenitors, *Kraken* and *Gaia-Enceladus*. The predictive accuracy of the ensemble increases with the threshold for the L^* test sample, the MW, and its two main progenitors (and increases slightly for the full test set).

To visualize the predictions, the first column of Fig. 17 shows the position-velocity diagram of the projected MW GCs coloured by their true origin, where each row corresponds to a different viewing angle. The other two columns show the origin labels (middle) and probabilities (right) predicted by the minimal classifier. To obtain the predictions we selected the model with the highest performance on the L^* test set, and further optimized its performance by tuning P_{thresh} to achieve a high accuracy and low ambiguous fraction (see Appendix C for details). Using only a single model from the ensemble may increase stochasticity (i.e. noise) in the results, but we checked explicitly that this is not the case when comparing to a voting ensemble of the 100 best models. Fig. 16 shows that selecting the model with the best performance on the L^* simulation test set guarantees a high accuracy on the MW system (blue line), without sacrificing the performance (i.e. due to overfitting) across the broad galaxy population (gray line).

The best-performing model predicts the origin of up to ∼9/10 of the MW GCs unambiguously with an accuracy of 85–87 per cent overall, and ≥80 and 100 per cent for the *Kraken* and *Gaia-Enceladus* GCs, respectively (adopting $P_{\text{thresh}} = 0.52$). Increasing the decision threshold to $P_{\text{thresh}} = 0.6$ improves the MW accuracy to 90 per cent and the ambiguous fraction to 0.4 (see Appendix C). For the baseline value $P_{\text{thresh}} = 0.5$ the performance is comparable to the accuracy obtained on the full test set drawn from the simulations (grey line in the right panel of Fig. 16), and on the 24 simulated galaxies with

masses $10^{10} < M_* < 10^{11}$ $M_\odot$ (x-axis of right panel of Fig. 16). The excellent performance on the MW GCs implies that the simulation training data accurately follows the physical processes that shape the observed properties of *in-situ* and accreted GC populations and their host galaxies in the real Universe, and that the ANN effectively learned this relation.

The first column of Fig. 17 also indicates the known galactic progenitors associated to each GC (from Kruijssen et al. 2020) using different symbols. Out of the five known progenitors that contributed accreted GCs, only the *Kraken* debris is located in the inner Galaxy, at galactocentric distances $r \lesssim 10$ kpc. This could potentially make the classification more challenging for the model, since it relies partly on the projected galactocentric distance (see Section 5.3). Despite this, we find that the model correctly identifies as accreted 8–10 out of the 13 known *Kraken* GCs (shown as squares), in addition to all the *Gaia-Enceladus* GCs, and at least a few GCs in each of the other three progenitors.

The success of the deep learning classifier in identifying debris from all the known MW progenitors has important implications for the observational reconstruction of the assembly histories of other galaxies, where only limited GC phase-space information is available. The accurate identification of accreted GCs by the model in this test shows that there is enough archaeological information in extragalactic GC observables to partially reconstruct the merger trees of galaxies in large surveys. We will investigate this intriguing possibility in future work.

5.6 Impact of uncertainties in observational data

The simulated observables used in training and evaluating the model so far assume measurements with perfect precision. Some GC and galaxy observables can include large uncertainties that arise either from the quality of the data, or from the methods used to infer the physical property from either the photometry or the spectra. Here we perform an analysis of the effect of uncertainties on the predictions to understand the sensitivity of the model, and to provide benchmarks for the expected behaviour of the model for given values of the uncertainties.

For this, we perform a Monte Carlo experiment. We first inject random noise following a normal distribution of width given by the uncertainty in each of the GC observables in the test set, [Fe/H], [α/Fe], $\log R_\mathrm{p}$, and V_p. For the metallicities, alpha-abundances, and projected positions we simply add a log-normal noise term, $\log X^\mathrm{obs} = \log X^\mathrm{true} + \mathcal{N}(0, \Delta X)$, where $X \in \{[\mathrm{Fe/H}], [\alpha/\mathrm{Fe}], \log \mathrm{R_p}\}$ and ΔX is the uncertainty. The velocity errors are calculated relative to the velocity dispersion of the galaxy, $V_\mathrm{p}^\mathrm{obs} = V_\mathrm{p}^\mathrm{true} + \sigma_\mathrm{gal}\mathcal{N}(0, \Delta V_\mathrm{p})$. For completeness we also inject Gaussian noise in the host galaxy observables, $\log M_*^\mathrm{gal}$, $\log R_\mathrm{e}^\mathrm{gal}$, $[\mathrm{Fe/H}]_\mathrm{gal}$ and $[\alpha/\mathrm{Fe}]_\mathrm{gal}$. We then use the fiducial NN classifier (trained on the unperturbed data) to obtain predictions for uncertainties in the range 0.0–0.5, equivalent to relative errors of up to a factor of 3 in [Fe/H], [α/Fe], R_p, M_*^gal, and $R_\mathrm{e}^\mathrm{gal}$, and absolute errors of up to 50 per cent of the galaxy velocity dispersion in V_p.

The resulting accuracy as a function of the relative observational uncertainty in each feature is shown in Fig. 18 for each of the GC and galaxy observables used in the fiducial feature set (see Table 2). The precision of the host galaxy alpha-element abundance dominates the prediction errors (as expected from its high importance in Fig. 13), followed by the galaxy metallicity and the GC alpha abundances. An uncertainty of 0.1 dex in $[\alpha/\mathrm{Fe}]_\mathrm{gal}$ reduces the accuracy by ∼5 per cent. The model is rather robust to large individual uncertainties in the all other GC and galaxy observables, with a decrease in accuracy of less than ∼1.5 per cent for individual relative errors as large as 0.2. The observational uncertainties in distances and LOS velocities of extragalactic GCs are typically smaller. Distances of galaxies within ∼40 Mpc can be determined to ∼10 per cent precision (e.g. Tonry et al. 2001; Blakeslee et al. 2009), and velocities to a precision $\lesssim 15$ km s^{-1} (e.g. Forbes et al. 2017), or about ∼12 per cent of the MW velocity dispersion. Uncertainties in metallicity determinations are larger, ∼0.15 dex (e.g. Caldwell & Romanowsky 2016), but still in the range where they would have a minimal effect on the accuracy of the model predictions. The results of this test indicate that the uncertainties in the host galaxy and GC alpha abundance (as well as the galaxy metallicity) will be the dominant observational sources of error in the model's predictions.

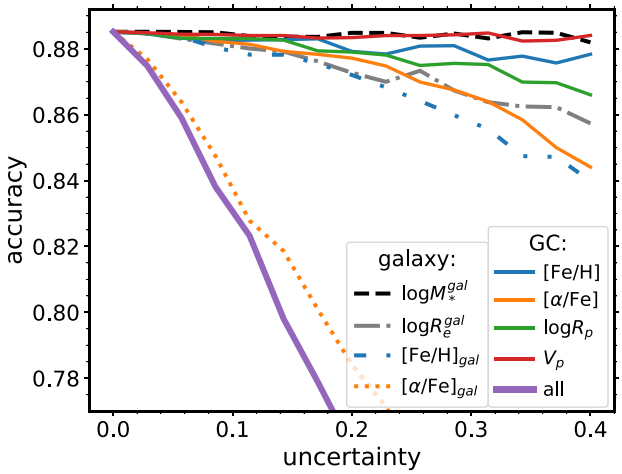

Figure 18. Impact of observational uncertainties on the accuracy of the GC origin predictions. Each line shows the accuracy as a function of the relative error in each of the GC and host galaxy observables: metallicity, alpha abundance, GC projected position and line-of-sight velocity, galaxy stellar mass, and effective radius. The bottom (thick) line shows the effect of uncertainties in all the observables combined. For V_p the x-axis represents fractional uncertainty with respect to the velocity dispersion of the host galaxy, while for all other quantities it represents order of magnitude uncertainties (i.e. in dex). The values are obtained by adding normally-distributed Monte Carlo errors to the test set drawn from the simulations. The accuracy is robust to relative uncertainties as large as ∼0.2 in the all observables except the galaxy alpha abundance (see Section 5.6 for the interpretation for each observable). The performance of the classifier is most sensitive to the precision of the alpha abundances and metallicities.

5.7 Including GC ages to improve performance

Due to limitations in the modelling of integrated spectra, GC ages are notoriously difficult to constrain beyond the Local Group (Worthey 1994). However, recent studies suggest that the precision of extragalactic GC ages can be improved significantly, reaching $\lesssim 0.1$ dex relative uncertainties (Usher et al. 2019; Cabrera-Ziri & Conroy 2022). High precision GC ages in the local Universe could therefore be within reach for wide spectroscopic surveys over the next decade. In this section we test the effect of including the ages of the simulated GCs in training the NN classifier. For this we add the precise GC age as an additional feature, and then evaluate the performance of the model on the test data from the simulation. We then run a Monte Carlo experiment to add random log-normal noise to the ages in the test data, and calculate the accuracy as a function of the uncertainty

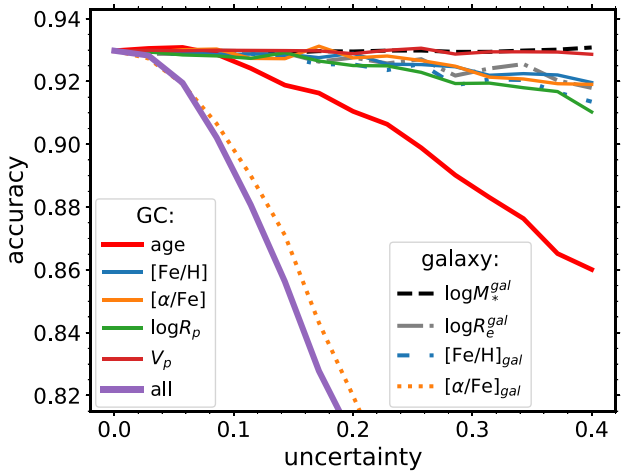

Figure 19. Accuracy of a NN classifier that includes GC ages in addition to all the features of the fiducial model. Each line shows the accuracy as a function of the relative error in the GC observables: metallicity, alpha abundances, projected position, and line-of-sight velocity. As in Fig. 18, the uncertainties in GC velocities are expressed as a fraction of the galaxy velocity dispersion, and for all other observables (including ages) in logarithmic units. The bottom line shows the combined effect of uncertainties in all the observables. The values are obtained by adding normally distributed Monte Carlo errors to the test set drawn from the simulations. Including GC ages significantly improves the accuracy of the predictions (to about 93 per cent), but the classifier becomes very sensitive to the precision of the ages for uncertainties >0.1 dex.

in the ages as well as in each of the other observables. As in the fiducial model, to remove ambiguous results we assume a decision threshold that predicts GC origin for ∼60 per cent of the test sample, $P_{\rm thresh} = 0.83$.

The impact of including GC ages on the predictions is shown in Fig. 19. Including ages increases the accuracy of the model with no uncertainties from ∼89 to ∼93 per cent. Relative uncertainties of up to ∼0.2 in the all other GC observables have almost no effect on the accuracy in this model. However, the performance of the classifier begins to drop significantly when the precision of the ages is reduced below ∼0.1 dex. This demonstrates the importance of the GC ages compared to all the other observables in shaping the model predictions. For reference, recent advances in stellar population modelling now make it possible to achieve this level of precision in the age determination of extragalactic clusters (see Cabrera-Ziri & Conroy 2022). For ages with a precision of $\lesssim 0.1$ dex (or about 25 per cent), the classifier reaches an accuracy of >92 per cent, suggesting that the current limiting precision of GC age measurements is already high enough to significantly improve the performance of the NN model.

6 DISCUSSION

The GC observables we select in this work have been found to be good indicators of GC origin in previous studies. Hughes et al. (2019b) found that the alpha-element abundance of recently accreted GCs is systematically lower at fixed [Fe/H] relative to *in-situ* GCs. Kruijssen et al. (2019a, b) showed that at a fixed age, the metallicity of GCs traces the metallicity of their galactic progenitor, and therefore GC age–metallicity relations can be used to reconstruct the assembly history of the Galaxy. Pfeffer et al. (2020) and Kruijssen et al. (2020) showed that adding 3D orbital information to the ages and metallicities allows the recovery of the masses and accretion redshifts of each progenitor.

We have extensively explored the space of GC and galaxy properties to use as features for reconstructing GC origin. In addition to our manual exploration, our choice of classifier model architecture (a densely layered NN) is meant to take advantage of the ability of these networks to capture highly non-linear relationships between the features and the output. It is therefore unlikely that we excluded a feature in the simulations that would dramatically improve the performance of the classifier. More sophisticated simulations that track the individual abundances of many isotopes (e.g. Reina-Campos et al. 2022) may capture additional information that could improve the predictions.

Another more subtle issue that arises in this type of machine learning problem is the completeness of the training set. Due to the steepness of the galaxy stellar mass function, our volume-limited simulated galaxy sample is dominated by low-mass galaxies and contains only a handful of massive elliptical galaxies. While this provides an unbiased representation of the galaxy population, it is not ideal for supervised learning. As shown in Section 5.1, the classifier has difficulties capturing the relation between GC observables and their origin in the most massive galaxies partly due to the small size of the galaxy training sample, which only includes four galaxies with $M_* > 10^{11}$ M$_\odot$ compared to 86 MW-mass ($10^{10} < M_*/{\rm M}_\odot < 10^{11}$) galaxies and 273 dwarfs in the mass range of the Magellanic Clouds ($4 \times 10^8 < M_*/{\rm M}_\odot < 3 \times 10^9$). Similarly, our results indicate that the loss of information when the phase-space distribution of the GC systems is observed in projection is one of the dominant limiting factors in the performance of our model, compared to one that takes as input the full 6D information. We have explicitly tested this hypothesis using the 'data augmentation' technique. This was done by retraining the model using an extended training set that includes three orthogonal projections of the simulation box, instead of the single projection used for the fiducial model. This procedure effectively yields a three-times larger training set. There was no significant improvement in the predictive accuracy, which suggests that the method is limited only by the lack of depth information, and not by the number of galaxies (or projections) in the training set. A further limitation of the model presented here is that the selection of the training sample implies that the results may apply only to central galaxies. Achieving a similarly good performance on satellites will likely require training specifically with satellite galaxies because their evolution is more sensitive to environmental processes.

There is also the possibility that the simulation used for training the algorithm does not capture certain aspects of the formation and evolution of galaxies and GCs, and this is the most difficult aspect of the uncertainties to quantify. As described in Section 2.1, E-MOSAICS reproduces many properties of observed galaxies and GCs. However, the EAGLE model produces L^* galaxies with stellar masses that are ∼0.1–0.2 dex below observations (Schaye et al. 2015). Furthermore, the lack of a cold interstellar medium in EAGLE results in the artificial survival of too many young, metal-rich clusters that should have otherwise disrupted [for a detailed discussion, see Pfeffer et al. (2018) and Kruijssen et al. (2019a)]. While the first problem is difficult to correct for in the training of the NN, Fig. 18 suggests that the predictions are robust to large errors in the stellar mass. We have also attempted to remove the underdisrupted GCs from the training and test samples using a metallicity selection (see Section 2.2). A new generation of simulations with better modelling of L^* galaxies and improved ISM physics will be needed to extend the origin predictions to metal-rich GCs with [Fe/H] > −0.5 (see

Reina-Campos et al. 2022), and will likely improve the identification of *in-situ* objects (see Section 5.1).

Lastly, the spread in Oxygen abundance due to the presence of multiple populations within individual GCs could introduce systematic biases in our classifier because the training data consists of simulations that do not model this phenomenology. However, while the spread within a cluster (from the lowest to the highest [O/Fe]) can be up to ∼1 dex, the scatter around the mean is typically only $\lesssim 0.25$ dex (Carretta et al. 2009). Fig. 18 shows that a 0.25 dex uncertainty in [α/Fe] would only decrease the predictive accuracy by ∼1 per cent. Relative to the field stars, GCs with multiple populations could have mean [O/Fe] values systematically lower by 0.1–0.25 dex. We have also tested this scenario and find that the accuracy drops by less than 1.5 per cent when assuming a bias of −0.25 dex in [α/Fe] for all GCs. This bias could be mitigated when applying the classifier to observations by measuring a different alpha element and then correcting to total [α/Fe] by assuming certain yields.

The reconstruction of the MW assembly history using *Gaia* and other spectroscopic surveys demonstrates that chemo-dynamical observations are a powerful tool. Thanks to these studies, the origin of the stellar halo of the MW has now been determined as a function of galactocentric radius (Naidu et al. 2020). This and other detailed observations like the radial profile of galactic components of different origin could become excellent tools to constrain cosmological hydrodynamical simulations. Simulations have already reached enough sophistication to reproduce many global galaxy observables, but still suffer from highly degenerate input physics, which limits their predictive power (for a review, see Naab & Ostriker 2017). The deep learning approach we demonstrate in this paper could in principle be extended to constrain the spatial distribution of *in-situ* and accreted stars and GCs in galaxy samples of up to millions of objects in the local Universe. Classifiers trained using observables that are independent of specific highly uncertain physical processes (i.e. stellar feedback) could determine the spatial distribution of *in-situ* and accreted material across the galaxy population. By comparing these constraints to the output of state-of-the-art cosmological simulations, their built-in hypotheses regarding the physics of star formation and feedback could be tested. Similar methods could be employed to constrain the physics of the DM particle using galaxy surveys.

7 CONCLUSIONS

In this work we use nearly a thousand simulated galaxies and their GC systems in the E-MOSAICS $(34.4\ \mathrm{Mpc})^3$ periodic volume to understand how the present day GC observables (e.g. metallicity, alpha abundances, projected distance, and velocity) can be used to infer the origin of specific GCs (i.e. *in-situ* vs. accreted). We first investigate how galaxy properties including halo mass and metallicity influence the fraction of GCs that are accreted from satellites across the galaxy mass spectrum, from dwarfs to giant ellipticals. In the second part we use supervised deep learning algorithms to model and understand the relation between GC observables in external galaxies and their *in-situ* or accreted origin. For this we exploit the success of the E-MOSAICS cluster formation and evolution physics in reproducing the observed properties of GCs in the local Universe. We train a Multilayer Perceptron NN on the mapping between 17 GC and host galaxy observable features (see Table 2), and their true origin labels (i.e. *in-situ* versus accreted). We test the performance of the classifier on an independent random subset comprised of ∼20 per cent of the simulated galaxies, and use the known origin of the MW GCs to benchmark the model for application on extragalactic GC systems. We investigate the importance of each observable for determining the predictions of the classifier, and the effect that uncertainties in the observations have on the accuracy of the predictions. Finally, we explore the benefits of including GC ages.

Our conclusions are summarised as follows:

(i) The balance of *in-situ* formation and accretion of GCs is strongly shaped by galaxy mass, in a similar way as for the field stars. The median accreted fraction of GCs increases with mass, such that dwarf galaxies are typically dominated by *in-situ* GCs, and massive ellipticals contain mostly accreted GCs (Fig. 2). Despite the large scatter in accreted GC fractions across the simulated galaxies, we find a weak trend with halo mass: at fixed stellar mass, galaxies in more massive haloes host larger fractions of accreted GCs (Fig. 2). Metal-poor galaxies also tend to have larger accreted GC fractions due to a larger contribution of relatively metal-poor satellites to their assembly, and the late formation of their DM haloes (Fig. 3).

(ii) There is a strong dependence of GC origin on GC metallicity. Metal-poor GCs are typically a mix of *in-situ* and accreted objects, whereas the origin of metal-rich GCs depends on stellar mass: in low-mass galaxies (with $M_* < 10^{10}\ \mathrm{M}_\odot$) they are almost entirely formed *in situ*, and in galaxies more massive than the MW they are mostly accreted (Figs 4 and 5).

(iii) A Multilayer Perceptron NN classifier trained on the observable properties of more than 50 000 GCs hosted by 736 simulated galaxies predicts the *in-situ*/accreted origin of GCs in a test sample drawn from the same simulation with an overall accuracy of ∼89 per cent for objects with unambiguous labels (with a completeness of 60 per cent; Section 4.2 and 5.1). The classifier is excellent at identifying accreted GCs (6 per cent false-positive rate), and less accurate for *in-situ* GCs (18 per cent false-positive rate; Fig. 8). The model performs generally well in low-mass galaxies (below the mass of the MW), but has more difficulty identifying *in-situ* GCs in the most massive galaxies (Figs 10 and 11). This is likely due to the similarity of the observables of *in-situ* and accreted populations in massive galaxies (Fig. 14), their low fraction of *in-situ* GCs, the small number of these galaxies in the simulated volume (∼6), and the exclusion of GCs with [Fe/H] > -0.5 from the sample.

(iv) The classifier uses only a few dominant observables to predict GC origin. These include the effective radius, stellar mass, and alpha-element abundance of the host galaxy, together with the GC metallicity and alpha-abundance relative to the galaxy, and its projected angular momentum and galactocentric radius (Fig. 13). The high predictive importance of the galaxy effective radius seems to originate from its correlation with the assembly time-scale of the galaxy and its effect on the GC accreted fraction (see Section 3). Simulated galaxies with larger effective radii formed later and in more massive DM haloes with larger accreted fractions.

(v) Using the simulated test data, we find a significant correlation between the mean prediction confidence (an output of the NN classifier) and the accuracy for each galaxy. This allows us to estimate the likelihood that predictions for GC origin in a real galaxy will reach a minimum desired accuracy (Fig. 15).

(vi) After removing observables that are either unimportant or difficult to obtain, we test a minimal version of the classifier on the MW GCs with known origin. Assuming that the Galaxy is observed in projection, the optimized model achieves excellent performance, with an accuracy of ∼85–90 per cent that is nearly independent of the inclination. The model identifies GCs associated to each of the five known GC-rich progenitor galaxies, including most of the GCs accreted from *Kraken*, and all of the *Gaia-Enceladus* GCs.

(vii) The classifier is robust to relatively large uncertainties in nearly all observables (i.e. larger than in currently available data). Relative uncertainties in the GC and host galaxy observables of up to ∼0.1 dex decrease the predictive accuracy on the test data by less than 5 per cent. The accuracy is most sensitive to the precision of the galaxy alpha abundance (Fig. 18).

(viii) Including GC ages as an additional feature in the model significantly increases the performance, with an accuracy on the simulation test data of ∼93 per cent. This increase in performance requires a precision of <0.1 dex (or ∼25 per cent) in the age measurements. Ages with lower than 0.1 dex precision produce a realtively steep decrease in accuracy (Fig. 19).

The NN classifier developed in this work can be readily used to make predictions for the origin of GCs in nearby galaxies for which metallicity, alpha-abundance, positions, and radial velocities have been measured. Over the next decade, wide-field space-based surveys will allow these data to be collected for very large samples of galaxies. The model developed in this work is the initial step in piecing together the assembly histories of galaxies beyond the MW as a function of mass and environment, leading to a detailed understanding of the process of galaxy formation. In future work we will explore efficient methods to constrain galaxy merger histories using GC observables. In a follow-up paper we apply the model to predict the origin of the GCs in M31 (Trujillo-Gomez et al., in preparation).

The python implementation of the fiducial and minimal classifiers in KERAS, along with an example of their use in an interactive JUPYTER notebook is available at https://github.com/sebastian-tg/GC-origin-ANNclassifier.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referee for their valuable contribution to the review process and for improving the quality of this article. STG is grateful to Kai Polsterer for useful discussions regarding ML methods. STG gratefully acknowledges the generous and invaluable support of the Klaus Tschira Foundation, as well as funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 138713538 – SFB 881 ('The Milky Way System', subproject A08). STG and JMDK gratefully acknowledge funding from the European Research Council (ERC-StG-714907, MUSTANG). JMDK gratefully acknowledges funding from the German Research Foundation (DFG – Emmy Noether Research Group KR4801/1-1). COOL Research DAO is a Decentralised Autonomous Organisation supporting research in astrophysics aimed at uncovering our cosmic origins. MRC gratefully acknowledges the Canadian Institute for Theoretical Astrophysics (CITA) National Fellowship for partial support. JP is supported by the Australian government through the Australian Research Council's Discovery Projects funding scheme (DP200102574). RAC is supported by the Royal Society. NB gratefully acknowledges financial support from the European Research Council (ERC-CoG-646928, Multi-Pop) as well as from from the Royal Society (University Research Fellowship). This study was supported by the Klaus Tschira Foundation. This work used the DiRAC Data Centric system at Durham University, operated by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). This equipment was funded by BIS National E-infrastructure capital grant ST/K00042X/1, STFC capital grants ST/H008519/1 and ST/K00087X/1, STFC DiRAC Operations grant ST/K003267/1 and Durham University. DiRAC is part of the National E-Infrastructure. The work also made use of high performance computing facilities at Liverpool John Moores University, partly funded by the Royal Society and LJMU's Faculty of Engineering and Technology.

This work made use of the following software packages: NUMPY (Oliphant 2006), SCIPY (Virtanen et al. 2019), MATPLOTLIB (Hunter 2007), PANDAS (McKinney 2010), SEABORN (Waskom 2021), JUPYTER (Kluyver et al. 2016), PYNBODY (Pontzen et al. 2013), SCIKIT-LEARN (Pedregosa et al. 2011), TENSORFLOW (Abadi et al. 2015), and KERAS (Chollet et al. 2015).

DATA AVAILABILITY

The data underlying this article will be made available upon reasonable request to the corresponding author.

REFERENCES

Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/
Abbott T. M. C. et al., 2018, ApJS, 239, 18
Bastian N., Pfeffer J., Kruijssen J. M. D., Crain R. A., Trujillo-Gomez S., Reina-Campos M., 2020, MNRAS, 498, 1050
Baumgardt H., Hilker M., Sollima A., Bellini A., 2019, MNRAS, 482, 5138
Belokurov V., 2013, New Astron. Rev., 57, 100
Belokurov V., Erkal D., Evans N. W., Koposov S. E., Deason A. J., 2018, MNRAS, 478, 611
Blakeslee J. P., Tonry J. L., Metzger M. R., 1997, AJ, 114, 482
Blakeslee J. P. et al., 2009, ApJ, 694, 556
Bland-Hawthorn J., Gerhard O., 2016, ARA&A, 54, 529
Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, 517
Boser B. E., Guyon I. M., Vapnik V. N., 1992, in Proceedings of the Fifth Annual Workshop on Computational Learning Theory. COLT'92. Association for Computing Machinery, New York, NY, USA, p. 144
Breiman L., 2001, Mach. Learn., 45, 5
Burkert A., Forbes D. A., 2020, AJ, 159, 56
Cabrera-Ziri I., Conroy C., 2022, MNRAS, 511, 341
Caldwell N., Romanowsky A. J., 2016, ApJ, 824, 42
Carretta E. et al., 2009, A&A, 505, 117
Cautun M. et al., 2020, MNRAS, 494, 4291
Chambers K. C. et al., 2016, preprint (arXiv:1612.05560)
Chollet F. et al., 2015, Keras. https://keras.io
Clauwens B., Schaye J., Franx M., Bower R. G., 2018, MNRAS, 478, 3994
Conroy C. et al., 2019, ApJ, 883, 107
Crain R. A., van de Voort F., 2023, ARA&A, 61, 473
Crain R. A. et al., 2015, MNRAS, 450, 1937
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Davison T. A., Norris M. A., Pfeffer J. L., Davies J. J., Crain R. A., 2020, MNRAS, 497, 81
Deason A. J., Belokurov V., Sanders J. L., 2019, MNRAS, 490, 3426
Dolag K., Borgani S., Murante G., Springel V., 2009, MNRAS, 399, 497
Eggen O. J., Lynden-Bell D., Sandage A. R., 1962, ApJ, 136, 748
Fix E., Hodges J. L., 1989, Int. Stat. Rev., 57, 238
Forbes D. A. et al., 2017, AJ, 153, 114
Forbes D. A., Read J. I., Gieles M., Collins M. L. M., 2018, MNRAS, 481, 5592
Furlong M. et al., 2015, MNRAS, 450, 4486
Furlong M. et al., 2016, MNRAS, 465, 722
Gaia Collaboration, 2018a, A&A, 616, A1
Gaia Collaboration, 2018b, A&A, 616, A12
Gallart C., Bernard E. J., Brook C. B., Ruiz-Lara T., Cassisi S., Hill V., Monelli M., 2019, Nat. Astron., 3, 932
Georgiev I. Y., Puzia T. H., Goudfrooij P., Hilker M., 2010, MNRAS, 406, 1967

Grillmair C. J., Carlin J. L., 2016, in Newberg H. J., Carlin J. L.eds, Astrophysics and Space Science Library, Vol. 420, Tidal Streams in the Local Group and Beyond. p. 87, preprint (arXiv:1603.08936),
Harris W. E., 1996, AJ, 112, 1487
Haywood M., Di Matteo P., Lehnert M. D., Snaith O., Khoperskov S., Gómez A., 2018, ApJ, 863, 113
Helmi A., 2020, ARA&A, 58
Helmi A., Babusiaux C., Koppelman H. H., Massari D., Veljanoski J., Brown A. G. A., 2018, Nature, 563, 85
Hoerl A. E., Kennard R. W., 1970, Technometrics, 12, 55
Horta D. et al., 2021, MNRAS, 500, 1385
Hughes M. E., Pfeffer J., Martig M., Bastian N., Crain R. A., Kruijssen J. M. D., Reina-Campos M., 2019a, MNRAS, 482, 2795
Hughes M. E., Pfeffer J. L., Martig M., Reina-Campos M., Bastian N., Crain R. A., Kruijssen J. M. D., 2019b, MNRAS, 491, 4012
Hughes M. E., Pfeffer J. L., Bastian N., Martig M., Kruijssen J. M. D., Crain R. A., Reina-Campos M., Trujillo-Gomez S., 2022, MNRAS, 510, 6190
Hunt E. B., Marin J., Stone P. J., 1966
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ibata R. A., Gilmore G., Irwin M. J., 1994, Nature, 370, 194
Iorio G., Belokurov V., 2019, MNRAS, 482, 3868
Jiang L., Helly J. C., Cole S., Frenk C. S., 2014, MNRAS, 440, 2115
Kingma D. P., Ba J., 2014, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings
Klement R. J., 2010, A&AR, 18, 567
Kluyver T. et al., 2016, in Loizides F., Schmidt B.eds, Positioning and Power in Academic Publishing: Players, Agents and Agendas. p. 87
Koppelman H. H., Helmi A., Massari D., Roelenga S., Bastian U., 2019a, A&A, 625, A5
Koppelman H. H., Helmi A., Massari D., Price-Whelan A. M., Starkenburg T. K., 2019b, A&A, 631, L9
Kruijssen J. M. D., 2012, MNRAS, 426, 3008
Kruijssen J. M. D., Pelupessy F. I., Lamers H. J. G. L. M., Portegies Zwart S. F., Icke V., 2011, MNRAS, 414, 1339
Kruijssen J. M. D., Pfeffer J. L., Crain R. A., Bastian N., 2019a, MNRAS, 486, 3134
Kruijssen J. M. D., Pfeffer J. L., Reina-Campos M., Crain R. A., Bastian N., 2019b, MNRAS, 486, 3180
Kruijssen J. M. D. et al., 2020, MNRAS, 498, 2472
Mackereth J. T., Crain R. A., Schiavon R. P., Schaye J., Theuns T., Schaller M., 2018, MNRAS, 477, 5072
Mackereth J. T. et al., 2019, MNRAS, 482, 3426
McKinney W., 2010, in van der Walt S., Millman J.eds, Proceedings of the 9th Python in Science Conference. p. 51
Malhan K. et al., 2022, ApJ, 926, 107
Massari D., Koppelman H. H., Helmi A., 2019, A&A, 630, L4
Moster B. P., Naab T., White S. D. M., 2020, MNRAS, 499, 4748
Myeong G. C., Evans N. W., Belokurov V., Amorisco N. C., Koposov S. E., 2018a, MNRAS, 475, 1537
Myeong G. C., Evans N. W., Belokurov V., Sanders J. L., Koposov S. E., 2018b, MNRAS, 478, 5449
Myeong G. C., Evans N. W., Belokurov V., Sanders J. L., Koposov S. E., 2018c, ApJ, 856, L26
Myeong G. C., Vasiliev E., Iorio G., Evans N. W., Belokurov V., 2019, MNRAS, 488, 1235
Naab T., Ostriker J. P., 2017, ARA&A, 55, 59
Naidu R. P., Conroy C., Bonaca A., Johnson B. D., Ting Y.-S., Caldwell N., Zaritsky D., Cargile P. A., 2020, ApJ, 901, 48
Navarro J. F., Frenk C. S., White S. D. M., 1995, MNRAS, 275, 56
Necib L. et al., 2020a, Nat. Astron., 4, 1078
Necib L., Ostdiek B., Lisanti M., Cohen T., Freytsis M., Garrison-Kimmel S., 2020b, ApJ, 903, 25
Oliphant T., 2006, NumPy: A guide to NumPy. Trelgol Publishing, USA, http://www.numpy.org/
Pearl R., Reed L. J., 1920, Proc. Natl. Acad. Sci., 6, 275
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Peng E. W. et al., 2008, ApJ, 681, 197
Pfeffer J., Kruijssen J. M. D., Crain R. A., Bastian N., 2018, MNRAS, 475, 4309
Pfeffer J. L., Trujillo-Gomez S., Kruijssen J. M. D., Crain R. A., Hughes M. E., Reina-Campos M., Bastian N., 2020, MNRAS, 499, 4863
Poci A. et al., 2021, A&A, 647, A145
Pontzen A., Roškar R., Stinson G. S., Woods R., Reed D. M., Coles J., Quinn T. R., 2013, pynbody: Astrophysics Simulation Analysis for Python
Qu Y. et al., 2017, MNRAS, 464, 1659
Reina-Campos M., Kruijssen J. M. D., 2017, MNRAS, 469, 1282
Reina-Campos M., Trujillo-Gomez S., Deason A. J., Kruijssen J. M. D., Pfeffer J. L., Crain R. A., Bastian N., Hughes M. E., 2022, MNRAS, 513, 3925
Reina-Campos M., Keller B. W., Kruijssen J. M. D., Gensior J., Trujillo-Gomez S., Jeffreson S. M. R., Pfeffer J. L., Sills A., 2022, MNRAS, 517, 3144
Rodriguez-Gomez V. et al., 2016, MNRAS, 458, 2371
Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Nature, 323, 533
Schaye J. et al., 2015, MNRAS, 446, 521
Searle L., Zinn R., 1978, ApJ, 225, 357
Smith M. C., 2016, in Newberg H. J., Carlin J. L.eds, Astrophysics and Space Science Library, Vol. 420, Tidal Streams in the Local Group and Beyond. p. 113
Springel V., Yoshida N., White S. D. M., 2001, New Astron., 6, 79
Springel V. et al., 2005, Nature, 435, 629
Tacchella S. et al., 2019, MNRAS, 487, 5416
Tonry J. L., Dressler A., Blakeslee J. P., Ajhar E. A., Fletcher A. B., Luppino G. A., Metzger M. R., Moore C. B., 2001, ApJ, 546, 681
Trayford J. W. et al., 2015, MNRAS, 452, 2879
Trujillo-Gomez S., Kruijssen J. M. D., Reina-Campos M., Pfeffer J. L., Keller B. W., Crain R. A., Bastian N., Hughes M. E., 2021, MNRAS, 503, 31
Usher C., Pfeffer J., Bastian N., Kruijssen J. M. D., Crain R. A., Reina-Campos M., 2018, MNRAS, 480, 3279
Usher C., Brodie J. P., Forbes D. A., Romanowsky A. J., Strader J., Pfeffer J., Bastian N., 2019, MNRAS, 490, 491
Vasiliev E., 2019, MNRAS, 484, 2832
Virtanen P. et al., 2020, NatMe, 17, 261
Waskom M. L., 2021, J. Open Source Softw., 6, 3021
Wei P., Lu Z., Song J., 2015, Reliab. Eng. Syst. Saf., 142, 399
Worthey G., 1994, ApJS, 95, 107
York D. G. et al., 2000, AJ, 120, 1579
Zhu L. et al., 2020, MNRAS, 496, 1579

APPENDIX A: CLUSTERING ANALYSIS FOR FEATURE IMPORTANCE

Fig. A1 shows the results of the clustering analysis for the features of the fiducial NN classifier in Section 5.3. A threshold of 0.25 was used to select representative features in each covariant group. A new classifier was then trained on the selected subset of independent features to obtain an unbiased estimate of the predictive importance of each one.

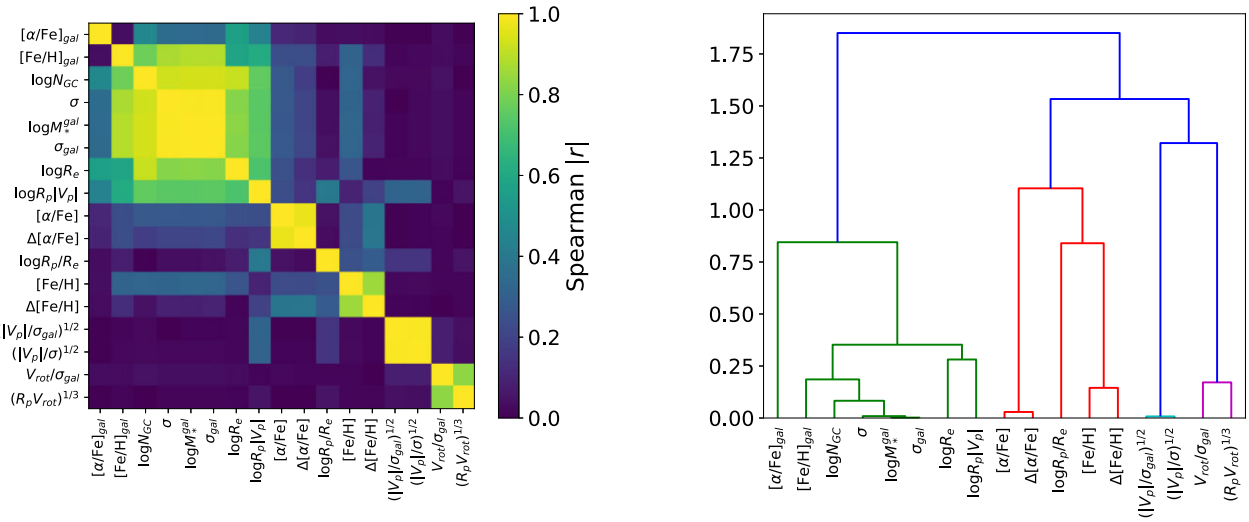

Figure A1. Clustering of GC and galaxy features. Left: Spearman correlation matrix. Right: Dendrogram of correlated feature clusters.

APPENDIX B: DISTRIBUTION OF GC OBSERVABLES WITH THE HIGHEST PREDICTIVE POWER

Fig. B1 shows the joint distribution of the seven most important GC origin predictors across the entire simulated GC sample, with colour indicating their true origin. The observables with the highest importance also show the most distinct separation in the distributions of *in-situ* and accreted GCs. This qualitatively confirms the result of the permutation importance analysis, and shows that the classifier effectively uses the GC and galaxy properties that correlate most with GC origin.

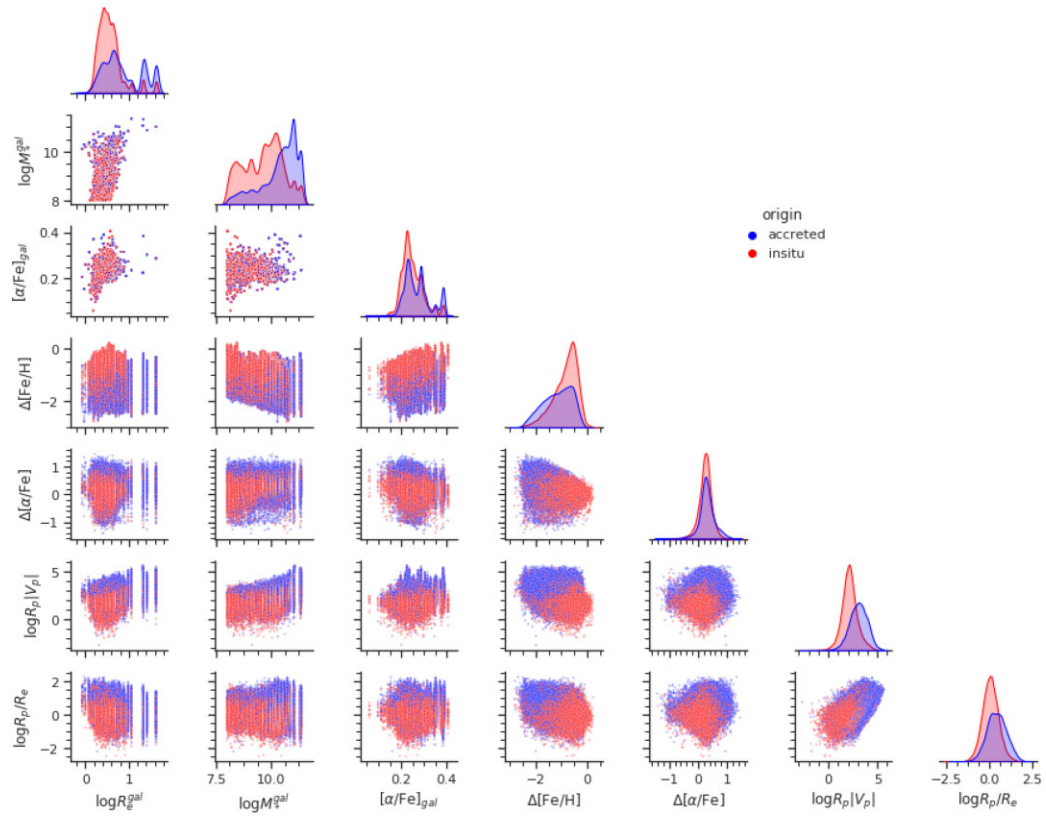

Figure B1. Joint and marginal distributions of GC origin across the galaxy and GC observables with the most predictive power. The panels show the distribution of *in-situ* and accreted GCs across the entire simulated sample in the space of the observables with the most predictive power (see Fig. 13 and Section 5.3). *In-situ* GCs are coloured red, while accreted GCs are shown in blue. Significant overlap of the two classes limits the predictive power of individual observables, but the NN classifier is able to combine them optimally.

APPENDIX C: PERFORMANCE OF THE MINIMAL CLASSIFIER

Fig. C1 shows the performance of the best minimal classifier and the fraction of unambiguous predictions as a function of the adopted decision threshold. The accuracy is shown for both the simulation GC test set and the MW GC system. These values can be used to obtain a rough estimate of the expected performance on observed galaxies.

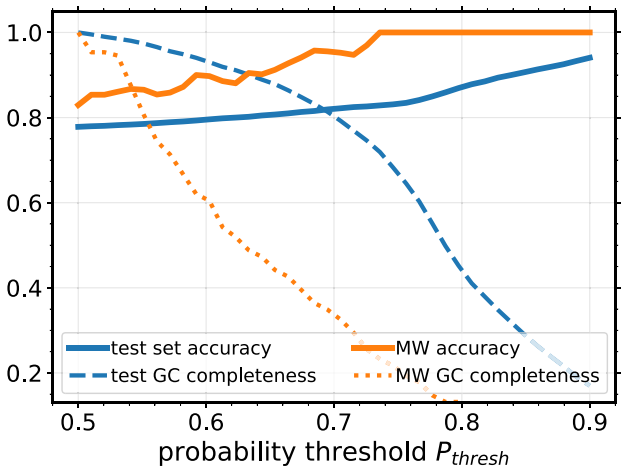

Figure C1. Accuracy and sample completeness of predictions as a function of the decision threshold adopted for the minimal classifier. The coloured solid lines indicate the performance on the simulation test set (blue), and on the MW GC system (orange).

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