Astronomy & Astrophysics manuscript no. main
June 2, 2026

Predicting Images for the Dynamics Of globular Clusters: π-DOC

Arn Marklund[1], Paolo Bianchini[1], Paul Magron[2], Abbas Askar[3,·], and Ariance Lançon[1]

[1] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, F-67000 Strasbourg, France
e-mail: arn.marklund@astro.unistra.fr
[2]

Received September 15, 1996; accepted March 16, 1997

ABSTRACT

Aims. Recent and upcoming wide-field photometric surveys will observe hundreds of thousands of globular clusters (GCs) in the next decade. This poses an emerging challenge: to derive reliable estimates of their properties using methods that scale to large volumes of data. We present a scalable framework that combines forward modelling with deep learning, allowing accurate and detailed analysis of GCs.
Methods. We develop `CREMANT`, a forward-modeling code that creates realistic mock images of GCs from state-of-the-art simulations by embedding them in real backgrounds. Using HST observations of M31 (PHAT and PHAST surveys) as a test case, we train a neural network, π-DOC, to decontaminate GC images from field stars, reconstruct 2D mass maps, and infer global properties (age, distance, ellipticity, and position angle) from multi-band photometric images and pixel colour-magnitude diagrams.
Results. We find that the training data covers a majority of the parameter space of M31's GCs. On synthetic data, π-DOC achieves mass uncertainties below a factor of two (comparable to typical discrepancies among Galactic GCs) and robustness to low signal-to-noise, maintaining mass accuracy within a factor of two and age accuracy within 2.5 Gyr for GCs up to five magnitudes fainter than the background. Applied to GCs in the PHAT and PHAST footprints, the model agrees well with the literature showing agreement within a factor of four for total masses, and furthermore scales to large volumes of data: $\sim$ 10 objects per second (full M31 sample in $\sim$ 30 s). We release a publicly available catalogue of π-DOC's estimates for 349 GCs in M31, as well as both `CREMANT` and π-DOC.
Conclusions. Out-of-distribution detection based on a principal component analysis highlight massive and metal rich clusters as the current biggest limitation in our training data, but also find that metallicity-related information is already encoded in the latent space of the network, despite not being a part of the outputs. We highlight limitations of the current architecture and discuss improvements necessary for extending the model to become more dynamic and survey-agnostic. Our results establish our framework as a scalable and physically motivated approach to inferring reliable GC properties from large multi-band photometric datasets.

Key words. methods: numerical, deep learning – globular clusters: general – galaxies: star clusters: general – objects: M31

1. Introduction

Globular clusters (GCs) are dense, gravitationally bound stellar systems composed of up to a few million stars. They are among the oldest known stellar populations in the Universe, with Galactic GCs showing ages of up to 13 Gyr (Leaman et al. 2013; Valcin et al. 2025; Valcin et al. 2026). Thanks to recent gravitationally lensed observations with the James Webb Space Telescope (JWST), the progenitors of present-day GCs have been identified at redshifts corresponding to only about 500 Myr after the Big Bang (Adamo et al. 2024; Mowla et al. 2024; Vanzella et al. 2023). While their exact origin still remains an open questions, their formation is linked to key phases of structure formation in the early Universe (e.g. high pressure star formation and galactic merger events Kruijssen 2015; Lahén et al. 2020, 2025; Taylor et al. 2025).

Over their roughly 13 Gyr of evolution, GCs are shaped by internal dynamical processes, such as two-body relaxation and stellar evolution, as well as by external effects, such as the coupling to the tidal field of their host galaxy. Related to the possible formation channel of GCs during a galaxy's assembly phase, some GCs are also thought to have been accreted during galaxy merger events (Forbes 2020; Forbes & Bridges 2010; Forbes et al. 2018; Massari et al. 2019; Myeong et al. 2019; Pfeffer et al. 2021; Renaud et al. 2017; Chen & Gnedin 2024), or even be the remnant nuclei of dwarf galaxies (Alfaro-Cuello et al. 2019; Ibata et al. 1995; Pagnini et al. 2025, 2026). The interplay between these mechanisms and the resulting complex evolution over a Hubble-time, make GCs useful testbeds for many different astrophysical topics: stellar evolution models, dynamical theories of collisional systems, and the long-term effects of galactic environments on stellar populations.

The development of high-precision N-body (Aarseth 2003) and Monte Carlo codes (such as the MOCCA and CMC codes Hypki & Giersz 2013; Giersz et al. 2013; Rodriguez et al. 2016), and particularly GPU-accelerated implementations such as `Nbody6++GPU` (Wang et al. 2015), has made it possible to simulate clusters containing up to and above a million stars, reaching parity with the mass regime of typical Galactic GCs (e.g. Bianchini et al. 2026; Arca sedda et al. 2024; Wang et al. 2016). These modern simulations also include ingredients such as time-dependent tidal fields (Renaud et al. 2011; Webb et al. 2024) and stellar evolution (Bissekenov et al. 2025; Kamlah et al. 2022; Wu et al. 2026), offering increasingly realistic representations of GC dynamics, stellar remnants, and internal rotation. Despite these advances, such direct simulations remain extremely demanding in terms of computational time and energy consumption, limiting their use for large-scale parameter studies or statistical comparisons (Bianchini et al. 2026).

In parallel, the growth of deep-field, as well as wide-field surveys, are pushing us into a new era for GC observations. Space-

and ground-based facilities such as HST, JWST (Gardner et al. 2023), Euclid (Euclid Collaboration et al. 2022), and the Vera C. Rubin Observatory (LSST Ivezić et al. 2019) are rapidly increasing both the number of known clusters and the redshift range over which they can be studied. Recent projections estimate that Euclid alone could detect several hundred thousand GCs across the local Universe (Euclid Collaboration et al. 2025). Upcoming missions, including the Roman space telescope (Mosby et al. 2020), are expected to continue this trend by delivering deep, high-resolution multi-band imaging over large areas of the sky. These developments underline an emerging challenge: the need for scalable, automated tools capable of analysing vast samples of clusters efficiently.

Determining fundamental GC properties such as total mass and age remains non-trivial, in particular due to their complex dynamical evolution (Baumgardt & Makino 2003; Bianchini et al. 2016; Gieles et al. 2011; Jindal et al. 2019; Kuzma et al. 2016; Malhan et al. 2018; Spitzer 1987; Trenti & van der Marel 2013; Vesperini & Heggie 1997; Watkins et al. 2015; Webb & Vesperini 2016) and because of degeneracies in simple stellar population modelling, notably the degeneracies between age, metallicity, and extinction (Usher et al. 2019, 2024; Worthey 1994, 1999) and systematic differences between SSP models (Conroy et al. 2009; Conroy & Gunn 2010; Fan & de Grijs 2012; Maraston 2005).

Typical approaches involve assuming a fixed mass-to-light (M/L) ratio, which is not strictly valid for GCs (as demonstrated in e.g. Baumgardt 2017; Bianchini et al. 2017); deriving dynamical masses from kinematic data (Bellini et al. 2017), which becomes increasingly difficult at extragalactic distances; or constructing large grids of N-body simulations (Baumgardt 2017), a procedure that is both computationally expensive and environmentally costly (Bianchini et al. 2026). Monte-Carlo based methods, such as MOCCA (Hypki & Giersz 2013; Giersz et al. 2013), however, are computationally cheaper and useful for large grid explorations, but relies on more stringent assumptions (such as spherical symmetry). Age and metallicity estimates are usually obtained through isochrone fitting of color–magnitude diagrams (CMDs; Dotter et al. 2010; Leaman et al. 2013; Rosenberg et al. 1999; Valcin et al. 2025; Valcin et al. 2026; Zoccali et al. 2003), a method that is relatively robust considering the overall difficulty in determining ages, but that is not easily scalable to the volume of data expected from next-generation surveys.

In this context, machine learning, and in particular deep learning (DL), offers promising alternatives (Bialopetravičius & Narbutis 2020a,b; Guiglion et al. 2024; Hiegel et al. 2023). DL approaches have already shown great potential in automating the classification of extragalactic star clusters (Zhang et al. 2025), estimating their ages (Boin et al. 2026; Viaña et al. 2026) and dynamical properties (Askar et al. 2019; Bialopetravičius et al. 2019; Pasquato & Chung 2016; Pasquato et al. 2024), and more recently, estimating global parameters of GCs directly from imaging data (Chardin & Bianchini 2021). The latter introduced the proof-of-concept algorithm π-DOC (Predicting Images for the Dynamics Of globular Clusters), trained on mock photometric observations generated from forward-modeled N-body simulations (Bianchini et al. 2017; Miholics et al. 2016). The study demonstrated that deep networks can infer underlying physical quantities of GCs, such as their internal mass distribution, age, and distance, directly from photometric images. This approach offers not only an efficient and scalable method, but is also one of the first tools capable of mapping a two-dimensional mass distribution without assuming a mass-to-light (M/L) ratio.

The initial proof-of-concept algorithm relied on idealised mock images constructed from low-mass GC simulations, lacking realistic observational uncertainties. Neither instrumental noise nor galactic field-star contaminants were modeled, and consequently the network cannot reliably be employed on observational images unless the GC has been isolated beforehand. Moreover, the limited parameter coverage restricted applicability to just five Galactic GCs. Such limitations reflect a synthetic gap; a mismatch between idealised training data and complex real-world observations that typically degrades performance. Nevertheless, these results demonstrated the strong potential of DL approaches for large-scale analyses of GCs.

In this study, we address the limitations of the previous proof-of-concept version of π-DOC. In particular, we extend the underlying data with more massive GC simulations from both the `ROLLIN'` suite (Bianchini et al. 2026; Marklund et al. in prep.) and the MOCCA simulations (Askar et al. 2025; Zhao et al. 2026).In addition, we combine our synthetic mock images with real observations of field stars, simultaneously modelling instrumental and physical uncertainties from observations. These steps ensure we narrow the synthetic gap. The underlying data and the construction of realistic mock images are described in more detail in Section 2.

The previous version of π-DOC was split into two parts: one autoencoder, which mapped the single input image into a two-dimensional mass distribution, and one classical convolutional neural network (CNN) which predicted the cluster's age and distance, from the same input image as the autoencoder. We have combined the two networks into one multi-task network, which now utilises multi-band photometric images, and in addition predicts a decontaminated version of the observed cluster (i.e. removal of field stars), as well as the cluster's ellipticity and corresponding position angle. A full description of the new architecture can be found in Section 3, and the code is available online[1].

As a concrete test case for developing an efficient algorithm intended for large-scale surveys, we apply π-DOC to GCs in the Andromeda galaxy (M31) using archival Hubble Space Telescope (HST) data. In this work, the network is therefore restricted to GCs that are at least partially resolved. This constitutes an initial step toward a more general tool that is scalable to the growing volume of observational data and applicable to detailed analyses of extragalactic GC systems. Future extensions will focus on quasi-resolved and unresolved GCs, which are expected to constitute the majority of clusters in current and upcoming large-scale surveys (see e.g. Euclid Collaboration et al. 2025).

The remainder of the paper is structured as follows. In Section 4, we assess the performance of π-DOC using realistic mock observations, exploring its robustness and reliability and deriving conservative confidence limits expected to apply to comparable real observations. In Section 5, we apply our algorithm to M31 GCs, present the main results, and compare π-DOC's estimates with the literature. We conclude with a brief discussion of future steps in Section 6 and a summary of our findings in Section 7.

2. Real and synthetic data

π-DOC represents a significant step towards automating the detailed analysis of GCs directly from multi-band photometric observations. Because we employ a supervised deep neural net-

[1] github link here

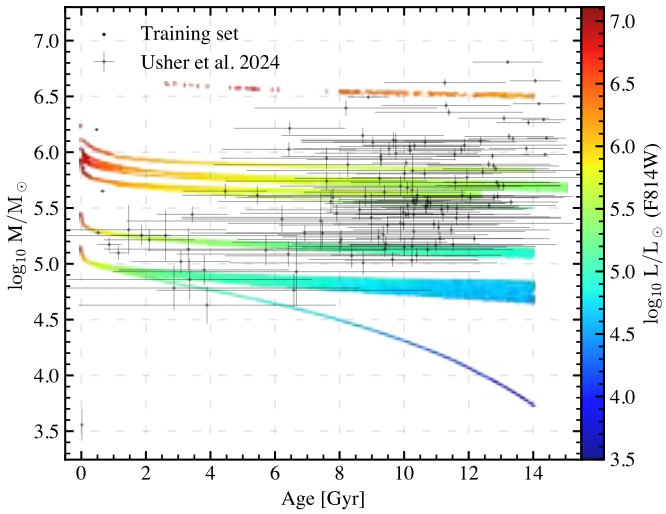

Fig. 1. Comparison of the mass–age relation for our mock images and the M31 globular clusters from (Usher et al. 2024) located within the PHAT and PHAST footprints. The mock images are colour-coded by their total luminosity in the F814W passband, while the observed GCs are shown as black points. Our mock sample broadly covers the parameter space of the M31 GC system, and, given the age uncertainties of the observed clusters, the most important aspect is a representative sampling of the mass distribution. Nonetheless, some gaps remain, particularly at the highest masses.

work, it is essential that the training data are both physically relevant and representative of real observations. In this section, we first describe the real observational data we consider (Section 2.1), and then explain how these observations guide the construction of realistic synthetic mock images (Sections 2.2–2.3), thereby reducing the synthetic gap. Finally, we summarise the datasets used throughout the remainder of this paper (Section 2.4).

2.1. Real GC data

In this work we consider GCs in M31, using archived HST data from the Panchromatic Hubble Andromeda Treasury (PHAT) survey[2] (Dalcanton et al. 2012; Williams et al. 2014, 2023), as well as the recent extension to the southern parts of M31 (PHAST, Chen et al. 2025). The PHAT and PHAST surveys each cover about a third of M31's star forming disk, divided into 23 and 13 bricks, respectively. The former is observed in six passbands: F110W, F160W, F275W, F336W, F475W and F814W, whereas the PHAST survey only has the latter two available for mosaic images. We consider all bricks, but due to low signal-to-noise (S/N, Williams et al. 2023) we choose to only work with the F336W (where applicable), F475W and F814W passbands. We use drizzled images (multiple exposures), and align them north to east, conserving the flux and assuming a common scale of 0.04 arcsec/pixel using the reproject package[3]. We select GCs from the Peacock et al. (2010a,b) catalogue that lie within the boundaries of the PHAT and PHAST footprints, amounting to a total of 349 GCs. Every image of a GC we consider has a 12" × 12" field of view (FoV), which is large enough to include typical GCs in M31 to a few times the half-light radii (typical half-light radii of M31 GCs are $\sim 1-2$" Barmby & Huchra 2001; Huxor et al. 2014), and found at a distance of $\sim$ 785 kpc (McConnachie et al. 2005).

In addition, we sample every brick and keep a total of 12264 images from the PHAT survey that are free of GCs and other star clusters, with the intention of using these as backgrounds[4] to our synthetic dataset (see Section 2.3 for details).

2.2. Mock data

Given the details of the real observational images, it is important that our synthetic data, which will be used to train π-DOC, reflect the same underlying features and parameter space present in the observations. In other words, we must ensure that the synthetic data share both the underlying physical properties and the observational characteristics. We address the first concern by considering two different suites from state-of-the-art simulations:

1. The first and main one, refers to the ROLLIN' suite (Bianchini et al. 2026; Marklund et al. in prep.), and consists of initially rotating, axisymmetric direct N-body simulations with an initial amount of 250k-1.5M stars, including prescriptions for stellar evolution and an external tidal field. These initially rotating models lead to a wide range of GC-morphologies due to internal rotation and interactions with the external tidal field (Marklund et al. in prep.), but remain limited in total mass and density due to the significant computational cost of direct N-body integrations. As such, they particularly resemble low-density GCs in M31 and the Milky Way. In total we use ten such simulations with an initial amount of 250k stars, three with 500k stars, and one with 1.5M stars.
2. The second group refers to MOCCA simulations, and include models of 47 Tuc, ω Centauri, and five additional high-density models which are described in more detail in (Askar et al. 2025; Zhao et al. 2026, also private communication Askar 2026). MOCCA simulations relaxes the computational constraints in the direct N-body method by assuming spherical symmetry. As such, while they do not account for a morphological variety, they allow us to run simulations with considerably more stars, higher total masses, as well as higher densities. In total we use one such model with an initial amount of 1M stars, two with 1.5M stars, five with 2M stars, one with 2.4M stars, and one with 12M stars (roughly resembling ω Centauri). These models correspond to high-density (except for ω Cen) and massive M31 and Galactic GCs.

These two groups of simulations therefore complement each other and ensure that the parameter space of our full dataset broadly matches that of real GCs in M31. Figure 1 shows the masses and ages of the mock images in our dataset (coloured points), together with the corresponding mass and age estimates for GCs in M31 from Usher et al. (2024). Although a few gaps remain, the dataset overall covers most of the relevant parameter space.

The underlying data to be used for each mock image corresponds to a snapshot from one of these simulations at different time steps. For the ROLLIN' simulations, the frequency of the snapshots vary between 1 – 15 Myr depending on the simulation, whereas for the MOCCA simulations it is either 50 or 100 Myr, depending on the simulation and the age of the GC. As such, different simulations will be more or less prominent in the overall

[2] http://dx.doi.org/10.17909/T91S30
[3] https://reproject.readthedocs.io/

[4] Note that we call these images backgrounds, but in principle they could actually be foreground images.

dataset, and in particular, there will be a strong bias towards the ROLLIN' simulations. This is also partly by construction as we are interested in exploring the morphology of GCs.

However, to mitigate biases within the dataset, we restrict the number of unique snapshots of simulations with 250k stars to min(N, 2000), while for the more massive ROLLIN' simulations (500k-1.5M stars initially) the total number of snapshots are set to min(N, 3000) where N is the total amount of available snapshots from a simulation. On the contrary, we consider all available snapshots from the MOCCA simulations. In total we are left with $\sim 40k$ unique snapshots from a total of 24 simulations.

As for the second concern, and in order to generate mock images from these simulated models of GCs that properly treat typical observational characteristics of HST M31 data, we have developed a forward modelling algorithm Cremant (Creating REalistic Mock imAges for Numerical simulaTions)[5]. The algorithm makes use of

(i) the flexible stellar population synthesis (FSPS) code (Conroy et al. 2009; Conroy & Gunn 2010) to compute individual stars' magnitudes in the passbands we consider for our mock images. The stellar parameters from the simulations that are passed to FSPS are mass, radius, luminosity, and metallicity, where the effective temperature and surface gravity are calculated in FSPS. These stellar properties (excluding metallicity which is passed as an input to the simulations) are, in turn, provided by the stellar and binary evolution from the SSE and BSE prescriptions of Hurley et al. (2000, 2002) incorporated within the star cluster evolution codes (we refer to Kamlah et al. 2022, for the stellar evolution implementation in our N-body simulations). We use PADOVA isochrones (Bressan et al. 2012) and the BASEL spectral library (Lejeune et al. 1997, 1998; Westera et al. 2002).

(ii) two angles and a distance, as free parameters, that define geometrical projections. The field of view of our images (12" × 12") and the pixelscale (0.04") are also free parameters[6] in the algorithm, and together define a two-dimensional pixel-grid onto which the particles from a simulation can be projected. To generate a projection, we first define a versor, which represents a rotation axis with respect to the coordinate axis of the simulation, and a rotation angle $\phi \in [0, \pi)$. These two together determine the inclination of the cluster with respect to the observer's line of sight. The projected cluster can subsequently be rotated in the image plane by an additional angle $\theta \in [0, 2\pi)$, applied around the axis perpendicular to the projection plane. These two rotations fully specify the apparent orientation of the simulated cluster. The cluster is then placed at a chosen distance which, together with the apparent orientation, determines the relative positions for each star within the pixel-grid.

(iii) point-spread functions (PSFs), derived from non-saturated bright stars in the outer regions of M31 (PHAT brick 22 with each passband aligned north-east) using DrizzlePac[7]. We include four distinct PSFs per passband, including STScI's empirical model[8], all accessible via Cremant. Multiple PSFs are included to prevent the network from memorising an imperfect PSF model with respect to the backgrounds' PSFs. We oversample each mock image and PSF with a factor of four before convolving the images to produce the final mock image.

2.3. Modelling noise

Our mock cluster images are idealised source-only realisations, effectively corresponding to a noise-free limit in which the GC flux is known without photon-counting noise and no background is present. To generate realistic observations, and to thereby reduce the synthetic gap, we combine our mock images with backgrounds from real observations following the steps outlined in Bialopetravičius et al. (2019). We first transform the mock images into units of counts per second, matching the background images, and then:

1. compute the median pixel value of the background and add it to each pixel of a mock cluster image;
2. draw a Poisson realisation for each pixel, in total counts, with the mean set to the corresponding pixel value;
3. subtract the median offset and add the real background image;
4. convert the pixel values to magnitudes (ST system).

This approach ensures that we model both the instrumental and background noise associated with an observation. Because the mock images represent noise-free source models, we temporarily shift them by the median background level of a real field before applying Poisson sampling to the total count rate. The temporary offset places the source model on a realistic count scale before Poisson sampling. We then remove the offset and add the full real background image, so that the final mock preserves both the observed background structure and realistic counting statistics.

We do not apply any magnitude cuts, since the background already sets the effective detection floor. During training (see Section 3.4), 10% of the mock images are also generated background-free to improve generalisation to real HST data with negligible contamination[9]. Since Cremant currently does not implement extinction, each background is corrected for Milky Way extinction using dustmaps[10] (Green 2018; Schlegel et al. 1998). This procedure is carried out on the fly during training, effectively extending our total dataset by four orders of magnitude, since each mock image can be paired with any of the 12264 backgrounds. A qualitative comparison between examples of our mock images and real GCs in M31 is shown in Fig. A.1.

2.4. Building the datasets

For each snapshot of a simulation, we create a total of 6 mock images, with three different projection angles, and each angle at two different distances. In total, we create approximately 210k mock images of GCs from the ROLLIN' simulations, and roughly 10k images from the MOCCA simulations.

[5] github link

[6] The pixel scale is set by the choice of telescope (HST), and the field of view is chosen to minimise image size while including as much of the GC as possible. In particular, we used the more massive, low-density and extended ROLLIN' models at the distance of M31 to set a reasonable FoV, and then cross-checked against observed M31 GCs, bearing in mind that their full spatial extent is not known exactly.

[7] https://github.com/spacetelescope/hst_notebooks/tree/main/notebooks/DrizzlePac

[8] https://www.stsci.edu/hst/instrumentation/wfc3/data-analysis/psf

[9] In such cases, we omit step 3 above, and stars fainter than the typical HST background therefore remain visible.

[10] https://dustmaps.readthedocs.io/en/latest/index.html

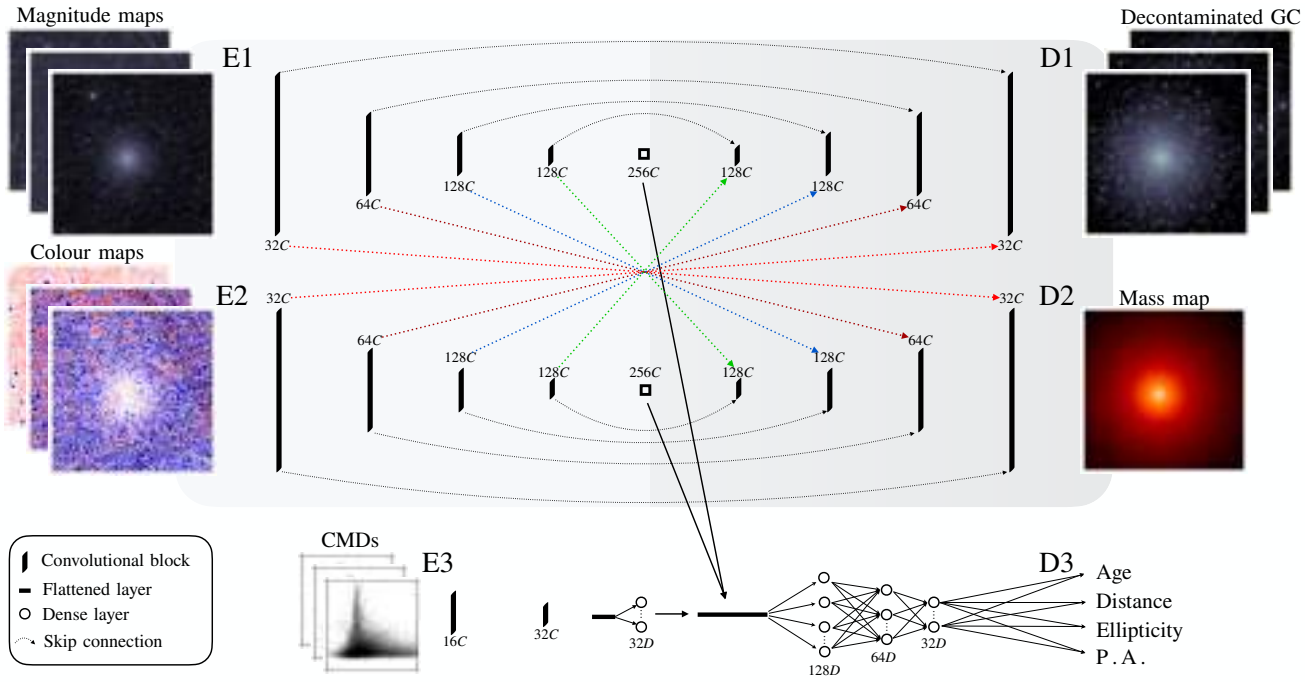

Fig. 2. Architecture of π-DOC and in particular the π-DOC$^{C}_{3}$ version. The different encoders are labeled E1-E3 and the respective decoders are labeled D1-D3. Convolutional and de-convolutional blocks consists of the layers in Table D.1 & D.2, respectively, and the corresponding number of filters (channels) are indicated below the blocks. Skip connections are indicated by dotted arrows, and skip connections shared between E1-D1 and E2-D2 are coloured. The regression head (D3) combines the latent features from each of the encoders, and the number of dense layers are indicated beneath the symbols.

The distributions of parameters for the dataset are chosen such that they cover the expected ranges of parameters for GCs in M31, where possible. Holland (1998) put most of the GCs at similar distances to M31 itself, and from the Galactic population of GCs we know they may extend as far as $\gtrsim$ 100 kpc from the Galactic Center (see e.g. Harris 2010). As such we choose distances to be sampled from a uniform distribution centred at the distance of M31, 785 ± 100 kpc (here using the distance estimate to M31 from McConnachie et al. 2005).

Some quantities are set directly by the simulations, such as the age (sampled uniformly in linear scale[11]), total luminosity, and total mass (see Bianchini et al. 2026; Marklund et al. in prep.). The projected images, the two-dimensional mass distribution, ellipticity, and the corresponding position angle (P.A.) for the minor axis[12], are also set by the simulations, but further depend on the two geometrical angles defined in `Cremant` (Section 2.2), as well as the distance. These angles and the versor are all sampled uniformly.

We compute a GC's apparent ellipticity $e = 1 - b/a$ using the second-moment tensor method on the stellar distribution within the FoV, where a and b are the major and minor axes, respectively. This robust approach reliably traces the cluster's intrinsic morphology (Fréour et al. 2026; Marklund et al. in prep.), facilitating clearer interpretation of our results. The position angle is measured east of north over $[0°, 180°)$.

We consider a total of three passbands, but also include versions with two passbands to incorporate the PHAST survey. From the difference between the three (two) passbands, it is also possible to construct three (one) colour maps. Furthermore, with both colour and magnitude maps, we construct pixel colour-magnitude diagrams (CMDs) that are 64×64 pixels[13]. In cases where three (two) passbands are used, we get three (one) possible pixel CMDs: `F475W` vs `F336W-F475W`, `F814W` vs `F336W-F814W`, and `F814W` vs `F475W-F814W`.

We split the entire datasets into three parts: training, validation, and testing. The three sets are constructed in such a way that no mock image exists in more than one of the sets. This also applies to the real backgrounds, which in turn are distributed among the three sets in proportion to that set's number of mock images (if X% of the mock images are in a given set, X% of the backgrounds are in that same set as well).

In order to get a diverse training dataset, it is customary to perform on-the-fly data augmentation. We combine the mock images and the real backgrounds on the go when loading each batch, and the two images are also rotated 0°, 45°, 90°, or 135°, irrespective of each other. This approach is powerful because we can obtain a completely new set of images for every epoch. It does however mean we rotate PSFs inconsistently!

[11] The distribution is only approximately uniform, as the largest N-body simulation was evolved only up to 12 Gyr, three MOCCA simulations were evolved up to 15 Gyr, the ω Cen equivalent model is additionally under-sampled at all evolutionary stages due to the computational cost of its magnitude calculations, and early snapshots are entirely absent, while all the remaining simulations were run up to 14 Gyr.

[12] In projection, the minor-axis position angle is equivalent to the major-axis angle shifted by 90°. We adopt the minor axis here as a matter of convention and for consistency with prior studies (e.g. Fréour et al. 2026; Marklund et al. in prep.).

[13] We note that using pixel CMDs with an increased resolution (128×128, 256×256) were not found to improve performance.

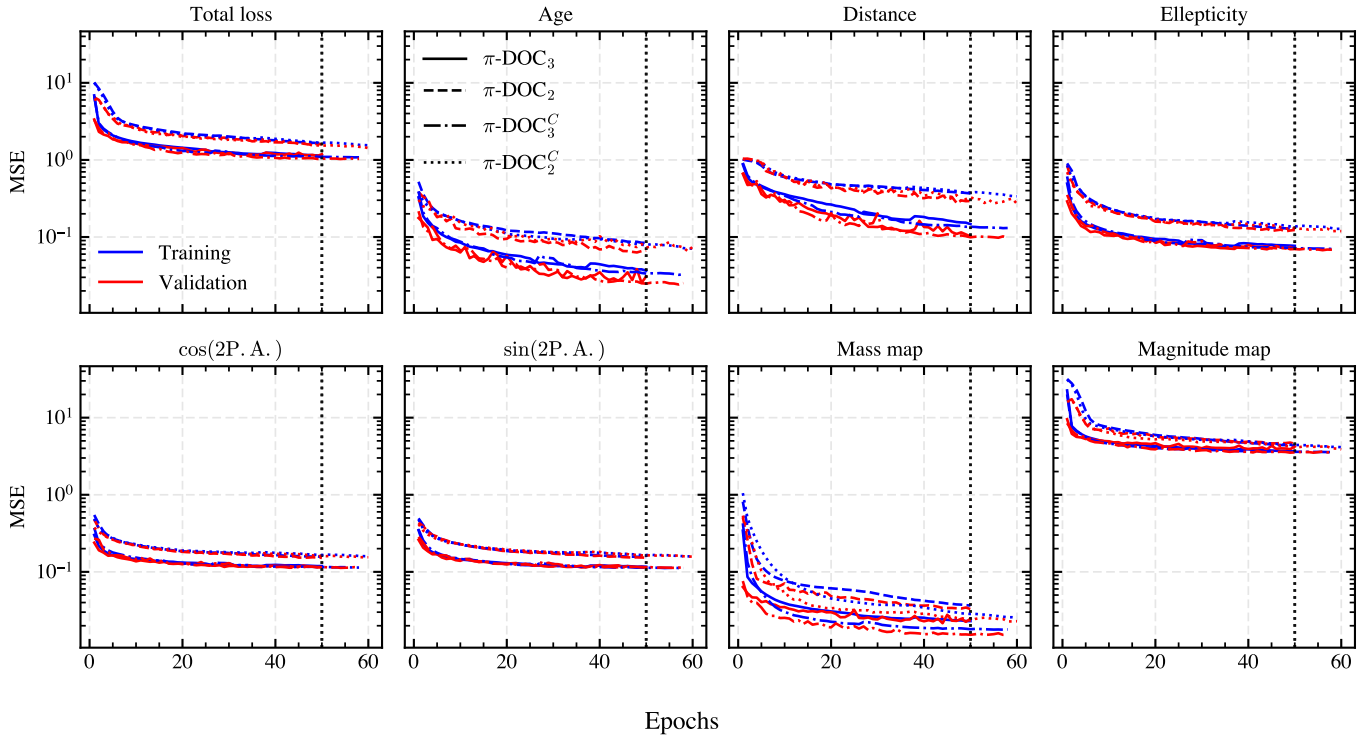

Fig. 3. Training curves for the four π-DOC variants considered in this study. Blue (red) lines correspond to the training (validation) MSE per output across epochs. The three-passband networks converge faster and achieve lower overall MSE than their two-passband counterparts. The addition of colour maps and an adapted architecture with two encoders and shared skip connections mainly improves the accuracy of the predicted mass maps. We interpret this as a result of the larger (but not deeper) network capacity. The vertical black dotted line marks the maximum 50 epochs used for network performance comparisons.

3. Network

In this section we go over the new architecture of π-DOC as well as the corresponding inputs and outputs. π-DOC is a convolutional autoencoder based on the U-net architecture (Ronneberger et al. 2015), but it has been modified to handle hybdrid-inputs and to produce multi-modal outputs. In particular, it is now capable of handling multi-band photometric images, and combines the main images with different types of input data such as colour maps and pixel CMDs, which we describe in more detail in Section 3.1. We note that the inclusion of pixel CMDs was found to significantly improve age and distance estimates in early testing of the new network architecture, in particular due to a more explicit encoding of the shape of the associated isochrones. There are also three completely new outputs: a decontamination of field stars from the input multi-band photometric image, and the GCs' ellipticity and corresponding position angle for the minor axis. We describe all outputs in more detail in Section 3.2.

The additional inputs and outputs have also resulted in a re-work of π-DOC's overall structure, and for this study we present a total of four different versions. We consider one version (π-DOC$_3^C$) that includes all of the aforementioned inputs (magnitude maps, colour maps, and CMDs), and one that does not include the colour maps (π-DOC$_3$). Furthermore, because the PHAST survey is available only in two passbands, each of these π-DOC versions exist for both three and two passbands (π-DOC$_2^C$ and π-DOC$_2$, respectively). Each of the four versions are tasked to predict the same outputs, and are also available online[14]. We describe the corresponding architecture(s) in more detail in Section 3.3.

3.1. Input data

The main inputs to the network consist of the multi-band photometric images alongside CMDs, but moreover also allow for the inclusion of colour maps. The inputs for the different versions of π-DOC that we consider are as follows:

1. π-DOC$_3$ uses three magnitude maps and three CMDs.
2. π-DOC$_2$ uses two magnitude maps and one CMD.
3. π-DOC$_3^C$ uses three magnitude maps, three colour maps, and three CMDs.
4. π-DOC$_2^C$ uses two magnitude maps, one colour map, and one CMD.

In contrast to the proof-of-concept version of π-DOC, the input images are not smoothed. See Section 2 for more details on the construction of the input data.

We standardise all our inputs, albeit slightly differently. The magnitude maps are standardised together to have a combined mean of zero, but we do not enforce a standard deviation of unity. This is to ensure that the magnitudes and colours remain consistent with respect to each other (e.g. even when standardised, the difference in magnitude corresponds to the colour). In practice, this should have little to no impact on the stability and performance of the training, because the standard deviation for the magnitude maps is already close to unity. Consequently, we

[14] github link here

do not standardise the colour at all. The number counts for the pixels of the CMDs are, however, standardised to have a mean of zero and a standard deviation of one. The standardisation constants (mean and std) are calculated by iterating over the training data for one epoch (where each mock image is combined with a background, see Section 2.3).

3.2. Output data

π-DOC is tasked to decontaminate the input images in all passbands from the Galactic background, map the corresponding two-dimensional mass distribution within the FoV (same pixelscale as the input images), and predict the cluster's age, distance, ellipticity, and position angle for the minor axis (P.A.). Instead of predicting the position angle directly, the network predicts cos(2P.A.), sin(2P.A.); the resulting position angle is then calculated using the two-argument arctangent (atan 2) as $\frac{1}{2}\,\mathrm{atan\,2}\,(\sin(2\mathrm{P.A.}), \cos(2\mathrm{P.A.}))$. This approach removes large errors at the periodic boundary where 0° and 180° are degenerate.

We also apply a Gaussian smoothing kernel to the mass maps ($\sigma = 4$). This is to avoid the network having to deal with predicting highly pixelated images, especially in the outskirts of the clusters where the stellar counts are sparse. Note that we have tested applying such a smoothing also on the input and output magnitude maps, but that this did not improve performance. The input images and the decontaminated output images therefore remain consistent with HST's pixelscale.

All of these outputs are also standardised during training except for the position angle because we use sin(2P.A.), cos(2P.A.). The clean magnitude images are standardised to be consistent with the input magnitude images (see Section 3.1), and the remaining outputs are standardised to have a mean of zero and a standard deviation of one.

3.3. Network architecture

The inclusion of additional input data, such as the CMDs and colour images, enables more complex operations within the network, encouraging it to extract both complementary and distinct features from multiple representations of the same underlying system. The new architecture reflects this, and we show a schematic overview in Fig. 2, which, in particular, corresponds to π-DOC$_3^C$. The convolutional and de-convolutional blocks, and their different layers are listed in Table D.1 & Table D.2, respectively. There are six different components, which can be split into three encoding parts (E1-E3) and three decoding parts (D1-D3)[15]:

E1 is an encoder applied on input magnitude maps (first row) that consists of five convolutional blocks (see Table D.1).
E2 is an encoder applied on input colour maps (second row) that is otherwise identical to E1.
E3 is a feature extractor (not an encoder in the usual sense) applied on CMDs (third row). The convolutional blocks are slightly different from those employed in E1 and E2 (see Table D.3).
D1 is a decoder (first row) consisting of five de-convolutional blocks (see Table D.2) and decontaminates the input magnitude maps, but uses skip connections from both encoders in E1 and E2
D2 is a decoder (second row) similar to D1 which also uses skip connections from both encoders in E1 and E2, but is tasked to predict the corresponding two-dimensional mass map.
D3 is the regression head of the CNN-like architecture in the third row and predicts the remaining scalar output (age, distance, ellipticity, and position angle) from the concatenated and flattened latent spaces of the three encoders E1-E3.

The corresponding architecture for π-DOC$_3$ is a bit simpler due to the removal of the colour maps as inputs. Consequently, the second row in Fig. 2 disappears and we only have one autoencoder (first row). The decoder part (D1), therefore, only uses skip connections from the input magnitude maps (E1), and is tasked to simultaneously predict the mass map along with the decontaminated GC[16]. Similarly, the regression head in row three (D3) only uses the concatenated features from the latent space of E1 and E3. For both architectures, there are versions for both two and three passbands, and we have set the dropout rate of each version to 0.1.

3.4. Training

For all versions of π-DOC, the training is governed by the same loss function, which we take to be the sum of each output's mean-square-error (MSE). Note that we also experimented with incorporating a perceptual loss function for predicting the mass and luminosity, as per Chardin & Bianchini (2021), but we did not observe it to yield any improvement. As a result, we consider the total loss function as

$$\mathcal{L} = \sum_{i=1}^{I} w_i \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2 \tag{1}$$

where I denotes the number of outputs, each $\mathbf{y}_i$ denotes one output: the decontaminated magnitude maps which is considered as a singular output regardless of the number of passbands, mass map, age, distance, ellipticity, and position angle. Consequently, $\hat{\mathbf{y}}_i$ is its estimate by the network. w_i denotes a scaling factor that is set to ensure each output MSE is in the same range[17]. We indicate the total MSE, and the output-specific MSEs, for the versions of π-DOC in Table 1. We find, in particular, that the inclusion of `F336W` significantly improves the performance for all outputs, but especially for the age and distance. The inclusion of colour maps for the input also improves performance.

The different π-DOC networks have been trained on NVIDIA H200 GPUs and Tesla V100-PCIE GPUs. We use the `ADAM` optimizer (Kingma & Ba 2014), and a learning rate of $\eta = 10^{-4}$, which we found to converge sufficiently fast while also keeping the training stable. Furthermore, we implemented a learning-rate scheduler (halving after 2 epochs without validation improvement) and early stopping (after 5 epochs). Training ran to 50 epochs (no early stopping triggered), yielding near-converged models with stable validation performance (> 1000 GPU hours).

Fig. 3 shows the training and validation performance per epoch for each network. Overall, training is stable, with com-

[15] The latter of these parts illustrate a feature extractor and a regression head rather than an encoder and decoder, but for simplicity we denote them as E3 and D3.

[16] We also tested an architecture with two decoders dedicated to decontamination and mapping the mass-distribution, respectively, even when colour images were not included. This did not yield significant improvements.

[17] The weights are set to unity for all outputs except for the magnitude and cos(2P.A.), sin(2P.A.), which are set to 0.2, 0.5, and 0.5, respectively.

Table 1. MSE of the validation set for the different outputs, for the best performing epoch (based on total validation loss), for the different versions of π-DOC. For the two networks we continue training with, perhaps indicate the best performance in paranthesis next to the current values for the best epoch?

MSE	π-DOC$_3$	π-DOC$_2$	π-DOC$_3^C$	π-DOC$_2^C$
Magnitude map	3.90	4.59	3.51 (3.45)	4.29 ()
Mass map	0.02	0.03	0.02 (0.02)	0.03 ()
Age	0.03	0.07	0.03 (0.02)	0.07 ()
Distance	0.11	0.29	0.10 (0.10)	0.28 ()
Ellipticity	0.07	0.15	0.07 (0.07)	0.13 ()
Position angle	0.12	0.16	0.11 (0.11)	0.17 ()
Total	1.13	1.59	1.03 (1.01)	1.53 ()
Epoch	49/50	50/50	49/50 (62/65)	48/50 ()

Notes. The MSE for the P.A. is taken to be the average between sin(2P.A.) and cos(2P.A.).

parable performance on the training and validation sets. The networks converge rapidly on the decontamination of the input images. Scalar outputs generally improve more slowly but steadily; further training focused on these outputs could therefore yield small additional improvements, however, we assess convergence primarily from the total loss.

Table 1 shows that π-DOC$_3^C$ consistently outperforms other architectures across all output parameters. Likewise, when restricted to two passbands, π-DOC$_2^C$ (with colour maps) significantly outperforms its counterpart (π-DOC$_2$). Hereafter, we focus exclusively on these two architectures and specify which one is used in each analysis. To ensure stable and reliable performance, we continued the training for the two models until convergence (epochs 62 and *XX*, respectively; see Fig. 3). The corresponding performance metrics of the converged models are reported in Table 1 and correspond to the versions used throughout the remainder of this paper. Tables and figures will be updated once they are more converged!

4. π-DOC's performance on mock images

For the assessment of the network's performance on mock observations, we begin by visually inspecting some of the predictions. Throughout this section, all predictions are obtained with π-DOC$_3^C$ unless explicitly mentioned otherwise. Fig. 4 shows a mosaic of four selected mock images from the test set, using each of the three passbands to construct a pseudo-RGB image. These mock images were chosen to be representative of different cases we expect to find: (a) an older, faint, and low-density cluster with low background contamination (high S/N and `ROLLIN'` model), (b) an old but fairly dense (bright and massive) cluster in a high background contamination (low S/N and MOCCA model), (c) a younger fainter cluster in a higher background contamination (medium S/N and `ROLLIN'` model), and (d) an old dense cluster in a low background contamination (high S/N and MOCCA model). The second and third row show the decontamination part of the network, and also includes the scalar regression output for the distance and age, as well as the total luminosity and half-light radius, which are derived directly from the decontaminated map.

The successful decontamination of the faint cluster in the brighter background (c), demonstrates that π-DOC can distinguish between stars that belong to the GC and those belonging to the background. Not all stars (or pixels) are successfully recovered where, in particular, stars far away from the centre of the GC are missing. This should not, however, affect global quantities significantly. On the other hand, many of the brighter stars (for example bright orange-red stars) that are recovered in all examples except for (b), mostly retain their colours as well. As such, demonstrating that the network efficiently decontaminates the images, and that this applies reasonably well across the different passbands.

The remaining two rows show the true and predicted 2D mass distributions, and the corresponding total mass and half-mass radius (derived directly from the mass maps). The total mass is constrained rather well for these four examples, with the worst case (b) amounting to an underestimation by a factor of ~ 2.5. In this example, the decontamination does not fully capture the outer parts of the cluster, likely due to its low S/N, and the corresponding mass-distribution is also not fully captured. This is also the case for the central parts (within the half-mass radius), where there is an underestimation of both the brightness and the total mass. To a lesser extent, this is true for the other three examples (a,c,d) as well, which in turn leads to an overestimation of the half-mass radius.

The ellipticity for the two denser and more massive (brighter) clusters in columns (b) and (d), are zero because MOCCA simulations assume sphericity. In both cases the ellipticity is recovered fairly accurately (small overestimation for b), while the position angle is not; this is not an issue since the position angle is not defined for $e = 0$. In contrast, both the ellipticity and position angle of the two `ROLLIN'` models are recovered, although the ellipticity is underestimated in both cases and more strongly for the lower S/N example (c).

The four mock images presented here, while only a small subset of our full test set, show how π-DOC performs under realistic observational conditions. Overall, the network successfully decontaminates fainter clusters even in bright, crowded fields, correctly distinguishing member stars from background contamination and preserving colors across passbands for brighter stars. However, limitations become apparent in low S/N regimes: case (b), which suffers from high contamination shows the largest discrepancies, with the largest underestimation of total mass and a significantly overestimated distance (30 kpc vs. the true value). In the other three cases, distances are recovered within ~ 10 kpc and age predictions are similarly accurate within $\lesssim 0.6$ Gyr, indicating that π-DOC reliably infers global cluster properties when S/N is sufficient.

4.1. Performance on test set

Qualitatively, the performance in Fig. 4 is representative of the full test sample. Fig. 5 shows the prediction distributions for all π-DOC outputs on the test set, alongside the corresponding train-

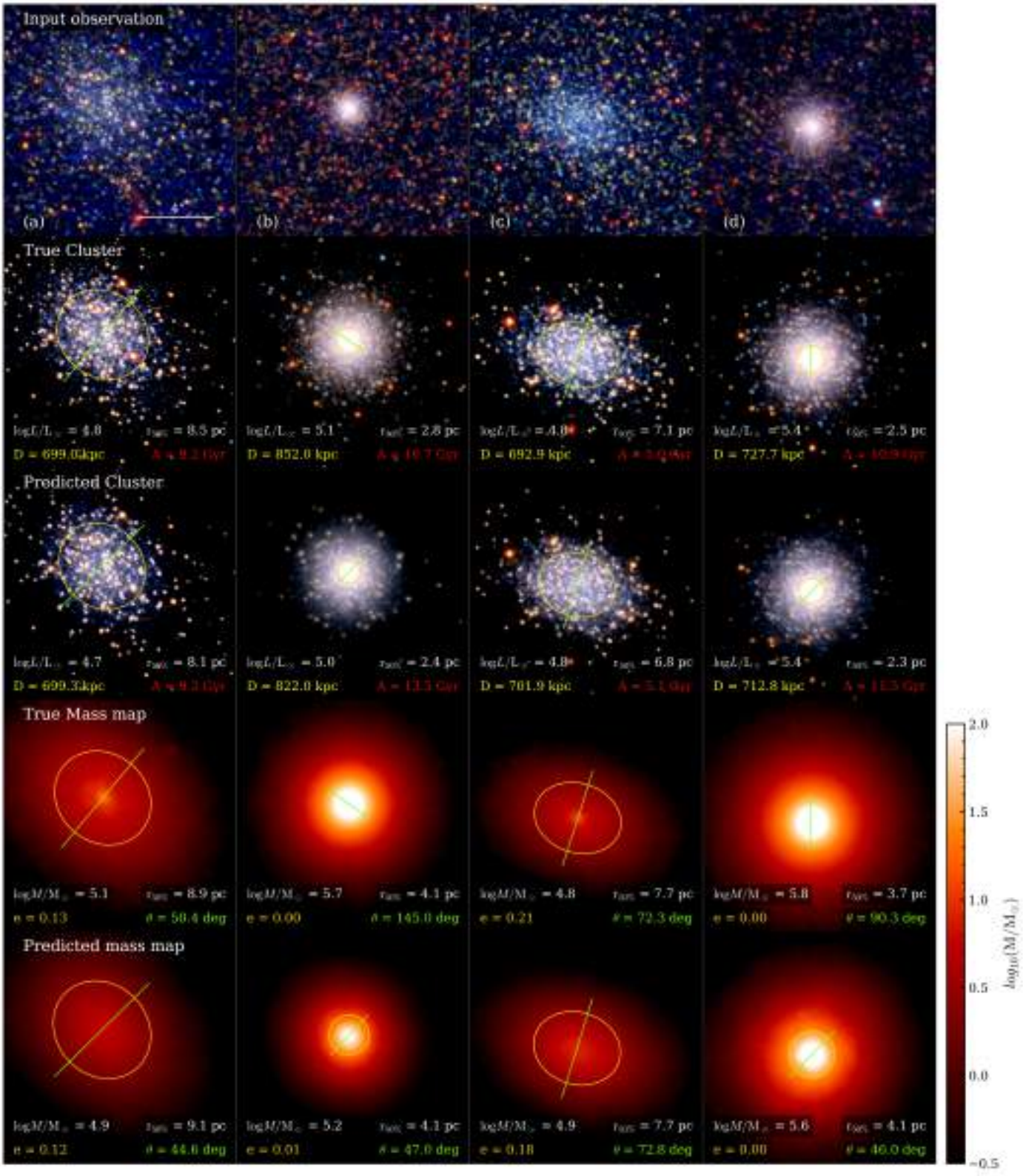

Fig. 4. Mosaic example of four different mock images (top row), the true decontaminated cluster (second row), the predicted decontaminated cluster (third row), the true mass distribution (fourth row), and the predicted mass distribution (bottom row). The decontaminated panels also indicate the true and predicted values for total luminosity and half-light radius within the FoV, distance, and age. Similarly, the mass distribution panels indicate the total mass and half-mass radius within the FoV, the ellipticity, and the position angle. The golden ellipse in each panels indicate the ellipticity and has its semi-major axis placed at the half-light or half-mass radius, whereas the green line indicates the position angle of the minor axis. The true and predicted decontaminated cluster images has a different contrast to the input observations to highlight the differences in decontamination across the full extent of the cluster. Change Cluster to GC, Mass to mass, and PA needs to be defined north to east 0-180 (to be consistent with later)

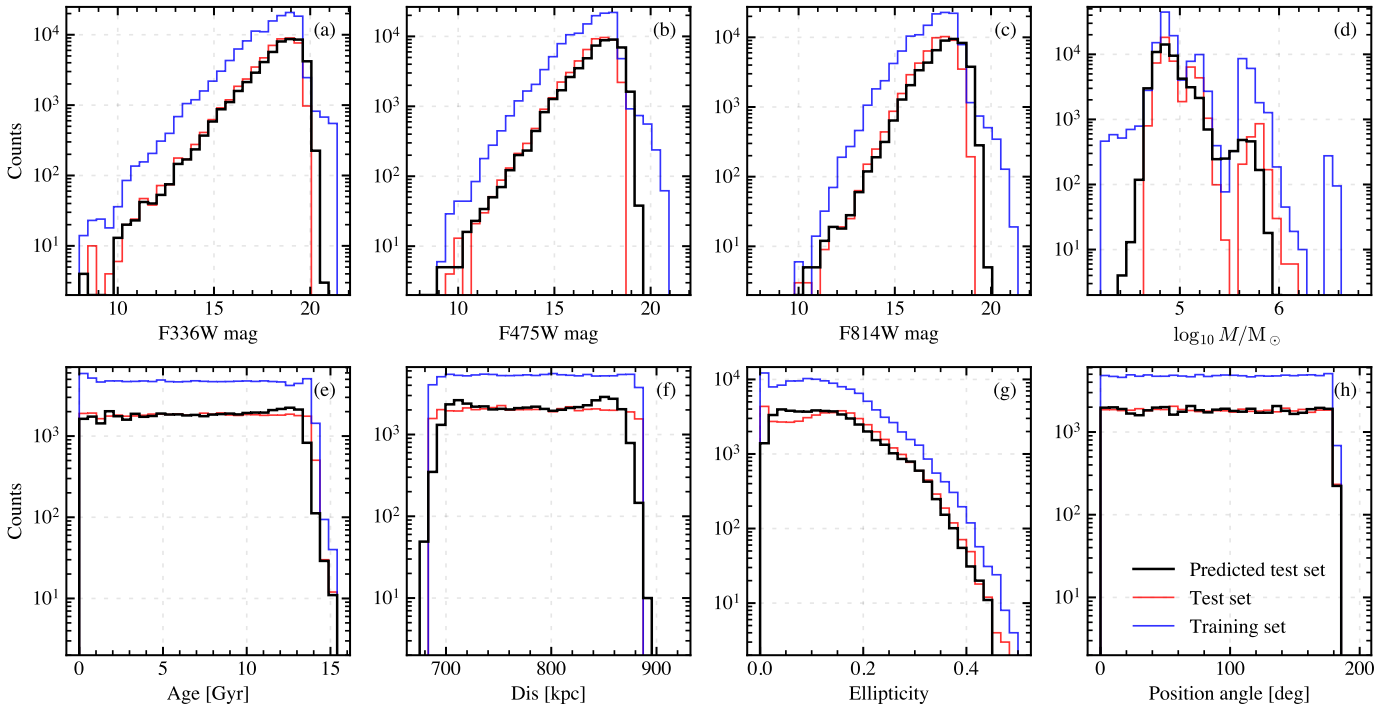

Fig. 5. Distribution of the predictions for the main outputs of π-DOC (black lines), including the corresponding distributions for the test data (red lines; ground truths to predictions) and the training data (blue lines).

ing and test parameter distributions. The training data cover the test set parameter space well, and the predictions closely match the test distributions across the full range, indicating the broad applicability of π-DOC.

The three first panels (a-c) show the performance for the decontamination part of the network in all three photometric passbands, respectively. The integrated magnitudes for the faintest clusters are overestimated (their brightness is underestimated, see tail of the distributions at large magnitudes). This is expected, since fainter clusters are more difficult to distinguish from the background. At the other end of the distribution, there is a tendency to underestimate the magnitudes (thus overestimating the brightness). Since there are not many very bright clusters (mag $\lesssim$ 12) within the training data, this may reflect the network not having had enough training material to properly process data in this regime.

In the case of the mass distribution in panel (d), we can notice two distinct peaks for the test set. The network appears to interpolate between these two peaks, thereby producing a more unimodal distribution, while also underestimating both the most massive clusters as well as the least massive clusters. This likely reflects on the accuracy of the predicted brightness for the clusters: clusters that are not properly recovered in the decontamination, are also underestimated in total mass. For the massive end of the distribution, this may also reflect a bias towards predicting lower masses, which likely is caused by a shortage of more massive simulations in the training data.

The training and test set distributions for age in panel (e) and distance in panel (f) are both uniform. The predictions are generally uniform as well, except near the boundaries of the parameter ranges. This behaviour is relatively common in neural networks (Kang et al. 2024), which typically perform less reliably close to the edges of the training domain. Because the network implicitly learns the limits of the parameter space, it avoids extrapolating beyond them and instead biases predictions toward interior values. Note that real GCs are unlikely to populate these boundary regions (see e.g. Harris 2010), so our subsequent analysis is unlikely to be affected by such biases.

The predictions of ellipticity in panel (g) and position angle in panel (h) almost match the ground truths across the entire range. Small values of ellipticity ($\lesssim$ 0.1) are overestimated, likely indicating a bias where the MOCCA models generally are not predicted to be zero (see Fig. 4). Additionally, for very small values of ellipticity, the position angle is either not defined ($e = 0$) or ill-defined as there is no clear major or minor axis; large errors for predictions of the position angle are therefore to be expected, but should not be considered an issue in this context.

We include a comparison between the performance of π-DOC$_3^C$ and π-DOC$_2^C$ on the test set in Appendix C. Will add dedicated figures of π-DOC$_2^C$ performance also

4.2. Biases within the network

4.2.1. Dependence on the brightness of clusters

The previous section explored how the network retains the overall distributions of the test set. Here, we explicitly aim to relate the cases where the predictions are most uncertain. Fig. 6 shows a direct comparison of the differences between the predictions and the ground truth of the test data, as well as the correlations between the errors. To facilitate the discussion below we keep the ground truth magnitude for the `F336W` passband in column (a) rather than the corresponding error.

Column (a) shows that the errors increase for most quantities the fainter the clusters are. As previously mentioned, this likely reflects an increased difficulty in distinguishing the GC from the background. Indeed, the majority of the GCs in our data

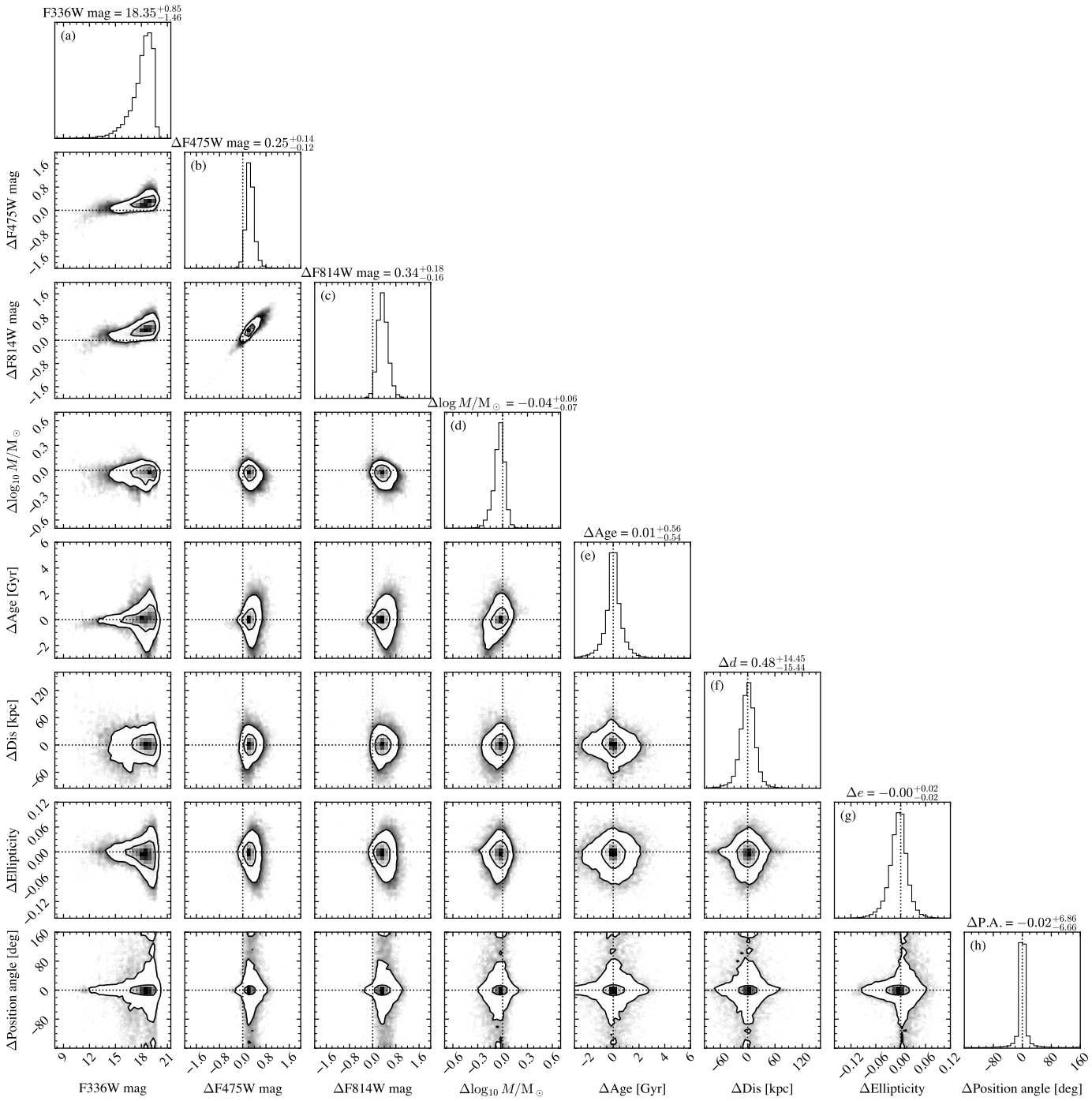

Fig. 6. Corner plot showing the distributions and correlated distribution of the errors of the main predicted outputs of π-DOC. The contours are set to σ, 2σ, and the dotted lines indicate the 0-axes (e.g. no discrepancy), where applicable.

set are relatively faint (see also Fig. 5), having an average integrated magnitude of 18.52, 17.46, 17.70 for `F336W`, `F475W`, and `F814W`, respectively. In comparison, the average integrated magnitudes of the added backgrounds for the three passbands are 16.51, 15.20, 15.28, respectively. In addition, most of our mock images correspond to N-body simulations, which are equivalent to low-density Galactic GCs (Bianchini et al. 2026). Their lower surface brightness, and consequently lower S/N, likely makes recovering their intrinsic properties more challenging.

For brighter clusters ($\sim$ 12–17 mag), the errors remain small, especially for mass, age, ellipticity, and position angle, while the distance errors are roughly independent of brightness. In contrast, the decontamination in `F475W` and `F814W` shows a nearly constant brightness underestimation which, on average is 0.25 mag and 0.35 mag in `F475W` and `F814W`, respectively. These discrepancies become more severe for both really faint and really bright clusters, but is nonetheless present across the entire range. We see a similar shift of 0.20 mag also for the `F336W` passband as indicated in Fig. 7. The overall uncertainty thereby increases with redder passbands: `F336W` is best constrained and `F814W` worst. Because the brightness in each passband is underestimated by different amounts, it also means that the colours of

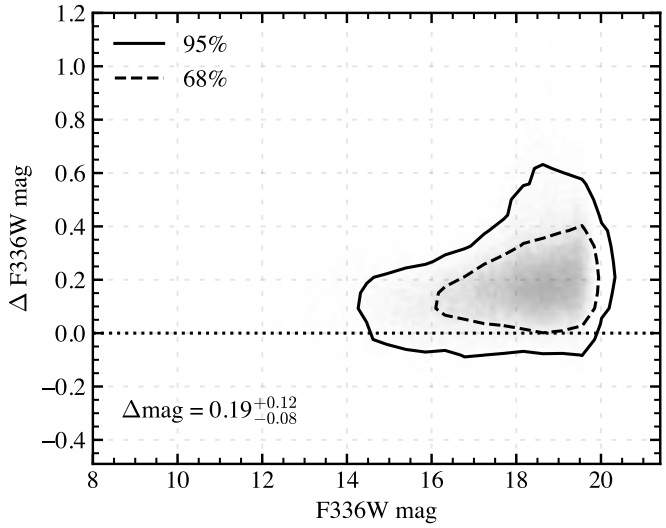

Fig. 7. Integrated F336W magnitude and the corresponding errors for the decontaminated predictions of the test set. The contours indicate the σ, 2σ regions. Similar to F475W and F814W, the predictions systematically underestimate the clusters' brightness.

the decontaminated GC are inconsistent, with π-DOC predicting bluer colours on average.

Values to be updated once converged model is included: The diagonal of Fig. 6 further shows that there are small biases in overestimating ages (~ 0.11 Gyr) and underestimating distances (~ -4.96 kpc). In both cases, however, this shift is a lot smaller than the overall spread in predictions (indicated in the figure), and should therefore not result in significant systematic errors relative to the intrinsic uncertainties.

These intrinsic errors are also comparable to limitations in the field, and we regard the 2σ limits as a conservative confidence range for π-DOC's predictions. For the F336W, F475W, and F814W passbands, the decontamination errors are on average $^{+0.56}_{-0.01}$, $^{+0.64}_{-0.09}$, and $^{+0.81}_{-0.11}$ mag, respectively. The mass ($^{+0.12}_{-0.19}$ $\log_{10} M/\mathrm{M}\odot$ dex?) and age ($^{+2.0}_{-1.7}$ Gyr) are particularly well constrained compared to typical uncertainties for M31 GCs (Chen et al. 2016; Fan et al. 2010; Usher et al. 2024), and the mass confidence matches that of fully resolved Galactic GCs (see e.g. Bellini et al. 2017; Zocchi et al. 2017). A handful of outliers lie beyond the 2σ boundary, particularly for the brightest clusters where masses are underestimated by factors exceeding four; these rare cases correspond to early snapshots ($\lesssim 200$ Myr) from the underrepresented high-mass regime ($\gtrsim 10^6, \mathrm{M}\odot$) of the MOCCA simulations. Although we do not expect such large errors for typical M31 clusters, exceedingly bright systems beyond the training distribution remain susceptible to similar failures. Distance ($^{+34.9}_{-50.8}$ kpc), ellipticity ($^{+0.04}_{-0.06}$), and particularly P.A. ($^{+89.0^\circ}_{-110.9^\circ}$) show larger uncertainties. The confidence range for the ellipticity and P.A. are primarily exacerbated by faint, low-S/N clusters which is evident from the elongated contours in column (a) of Fig. 6. Furthermore, excluding clusters with a predicted ellipticity below 0.05 (where the P.A. is rather ill-defined) yields a more reasonable confidence range ($^{+XX}_{-XX}$).

4.2.2. Robustness to noise

We have shown above that most of our clusters are faint, whereas most of the backgrounds are, on average, a few magnitudes brighter. The accuracy of the predictions should therefore depend on the relative brightness of the GC and the background. In Fig. 8, we show the two-dimensional distributions of the errors for each quantity as a function of the magnitude difference between the GC and the corresponding background in the F814W passband. Negative values indicate that the GC is brighter, while positive values indicate that the background is brighter. Qualitatively, we find that the errors depend on the relative GC-to-background brightness in much the same way as they do on GC brightness alone (see Fig. 6): fainter clusters generally lead to larger discrepancies, and the spread in the predictions increases substantially when the background is much brighter than the cluster. Nevertheless, the distributions remain broadly symmetric around zero, indicating that the predictions are unbiased.

This is not equally true for all quantities, however. For the ellipticity, we find a slight tendency to underestimate it in background-dominated images. In such cases, the stellar distribution becomes more diffuse and the cluster therefore appears more circular. This effect is at most 0.02, which is small compared with the overall spread.

The decontamination shows a systematic shift away from zero in all passbands, with the bias increasing for redder bands and in all cases corresponds to an underestimation of the cluster brightness. The decontaminated clusters and their integrated magnitudes should therefore be regarded as upper limits, implying that in most cases the clusters are in fact brighter than predicted by π-DOC. The shifts in the F336W, F475W, and F814W passbands are at most $\lesssim 0.3$, $\lesssim 0.5$, and $\lesssim 0.6$ magnitudes, respectively.

Values to be updated: We further examine the visual appearance of the decontaminated images in this regime. The low-S/N mock images in Fig. 4, although less accurate than the corresponding high-S/N examples, already showed that the network is reasonably robust to noise. To explore this further, we investigate how the predictions change when the same mock image of a relatively faint cluster (F814W $\sim$ 17.2) is combined with backgrounds of different total integrated magnitudes in Fig. 9. Overall, the predictions remain fairly robust to the varying noise levels. The total mass varies within a factor of 1.5, the position angle within 5°, the ellipticity within 0.03, the distance within 42 kpc, the age within 2.2 Gyr, and the integrated magnitude within 0.4 mag in F814W. These variations lie within the confidence range previously defined for each quantity in Sect. 4.2.1, and also show that more reliable estimates can be expected for GCs in less crowded and less bright regions.. Although this is beyond the purposes of this study, these results also suggest π-DOC could be adapted to simultaneously detect GCs (Barbisan et al. 2022; Dold & Fahrion 2022; Singlow et al. 2022) in addition to inferring their properties.

Moreover, while most predictions remain relatively accurate and scatter around the ground truth, the magnitudes are in all cases overestimated, as expected from the discussion above (see Fig. 8).

4.3. Reliability of predictions

Deep neural networks provide powerful nonlinear mappings between observational data and physical parameters. Estimating a network's uncertainty for individual predictions in a reliable manner, however, remains challenging. To quantify a predictive uncertainty in our model, we adopt Monte Carlo (MC) dropout, a computationally efficient approximation to Bayesian inference (see Gal & Ghahramani 2015a,b; Gal et al. 2017). During training, dropout layers randomly deactivate subsets of neurons to

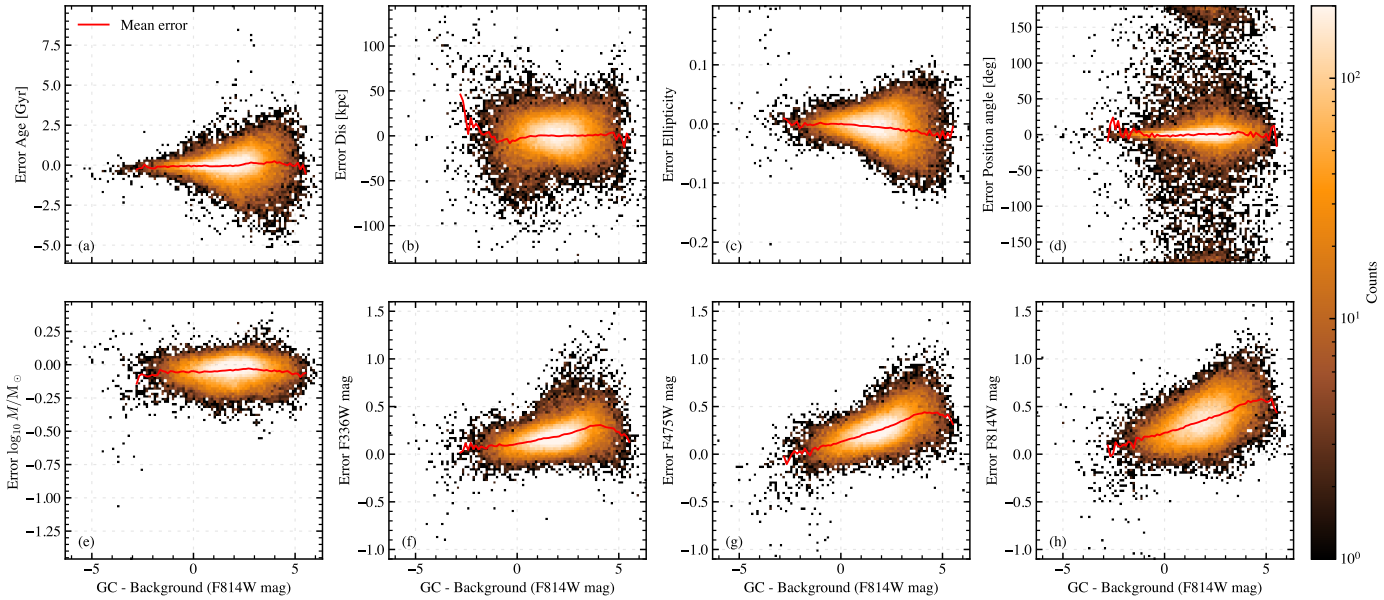

Fig. 8. Errors for each predicted quantity and how they depend on the relative brightness between a cluster and its associated background. π-DOC performs worse for background dominated images for each predicted quantity, indicating that low S/N, in particular, is a limitation.

prevent overfitting. In MC dropout, this stochastic behavior is retained during inference by performing multiple forward passes with dropout enabled. Each forward pass samples a different realisation of the network weights, effectively drawing from an approximate posterior distribution over models. For a given input, we obtain an ensemble of predictions that provide a measure of the model's so-called epistemic uncertainty. These values should not be regarded as physical uncertainties, but rather as the intrinsic confidence level of π-DOC. More details on our implementation can be found in Appendix B.

Because π-DOC is trained exclusively on synthetic data generated under controlled physical assumptions, when applied to real astronomical observations, there is no guarantee that all inputs lie within the same statistical distribution as the training set. Predictions made on inputs that differ significantly from the training distribution may therefore be unreliable, even if the model reports low internal uncertainty. To address this limitation, we implement an explicit out-of-distribution (OOD) detection framework.

OOD detection is based on the premise that inputs similar to the training data occupy a well-defined region of the network's internal feature space. We extract intermediate feature representations from the latent space [Not completely accurate, should maybe indicate the exact layer with a symbol in the figure] (see Fig. 2), keeping the features agnostic to specific output heads. These feature vectors are projected onto a lower-dimensional subspace using incremental principal component analysis (PCA) to improve numerical stability and reduce redundancy. Assuming that the feature distribution of the training data can be approximated by a multivariate Gaussian in this low-dimensional space, we characterise it by its mean vector and covariance matrix. Given a new input, we compute the Mahalanobis distance between its feature representation and the training distribution. This distance provides a scalar measure of how atypical the input is relative to the data seen during training. A threshold on the Mahalanobis distance is calibrated using the validation set, corresponding to a chosen (and assumed) false-positive rate set to the 90th perecentile. Inputs exceeding this threshold are flagged as OOD. An OOD flag does not imply that a prediction is incorrect, but rather that it is not supported by the training distribution and should be treated with caution. In particular, OOD detection complements MC dropout uncertainty: while MC dropout quantifies uncertainty within the learned model, OOD detection identifies cases where the model itself may not be directly applicable.

For the application of π-DOC on real observations, we provide both an epistemic uncertainty and an OOD flag for quality assessment.

5. Application to M31 GCs

We have so far demonstrated π-DOC's potential on synthetic data only; although the specific clusters in the testset are unique to that set, they correspond to simulations and therefore also share features that could deviate from real observations. The training data encompasses the majority of M31's GCs, as shown in Fig. 1, implying that π-DOC is broadly applicable in this context. We only have a few mock images for the most massive simulation, which ideally could help the network deal with the most massive of GCs. However, predictions at the upper end of the mass distribution are less accurate for synthetic data, and likely also exacerbated by a significant gap to the next most massive simulation. Consequently, we do not expect reliable mass recovery for $\gtrsim 10^6$ $M_\odot$.

5.1. Demonstrating π-DOC's performance on real observations

Following the outline of Section 4, we begin by demonstrating π-DOC's performance on four example GCs. These clusters: `Bol 188`, `Bol 224`, `Bol 232`, and `Bol 386`, are located within the PHAT footprint and as such we employ π-DOC$^C_3$. Fig. 10 shows the input observations and all corresponding estimates for these GCs.

Qualitatively, the decontamination of these GCs (middle row) resembles that obtained for the mock images in Figs. 4

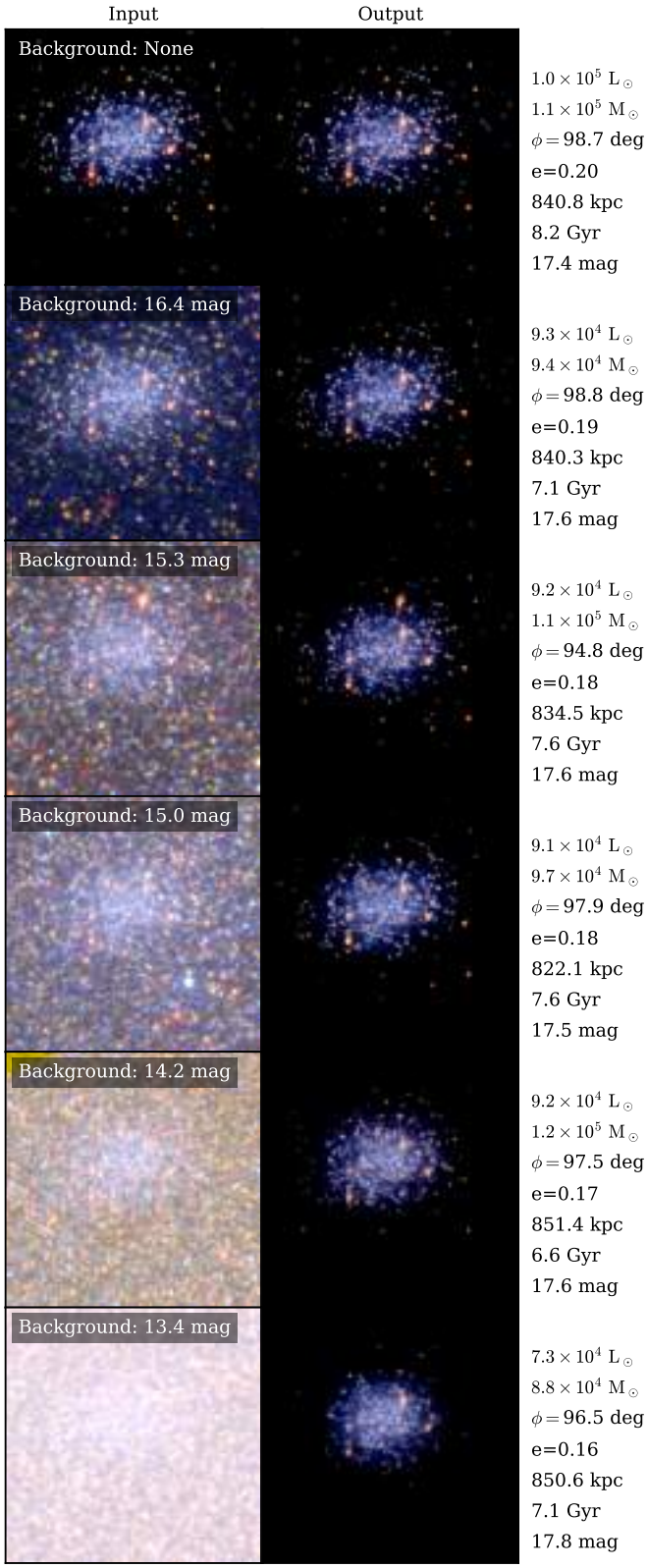

Fig. 9. One mock GC combined with different backgrounds (left column), and π-DOC's predictions (right column). Each mock image's associated predictions are indicated to the right of the decontaminated image where we denote the P.A. as ϕ. The ground truth, in the same order as the predictions from top to bottom, are: 1.4×10^5 M$_\odot$, 1.3×10^5 L$_\odot$, $\phi = 97.9°$, $e = 0.20$, 842.8 kpc, 8.0 Gyr, and the total integrated magnitude in the `F814W` passband is 17.2 mag. The predictions remain within the corresponding confidence limits, and the decontamination is fairly robust to decreasing S/N.

and 9. In each case, a limited colour spread is present, with orange–red stars surrounding the bulk of the cluster, while most such stars exhibit minimal variation with respect to each other. This behaviour suggests that the decontamination is not uniformly accurate across all passbands, consistent with the behaviour observed in panel b of Fig. 4. Nevertheless, the overall similarity indicates that π-DOC generalises effectively from mock to real observations. Together with the estimated distances, the decontamination is also associated with the total luminosity and the half-light radius of each GC within the FoV. Except for `Bol 188`, the GCs are estimated to be 20 – 70kpc away from the central distance to M31 (McConnachie et al. 2005), and are also rather luminous ($L/L_\odot \gtrsim 10^5$). These GCs, in particular, are very bright in their central regions, and while the decontaminated maps show this, the saturated part (really bright white region) do not extend as far out as it does in the input images. This is reminiscent of the decontamination of the two MOCCA models in Fig. 4, which indeed show an underestimation of both their central and total brightness. This bias, which was also seen in Figs. 6–9, suggests that their inferred brightness likely are lower estimates.

Age estimates are notoriously difficult to constrain, as for example seen from the corresponding errorbars in Fig. 1 (see also Usher et al. 2024). The epistemic uncertainties indicated in Fig. 10 should not be considered physical uncertainties, but rather the intrinsic confidence level of π-DOC's estimates (see Appendix B). For these four GCs (a)–(d), however, we do find a fairly good agreement with the literature: Usher et al. (2024) find the corresponding ages to be $9.4^{+3.8}_{-2.7}$ Gyr, $11.8^{+1.3}_{-2.0}$ Gyr, $10.1^{+1.1}_{-1.0}$ Gyr, and $8.3^{+4.8}_{-3.5}$ Gyr, respectively.

Similar to the brightness, both the total and central masses for the two MOCCA models were underestimated in Fig. 4. Since at least two of the M31 GCs in Fig. 10 appear both massive and centrally dense, characteristics primarily represented by our MOCCA simulations, their estimated mass maps (bottom row) likely reflect a lower limit as well. Indeed, compared to literature values from Usher et al. (2024), π-DOC's estimates for the four GCs in panels (a)-(d) are a factor of 2.0, 3.3, 2.5, and 2.4, lower, respectively. These underestimates likely also reflect that π-DOC's estimates are confined to the FoV; we are not necessarily reporting the GCs' total mass.

The clusters also exhibit morphological diversity: `Bol 386` is nearly spherical which primarily is evident from the large epistemic uncertainty for the position angle, while the others have ellipticities $\gtrsim 0.05$ (`Bol 232` being the most obvious with $e = 0.14$). `Bol 232` furthermore shows a clear direction for its flattening which also is captured by π-DOC's position angle estimate, and in good agreement with Staneva et al. (1996, with $e \sim 0.1$ and a P.A. of 120°).

We provide an online repository[18] containing the input magnitude maps as well as estimates of the decontaminated GC maps, mass maps, integrated magnitudes in each passband (within the FoV), total luminosities (within the FoV), half-light radii (within the FoV), total mass (within the FoV), half-mass radii (within the FoV), ages, distances, ellipticities, and position angles, for all 349 GCs within the footprints of the PHAT and PHAST surveys. The repository also provides OOD flags and epistemic uncertainties for the full sample. In some cases—particularly for OOD clusters in very low signal-to-noise regions—we obtain unphysical estimates (e.g., negative ellipticities), which are excluded from the tables and marked as N/A.

[18] List it here, or in the beginning of the paper?

A subset of the full sample using both versions of π-DOC is reported in Table D.4.

5.2. Inference of GC properties within the PHAT and PHAST footprints

The decontamination of the GCs shown in Fig. 10 exhibits qualitatively similar performance to that of the four mock images in Fig. 4. This suggests that the network generalises reasonably well from the synthetic to the real-data domain. We therefore proceed to infer the properties of the GCs within the PHAT and PHAST footprints. We use π-DOC$_3^C$ for the full PHAT sample, whereas the PHAST survey, which is available only in the `F475W` and `F814W` passbands, is analysed with π-DOC$_2^C$. A comparison between the two networks for GCs within the PHAT footprint is presented in Appendix C. The inference speed is $\sim$ 10 GCs s^{-1}, so the full M31 sample required only about half a minute[19].

Figure 11 shows the GCs from the Peacock et al. (2010a) catalogue within the PHAT and PHAST footprints, colour-coded by the mass (a), age (b), ellipticity (c), and distance (d), as estimated by π-DOC. We find more massive and older clusters to lie closer to the central parts of M31 which could be interpreted as a consequence of orbital decay driven by dynamical friction, but it may also reflect the intrinsic uncertainties of π-DOC and the aforementioned biases. Indeed, OOD GCs (points with white edges), which constitute approximately 30% of the sample, are located preferentially near the centre of M31, where contamination is the highest and S/N the lowest. Given the sensitivity of π-DOC to low-S/N images in the synthetic test set, and the OOD classification, we do not expect these estimates to be reliable.

Most of the GCs have very low ellipticities, with only a few exceeding 0.1; these are predominantly located well outside the central regions of M31, possibly reflecting a bias of underestimating ellipticity measurements at low S/N (see Section 4.2). We find that the average distance to the GCs in M31 is 818 ± 42 kpc. Moreover, the distance estimates place the GCs in the north-east region further away than those in the central and south-west regions, with differences of order $\lesssim$ 100 kpc, which is significantly larger than the physical extent of M31's disk (Chemin et al. 2009; Li et al. 2021; van der Marel et al. 2012, 2019). The typical uncertainty in distance (derived confidence limits; Section 4.2) is smaller than this shift, but remain comparable in magnitude. These results therefore do not reflect the geometry of M31, and are likely driven by intrinsic biases in π-DOC related to extinction and age–metallicity gradients across M31 (Lee et al. 2025). The trend is further accentuated by combining estimates from π-DOC$_3^C$ (PHAT) and π-DOC$_2^C$ (PHAST), the latter being less accurate in distance inference (see, e.g., Table 1). A direct comparison of their estimates for the PHAT GCs is presented in Appendix C, which confirms overall agreement while showing that π-DOC$_2^C$ systematically estimates larger distances when π-DOC$_3^C$ yields smaller ones.

5.2.1. OOD GCs

In order to identify features present in the sample of real GCs but underrepresented in the synthetic training set, we compare the GCs classified as OOD with those that are not. The OOD framework is designed to flag data points that lie outside the region of parameter space well supported by the training sample. In our case, this affects approximately 30% of the GCs within the PHAT and PHAST footprints. The OOD flag is raised predominantly for low-S/N clusters near the centre of M31, which may reflect the absence of internal extinction in our mock images. Spatially varying extinction within M31 could induce colour differences between clusters and backgrounds that are not consistently represented during training, thereby shifting sources in feature space and triggering OOD classifications. This interpretation is plausible, particularly because the training sample does include backgrounds from the central regions of M31. However, when we compare the extinction estimates A_V from Usher et al. (2024) for OOD and non-OOD clusters (Fig. XXX), we find no significant difference between the two distributions. We therefore conclude that extinction can account for only a small fraction of the OOD sample.

A more important factor, as it turns out, is the metallicity distribution. Our training set is limited to clusters with metallicities [Fe/H] ≤ -1.3 (Askar et al. 2025; Bianchini et al. 2026; Zhao et al. 2026), which does not fully represent the metallicity range of M31 GCs (see Chen et al. 2016; Fan et al. 2010; Usher et al. 2024). Since age and metallicity are degenerate and both primarily affect the colours of GCs, it is plausible that metallicity information is encoded in the latent space used by the OOD classifier. Figure 12 shows a comparison of the metallicities (using estimates from Usher et al. 2024) of OOD and in-distribution clusters, revealing a shift between the two populations. In particular, the OOD clusters are shifted toward significantly higher metallicities, with most of them having [Fe/H] $\gtrsim -1$.

The fact that the OOD sample is preferentially associated with high metallicity is particularly informative, as it suggests that the latent-space is sensitive to physical properties that are not explicitly part of the network outputs. In practice, this means that the OOD flag is not merely a technical warning about poor reconstruction, but also a diagnostic of which regions of parameter space that are missing from the training set. This is useful for both interpreting failures on real clusters and guiding future extensions of the training data toward more metal-rich systems and, more generally, toward a more complete representation of GC populations.

5.2.2. Comparisons with other studies

Having identified the GCs for which π-DOC is expected to be most reliable, we compare the network's estimates for the in-distribution sample with independent estimates from the literature. This allows us to assess the performance of the model in a more controlled way, separating genuine discrepancies from cases that are likely driven by out-of-distribution behaviour. In the following, we therefore focus on the non-OOD clusters and examine how π-DOC's mass and age estimates compare with estimates from Chen et al. (2016); Fan et al. (2010); Usher et al. (2024), and how its ellipticity and position angle estimates compare to estimates from (Staneva et al. 1996).

As seen from Fig. 13, π-DOC's mass estimates generally agree with the literature values, which themselves are mutually consistent (see also Usher et al. 2024). However, we observe a systematic trend: π-DOC infers higher masses for clusters with lower literature masses and systematically lower masses for those above $\gtrsim 10^{5.5}$ M$_\odot$, at least with respect to Chen et al. (2016); Usher et al. (2024). This behaviour mirrors the test set performance in Section 4, where high-mass clusters, in particular, were underestimated (see Fig. 5).

Usher et al. (2024) combined ground-based photometry from the Sloan Digital Sky Survey (SDSS; York et al. 2000) and the Pan-Andromeda Archaeological Survey (PAndAS; Huxor et al.

[19] With MC dropout enabled, the corresponding speed is $\sim$ 0.7 s per GC for π-DOC$_3^C$ and $\sim$ 0.5 s per GC for π-DOC$_2^C$.

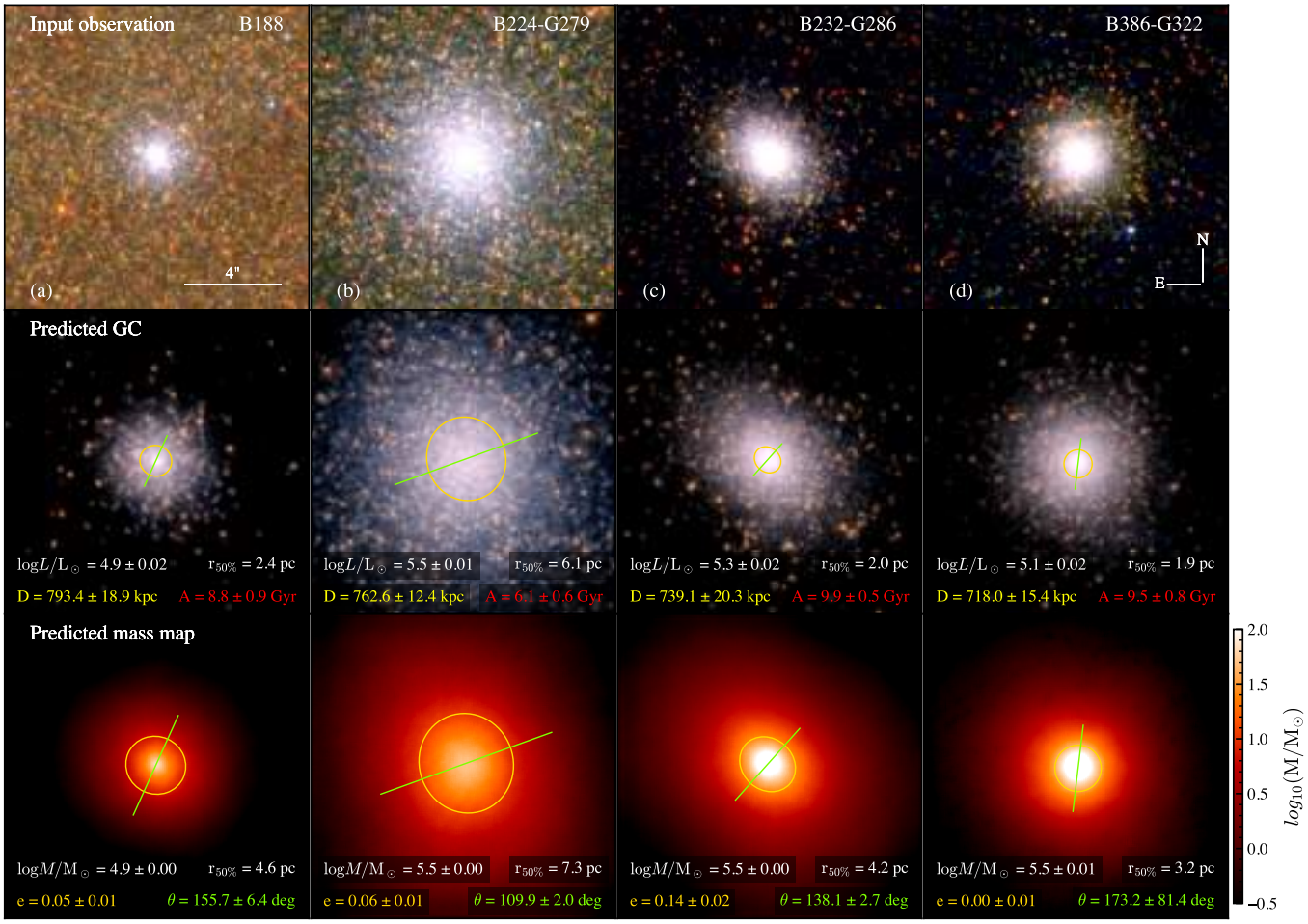

Fig. 10. Mosaic of four selected M31 GCs in the PHAT survey from left to right: Bol 188, Bol 224, Bol 232, and Bol 386. The first row shows the input observations as RGB images, the second row shows the corresponding decontaminated GC maps, and the third row shows the estimated mass maps. The total estimated luminosities and half-light radii in the middle row corresponds to the `F814W` passband and are obtained from the corresponding decontaminated image. The same applies to the estimated mass maps and half-mass radii. The golden ellipse indicates the estimated ellipticity and its major-axis extends to the half-light or half-mass radius, while the green line cutting through the ellipse indicates the position angle of the minor axis. The estimated values for the ellipticity and the position angle are indicated in the bottom row, while the estimates for age and distance are indicated in the middle row. The epistemic uncertainty for each scalar quantity is also indicated. Need to compute the epistemic uncertainty for the half-light and half-mass radii...

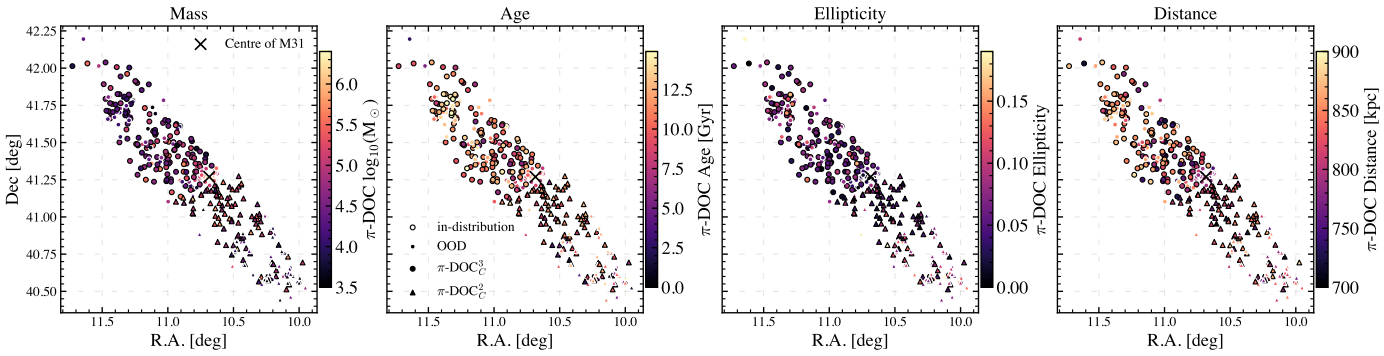

Fig. 11. GCs within the footprints of the PHAT (circles) and PHAST (triangles) surveys, colour-coded by π-DOC estimates of total mass (a), age (b), ellipticity (c), and distance (d). OOD clusters are outlined in white. There is a tentative concentration of higher masses toward the centre, but OOD flags necessitate cautious interpretation. The oldest GCs according to π-DOC lie in the north-east region of M31, which also corresponds to larger inferred distances; this trend lacks a clear physical explanation and likely reflects biases in π-DOC and observational systematics such as extinction and age–metallicity gradients. Most GCs appear nearly circular, with the highest ellipticities typically found at larger galactocentric distances, likely reflecting S/N.

2014) with spectroscopy targeting the CaT region from Caldwell et al. (2009). Because our images cover relatively small FoVs and may not encompass the full spatial extent of the GCs, their wider-field photometry includes more extended cluster light, likely yielding higher mass estimates. Using Markov Chain Monte Carlo (MCMC), they simultaneously fitted ages, metallicities, present-day stellar masses, and reddening with FSPS models based on MIST isochrones. Explicit reddening modeling systematically increases the inferred total brightness and thus the associated mass estimates compared to our values, since we only account for Milky Way reddening. At the same time, their photometric masses trace the current surviving stellar populations and do not include dynamical mass loss (e.g., evaporation, stripping), which π-DOC does; photometric and dynamical mass estimates are known to show discrepancies (Sollima et al. 2017).

Chen et al. (2016) also combined ground-based photometry from Peacock et al. (2010a) with spectroscopy using the Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST Liu et al. 2015; Yuan et al. 2015). Ages and metallicities were derived via pixel-by-pixel fitting of integrated spectra to single stellar population (SSP) models, with masses inferred from multi-band photometric spectral energy distributions (SEDs) scaled to those parameters. Similarly, Fan et al. (2010) also derived age and mass estimates by comparing their observed SEDs with SSP models. Their estimates thus yield photometric masses and are subject to the same biases as Usher et al. (2024). These comparisons therefore validate π-DOC's mass estimates, as dynamical-photometric tensions are expected and because our values otherwise agree.

In contrast, the age estimates show poor overall agreement with large discrepancies across the entire range. Given that our training set is limited to metallicities [Fe/H] ≤ -1.3, while M31 GCs span a much wider range, accurate age recovery is not expected except for clusters near this metallicity. This is partially also the case, where we see a better agreement with low-metallicity clusters and, in particular, with Usher et al. (2024). The age-metallicity degeneracy would furthermore likely lead to overestimated ages for metal-rich systems, but even studies that agree on metallicity, such as Chen et al. (2016) and Usher et al. (2024), exhibit comparable scatter to that seen with π-DOC, demonstrating that all literature estimates show discrepancies (see Usher et al. 2024). The presence of ages near 0 Gyr or 20 Gyr in Fan et al. (2010) and Chen et al. (2016) further underscores the intrinsic challenges of GC age determination. These comparisons therefore demonstrate that age estimation remains notoriously difficult across methods, rather than indicating specific shortcomings of π-DOC. Importantly, the metallicity signal we identified in π-DOC's latent space (Section 5.2.1) suggests that metallicity information is already encoded in the network's features. This opens the possibility of extending future versions of π-DOC to jointly predict age and metallicity as explicit outputs, thereby providing a new approach to breaking the age-metallicity degeneracy.

Figure 14 compares our ellipticity and position angle estimates with those reported by Staneva et al. (1996). We find no clear agreement for either of the quantities. The overall distribution for ellipticity is similar, however, but π-DOC systematically infers lower values. Discrepancies are to be expected given the differences in methodology: Staneva et al. (1996) derived ellipticities from iso-density contours, whereas π-DOC was trained on the intrinsic stellar distributions within the field of view, quantified using the second-moment tensor method. These two approaches are known to produce systematically different results even for identical stellar distributions (Fréour et al. 2026; Marklund et al. in prep.). Furthermore, Fréour et al. (2026) demonstrated that ellipticity measurements for more spherical distributions ($e \lesssim 0.1$) are known to produce large systematical overestimations and that these need to be bias corrected, which is not the case for (Staneva et al. 1996).

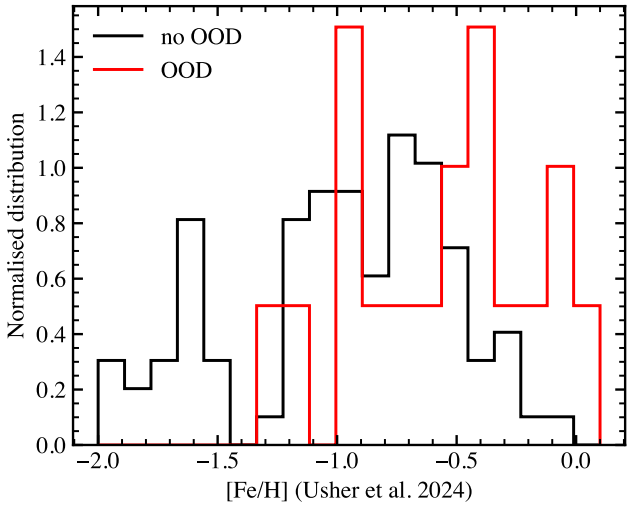

Fig. 12. Distribution of the M31 GCs metallicities based on the OOD classification. In-distribution GCs are typically more metal poor and thus closer to the metallicity for the training data, while OOD clusters typically are more metal-rich, highlighting a limitation of the current dataset.

A second important factor is spatial resolution and coverage. Ground-based observations inevitably have lower resolution and typically a wider FoV than HST imaging. Outer cluster regions, which dominate ground-based measurements, are generally more elliptical than inner regions (Fréour et al. 2026; Marklund et al. in prep.). For spatially resolved clusters in the Milky Way, Fréour et al. (2026) found that there typically is not a significant difference between the measured flattening within $3r_{50\%}$ and $19r_{50\%}$. On the contrary, the intrinsic flattening within and outside $r_{50\%}$ may show significant differences (Marklund et al. in prep.); while our mock images extend beyond the half-mass radius of the clusters, they remain confined to radii of approximately 20 − 30pc (set by the chosen FoV of our images and subject to the distance). This difference in spatial resolution and radial coverage may naturally contribute to the observed offsets.

Although direct one-to-one agreement in ellipticity across different studies should not be expected (Fréour et al. 2026), one might at least anticipate reasonable agreement in position angles for clusters that both Staneva et al. (1996) and π-DOC identify as elliptical. As shown by the scatter in the bottom panel of Fig. 14, this is only partially the case. The differences likely reflect the difficulty of robustly measuring both ellipticity and position angle for clusters that are only mildly flattened, and that lie in crowded or low-S/N regions. In such regimes, choices of isophote level, background subtraction, and fitting method can all shift the inferred orientation, even when the underlying structure is similar. In particularly clear examples, such as `Bol 232` (see Sect. 5.1), both morphology and orientation agree qualitatively with (Staneva et al. 1996), indicating that the method is reliable when data are favourable.

These comparisons therefore provide only a limited way to assess model performance, and the significant scatter between independent studies (including our own) suggests that simple visual or method-to-method agreement should not be

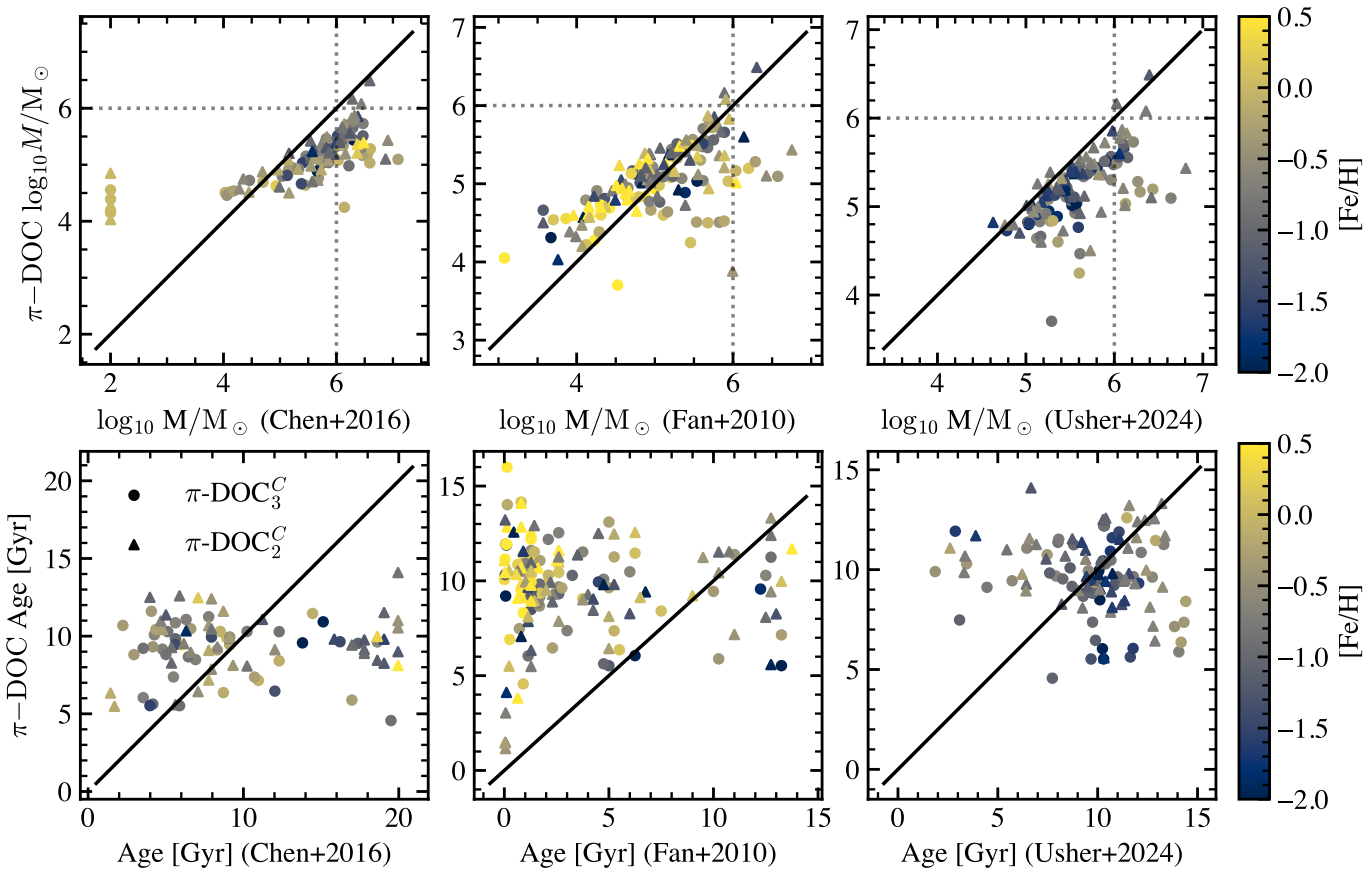

Fig. 13. Comparison of the π-DOC mass and age estimates with those of Chen et al. (2016); Fan et al. (2010); Usher et al. (2024) for 109, 150, and 113, GCs, respectively. The solid black line indicates a 1:1 relation, and the dotted lines mark 10^6 $M_\odot$, above which reliable coverage is not expected (see Fig. 1). Points are colour-coded by the metallicity estimates from the respective studies. We find overall agreement in total mass to within a factor of four, while acceptable age agreement is obtained only for low-metallicity GCs with Usher et al. (2024).

over-interpreted. More telling constraints on π-DOC therefore come from the controlled synthetic tests presented in Sect. 4, where both ellipticity and position angles are recovered accurately over a broad range of configurations.

6. Discussion

In this section, we explore a few preliminary avenues for improving π-DOC and assessing the robustness of our results. In particular, we test whether restricting the training set to higher-S/N clusters or rebalancing the training towards specific parts of the parameter space can yield measurable gains.

6.1. Limiting the dataset to brighter clusters

The majority of the simulations from the `ROLLIN'` suite are both low-mass and low density. At later times, after significant mass loss due to both stellar evolution and dynamical processes, the clusters become very faint (F814W mag > 18) and diffuse, making it increasingly more difficult to distinguish them from the background. Even though π-DOC show an impressive robustness to noise as, for example, demonstrated in Fig. 9, the estimates become increasingly less accurate and more biased towards underestimating brightness (see also Fig. 8).

We therefore trained a smaller version of π-DOC on a subset ($\sim$ 50%) of the total training data where, in particular, we removed the low-mass and low-density `ROLLIN'` simulations (250k stars). Thus reducing the number of low-S/N images while simultaneously increasing the prominence of the more massive clusters (i.e. MOCCA models). We applied the same filtering to the test set, creating a higher-S/N subset that, in principle, should yield improved performance. However, no significant improvements were observed for the estimated quantities[20], nor did the known biases (brightness underestimation, high-mass deficits) diminish. This confirms that removing challenging low-S/N cases from training does not enhance performance on the remaining higher-S/N cases, indicating that the network's limitations reflect the intrinsic difficulty of the task rather than training-set composition.

6.2. Task-specific improvements

The most massive clusters ($\gtrsim 10^{5.8}$ $M_\odot$) represent only a small proportion of the training data, amounting to $\sim$ 5% of the entire training set. Additionally, because of mass loss (in particular related to stellar evolution, see Bianchini et al. 2026), the most massive clusters are exclusively populated by the MOCCA sim-

[20] The corresponding mean errors as those shown in Fig. 6 are: ΔF336W mag $\sim$ 0.18 $\pm$ 0.15, ΔF475W mag $\sim$ 0.24 $\pm$ 0.20, ΔF814W mag $\sim$ 0.43 $\pm$ 0.24, $\Delta \log M/M_\odot \sim$ 0.01 $\pm$ 0.12, ΔAge $\sim$ -0.21 ± 1.1 Gyr, ΔDis $\sim$ 15 $\pm$ 37 kpc, Δe $\sim$ -0.01 ± 0.02, and ΔP.A. $\sim -0.21^\circ \pm 9^\circ$, respectively.

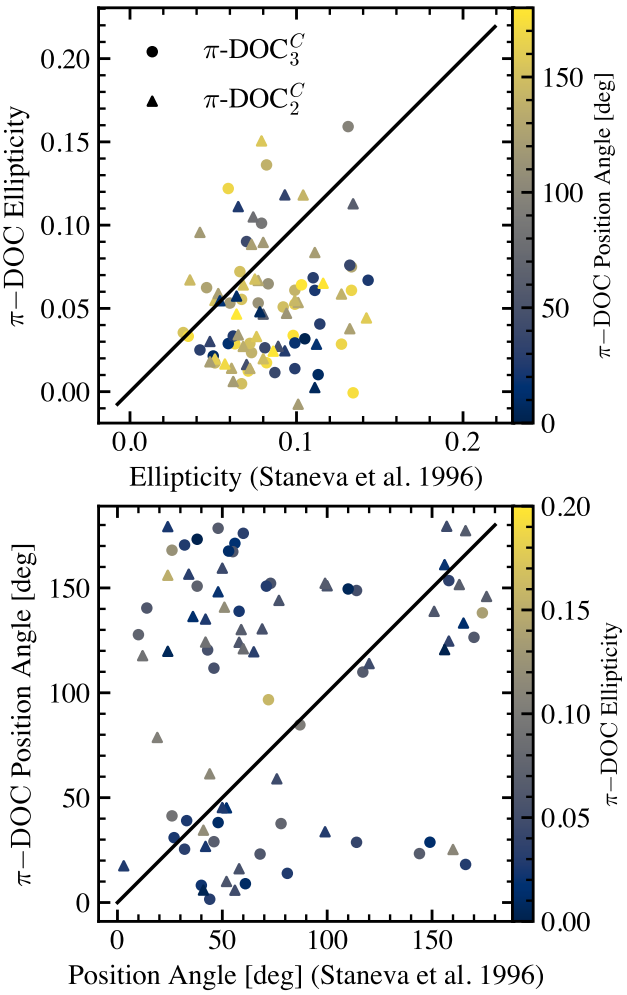

Fig. 14. Comparison between π-DOC's ellipticity and position angles estimates for 105 GCs to those in Staneva et al. (1996). π-DOC's ellipticity estimates are systematically lower than those of (Staneva et al. 1996), and the position angles show a poor overall agreement. This can partially be explained by the degeneracy at 0°, 180° and overall small values for ellipticity.

ulations for $\gtrsim$ 1 Gyr. Given the limited number of mock images for these models, and in order for the network to see a more balanced representation of the total parameter space, we conducted some preliminary tests by training alternative versions of π-DOC that were forced to see batches where at least 3%, 20% of the images were associated to MOCCA models. While this improved the massive end of the distribution, it significantly worsened other estimates, particularly for ellipticity and the position angle. This suggests that π-DOC can be fine-tuned towards specific tasks, such as improving reliability at the high-mass end, and that future versions can exploit this. For the purposes of this paper, however, we prioritise a more general model with good performance across all outputs.

7. Conclusions

Upcoming and ongoing large-scale photometric surveys are rapidly increasing both the volume and diversity of observations of GCs, spanning the local Universe and extending to higher redshifts. Fully exploiting the scientific potential of these datasets requires robust and scalable methods, which can greatly benefit from linking observational data to the underlying physical processes typically captured by numerical simulations. Establishing this connection between simulations and observations is therefore becoming increasingly important.

In this work, we introduced two complementary approaches aimed at addressing this challenge. First, we developed the forward-modelling framework `Cremant`, which transforms numerical simulation snapshots into realistic mock observations. Second, we presented the neural network π-DOC, designed to infer intrinsic dynamical and morphological properties of GCs directly from multi-band photometric images. In particular, π-DOC performs field-star decontamination, reconstructs the underlying 2D mass distribution, and infers the age, distance, ellipticity, and position angle of the GC. We found, in particular, that the inclusion of pixel-CMDs significantly improves inference accuracy for ages and distances, likely due to a more explicit encoding of the shape of the associated isochrones. In addition, we found that the inclusion of multiple passbands (a change from the original proof-of-concept algorithm) improves both accuracy across all outputs and the time to reach convergence (see Section 3).

In the synthetic tests (Section 4), representing a best-case scenario in which the data closely match the training set, π-DOC achieves strong performance. Conservative confidence limits, derived from the 2σ scatter of the residual distributions, indicate that cluster ages are recovered within ~2.5 Gyr even at low S/N, with substantially improved precision at high S/N, while masses are typically constrained within a factor of two across the full S/N range, broadly consistent with estimates for fully resolved Galactic GCs. A slight systematic underestimation of mass is nevertheless present, particularly for massive clusters ($\gtrsim 10^6$ M$_\odot$), likely owing to their limited representation in the training set. Ellipticity and position angle estimates are well constrained for genuinely elliptical clusters, while lower ellipticities ($\lesssim$ 0.1) show a considerable spread. Distance estimates remain broadly reliable, albeit with larger formal uncertainties. The decontamination performs more reliably at high S/N, but total integrated magnitudes are systematically underestimated by ~0.56–0.81 mag from blue to red passbands, producing a non-negligible colour bias of up to ~0.3 mag, particularly at low S/N. This performance at low signal-to-noise ratios, suggests that future developments of π-DOC could extend the framework toward simultaneous GC detection and parameter inference.

Applied to the PHAT and PHAST GC samples in M31, our pipeline produces mass estimates broadly consistent with literature values. Age estimates exhibit the expected scatter arising from the age–metallicity degeneracy and from the metal-poor nature of the training set ([Fe/H] ≤ -1.3), while showing improved agreement for low-metallicity systems, within typical literature uncertainties. Although the overall ellipticity distribution of M31 GCs is consistent with that reported by Staneva et al. (1996), with most clusters being only mildly elliptical ($\lesssim$ 0.1), this agreement does not extend to individual clusters. These differences most likely arise from a systematic overestimation of ellipticities by Staneva et al. (1996), since methods applied to only mildly elliptical clusters are known to be biased in this regime, as demonstrated by Fréour et al. (2026), as well as from differences in the observational data.

We also defined an explicit out-of-distribution classifier which identified metal-rich GCs as one of the main current limitations of the model. At the same time, the latent-space representations already encode information related to metallicity, suggesting that future extensions could jointly constrain age and

metallicity while enabling a deeper exploration of the underlying physical correlation encoded in the feature space by the network.

The diversity of observed GC properties further highlights the need for more extensive simulation datasets, particularly at high masses, high densities, and high metallicities, which remain underrepresented in the current training sample. Preliminary tests also indicate that π-DOC can be optimised for specific inference tasks, for example by improving robustness at the high-mass end, and similar gains may be achievable for other parameters such as ellipticity as well. Such targeted approaches may help reduce the need for extremely large simulation grids and extensive training datasets. Recent studies, such as Thuruthipilly et al. (2026), have also shown the benefits of employing domain-adaptation techniques to re-train neural networks on different photometric studies, which further suggests that vast training datasets tailored to specific surveys may not be necessary. Future extensions of π-DOC should attempt utilising such techniques.

More generally, this work illustrates how forward modelling combined with deep learning provides a scalable framework for interpreting the next generation of large photometric surveys. The inference speed of π-DOC for the current set of input image properties exceeds $\sim$ 10 inferences per second, such that processing all GCs within the footprints of the PHAT and PHAST surveys requires less than half a minute. The current version of π-DOC is limited to M31 and HST data (partly resolved GCs observed in `F336W-F475W-F814W` or `F475W-F814W`) unless special care is taken for fine-tuning or re-training. Upcoming extensions should therefore focus on a more dynamic approach that is robust across different surveys and telescopes, is invariant to combinations of passbands (e.g. by randomly disabling certain bands during training), and can handle a range of image sizes (a necessity for more distant GCs). Nevertheless, we expect that re-training such a variant of π-DOC for other observational setups will yield comparable performance for forthcoming wide-field photometric surveys.

Acknowledgements. The authors would like to acknowledge the High Performance Computing Center of the University of Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data.

References

Aarseth, S. J. 2003, Gravitational N-Body Simulations
Adamo, A., Bradley, L. D., Vanzella, E., et al. 2024, Nature, 632, 513
Alfaro-Cuello, M., Kacharov, N., Neumayer, N., et al. 2019, ApJ, 886, 57
Arca sedda, M., Kamlah, A. W. H., Spurzem, R., et al. 2024, MNRAS, 528, 5140
Askar, A., Askar, A., Pasquato, M., & Giersz, M. 2019, MNRAS, 485, 5345
Askar, A., Vergara, M. C., & Ali, S. 2025, arXiv e-prints, arXiv:2510.03766
Barbisan, E., Huang, J., Dage, K. C., et al. 2022, MNRAS, 514, 943
Barmby, P. & Huchra, J. P. 2001, The Astronomical Journal, 122, 2458–2468
Baumgardt, H. 2017, MNRAS, 464, 2174
Baumgardt, H. & Makino, J. 2003, MNRAS, 340, 227
Bellini, A., Bianchini, P., Varri, A. L., et al. 2017, ApJ, 844, 167
Bialopetravičius, J. & Narbutis, D. 2020a, A&A, 633, A148
Bialopetravičius, J. & Narbutis, D. 2020b, AJ, 160, 264
Bialopetravičius, J., Narbutis, D., & Vansevičius, V. 2019, A&A, 621, A103
Bianchini, P., Sills, A., & Miholics, M. 2017, Monthly Notices of the Royal Astronomical Society, 471, 1181
Bianchini, P., Sills, A., van de Ven, G., & Sippel, A. C. 2017, MNRAS, 469, 4359
Bianchini, P., van de Ven, G., Norris, M. A., Schinnerer, E., & Varri, A. L. 2016, MNRAS, 458, 3644
Bianchini, P., Varri, A. L., Askar, A., Marklund, A., & Mastrobuono-Battisti, A. 2026, ROLLIN': Rotating globular cluster simulations. I. The kinematic evolution of realistic direct N-body models
Bissekenov, A., Pang, X., Kamlah, A., et al. 2025, A&A, 699, A196
Boin, T., Casamiquela, L., Haywood, M., et al. 2026, A&A, 708, A215
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Caldwell, N., Harding, P., Morrison, H., et al. 2009, AJ, 137, 94
Chardin, J. & Bianchini, P. 2021, MNRAS, 504, 5656
Chemin, L., Carignan, C., & Foster, T. 2009, ApJ, 705, 1395
Chen, B., Liu, X., Xiang, M., et al. 2016, AJ, 152, 45
Chen, Y. & Gnedin, O. Y. 2024, The Open Journal of Astrophysics, 7, 23
Chen, Z., Williams, B., Lang, D., et al. 2025, ApJ, 979, 35
Conroy, C. & Gunn, J. E. 2010, ApJ, 712, 833
Conroy, C., Gunn, J. E., & White, M. 2009, ApJ, 699, 486
Dalcanton, J. J., Williams, B. F., Lang, D., et al. 2012, ApJS, 200, 18
Dold, D. & Fahrion, K. 2022, Astronomy & Astrophysics, 663, A81
Dotter, A., Sarajedini, A., Anderson, J., et al. 2010, ApJ, 708, 698
Euclid Collaboration, Scaramella, R., Amiaux, J., et al. 2022, A&A, 662, A112
Euclid Collaboration, Voggel, K., Lancon, A., et al. 2025, A&A, 693, A251
Fan, Z. & de Grijs, R. 2012, MNRAS, 424, 2009
Fan, Z., de Grijs, R., & Zhou, X. 2010, ApJ, 725, 200
Forbes, D. A. 2020, MNRAS, 493, 847
Forbes, D. A., Bastian, N., Gieles, M., et al. 2018, Proceedings of the Royal Society of London Series A, 474, 20170616
Forbes, D. A. & Bridges, T. 2010, MNRAS, 404, 1203
Fréour, L., Leitinger, E., Pancino, E., Zocchi, A., & van de Ven, G. 2026, arXiv e-prints, arXiv:2603.08432
Gal, Y. & Ghahramani, Z. 2015a, arXiv e-prints, arXiv:1506.02157
Gal, Y. & Ghahramani, Z. 2015b, arXiv e-prints, arXiv:1506.02142
Gal, Y., Islam, R., & Ghahramani, Z. 2017, arXiv e-prints, arXiv:1703.02910
Gardner, J. P., Mather, J. C., Abbott, R., et al. 2023, PASP, 135, 068001
Gieles, M., Heggie, D. C., & Zhao, H. 2011, MNRAS, 413, 2509
Giersz, M., Heggie, D. C., Hurley, J. R., & Hypki, A. 2013, Monthly Notices of the Royal Astronomical Society, 431, 2184
Green, G. 2018, The Journal of Open Source Software, 3, 695
Guiglion, G., Nepal, S., Chiappini, C., et al. 2024, A&A, 682, A9
Harris, W. E. 2010, arXiv e-prints, arXiv:1012.3224
Hiegel, J., Thélie, É., Aubert, D., et al. 2023, A&A, 679, A125
Holland, S. 1998, AJ, 115, 1916
Hurley, J. R., Pols, O. R., & Tout, C. A. 2000, MNRAS, 315, 543
Hurley, J. R., Tout, C. A., & Pols, O. R. 2002, MNRAS, 329, 897
Huxor, A. P., Mackey, A. D., Ferguson, A. M. N., et al. 2014, MNRAS, 442, 2165
Hypki, A. & Giersz, M. 2013, MNRAS, 429, 1221
Ibata, R. A., Gilmore, G., & Irwin, M. J. 1995, MNRAS, 277, 781
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jindal, A., Webb, J. J., & Bovy, J. 2019, MNRAS, 487, 3693
Kamlah, A. W. H., Leveque, A., Spurzem, R., et al. 2022, MNRAS, 511, 4060
Kang, K., Setlur, A., Tomlin, C., & Levine, S. 2024, Deep Neural Networks Tend To Extrapolate Predictably
Kingma, D. P. & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980
Kruijssen, J. M. D. 2015, Monthly Notices of the Royal Astronomical Society, 454, 1658
Kuzma, P. B., Da Costa, G. S., Mackey, A. D., & Roderick, T. A. 2016, MNRAS, 461, 3639
Lahén, N., Naab, T., Johansson, P. H., et al. 2020, ApJ, 891, 2
Lahén, N., Rantala, A., Naab, T., et al. 2025, MNRAS, 538, 2129
Leaman, R., VandenBerg, D. A., & Mendel, J. T. 2013, Monthly Notices of the Royal Astronomical Society, 436, 122
Leaman, R., VandenBerg, D. A., & Mendel, J. T. 2013, MNRAS, 436, 122
Lee, A. J., Weisz, D. R., Ren, Y., Savino, A., & Dolphin, A. E. 2025, ApJ, 995, 135
Lejeune, T., Cuisinier, F., & Buser, R. 1997, A&AS, 125, 229
Lejeune, T., Cuisinier, F., & Buser, R. 1998, A&AS, 130, 65
Li, S., Riess, A. G., Busch, M. P., et al. 2021, ApJ, 920, 84
Liu, X.-W., Zhao, G., & Hou, J.-L. 2015, Research in Astronomy and Astrophysics, 15, 1089
Malhan, K., Ibata, R. A., & Martin, N. F. 2018, MNRAS, 481, 3442
Maraston, C. 2005, MNRAS, 362, 799
Marklund, A., Bianchini, P., Varri, A. L., Kraljic, K., & Pagnini, G. in prep., in preparation
Massari, D., Koppelman, H. H., & Helmi, A. 2019, A&A, 630, L4
McConnachie, A. W., Irwin, M. J., Ferguson, A. M. N., et al. 2005, MNRAS, 356, 979
Miholics, M., Webb, J. J., & Sills, A. 2016, MNRAS, 456, 240
Mosby, G., Rauscher, B. J., Bennett, C., et al. 2020, Journal of Astronomical Telescopes, Instruments, and Systems, 6, 046001
Mowla, L., Iyer, K., Asada, Y., et al. 2024, Nature, 636, 332
Myeong, G. C., Vasiliev, E., Iorio, G., Evans, N. W., & Belokurov, V. 2019, MNRAS, 488, 1235
Pagnini, G., Di Matteo, P., Haywood, M., et al. 2026, A&A, 708, A161
Pagnini, G., Di Matteo, P., Haywood, M., et al. 2025, A&A, 693, A155
Pasquato, M. & Chung, C. 2016, A&A, 589, A95
Pasquato, M., Trevisan, P., Askar, A., et al. 2024, ApJ, 965, 89

Peacock, M. B., Maccarone, T. J., Knigge, C., et al. 2010a, MNRAS, 402, 803
Peacock, M. B., Maccarone, T. J., Knigge, C., et al. 2010b, VizieR Online Data Catalog: M31 globular cluster system (Peacock+, 2010), VizieR Online Data Catalog: J/MNRAS/402/803. Originally published in: 2010MNRAS.402..803P
Pfeffer, J., Lardo, C., Bastian, N., Saracino, S., & Kamann, S. 2021, MNRAS, 500, 2514
Renaud, F., Agertz, O., & Gieles, M. 2017, MNRAS, 465, 3622
Renaud, F., Gieles, M., & Boily, C. M. 2011, MNRAS, 418, 759
Rodriguez, C. L., Morscher, M., Wang, L., et al. 2016, MNRAS, 463, 2109
Ronneberger, O., Fischer, P., & Brox, T. 2015, arXiv e-prints, arXiv:1505.04597
Rosenberg, A., Saviane, I., Piotto, G., & Aparicio, A. 1999, AJ, 118, 2306
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Singlow, T., Techa-Angkoon, P., Bootkrajang, J., Suwannajak, C., & Tanakul, N. 2022, in 2022 19th International Conference on Electrical Engineering/Electronics, 138
Sollima, A., Dalessandro, E., Beccari, G., & Pallanca, C. 2017, Monthly Notices of the Royal Astronomical Society, 464, 3871
Spitzer, L. 1987, Dynamical evolution of globular clusters
Staneva, A., Spassova, N., & Golev, V. 1996, A&AS, 116, 447
Taylor, E. D., Read, J. I., Orkney, M. D. A., et al. 2025, Nature[Arxiv:2509.09582v1]
Thuruthipilly, H., Lisiecki, K., Junais, et al. 2026 [Arxiv:2605.13842v1]
Trenti, M. & van der Marel, R. 2013, MNRAS, 435, 3272
Usher, C., Brodie, J. P., Forbes, D. A., et al. 2019, MNRAS, 490, 491
Usher, C., Caldwell, N., & Cabrera-Ziri, I. 2024, MNRAS, 528, 6010
Valcin, D., Jimenez, R., Lardo, C., Seljak, U., & Verde, L. 2026 [Arxiv:2603.04872v1]
Valcin, D., Jimenez, R., Seljak, U., & Verde, L. 2025, arXiv e-prints, arXiv:2503.19481
van der Marel, R. P., Fardal, M., Besla, G., et al. 2012, ApJ, 753, 8
van der Marel, R. P., Fardal, M. A., Sohn, S. T., et al. 2019, ApJ, 872, 24
Vanzella, E., Claeyssens, A., Welch, B., et al. 2023, ApJ, 945, 53
Vesperini, E. & Heggie, D. C. 1997, MNRAS, 289, 898
Viaña, J., Lee, J. C., Vanderburg, A., et al. 2026 [Arxiv:2603.07289v1]
Wang, L., Spurzem, R., Aarseth, S., et al. 2016, Monthly Notices of the Royal Astronomical Society, 458, 1450–1465
Wang, L., Spurzem, R., Aarseth, S., et al. 2015, MNRAS, 450, 4070
Watkins, L. L., van der Marel, R. P., Bellini, A., & Anderson, J. 2015, ApJ, 803, 29
Webb, J. J., Reina-Campos, M., & Kruijssen, J. M. D. 2024, ApJ, 975, 242
Webb, J. J. & Vesperini, E. 2016, MNRAS, 463, 2383
Westera, P., Lejeune, T., Buser, R., Cuisinier, F., & Bruzual, G. 2002, A&A, 381, 524
Williams, B. F., Durbin, M., Lang, D., et al. 2023, ApJS, 268, 48
Williams, B. F., Lang, D., Dalcanton, J. J., et al. 2014, ApJS, 215, 9
Worthey, G. 1994, ApJS, 95, 107
Worthey, G. 1999, in Astronomical Society of the Pacific Conference Series, Vol. 192, Spectrophotometric Dating of Stars and Galaxies, ed. I. Hubeny, S. Heap, & R. Cornett, 283
Wu, K., Tanikawa, A., Flammini Dotti, F., et al. 2026, A&A, 709, A142
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579
Yuan, H.-B., Liu, X.-W., Huo, Z.-Y., et al. 2015, Monthly Notices of the Royal Astronomical Society, 448, 855
Zhang, B., Chen, B., Yuan, H., et al. 2025, ApJS, 278, 16
Zhao, Y., Askar, A., Lu, Y., et al. 2026, ApJ, 998, 328
Zoccali, M., Renzini, A., Ortolani, S., et al. 2003, A&A, 399, 931
Zocchi, A., Gieles, M., & Hénault-Brunet, V. 2017, MNRAS, 468, 4429

Appendix A: Realism of mock images

Fig. A.1 shows a qualitative comparison between five mock images (top row) and five GCs from the PHAT survey (middle row). The bottom row shows the corresponding pixel-CMDs.

Appendix B: Epistemic Uncertainty

For each input image, we perform 20 stochastic forward passes (using a dropout rate of 0.1) and compute the mean and standard deviation of the resulting predictions. We adopt the mean prediction as the final model output, while the standard deviation is used to quantify the model's confidence. This procedure naturally yields pixel-wise uncertainty maps for spatially resolved outputs (e.g. mass and magnitude maps), as well as uncertainties for our scalar regression targets: age, distance, ellipticity, and position angle. Because we predict the cos(2P.A.) and sin(2P.A.) components rather than the position angle directly, and to account for the circular nature of angular quantities, the epistemic uncertainty is estimated using the circular standard deviation. Specifically, we convert each sampled angle to its unit-vector representation (cos P.A., sin P.A.), compute the mean resultant vector over stochastic forward passes, and derive the circular standard deviation as $\sigma_{\rm circ} = \sqrt{-2 \ln R}$, where R is the magnitude of the mean resultant vector. The mean position angle is obtained from the averaged P.A. $= \frac{1}{2} \arctan 2(\sin(2\text{P.A.}), \cos(2\text{P.A.}))$.

The resulting epistemic uncertainties are shown in Fig. B.1 against the absolute error of the predictions for the test set. We see that the epistemic uncertainties, for all predicted quantities, significantly underestimates the true error. Consequently, it does not tell us much about the physical uncertainty in the prediction of a cluster, but rather indicates the confidence level of π-DOC's prediction.

Appendix C: π-DOC in two versus three passbands

We use two versions of π-DOC for the survey analyses: π-DOC$_3^C$ for PHAT and π-DOC$_2^C$ for PHAST, the latter being limited to two passbands. To assess the effect of the missing `F336W` passband, we directly compare the two networks and their predictions for the mock images in our test set. Figure C.1 shows that the networks generally agree within the typical confidence limits (Section 4.2). However, π-DOC$_2^C$ systematically predicts higher masses above 10^5 M$_\odot$, higher ages below 8 Gyr, and larger distances.

A similar comparison for the PHAT GCs reveals the same overall trends. In general, the two networks agree reasonably well for most quantities, although with non-negligible scatter. The mean absolute discrepancies are approximately 0.36 dex for the luminosity in both passbands (`F475W` and `F814W`), 0.35 dex for mass, 1.75 Gyr for age, 25.1 kpc for distance, 0.03 for ellipticity, and 50° for position angle.

Agreement is particularly good for age across the full range of values. For luminosity, however, π-DOC$_3^C$ tends to predict brighter values for the faintest objects, suggesting that the third passband is especially informative for low-luminosity GCs; note that this regime is not covered by the test set. At the bright end, the opposite trend is seen, with π-DOC$_2^C$ predicting luminosities higher by a factor of ~ 2. A similar trend is seen in mass, although π-DOC$_2^C$ also shows a systematic shift toward higher masses. As a result, its estimates are in closer agreement with those of (Chen et al. 2016; Usher et al. 2024). Distances also agree well overall, although π-DOC$_3^C$ more often predicts shorter distances, possibly because the additional pixel CMDs improve the distance inference. Ellipticity is concentrated near $e \sim 0.1$, indicating that it is only weakly constrained; this is consistent with the synthetic tests, where nearly spherical clusters tend to be overestimated and objects with $e < 0.1$ show substantial scatter. Position angle shows the largest discrepancies, likely because most clusters in the sample are only mildly elliptical. By contrast, the synthetic tests in Section 4 showed that more elliptical clusters ($e \gtrsim 0.1$–0.15) are well constrained, implying that such objects would likely have produced a clearer signal in the real-sample predictions.

Table D.1. Convolutional block used in E1 and E2.

Layer	Activation	Input	Output
`Conv2D(`N_F`, (3,3))`	ReLu	x_0	x_1
`BatchNorm`	–	x_1	x_2
`SpatialDropout2D`	–	x_2	x_3
`Conv2D(`N_F`, (3,3))`	ReLu	x_3	x_4
`BatchNorm`	–	x_4	x_5
`SpatialDropout2D`	–	x_5	x_6
`Concatenate`	–	x_3,x_6	x_7
`MaxPooling2D(2,2)`	–	x_7	x_8
`Block output`	–	–	x_3, x_8

Notes. N_F indicates number of filters (see Fig. 2).

Table D.2. De-convolutional block used in D1 and D2.

Layer	Activation	Input	Output
`Conv2D(`N_F`, (3,3))`	ReLu	x_8	y_1
`BatchNorm`	–	y_1	y_2
`SpatialDropout2D`	–	y_2	y_3
`UpSampling2D(2,2)`	–	y_3	y_4
`Concatenate`	–	y_4, x_3,x_3^C	y_5
`Conv2D(`N_F`, (3,3))`	ReLu	y_5	y_6
`BatchNorm`	–	y_6	y_7
`SpatialDropout2D`	–	y_7	y_8
`Concatenate`	–	y_8, x_8,x_8^C	y_9

Notes. ^{C} indicates outputs from the colour-branch of the network (E2).

Table D.3. The two convolutional blocks of the feature extractor (E3).

Layer	Activation	Input	Output
`Conv2D(16, (5,5))`	ReLu	CMD	x_1
`BatchNorm`	–	x_1	x_2
`MaxPooling2D(2,2)`	–	x_2	x_3
`SpatialDropout2D`	–	x_3	x_4
`Conv2D(32, (3,3))`	ReLu	x_4	x_5
`BatchNorm`	–	x_5	x_6
`MaxPooling2D(2,2)`	–	x_6	x_7
`SpatialDropout2D`	–	x_7	x_8
`Block output`	–	–	x_8

Appendix D: Tables

Man that table be ugly... need to include which version of π-DOC the prediction is associated to also. Perhaps to make space, for this table we remove the OOD column and just show in-distribution examples? Maybe I can remove the Ra and Dec here, keep only magnitudes (luminosities are a bit redundant, or alternatively vice versa)

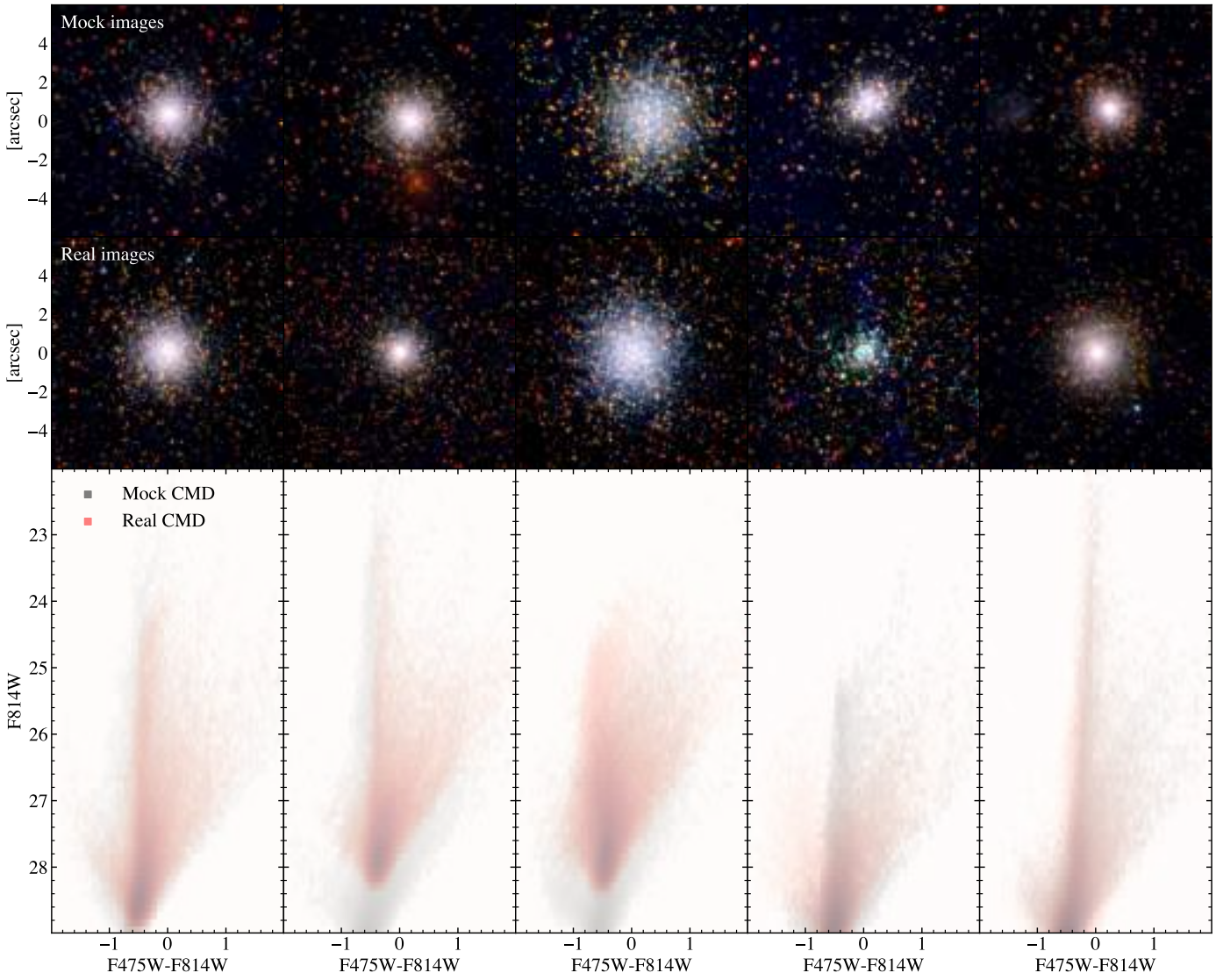

Fig. A.1. Qualitative comparison of mock images (top row), real images (middle row), and the corresponding pixel-CMDs (bottom row) for M31 GCs. The clusters resemble each other in appearance: similar sizes, shapes, and brightness. Although there are differences, especially noticeable in the pixel-CMDs, each respective image's background brightness, distance, age, metallicity, likely contribute to the shifts in magnitude and colour.

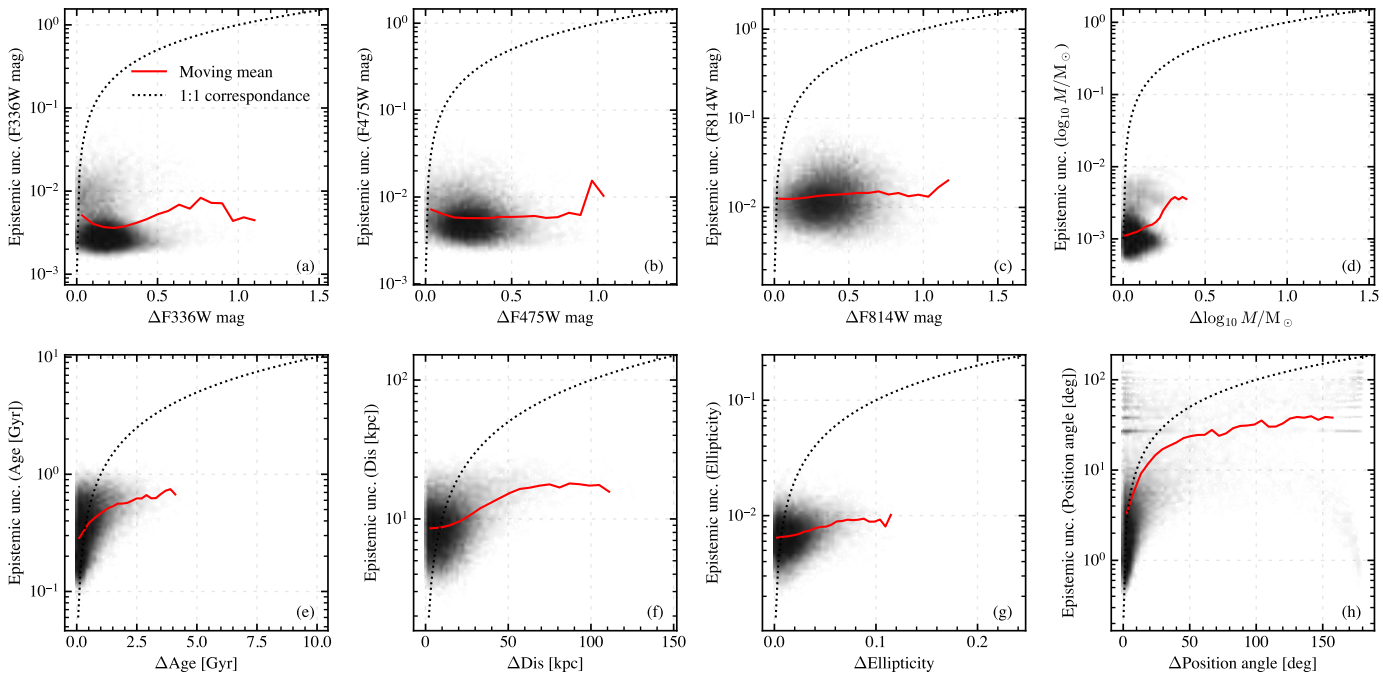

Fig. B.1. Epistemic uncertainty compared to actual errors for each predicted quantity. In all cases, the epistemic uncertainty significantly underestimates the error.

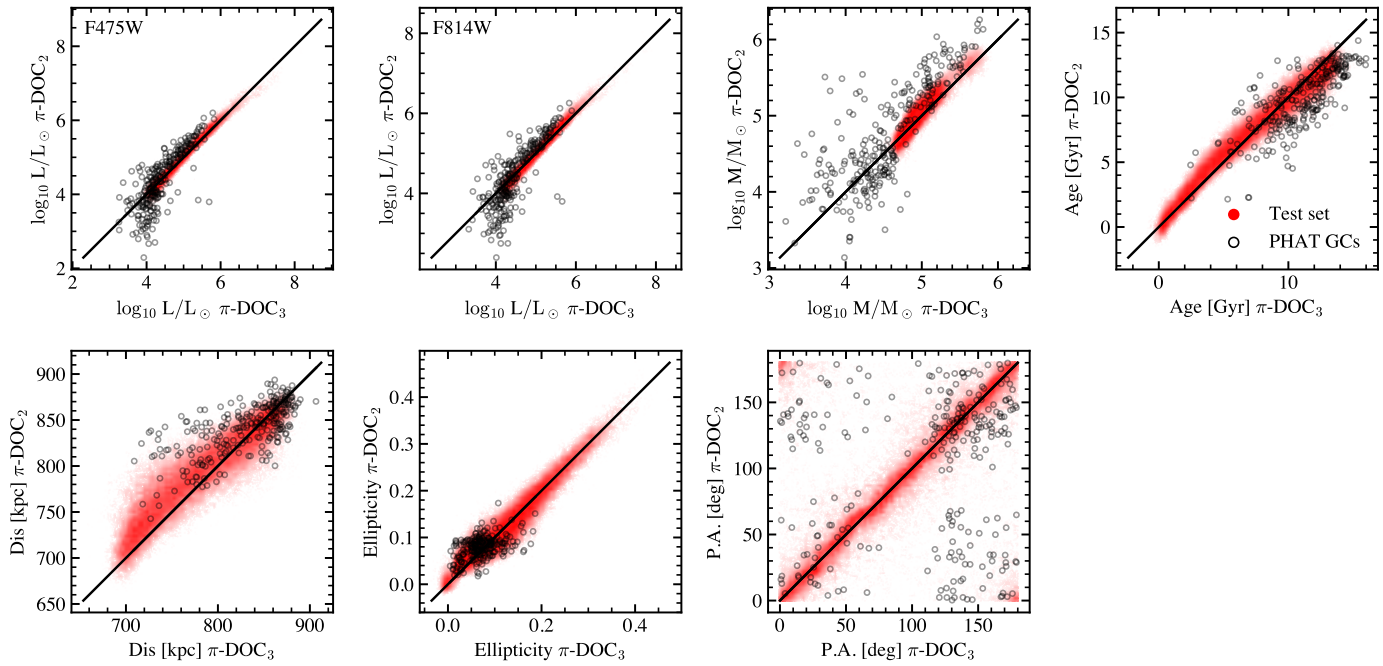

Fig. C.1. Comparisons of estimates between π-DOC$_3^C$ (x-axis) and π-DOC$_2^C$ (y-axis) for the total luminosities (within FoV) in `F475W` and `F814W` passbands, total mass (within FoV), age, distance, ellipticity, and position angle. The networks typically agree (black line shows 1:1 correspondence), however, they show some intrinsic scatter with respect to each other for the test set (red points). This scatter is also seen for the estimates of M31's GC properties which otherwise follow the distributions of the test set.

Table D.4. Globular cluster catalogue.

Name	Ra	Dec	F336W	F475W	F814W	$\log M/M_\odot$	Age	Distance	Ellipticity	P.A.	OOD
	deg	deg	mag	mag	mag	dex	Gyr	kpc		deg	
2MASS J00431591+4130329	10.82	41.51	20.38 ± 0.01	19.41 ± 0.01	19.76 ± 0.01	3.97 ± 0.00	14.39 ± 0.77	864.59 ± 9.52	0.09 ± 0.01	22.65 ± 0.13	0
2MASS J00440790+4147023	11.03	41.78	19.15 ± 0.02	18.33 ± 0.02	18.58 ± 0.12	4.56 ± 0.00	11.99 ± 0.55	800.66 ± 15.35	0.05 ± 0.01	91.40 ± 1.72	1
2MASS J00444289+4133261	11.18	41.56	20.30 ± 0.01	19.40 ± 0.01	19.82 ± 0.02	3.77 ± 0.00	13.45 ± 0.88	861.46 ± 9.74	0.06 ± 0.01	42.05 ± 0.10	1
ACH 12	10.76	41.21	20.03 ± 0.01	19.04 ± 0.01	19.40 ± 0.01	4.22 ± 0.01	10.88 ± 1.57	845.12 ± 26.60	0.05 ± 0.03	141.26 ± 0.43	1
ACH 14	10.64	41.29	16.84 ± 0.00	15.95 ± 0.00	16.41 ± 0.00	5.26 ± 0.00	9.99 ± 1.04	788.24 ± 13.34	0.06 ± 0.02	46.91 ± 0.21	1
ACH 15	10.74	41.25	19.69 ± 0.01	18.81 ± 0.01	19.21 ± 0.01	4.55 ± 0.00	12.28 ± 1.22	822.73 ± 10.71	0.04 ± 0.02	27.42 ± 0.27	1
ACH 4	10.64	41.31	18.99 ± 0.01	18.02 ± 0.01	18.38 ± 0.01	4.82 ± 0.00	11.94 ± 1.03	801.89 ± 15.93	0.07 ± 0.02	58.96 ± 1.11	1
ACH 5	10.66	41.27	14.57 ± 0.00	13.69 ± 0.00	14.18 ± 0.00	5.97 ± 0.00	16.93 ± 3.05	738.08 ± 27.39	-0.16 ± 0.06	95.01 ± 0.91	1
AP J00454469+4151594	11.44	41.87	18.10 ± 0.01	17.13 ± 0.01	17.52 ± 0.01	5.13 ± 0.00	11.89 ± 1.01	852.73 ± 10.29	0.02 ± 0.01	78.83 ± 1.39	0
BA 4-20	11.43	41.96	20.40 ± 0.01	19.33 ± 0.02	19.57 ± 0.02	4.05 ± 0.00	10.59 ± 0.61	883.22 ± 6.91	0.14 ± 0.02	64.43 ± 0.08	0
B103	10.62	41.30	16.96 ± 0.01	16.00 ± 0.01	16.34 ± 0.02	5.54 ± 0.00	9.83 ± 0.99	767.09 ± 7.13	0.03 ± 0.02	118.12 ± 1.23	1
B104	10.62	41.29	18.03 ± 0.01	17.15 ± 0.01	17.60 ± 0.01	5.09 ± 0.00	8.16 ± 0.83	780.35 ± 14.15	0.05 ± 0.02	84.48 ± 0.31	1
B107	10.63	41.33	16.98 ± 0.03	16.09 ± 0.02	16.57 ± 0.02	5.50 ± 0.00	7.72 ± 0.94	745.93 ± 7.93	0.05 ± 0.01	104.22 ± 0.15	1
B112	10.64	41.30	18.00 ± 0.01	17.03 ± 0.01	17.37 ± 0.02	5.08 ± 0.00	10.83 ± 1.06	777.03 ± 21.96	0.03 ± 0.02	42.05 ± 0.30	1
B115	10.64	41.23	17.41 ± 0.01	16.47 ± 0.01	16.83 ± 0.02	5.19 ± 0.00	8.20 ± 1.27	776.13 ± 16.10	0.03 ± 0.02	86.30 ± 0.59	1
B119	10.65	41.29	16.84 ± 0.00	15.98 ± 0.00	16.46 ± 0.00	5.36 ± 0.00	8.70 ± 1.07	798.40 ± 16.23	0.05 ± 0.02	57.33 ± 0.37	1
B124	10.67	41.26	14.43 ± 0.00	13.53 ± 0.00	14.01 ± 0.01	6.27 ± 0.00	17.24 ± 7.43	815.45 ± 83.26	-0.32 ± 0.19	92.38 ± 0.97	1
B126	10.68	41.21	18.22 ± 0.01	17.31 ± 0.01	17.77 ± 0.01	5.06 ± 0.00	8.53 ± 1.10	787.33 ± 16.91	0.08 ± 0.02	119.01 ± 0.14	1
B129	10.70	41.42	22.30 ± 0.02	21.25 ± 0.02	21.42 ± 0.02	3.37 ± 0.00	13.84 ± 0.50	842.47 ± 20.60	0.11 ± 0.02	109.41 ± 1.25	1
B131	10.71	41.29	15.08 ± 0.01	14.23 ± 0.01	14.72 ± 0.01	5.85 ± 0.00	16.35 ± 4.10	758.76 ± 20.55	-0.14 ± 0.08	95.99 ± 0.93	1
...	...	...	...	...	...	...	...	...	...	...	...
[KLG2007 GC3 161	11.24	41.92	16.34 ± 0.24	15.23 ± 0.27	15.88 ± 0.27	3.86 ± 0.00	12.89 ± 0.38	859.11 ± 11.86	0.10 ± 0.01	77.45 ± 0.11	1
[KLG2007 GC3 213	11.74	42.30	20.26 ± 0.00	19.26 ± 0.01	19.62 ± 0.01	4.38 ± 0.00	11.74 ± 0.68	884.37 ± 10.40	0.06 ± 0.01	140.69 ± 0.46	1
[M93b 317	10.95	41.46	20.87 ± 0.01	19.54 ± 0.04	19.16 ± 0.18	3.99 ± 0.00	13.43 ± 0.67	878.55 ± 6.17	0.09 ± 0.01	90.31 ± 1.24	1
[MKK98 M 31 69	11.30	41.82	19.73 ± 0.00	18.86 ± 0.01	19.31 ± 0.01	4.47 ± 0.00	13.77 ± 0.50	816.78 ± 20.77	0.16 ± 0.02	141.72 ± 0.72	0
[TIC99 10	10.89	41.24	19.79 ± 0.01	18.82 ± 0.01	19.17 ± 0.01	4.12 ± 0.00	12.85 ± 1.67	870.56 ± 13.15	0.08 ± 0.02	62.39 ± 0.77	1