Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **519,** 4384–4396 (2023) https://doi.org/10.1093/mnras/stac3799
Advance Access publication 2022 December 28

A probabilistic deep learning model to distinguish cusps and cores in dwarf galaxies

J. Expósito-Márquez,[1,2]★ C. B. Brook,[1,2] M. Huertas-Company,[1,2,3] A. Di Cintio,[1,2] A. V. Macciò,[4,5,6] R. J. J. Grand,[1,2] G. Battaglia[1,2] and E. Arjona-Gálvez[1,2]

[1]*Universidad de La Laguna. Avda. Astrofísico Fco. Sánchez, La Laguna, Tenerife 38206, Spain*
[2]*Instituto de Astrofísica de Canarias, Calle Via Láctea s/n, E-38206 La Laguna, Tenerife, Spain*
[3]*LERMA, Observatoire de Paris, CNRS, PSL, Université de Paris, 75014, Paris, France*
[4]*New York University Abu Dhabi, PO Box 129188, Abu Dhabi, United Arab Emirates*
[5]*Center for Astro, Particle and Planetary Physics (CAP^3), New York University, Abu Dhabi , PO Box 129188, United Arab Emirates*
[6]*Max Planck Institute für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany*

Accepted 2022 December 2. Received 2022 November 24; in original form 2022 September 13

ABSTRACT

Numerical simulations within a cold dark matter (DM) cosmology form haloes whose density profiles have a steep inner slope ('cusp'), yet observations of galaxies often point towards a flat central 'core'. We develop a convolutional mixture density neural network model to derive a probability density function (PDF) of the inner density slopes of DM haloes. We train the network on simulated dwarf galaxies from the NIHAO and AURIGA projects, which include both DM cusps and cores: line-of-sight velocities and 2D spatial distributions of their stars are used as inputs to obtain a PDF representing the probability of predicting a specific inner slope. The model recovers accurately the expected DM profiles: ~ 82 per cent of the galaxies have a derived inner slope within ± 0.1 of their true value, while ~ 98 per cent within ± 0.3. We apply our model to four Local Group dwarf spheroidal galaxies and find results consistent with those obtained with the Jeans modelling based code GravSphere: the Fornax dSph has a strong indication of possessing a central DM core, Carina and Sextans have cusps (although the latter with large uncertainties), while Sculptor shows a double peaked PDF indicating that a cusp is preferred, but a core cannot be ruled out. Our results show that simulation-based inference with neural networks provide a innovative and complementary method for the determination of the inner matter density profiles in galaxies, which in turn can help constrain the properties of the elusive DM.

Key words: galaxies: dwarf – galaxies: evolution – galaxies: formation – galaxies: haloes – (cosmology:) dark matter.

1 INTRODUCTION

Dark matter (DM) haloes that form in simulations within a Lambda cold dark matter (ΛCDM) cosmological context have a characteristic density profile, which has a logarithmic inner slope of -1 (the NFW profile Navarro, Frenk & White 1996). Such a steep inner density profile has been referred to as a 'cusp'. Nevertheless, observations of dwarf galaxies inhabiting these haloes have shown discrepancies with the predictions of the model, showing significant evidence that several of these galaxies have a flat inner density profile, with slope approaching zero, referred to as a 'cored' profile (Moore 1994). The discrepancy between theory and observations has been referred to as the 'core-cusp' problem (e.g. Simon et al. 2005; de Blok et al. 2008; Bullock & Boylan-Kolchin 2017).

While over the years several alternative DM models have been proposed to tackle this issue (e.g. Spergel & Steinhardt 2000; Kaplinghat, Tulin & Yu 2016; Schneider et al. 2017), it has been also shown that cores can be explained within ΛCDM considering the effect that baryons have on DM matter. Navarro, Eke & Frenk (1996) showed that if gas is slowly accreted on to a dwarf galaxy and then suddenly removed through processes such as stellar winds or supernovae feedback, the DM distribution can expand, lowering the central density of the halo. This effect of DM heating is small in realistic conditions (Gnedin & Zhao 2002), but Read et al. (2006) showed that if the effect repeats over several cycles of star formation, it accumulates leading to a complete core formation. This core can be permanent if the outflows are sufficiently rapid (Pontzen & Governato 2012). Modern hydrodynamical simulations of dwarf galaxies that take into consideration baryonic feedback and have a sufficiently high density threshold for star formation have indeed succeeded at creating DM cores (e.g. Governato et al. 2010; Di Cintio et al. 2014a; Tollet et al. 2016; Chan et al. 2015). Still, the 'cusp-core' problem is far from being completely solved, due to the difficulties of uncovering the underlying DM distribution in observed dwarf galaxies, and significant effort has gone into the development and improvement of methods to infer the inner DM density profile of such galaxies.

Analysis of the rotation velocity of gas in low surface brightness galaxies, for example, allow to derive and fit their underlying DM distribution suggesting the presence of a DM core in such systems (e.g. Moore 1994; Gentile et al. 2004; de Blok et al. 2008; Lelli, McGaugh & Schombert 2016). On the other side, in pressure-supported galaxies that are devoid of gas, such as the dwarf spheroidal

★ E-mail: expox7@gmail.com

galaxies (dSphs) found within the Local Group, the kinematic information on which dynamical modelling relies on, comes from the line-of-sight velocity distribution of their stellar component. A variety of methods have been employed on dwarf galaxies to derive their central DM density, such as Jeans (e.g. van der Marel 1994; Kleyna et al. 2001; Battaglia et al. 2008; Read, Walker & Steger 2019; Collins et al. 2021) or Schwarzschild (e.g. Schwarzschild 1979; Cappellari et al. 2006; van den Bosch & de Zeeuw 2010; Breddels et al. 2013; Breddels & Helmi 2013) modelling. The results in the literature seem to point to cored DM profiles being favoured over cuspy ones in the Fornax dSph (e.g. Geha et al. 2006; Walker & Peñarrubia 2011; Brook & Di Cintio 2015; Pascale et al. 2018), while in the case of Sculptor, another very well studied system, it is still very much debated if its DM halo is cored or cuspy, perhaps pointing to the presence of a mild cusp (e.g. Breddels & Helmi 2013; Zhu et al. 2016; Hayashi, Chiba & Ishiyama 2020) (for a review on these topics, see Battaglia et al. 2022 and references therein). A central limitation of the previously mentioned models, however, comes from the uncertainty in the anisotropy of the stellar orbits, in the case of Jeans modelling, which causes a degeneracy with the underlying mass profile (Binney & Mamon 1982); Schwarzschild modelling, on the other end, is hampered by its sensitivity to the available data (Kowalczyk, Łokas & Valluri 2017).

In this work, we present an alternative and innovative method to discriminate between cusps and cores in dwarf galaxies based on machine learning techniques. Namely, we use convolutional mixture density neural networks to determine a posterior distribution of the inner profile of DM haloes. This general approach has been successfully implemented for measuring cluster masses from galaxy dynamics (e.g. Ho et al. 2019; Kodi Ramanah et al. 2020; Kodi Ramanah, Wojtak & Arendse 2021). The neural network uses as inputs the phase-space mappings of positional and dynamical distributions of stars within galaxies. We use a suite of 171 dwarf galaxies from the NIHAO project with different initial conditions and parameters (Wang et al. 2015; Dutton et al. 2020) and 12 dwarf galaxies from the AURIGA project (Grand et al. 2017) as a training set for the network. We then apply our novel model to four dwarf spheroidal galaxies satellites of the Milky Way to infer the inner slope of their DM density profiles.

The paper is organized as follows. In Section 2, we present the simulation data set and the machine learning architecture. In Section 3, we show the results of the trained model on the test set. We then apply the model to observed dwarf galaxies in Section 4. The conclusions are discussed in Section 5.

2 METHODS

2.1 The training set

To train our model, we need a large set of simulated dwarf galaxies with well-known density profiles. We use fully cosmological simulations from NIHAO (Wang et al. 2015) and AURIGA (Grand et al. 2017) projects, in which DM and baryonic matter evolve together, making our training set as realistic as possible.

Importantly, we need to include simulations of galaxies with both cusps and cores in their central region, and with various stellar masses, in order to minimize any systematic dependence of cusp and core on properties such as mass. Indeed, the fiducial NIHAO galaxies have a density profile highly correlated with mass (Di Cintio et al. 2014b; Macciò et al. 2020), which could possibly allow the machine learning code to predict cusp or core based on any indicator of total mass, rather than by the details of the stellar velocities and positions. To maximize the neural network's ability to find and differentiate input data features directly related to their inner slope, it is prudent to avoid any non-physical correlation in the data set between the inner slope and other galaxy features. We therefore use simulations that have a range of different physical and/or parametric inputs, meaning that our final suite of simulations includes a range of inner slopes at various masses and sizes. We firstly include dwarf galaxies within the fiducial NIHAO model, ranging in halo mass from $\sim 10^9$ $M_\odot$ to $10^{11.5}$ $M_\odot$ and stellar mass from an order of 10^5 $M_\odot$ to $10^{9.5}$ $M_\odot$. This model includes energy feedback from massive stars and supernovae (Stinson et al. 2006), which has been shown to be able to modify the inner density profile and result in cores, particularly in simulated galaxies with stellar mass between 10^7 and 10^9 $M_\odot$ (Di Cintio et al. 2014a). We also use simulations of dwarfs from Dutton et al. (2020) that employs the same model as the fiducial NIHAO ones, but with different star formation thresholds, ranging from $\rho_{\rm thresh} = 0.1$ to 100 particles per cm^{-3}: this translates into galaxies of a similar stellar mass ending up with different density profiles, as the star formation density threshold has been shown to be one of the most important parameter for core formation in baryonic simulations (see Benítez-Llambay et al. 2019; Dutton et al. 2020). We further add a set of simulations with no stellar feedback run from the same initial conditions as fiducial NIHAO (Wang et al. 2015). The lower total feedback energy results in different inner density profiles than simulations in which the stellar feedback is included, for the same initial conditions, therefore further increasing the desired diversity of central DM profiles at a given galaxy mass. Finally, we include 12 simulated dwarf galaxies from the AURIGA project (Grand et al. 2017), all of which have a central DM cusp.

We have 183 simulated dwarf galaxies in total: 60 simulations from the fiducial NIHAO suite (Wang et al. 2015), 101 simulations from Dutton et al. (2020) with varying density thresholds and varying density profile, 10 simulations without stellar feedback also from Wang et al. (2015) and 12 simulations from Grand et al. (2017). All together, these simulations have a range in halo mass between $M_{\rm halo} = 3 \times 10^9$ $M_\odot$ and $M_{\rm halo} = 4 \times 10^{11}$ $M_\odot$. NIHAO simulations resolve the mass profile of galaxies to below 1 per cent of their virial radius at all masses, while AURIGA simulations are constructed to have a maximum physical softening of ∼370 pc.

We define the DM inner slope value of the simulated galaxies as the slope at 150 pc of the DM density profile of each galaxy. This value is extrapolated from the fit of the density profile to a double-power law profile (Di Cintio et al. 2014b) in order to avoid the noise effect of the computed density profile of the simulations in inner regions very close to the softening length.[1] We end up with a set of simulated dwarfs exhibiting a range of density profiles: the relationship between stellar mass and inner slope of DM halo for our full data set can be seen in Fig. 1.

To increase the size of our training set, we use three different output time-steps for each galaxy: $z = 0$, $z = 0.112$, and $z = 0.226$. Each simulated galaxy is already virialized at these redshifts, and it is therefore possible to take different snapshots of the dwarf. For the cosmological parameters from Planck Collaboration XIII (2016), the time between $z = 0$ and $z = 0.112$ is roughly 1.46 Gyr and between $z = 0.112$ and $z = 0.226$ is 1.27 Gyr. These time differences between snapshots correspond to multiple dynamical times of the galaxies from the data set, typically of the order of 10^{-2} or 10^{-1}

[1]This extrapolation is reasonable considering that AURIGA galaxies, although not resolved at $r < 370$ pc, consistently show a cuspy inner density, i.e. there is no sign of an artificial central DM core.

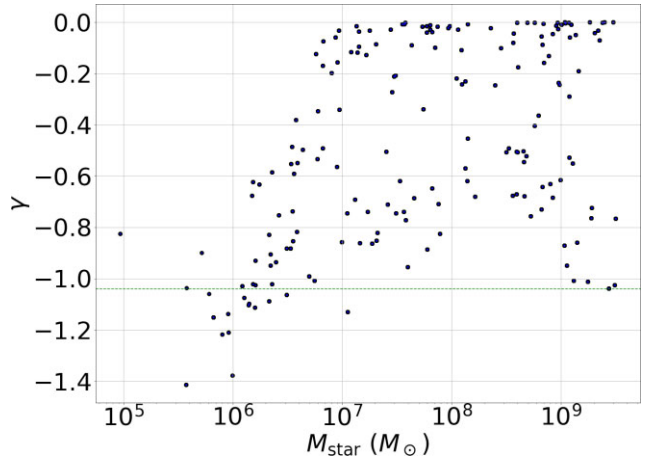

Figure 1. Relationship between the inner slope of the DM density profiles γ (defined as the logarithmic slope at 150 pc) and the stellar mass of the simulated galaxies in our data set. The green horizontal line marks the value of γ for a NWF profile.

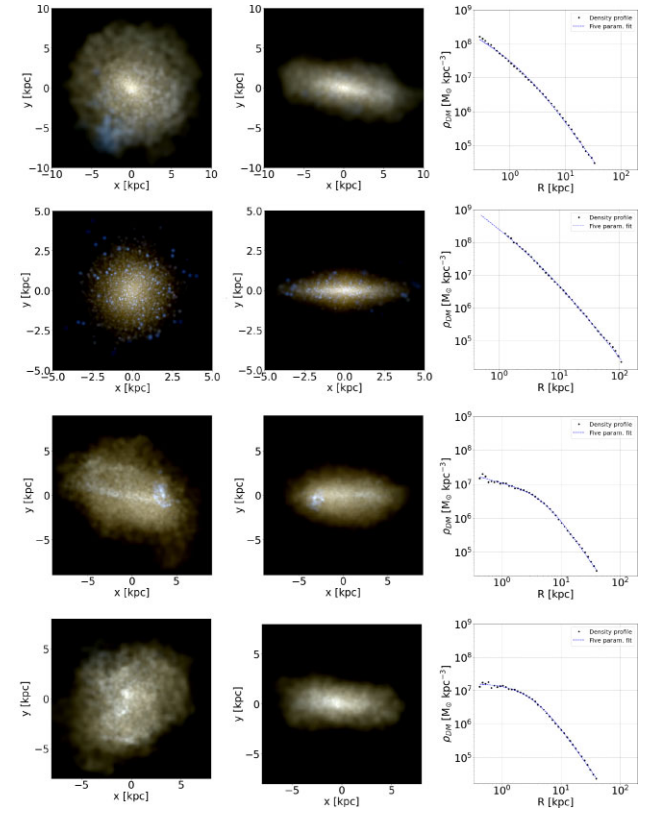

Figure 2. Example of cored and cuspy galaxies from our simulation dataset. Here, each row represents a different galaxy. Left columns: Rendering of the stars in a face-on view. Central columns: Rendering of the stars with an edge-on orientation. Right columns: DM density profiles and fit to a double-power law model (Jaffe 1983; Merritt et al. 2006).

Gyr for stars at distances to the centre under which 90 per cent of the stars of the galaxies are found. While this procedure does not change greatly the range of the obtained density profiles, it does change the position and velocities of the stars within each galaxy. We end up with a sample of 549 galaxy snapshots which we will use as training set for our method. We show in Fig. 2 examples of stellar renderings of cored and cuspy simulated galaxies together with their corresponding DM profiles. We then proceed to select stars within each galaxy snapshot. Typically, the number of stars for which spectroscopic data is available for Local Group dwarf galaxies is the order of hundreds or thousands, while the number of star particles available in our simulated galaxies range from a few hundred to several million, with a mean number of about 10^5 stellar particles in each galaxy.

Therefore, in order to simulate an observational sample of stars, and to further expand our training set, we have divided each simulated galaxy's complete sample of stars into a minimum of 20 subsets, each made of randomly selected stars. The number of stars within each subset of a given galaxy is dependent on the total number of star particles in the simulation, with an upper limit of 10^4 stars and a lower limit of 200 stars. The stars of each subset are then projected in arbitrary sky planes to simulate galaxies observed from different viewing angles. These projected stars are defined by their position $(x_{\rm proj}, y_{\rm proj})$ and their line-of-sight velocity $v_{\rm LOS}$. We oversample some galaxies by making multiple projections to each of their subset, and undersample some galaxies, with the objective of making the training set have a uniform distribution of inner slopes: this avoids biases in the model during training. We end up with a total of 10 273 data sets to train our model, each composed of randomly selected stars within different simulated galaxies and at different viewing angles, for which we stored information about their positions $(x_{\rm proj}, y_{\rm proj})$ and line-of-sight velocities $v_{\rm LOS}$.

2.2 The information inputs

The inputs of our deep neural network model are continuous 2D probability density functions (PDFs) of the distribution of stars in projected phase spaces, constructed with bivariate kernel density estimations (KDEs). The mapping generated with KDEs allows us to encapsulate the features of the original discrete distributions in the same form even if each galaxy subset is represented by a different number of stars.

2.2.1 *Kernel density estimation*

Let $\boldsymbol{X_1}$, $\boldsymbol{X_2}$,..., $\boldsymbol{X_n}$ denote a sample of size n from a random variable with density f, each variable being a two-dimensional vector for the case of a bivariate KDE. The kernel density estimate of f at the point $\boldsymbol{x}$ is given by

$$f_h(\boldsymbol{x}) = \frac{1}{n|\boldsymbol{H}|^{1/2}} \sum_{i=1}^{n} K\left[\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{X}_i)\right], \quad (1)$$

where K is a kernel function and $\boldsymbol{H}$ is a 2x2 bandwidth matrix.

The KDE sums up the density contributions from the collection of data points at the evaluation point $\boldsymbol{x}$, so that data points close to $\boldsymbol{x}$ contribute significantly to the total density, while data points further away from $\boldsymbol{x}$ contribute less. The shape of those contributions is determined by K, and their dimensions and orientation by $\boldsymbol{H}$. Usually the kernel function K is chosen to be a probability density symmetric about zero (Sheather 2004). In this work, we use a 2D Gaussian kernel:

$$K(\boldsymbol{u}) = (2\pi)^{-3/2}|\boldsymbol{H}|^{1/2} exp\left(-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{H}^{-1}\boldsymbol{u}\right), \quad (2)$$

where $\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{X_i}$. For the bandwidth matrix, a scaling factor κ is multiplied by the covariance matrix of the data. For the selection of κ, we use Scott's Rule (Scott 1992), which, for equally weighted

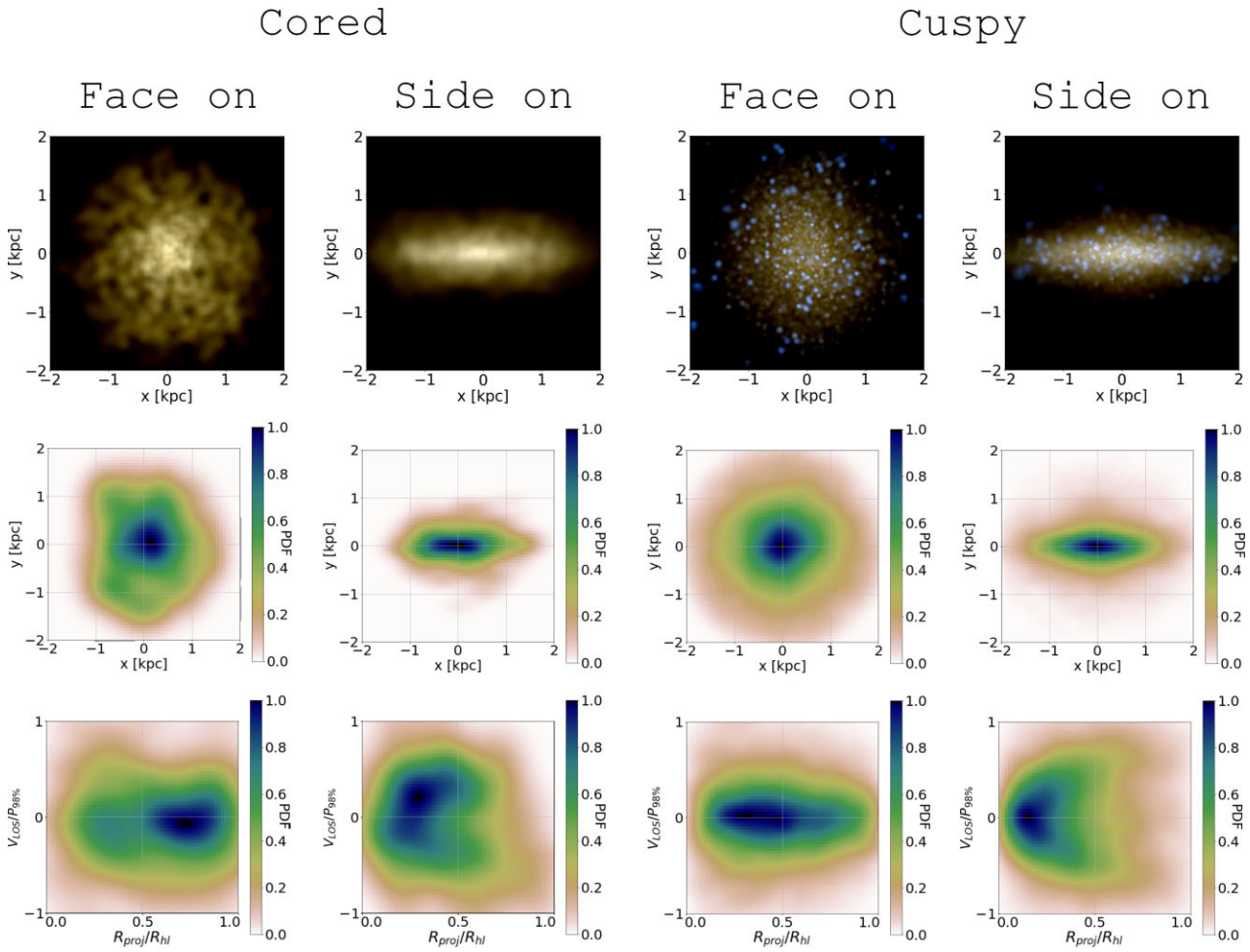

Figure 3. Model inputs for a cored and a cuspy galaxy, each one represented face-on and edge-on. The logarithmic slope at 150 pc is $\gamma = -0.20$ for the cored galaxy and $\gamma = -1.32$ for the cuspy galaxy. From top to bottom: A 3-colour image of the stars in the galaxy; the PDF in the $\{x,y\}$ phase space; and the PDF in the $\{\hat{R}_{\rm proj}, \hat{v}_{\rm LOS}\}$ phase space.

points and two dimensions is $\kappa = n^{\frac{-1}{6}}$, where n is the number of data points. This leads to a fairly strong smoothing, which interested us to reduce the relevance of the number of stars and strengthen the overall evaluation of the data as opposed to individual stars.

2.2.2 *Model inputs*

From the projected information (positions in the x–y plane and $v_{\rm LOS}$) of the sample of stars representing each galaxy we have made two maps:

(i) A PDF sampled at 64 x 64 points with the distribution of stars in $\{x,y\}$ phase space, between -2 kpc and 2 kpc in each coordinate, in the reference system where $(x,y) = (0,0)$ is the centre of the galaxy.

(ii) A PDF sampled at 64 x 64 points with the distribution of stars in $\{\hat{R}_{\rm proj}, \hat{v}_{\rm LOS}\}$ phase space, where $\hat{R}_{\rm proj} = \sqrt{x^2 + y^2}/R_{\rm hlr}$ is the radial position normalized by the half-light radius $R_{\rm hlr}$ and $\hat{v}_{\rm LOS} = v_{\rm LOS}/P_{98\ \rm per\ cent}$ is the line-of-sight velocity normalized by the 98 per cent percentile of the absolute value of $v_{\rm LOS}$ of all stars of the sample. $\hat{R}_{\rm proj}$ ranges from 0 to 1, and $\hat{v}_{\rm LOS}$ ranges from -1 to 1.

Note that both the 2 kpc bounds in the positional data PDF and the limit up to R_{hl} in the velocity PDF imply ignoring star data outside these regions. During the testing phase many bounds and normalization methods were tested. With the current data set, the limits used in the work are the ones that gave the best results. A likely explanation is that the information provided by stars outside these limits is negligible and their presence in the PDFs only detracts from the stars closer to the centre of the galaxy, where the key information for determining the internal slope is found.

In Fig. 3, we show our model inputs, as the PDFs corresponding to both maps, for a cored (left) and cuspy (right) galaxy.

2.3 The model

In this work, we use mixture density convolutional neural networks (MDCNNs) to map the input data composed of the two PDFs described in Section 2.2.1 into the inner slopes of the DM profiles of the galaxy associated to those two PDFs. We approximate the posterior distribution of the slopes with the sum of two Gaussian distribution whose parameters are estimated by the neural network.[2] Our model takes as input a two channel image consisting of the PDFs on the $\{\hat{R}_{\rm proj}, \hat{v}_{\rm LOS}\}$ phase space and the $\{x,y\}$ phase space separately.

[2]The use of a double Gaussian yields more accurate predictions than using a single one. On the other hand, using more than two Gaussians does not lead to more accurate slope predictions.

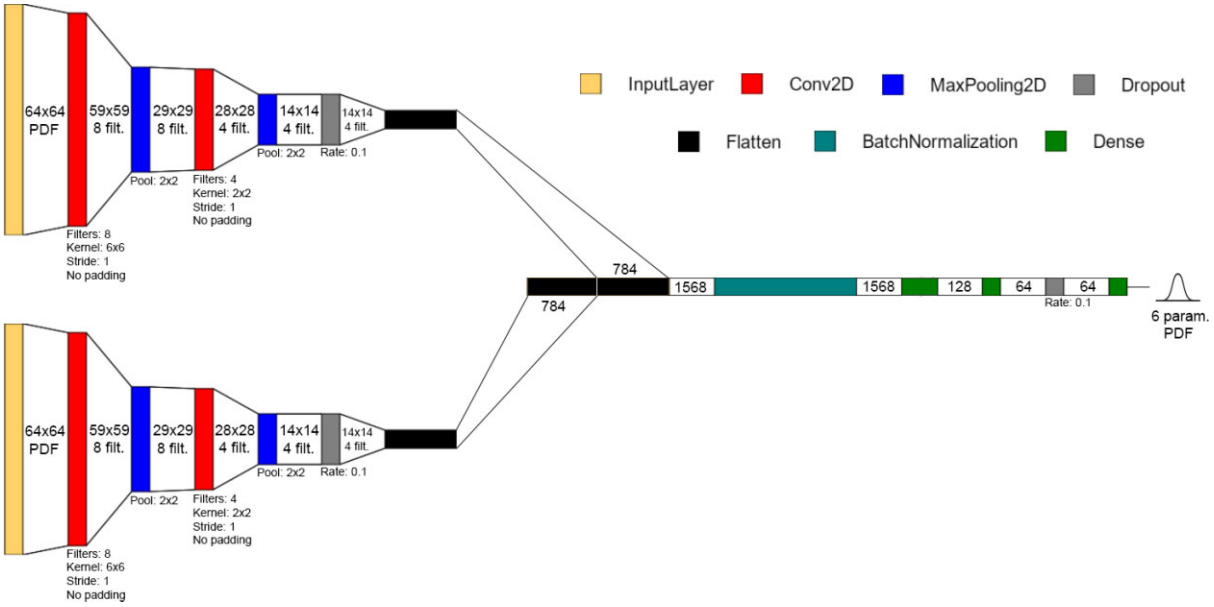

Figure 4. Schematic representation of our double channel MDCNN architecture to infer inner slopes of the DM profiles (slope at 150 pc) of galaxies from their 2D phase-space mappings of positional and dynamical distributions of stars. The MDCNN extracts the spatial features from the phase-space mappings and gradually compresses into high-order features until describing the input with only five parameters, which are used as parameters of a double Gaussian corresponding to the probability density distribution of the inner slopes values.

The images are passed through two convolutional sequential layers. The outputs of the two convolutional branches are then concatenated and fed into a three layer fully connected network. The final output consists of five parameters that parametrize the joint double Gaussian posterior.

A schematic view of the architecture used in this work can be seen in Fig. 4, while a more in-depth description of the different layers and neural network methods can be found in the Appendix.

2.3.1 *Training and evaluation*

The training is done over a training set consisting of 10 273 galaxy subsets with their respective inner slopes, which act as targets. The loss function to minimize during the training is the negative logarithmic likelihood of the training sample, defined as

$$L = -\ln L* = -\sum_{i=1}^{N} \ln\left[p_g(t_i|\theta)\right], \tag{3}$$

where t_i is the inner slope of the galaxy subset i and θ the set of parameters of the distribution p_g. For a certain galaxy subset, the likelihood is the value of the PDF (defined with a double Gaussian distribution as the output of the last layer) in its real inner slope value; i.e. the probability the model predicts for the inner slope of the galaxy to be its correct value:

$$p_g(x|\theta) = \sum_{j=1}^{2} \phi_j \boldsymbol{N}(x, \mu_j, \sigma_j), \tag{4}$$

where $\boldsymbol{N}(x, \mu_j, \sigma_j)$ is the j Gaussian with mean μ_j and standard deviation σ_j, ϕ_j is the weight of the j Gaussian, so that $\sum_{j=1}^{n_g} \phi_j = 1$, and θ is then a set of six parameters (mean, standard deviation, and weight of the two Gaussians), one of which is not independent due to the normalization criterion.

The minimization of the loss function is done with the adaptative moment estimation (ADAM) optimizer, an algorithm for optimization that uses the gradient descent iterative technique. Between the popular learning-method algorithms, ADAM is shown to compare favourably in performance and computational cost (Kingma & Ba 2015). After training, the evaluation of the model outputs a double Gaussian distribution that can be understood as an approximation to the true posterior distribution of the inner slope of a given input, given the prior distribution of the inner slopes in the training data set. This posterior then represents the probability that the model assigns a certain value of the inner slope, given the set of observables under the prior of the training set.

Usually, the test data set for the final evaluation of the converged model is constructed by randomly taking a sufficient number of elements from the complete data set to correctly represent all feature variety in the data. In this work, due to the limited number of galaxies available, removing too many galaxies with varying characteristics from the training data set is expected to worsen the performance of the model, since we do not have many different examples of galaxies with similar characteristics to each other. To properly evaluate the model, we have performed multiple complete training runs using only 10 galaxies as validation and test data sets in each one, changing the galaxies that would come out of the training data set in each of the training runs to evaluate the network in several projections of every galaxy. This allows us to analyse the consistency of the model training and its performance in a large number of galaxies without compromising the training data set.

2.3.2 *Representing uncertainties*

The output posterior distribution represents the random or aleatoric uncertainty in the slope prediction of the final model, but it does not represent the uncertainty due to the stochastic nature of the weight determination while training the neural network (epistemic uncertainty), which can lead to different models for the same training conditions when dealing with limited data. We use the Monte Carlo dropout method (MC-Dropout) (Gal & Ghahramani 2015) to

approximate the epistemic uncertainty that is based on the repeated evaluation of the same input, randomly setting to 0 the weights on some layers while doing each inference, to construct a final evaluation with statistical information about the epistemic uncertainty. Gal & Ghahramani (2015) showed that applying dropout during inference is equivalent to an approximation to a probabilistic Deep Gaussian process. It means we can measure the epistemic uncertainty by applying the dropout layer during inference for a statistically relevant number of them, acquiring a predictive mean and variance for each point of the posterior distribution. The constructed final posterior for each galaxy projection is the normalized mean of 100 double Gaussian posteriors inferred by the model with active dropout layers.

3 RESULTS

The goal of our work is to infer the logarithmic inner slope of the mass density profile in the central region of a galaxy (from now on: inner slope) from spectroscopic data of a random sample of its stars. To do so, all simulated galaxies and their subsets of stars are randomly projected in several sky planes, to simulate several viewing angles, and the neural network is trained to infer the inner slope of the galaxy from the positions and line-of-sight velocities of its stars. For each galaxy the neural network outputs, a PDF which approximates the posterior probability of obtaining a specific inner slope given the inputs.

3.1 Predicting DM inner slopes

We define two different methods to construct the predicted slope value γ from the posteriors:

(i) by using the mode of the posterior distribution (i.e. the maximum of the PDF): $\gamma_{\rm Pred,\,mode}$.

(ii) by using the mean of the normalized posterior distribution: $\gamma_{\rm Pred,\,mean}$.

The deviation ϵ of a prediction from its true value is defined as $\epsilon_i = \gamma_{\rm Real} - \gamma_{\rm Pred,\,i}$, where $\gamma_{\rm Real}$ is the real slope at 150 pc of the DM profile of a galaxy simulation. The results for the mode method can be seen in Fig. 5, which shows the difference between the real and predicted slopes of our simulated dwarf galaxies, $\gamma_{\rm Real} - \gamma_{\rm Pred}$, as a function of the real slope. Each point represents the mean deviation for every projection of each individual galaxy, while the deviation bars indicate the minimum and the maximum value amongst every possible projection of each galaxy. Shaded-coloured horizontal areas represent increasing uncertainty ranges, from ± 0.05 to ± 0.4.

The mean global absolute deviation on the predicted inner slope, for all the galaxies in our set, is of $\mu_\epsilon = 0.056$ for the mode method and of $\mu_\epsilon = 0.068$ for the second method. Note that while cuspy and 'in between' galaxies are scattered around $\gamma_{\rm Real} - \gamma_{\rm Pred} = 0$, cored galaxies tending towards $\gamma = 0$ are necessarily only scattered at $\gamma_{\rm Real} - \gamma_{\rm Pred} \geq 0$, since by construction the maximum possible inner slope is 0.

In Table 1, we can see the percentages of correctly predicted inner slopes, taking into account all the projections of every galaxy (middle column) and each galaxy individually (right column), for our complete test data set, within several uncertainty ranges. Roughly, 82 per cent of the galaxies recover the correct, real inner slope within ± 0.1, while 98 per cent of them lie within $|\gamma_{\rm Real} - \gamma_{\rm Pred}| \leq 0.3$. These ranges are clearly small enough to shed light on the discussion regarding the presence or not of cores in dwarf galaxies.

Finally, a histogram of the deviation distribution for every projection of each galaxy (i.e. 10 273 in total) can be seen in Fig. 6,

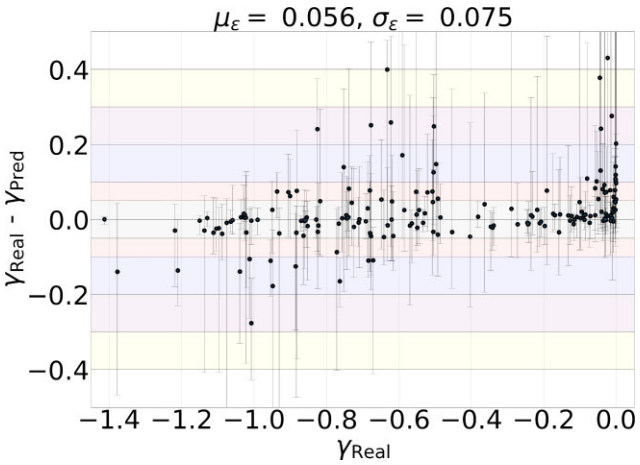

Figure 5. Difference between real and predicted value of DM profiles inner slopes (defined at 150 pc) versus real inner slope, for the simulated galaxies used in this work, defining the predicted value as the mode of the posterior distribution. Each point represents the mean $\gamma_{\rm Real} - \gamma_{\rm Pred}$ for all the projections of each individual galaxy, while error bars span the range between the minimum and the maximum deviation amongst every possible projection of each galaxy. Coloured areas represent increasing deviation ranges, from 0.05 to 0.4.

Table 1. Percentage of all the projections (central column) and of individual galaxies (right column) whose predicted inner slope lies within a given deviation range X, i.e. $|\epsilon| = |\gamma_{\rm Real} - \gamma_{\rm Pred,\,mode}| \leq X$. Here, we used the mode of the posteriors method to derive the inner slopes.

Deviation range ($\pm$ X)	Percentage of projections with $\lvert\epsilon\rvert \leq X$	Percentage of galaxies with $\lvert\epsilon\rvert \leq X$
0.05	66.67	67.80
0.1	80.79	81.92
0.2	94.35	94.35
0.3	98.31	98.31
0.4	99.44	98.87

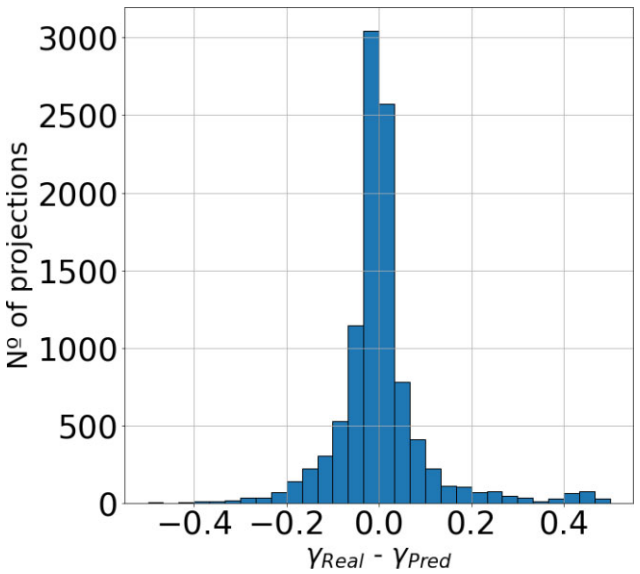

Figure 6. Distribution of $\gamma_{\rm Real} - \gamma_{\rm Pred,\,mode}$ for every projection of each galaxy in our training set.

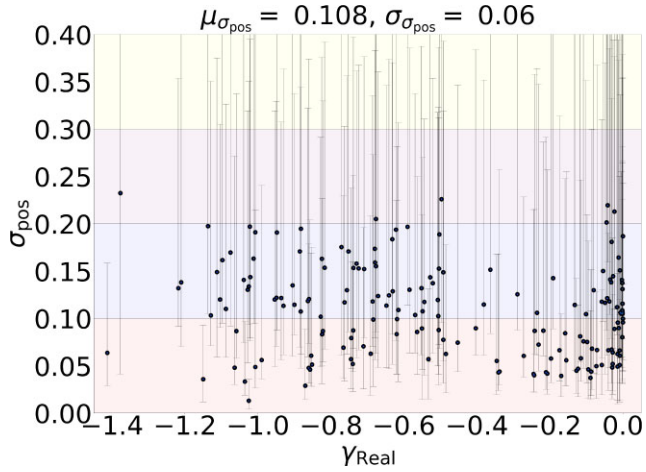

Figure 7. Standard deviation σ_{pos} of each posterior PDF versus γ_{Real}, for the simulated galaxies used in this work. Each point represents the mean standard deviation of each posterior, for every projection of an individual galaxy. The error bars range between the minimum and maximum standard deviation value of the posteriors of all the projections of that galaxy.

indicating that the values of $\gamma_{Real}-\gamma_{Pred}$ are peaked at and roughly symmetrically distributed around 0, except for very cored galaxies that have by definition $\gamma_{Real}-\gamma_{Pred} \geq 0$, as already stated, and a small asymmetry towards predicting stronger cores in galaxies in the range of small deviations. We showed that our method predicts accurately the expected inner slope of galaxies regardless of their actual real slope, with a mostly uniform scatter of $\sigma_\epsilon = 0.075$.

3.2 Uncertainty in the inference

In Fig. 7, we show the standard deviation σ_{pos} of each posterior PDF from every galaxy in the test data set, defined as the square root of the variance of the normalized posterior:

$$\sigma^2_{pos} = \int_{-\infty}^{\infty} (\gamma - \mu_{pos})^2 P(\gamma)\, d\gamma, \quad (5)$$

where $P(\gamma)$ is the normalized posterior distribution and μ_{pos} is the mean of the distribution:

$$\mu_{pos} = \int_{-\infty}^{\infty} \gamma P(\gamma)\, d\gamma. \quad (6)$$

The mean of all the σ_{pos} of the data set, $\mu_{\sigma_{pos}}$, is around 0.1 and only 8.99 per cent of the projections have values of σ_{pos} greater than 0.2, uncertainties that are small enough to clearly distinguish between cores and cusps in the vast majority of cases. Fig. 7 shows that the standard deviation σ_{pos} of each posterior PDF is uniform across the inner slopes values, i.e. the width of the PDFs does not depend on the inner slope of galaxies, such that the model is not biased towards recovering with higher accuracy either cusps or cores. Most galaxies show a significant variation in the size of their uncertainties depending on the projection, indicating that the amplitude of the uncertainty is strongly correlated with the angle of observation.

Table 2 shows the percentage of the test data set projections for which the true value of their inner slope is recovered within different multiples of σ_{pos}. If we approximate the posteriors to single Gaussians (which is a proper approximation for roughly 90 per cent of the projections), a well-calibrated uncertainty should provide around 68 per cent of the outputs within a confidence level of 1 $-\sigma_{pos}$. Our greater percentage ($\sim$ 86 per cent) of projections within the confidence level of $1-\sigma_{pos}$ indicates that the model is over-predicting the uncertainties σ_{pos}, yielding broader posteriors than it should. This can be an effect of a too high dropout rate (see Section A) during training, which has been shown to have such a outcome on the results of probabilistic neural network models (Ghosh et al. 2022). As it is, our model should be interpreted as conservative, since a future, better calibrated MDCNN would provide even tighter uncertainties in recovering the true inner slope of a galaxy.

Table 2. Percentage of predictions within increasing σ_{pos} ranges X, defined as $|\gamma_{Real} - \gamma_{Pred}| \leq X$.

Region	Percentage of projections for which γ_{Real} is within region
$1-\sigma_{pos}$	86.29
$2-\sigma_{pos}$	97.57
$3-\sigma_{pos}$	99.73

3.3 Effect of viewing angle on the inference of DM slopes

Most of the posteriors for the different projections have an approximately normal distribution (the second Gaussian disappearing or constituting a skewness correction to the main Gaussian), but several of them have two distinct peaks. Specifically, around 30 per cent of the galaxies have double peaks in more than 10 per cent of their posteriors. 54 per cent of these galaxies are cored while 46 per cent are cuspy, indicating that the appearance of double peaks in the PDFs arises in both scenarios (here, we define as cored galaxies those with inner slope $-0.6 < \gamma < 0$, and cuspy any galaxy with $\gamma < -0.6$). In Figs 8 and 9, we show the PDFs and posteriors of two galaxies at different observation angles, spanning the range between a face-on and a edge-on view. Strikingly, these images show that the width of the PDFs as well as the appearance of double peaks are strongly related to the viewing angle of the galaxy. This indicates that the appearance of double peaks is a consequence of the fact that some information on the underlying DM profiles is hidden when viewing the galaxy at some particular angle, while it is released and efficiently passed to the network when looking at the galaxy from other angles: this finding has profound consequences for the interpretation of 'cusp-cores' in dwarfs. For example, in Fig. 9 we observe that the double peaks in the posterior distribution disappear when the galaxy is seen edge-on, while a face-on configuration provides a second peak that mimics the presence of a cusp.

However, this is just an example, and we have several cases of galaxies in which the double peaks appear in edge-on view and disappear in face-on, so that the appearance of these multiple peaks is not related to a specific edge-on or face-on configuration: indeed, the distribution of angles for those PDFs showing double peaks is uniform throughout the complete data set. The occurrence, significance and widths of the double peaked PDFs will be explored in future works, as it goes beyond the scope of this paper.

4 APPLICATION TO OBSERVED GALAXIES

We proceed to test our model with real observed galaxies, in order to ensure the applicability of the model and to verify that the neural network is not detecting features of simulated galaxies that do not correspond to any real physical system. We selected four dSphs for

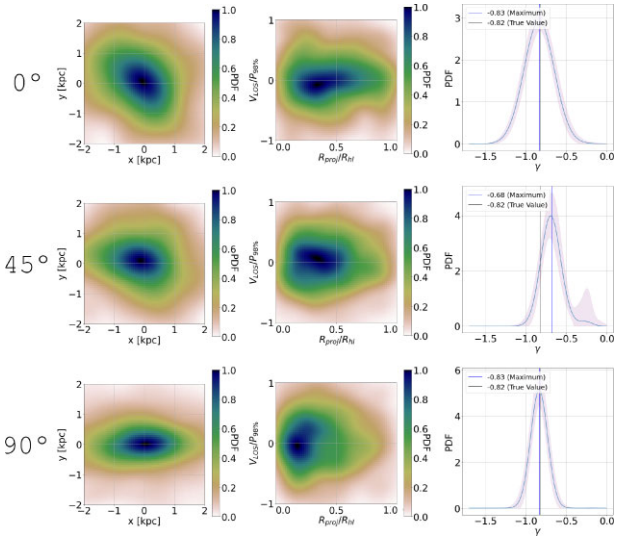

Figure 8. Probability density distributions used by the neural network as input in the case of one simulated galaxy subset seen at 0° (face-on), 45° and 90° (side-on), alongside with the Bayesian posteriors predicted by the model. Left columns: PDFs in the $\{x,y\}$ phase space. Central columns: PDFs in the $\{\hat{R}_{\rm proj}, \hat{v}_{\rm LOS}\}$ phase space. Right columns: predicted Bayesian posterior in the space of inner slope of the DM profile (slope at 150 pc); shaded regions represent the standard deviation of the posterior values for the MC-Dropout inferences at each slope point, while the blue vertical line shows the mode (maximum) of the posterior distribution and the black one the true value of the inner DM slope.

which detailed spectroscopic samples of stellar-kinematic data have been published. At this stage, we adopt the catalogues by Walker, Mateo & Olszewski (2009) to directly compare our results with those obtained using the code GRAVSPHERE, as in Read et al. (2019). The selected galaxies are Carina, Sextans, Fornax and Sculptor, for which we further use the center position, velocity, ellipticity and half-light radius as compiled in Battaglia et al. (2022).

To build our input PDFs, we considered only those stars with a 90 per cent or higher probability of being part of the galaxy and we took the mean value of the line-of-sight velocity for those stars with multiple measurements. We do not take observational uncertainties into account, since adding noise to the data by making use of uncertainties in the line-of-sight velocity only goes so far as to alter the mean, maximum, and width of the posteriors by an order of 10^{-2} over multiple iterations for these four galaxies. This may change in the future if more data sources with less accurate measurements, such as proper motion, are added. A full and formal treatment of the effect of observational uncertainties will be included in future work, but their inclusion does not affect the results presented here. In total, we considered 460 stars for Carina, 1353 for Fornax, 809 for Sculptor and 327 for Sextans, and we used their projected x–y positions and line-of-sight velocities. The x–y positions are normalized using the circularized half-light radius $R'_{\rm hlr} = R_{\rm hlr}\sqrt{1-{\rm ell}}$, where $R_{\rm hlr}$ and ell are the half-light radius and ellipticity from Battaglia et al. (2022).

4.1 Deriving central DM density slopes of dSphs with CNNs

We now infer the inner slope of the observed dwarfs. Fig. 10 shows the posterior distributions constructed by the model for each observed galaxy. Fornax presents a very narrow peak around $\gamma = -0.38$, indicating that this galaxy has a strong central DM core, while a secondary peak would give a 12 per cent probability that the inner slope is around $\gamma = -0.81$. This is consistent with several previous works that predict a cored profile for Fornax (see Goerdt et al. 2006; Walker & Peñarrubia 2011; Brook & Di Cintio 2015; Pascale et al. 2018, amongst others). For the other three galaxies, a cusp is predicted with varying degrees of certainty. The model has a clear peak around $\gamma = -1.06$ for Carina, which roughly corresponds to the slope of an NFW profile at 150 pc.

Sextans presents a relatively large uncertainty in the inner slope value, as depicted by the quite broad PDFs, with a broad peak around $\gamma = -1.25$ and a strong right wing that does not fall below 10 per cent of the peak value until it reaches $\gamma = -0.68$. Finally, Sculptor peaks at $\gamma = -1.08$, but it has a wide secondary peak, predicting a 18 per cent probability of having a mild core with $\gamma = -0.75$. A small core was derived for Sculptor by using kinematical data and a mass-dependent profile fit in Brook & Di Cintio (2015), in agreement with the Walker & Peñarrubia (2011) and Agnello & Evans (2012) methods that, employing multiple stellar populations within a galaxy, also predicted a core in such dwarf (see also Zhu et al. 2016; Breddels et al. 2013; Hayashi et al. 2020). Other studies, however, surprisingly predict a cusp for Sculptor after all (Richardson & Fairbairn 2014), highlighting the importance of deriving the DM density of this dSphs with several different methods. Our derived posterior distributions offer great versatility in interpreting the results, allowing for a more complex analysis compared to models that only allow for uncertainty ranges around the inferred value.

We compare the results of our model with the inner slopes inferred for these same galaxies at 150 pc using GRAVSPHERE, a non-parametric spherical Jeans analysis code, which make use of photometric and kinematic data from the galaxies (Read et al. 2019). The inferred values, along with their 68 per cent confidence intervals (in our case, taking the primary maximum as reference), are listed in Table 3. The derived values are consistent between the two models, within their respective uncertainty ranges, indicating that our neural network model is making predictions similar to those obtained by Jeans analysis. Furthermore, the accuracy of our neural network is greater, with errors roughly an order of magnitude smaller than those of GRAVSPHERE: this preliminary finding will be expanded and explored in more detail in future work.

Compared to GRAVSPHERE and similar codes, the neural network approach is significantly faster. In a modern laptop, GRAVSPHERE will need about half a day to run an analysis of one of these galaxies, whereas the neural network can be trained with the amount of data used in this work in less than half an hour on a standard GPU. Furthermore, the training and the evaluation are independent calculation in a neural network model, which means that, once the model has been trained, its application to any input data to construct the posterior distribution is nearly instantaneous. This feature will not change no matter how much the model is expanded and complexified to perform more complete analyses of the galaxy of interest.

4.2 Testing the similarity of training versus observational data

When training a neural network with simulations to then perform inference on real data, there is always the risk that the network will detect and learn from specific features of the simulation code that do not correspond to reality, and this would cause issues when interpreting observational data, as they have different qualities than those used in the training set. We can test the degree to which our

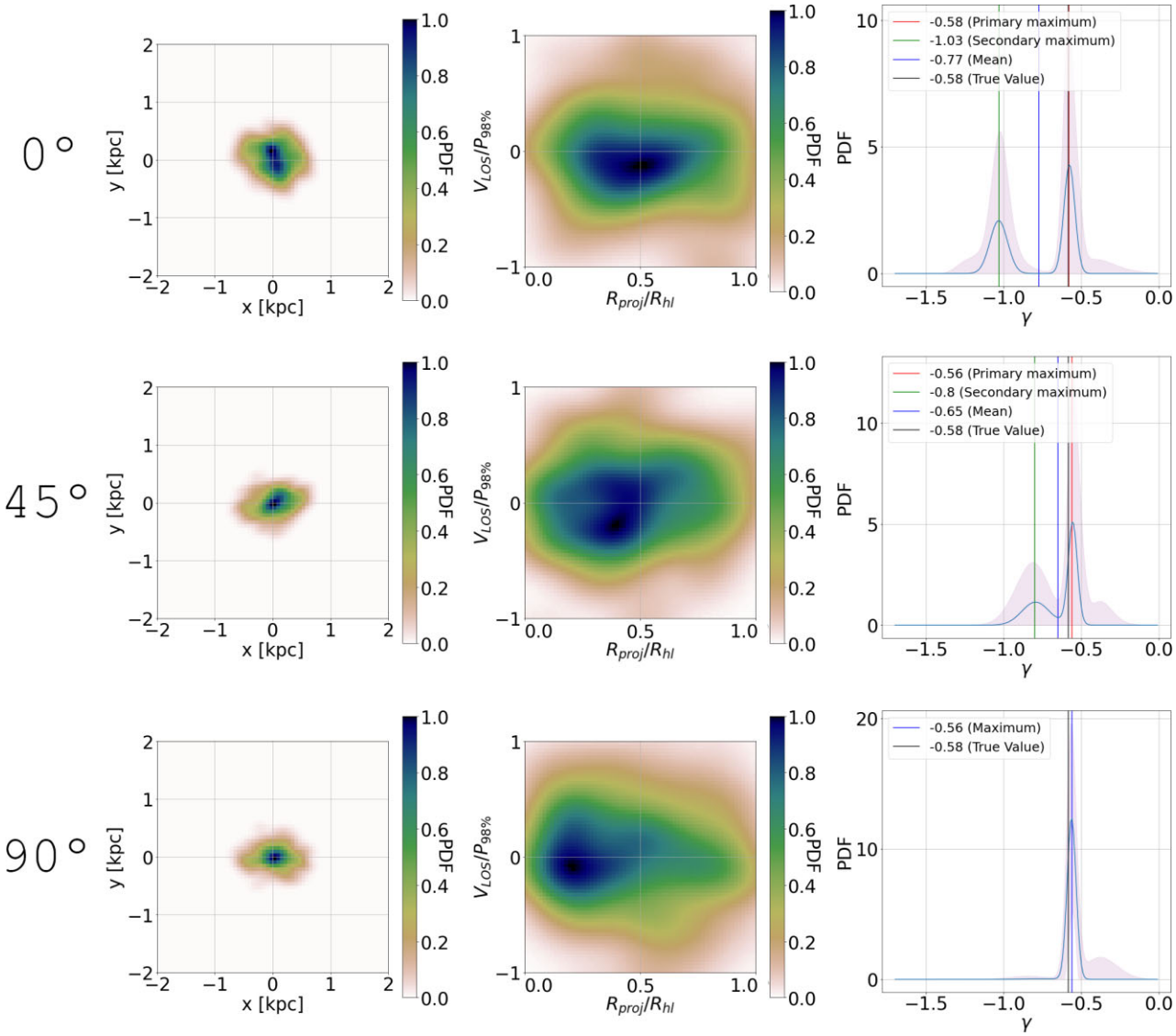

Figure 9. Probability density distributions used by the neural network as input in the case of one simulated galaxy subset seen at 0° (face-on), 45° and 90° (side-on), alongside with the Bayesian posteriors predicted by the model. Left columns: PDFs in the $\{x,y\}$ phase space. Central columns: PDFs in the $\{\hat{R}_{\rm proj}, \hat{v}_{\rm LOS}\}$ phase space. Right columns: predicted Bayesian posterior in the space of inner slope of the DM profile (slope at 150 pc); shaded regions represent the standard deviation of the posterior values for the MC-Dropout inferences at each slope point. The red vertical line shows the primary maximum of the posterior distribution (the mode), the green one the secondary maximum and the black one the true value of the inner DM slope. As a blue line, the mean between primary and secondary maximum is shown, when two peaks exist (in the bottom panel, instead, the blue line represents the unique maximum). This example shows how the appearance of double peaks in the posterior distributions is strongly related to the viewing angle.

network sees observational data as equivalent to the data it has been trained on by observing the parameter space of the test data set, defined as the set of all combinations of the six parameters corresponding to each element of such data set. Namely, our outputs are defined by the mean, standard deviation and weight of two Gaussians: this 6D parameter space will have regions populated with points and regions completely empty, corresponding to the combinations of parameters that do not parametrize the characteristics of any physical system found in the data set. If the neural network does not see differences in the input with respect to the data it has been trained on, the resulting parameters, coming from the evaluation of observational data with our model, will fall within the populated regions of the parameter space of the simulation data set.

To be able to visualize the 6D parameter space and test if this is the case, we use the Uniform Manifold Approximation and Projection (UMAP) for Dimension Reduction technique from McInnes, Healy & Melville (2018) to perform a dimension reduction from 6D to 2D, thus mapping each combination of means, standard deviations and weights to only two adimensional parameters representing such 'contraction', preserving the global structure of the original parameter space. This allows to visualize the parameter space in 2D. The result of the dimension reduction process from the complete test sample can be seen in Fig. 11, alongside with the position of the four observed dwarf galaxies shown in the same parameter space, each indicated as coloured star. As expected, the spatial location of the points in the parameter space is strongly linked to the value of their inner slope: points with a similar inner slope cluster together,

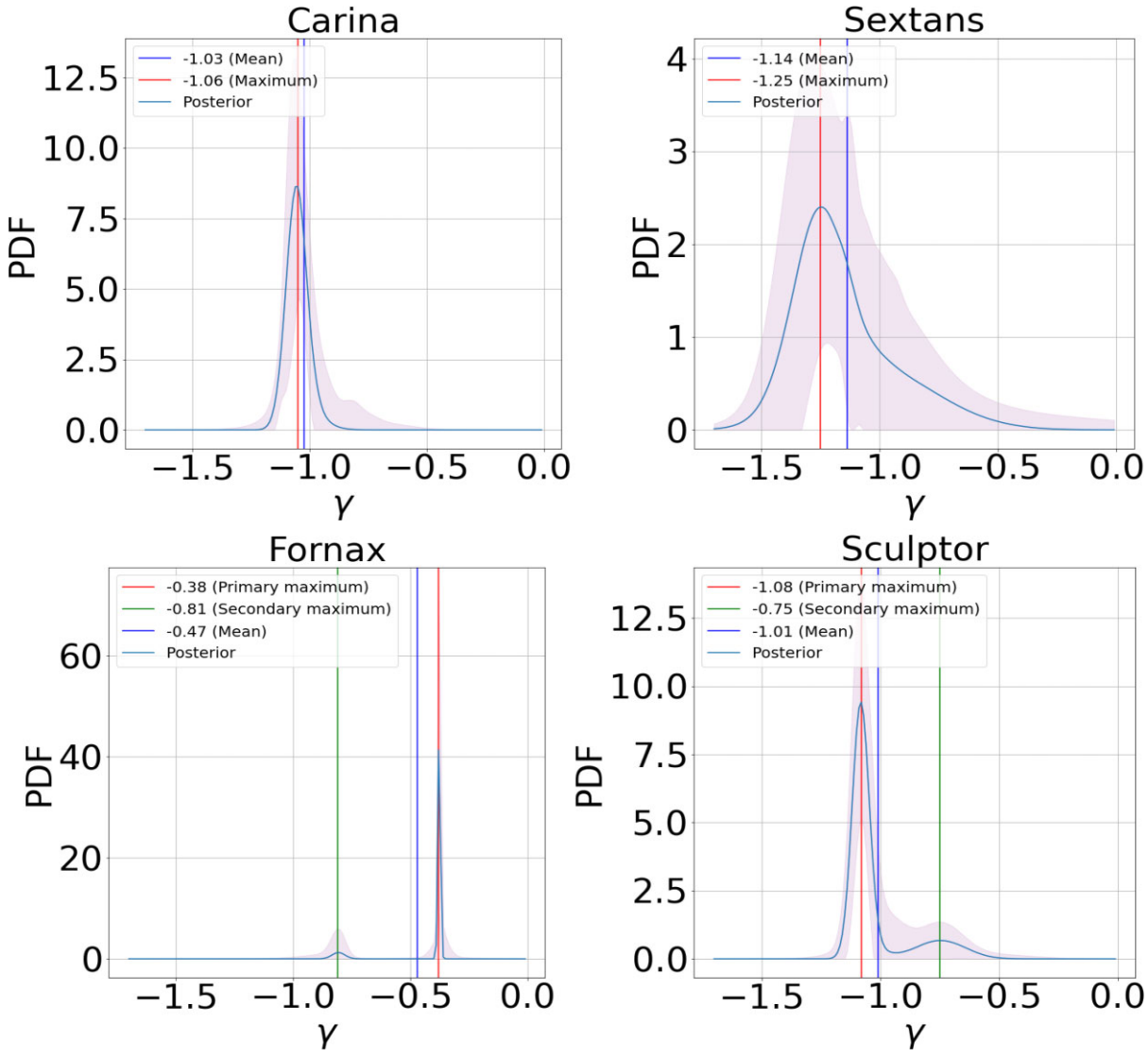

Figure 10. Bayesian posterior distributions in the space of inner slope of DM profiles (slope at 150 pc) predicted by our neural network model for the observed dSphs Carina, Sextans, Sculptor, and Fornax. Shaded regions represent the standard deviation of the posterior values for the MC-Dropout inferences at each slope point. In each panel, the global maximum of the posterior distribution as well as the mean value are indicated, together with primary and secondary peaks when they exist. Fornax has the strongest signature of a central DM density core, while Carina has the strongest signature of having an NFW profile. Sextans is cuspy, though with a large uncertainty, while Sculptor is cuspy with a secondary peak indicating a mild core.

Table 3. Inner slope of the DM profile (at 150 pc) for Carina, Sextans, Sculptor and Fornax galaxies predicted by GRAVSPHERE (γ_{GS}) and our neural network (γ_{NN}), with their 68 per cent confidence intervals (for the neural network posterior, taking the primary maximum as reference). The agreement between the two methods is encouraging.

	γ_{GS}	γ_{NN}
Carina	$-1.23^{+0.39}_{-0.35}$	$-1.06^{+0.05}_{-0.04}$
Sextans	$-0.95^{+0.25}_{-0.25}$	$-1.25^{+0.25}_{-0.09}$
Fornax	$-0.30^{+0.21}_{-0.28}$	$-0.38^{+0.01}_{-0.02}$
Sculptor	$-0.83^{+0.30}_{-0.25}$	$-1.08^{+0.08}_{-0.04}$

showing that the network is properly parametrizing the inner slope of galaxies during training. Interestingly, the four observed galaxies fall into the regions occupied by the simulated ones, which indicates that the model is considering them as data of equivalent nature as the test data. However, the fact that all four are close to the edges of the simulation input parameters could indicate the presence of some features that the model has not found in the simulations. The possible causes of this will be explored in future work employing a larger observational sample.

5 CONCLUSIONS

We present a novel model for determining the slope of the inner density profile of DM haloes with robust uncertainty quantification using machine learning techniques. The goal of this work is to be able to infer such density slopes (γ) by simply using positions and velocities of stars within galaxies. Our method uses mixture density convolutional neural networks with a Gaussian density layer backend to model complex galaxy substructure. We use line-of-sight velocities and positions of stars projected on the sky within simulated dwarf galaxies, employing Kernel Density Estimations (KDEs) to construct

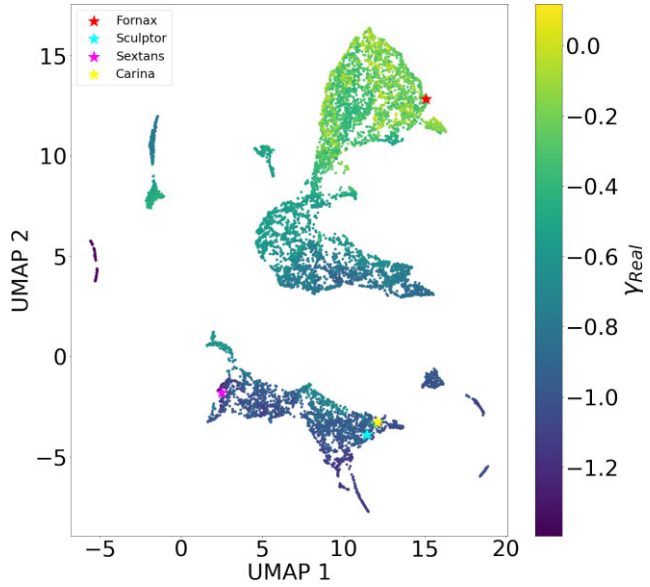

Figure 11. Representation of the parameter space for the test data from simulated galaxies reduced to two-dimension with a UMAP, colour coded by the real expected inner DM slope. Plotted as coloured stars are the locations in the reduced parameter space of Carina, Sextans, Sculptor and Fornax galaxies. Note that different inner slopes occupy different areas of the plot and, importantly, observed dwarf galaxies fall well within the simulation region, indicating that the neural network model is not seeing relevant differences between the simulated data with which we have fed it and the observational data.

continuous 2D PDFs of the distribution of such stars in $\{\hat{R}_{\rm proj}, \hat{v}_{\rm LOS}\}$ and $\{x,y\}$ phase space, which serve as input to our neural network using a double channel architecture (Figs 3 and 4).

We train and evaluate our model using a large set of fully cosmological simulations of dwarf galaxies with halo masses of 10^9 to $10^{11.5}$ $M_\odot$, and stellar masses of 10^5 to $10^{9.5}$ $M_\odot$, from the NIHAO and AURIGA projects (Wang et al. 2015; Dutton et al. 2020; Grand et al. 2017). The use of different physical models employed in these simulations allows us to have a range of density profiles at each particular galaxy mass, including both cores and cusps (Fig. 1). All simulated galaxies and their subsets of stars are randomly projected in several sky planes, to simulate several viewing angles.

The loss function to minimize during the training is the negative logarithmic likelihood of the training sample, defined as a double Gaussian probability distribution, which is the output of our Gaussian density layer backend. This allows a flexible probabilistic representation of the results, which yields accurate and statistically consistent uncertainties. For each galaxy, the neural network outputs a PDF that gives the posterior probability of a certain slope to be the inner slope of the galaxy.

The main results of this work are listed here:

(i) The inner slope of simulated galaxies is predicted with a mean absolute deviation of $\mu_\epsilon = 0.056$ (where the deviation is defined as $\epsilon = \gamma_{\rm Real} - \gamma_{\rm Pred}$, and the predicted inner slope, $\gamma_{\rm Pred}$, is obtained from the mode of the PDFs) and a standard deviation of $\sigma_\epsilon = 0.075$ for the whole sample (Figs 5 and 6).

(ii) 82 per cent (98 per cent) of the galaxies have their inner slope correctly determined within $\pm$ 0.1 (0.3) of their true value (Table 1).

(iii) The posteriors PDFs have a mean standard deviation of $\sigma_{\rm pos} = 0.108$, showing no bias towards more accuracy for cuspy or cored galaxies (Fig. 7).

(iv) While in most cases the output of the model is a single peaked PDF, in $\sim$ 30 per cent of the galaxies some of their projections show a double peak: we demonstrated that this is related to some viewing angles, indicating the importance of properly determining the inclination of galaxies (Figs 8 and 9).

(v) When applied to a set of four observed dSphs, our model recovers their inner slopes yielding values consistent with those obtained with the Jeans modelling based code GravSphere as in Read et al. (2019) (Table 3).

(vi) We found that the Fornax dSph has a strong indication of having a central DM core, Carina and Sextans have cusps (although the latter with a large uncertainty), while Sculptor shows a double peaked PDF indicating that a cusp is preferred, but a core cannot be ruled out (Fig. 10). These results are in agreement with several previously derived inner slopes for these galaxies.

The current architecture could be used as a basis for building models that provide a more complete output, such as a prediction of the full density profile of galaxies. The nature of the neural network allows it to be constantly extended and improved. While we have implemented a network of relatively low complexity, there are a series of interesting possibilities with a further level of sophistication that are worth exploring. For example, the use of normalizing flows may yield to more robust results (Kodi Ramanah et al. 2020) while the use of a 3D convolutional network applied to PDFs defined in the $\{x, y, \hat{v}_{\rm LOS}\}$ phase space has given good results in galaxy cluster masses inference (Kodi Ramanah et al. 2021).

In the future, the architecture of this model could be expanded by including more input data, such as surface brightness profiles or proper motion of stars from missions like *GAIA* (Gaia Collaboration 2021). Furthermore, the inclusion of other spectroscopic samples present in the literature, as well as of those soon to be acquired with upcoming facilities, will certainly be beneficial for this analysis. Adapting the architecture and introducing more information may enable the network to improve accuracy and reduce the range of variability of the results with respect to the angle of observation, an avenue that will be explored in future works.

We have shown that deep learning techniques provide an innovative method for the determination of the inner DM profile in dwarf galaxies, complementary to the use of Jeans and Schwarzschild modelling, achieving great accuracy and offering a complex representation of uncertainties.

Our newly developed neural network method is a promising tool for the study of the mass distribution within dwarf galaxies, which in turn can help discriminate between different models and, in such, constraining the properties of the elusive DM.

ACKNOWLEDGEMENTS

CBB led the project. JEM performed the analysis. JEM and ADC wrote the manuscript. JEM and MHC developed the machine learning architecture. CBB, MHC and ADC supervised the student during the project. AVM and RJJG provided the simulation data. EAG helped with the treatment of AURIGA simulations. GB helped with the selection of the observational data set. All authors provided feedback on the draft.

CB is supported by the Spanish Ministry of Science and Innovation (MICIU/FEDER) through research grant PID2021-122603NB-C22. ADC is supported by a *Junior Leader fellowship* from 'La Caixa' Foundation (ID 100010434), code LCF/BQ/PR20/11770010. She further acknowledges Macquarie University for the hospitality during the preparation of this work as a Honorary Visiting Fellow. MHC

e acknowledge financial support from the State Research Agency (AEIMCINN) of the Spanish Ministry of Science and Innovation under the grant and 'Galaxy Evolution with Artificial Intelligence' with reference PGC2018-100852-A-I00, from the ACIISI, Consejería de Economía, Conocimiento y Empleo del Gobierno de Canarias and the European Regional Development Fund (ERDF) under grant with reference PROID2020010057, and from IAC project P/301802, financed by the Ministry of Science and Innovation, through the State Budget and by the Canary Islands Department of Economy, Knowledge and Employment, through the Regional Budget of the Autonomous Community. GB acknowledges support from the Agencia Estatal de Investigación del Ministerio de Ciencia en Innovación (AEI-MICIN) under grant references PID2020-118778GB-I00/10.13039/501100011033 and grant number CEX2019-000920-S. Fellow RG acknowledges financial support from the Spanish Ministry of Science and Innovation (MICINN) through the Spanish State Research Agency, under the Severo Ochoa Program 2020–2023 (CEX2019-000920-S). Part of this research was carried out on the High Performance Computing resources at New York University Abu Dhabi (UAE).

DATA AVAILABILITY

The data used in this work are available upon reasonable request to the corresponding author and to the PIs of the NIHAO and AURIGA projects.

REFERENCES

Agnello A., Evans N. W., 2012, ApJ, 754, L39
Battaglia G., Helmi A., Tolstoy E., Irwin M., Hill V., Jablonka P., 2008, ApJ, 681, L13
Battaglia G., Taibi S., Thomas G. F., Fritz T. K., 2022, A&A, 657, A54
Benítez-Llambay A., Frenk C. S., Ludlow A. D., Navarro J. F., 2019, MNRAS, 488, 2387
Binney J., Mamon G. A., 1982, MNRAS, 200, 361
Breddels M. A., Helmi A., 2013, A&A, 558, A35
Breddels M. A., Helmi A., van den Bosch R. C. E., van de Ven G., Battaglia G., 2013, MNRAS, 433, 3173
Brook C. B., Di Cintio A., 2015, MNRAS, 450, 3920
Bullock J. S., Boylan-Kolchin M., 2017, ARA&A, 55, 343
Cappellari M. et al., 2006, MNRAS, 366, 1126
Chan T. K., Kereš D., Oñorbe J., Hopkins P. F., Muratov A. L., Faucher-Giguère C.-A., Quataert E., 2015, MNRAS, 454, 2981
Collins M. L. M. et al., 2021, MNRAS, 505, 5686
de Blok W. J. G., Walter F., Brinks E., Trachternach C., Oh S.-H., Kennicutt R. C., Jr, 2008, AJ, 136, 2648
Di Cintio A., Brook C. B., Macciò A. V., Stinson G. S., Knebe A., Dutton A. A., Wadsley J., 2014a, MNRAS, 437, 415
Di Cintio A., Brook C. B., Dutton A. A., Macciò A. V., Stinson G. S., Knebe A., 2014b, MNRAS, 441, 2986
Dutton A. A., Buck T., Macciò A. V., Dixon K. L., Blank M., Obreja A., 2020, MNRAS, 499, 2648
Gaia Collaboration, 2021, A&A, 649, A1
Gal Y., Ghahramani Z., 2015, Proceedings of The 33rd International Conference on Machine Learning
Geha M. C., Guhathakurta P., Rich R. M., Cooper M. C., 2006, AJ, 131, 332
Gentile G., Salucci P., Klein U., Vergani D., Kalberla P., 2004, MNRAS, 351, 903
Ghosh A. et al., 2022, ApJ, 935, 2
Gnedin O. Y., Zhao H., 2002, MNRAS, 333, 299
Goerdt T., Moore B., Read J. I., Stadel J., Zemp M., 2006, MNRAS, 368, 1073
Governato F. et al., 2010, Nature, 463, 203
Grand R. J. J. et al., 2017, MNRAS, 467, 179
Hayashi K., Chiba M., Ishiyama T., 2020, ApJ, 904, 45
Ho M., Rau M. M., Ntampaka M., Farahi A., Trac H., Póczos B., 2019, ApJ, 887, 25
Jaffe W., 1983, MNRAS, 202, 995
Kaplinghat M., Tulin S., Yu H.-B., 2016, Phys. Rev. Lett., 116, 041302
Kingma D. P., Ba J., 2014, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9
Kleyna J. T., Wilkinson M. I., Evans N. W., Gilmore G., 2001, ApJ, 563, L115
Kodi Ramanah D., Wojtak R., Ansari Z., Gall C., Hjorth J., 2020, MNRAS, 499, 1985
Kodi Ramanah D., Wojtak R., Arendse N., 2021, MNRAS, 501, 4080
Kowalczyk K., Łokas E. L., Valluri M., 2017, MNRAS, 470, 3959
LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436
Lelli F., McGaugh S. S., Schombert J. M., 2016, AJ, 152, 157
Macciò A. V., Crespi S., Blank M., Kang X., 2020, MNRAS, 495, L46
McInnes L., Healy J., Melville J., 2018, Journal of Open Source Software, 3, 861
Merritt D., Graham A. W., Moore B., Diemand J., Terzić B., 2006, AJ, 132, 2685
Moore B., 1994, Nature, 370, 629
Navarro J. F., Eke V. R., Frenk C. S., 1996, MNRAS, 283, L72
Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563
Pascale R., Posti L., Nipoti C., Binney J., 2018, MNRAS, 480, 927
Planck Collaboration XIII, 2016, A&A, 594, A13
Pontzen A., Governato F., 2012, MNRAS, 421, 3464
Read J. I., Wilkinson M. I., Evans N. W., Gilmore G., Kleyna J. T., 2006, MNRAS, 367, 387
Read J. I., Walker M. G., Steger P., 2019, MNRAS, 484, 1401
Richardson T., Fairbairn M., 2014, MNRAS, 441, 1584
Schneider A., Trujillo-Gomez S., Papastergis E., Reed D., Lake G., 2017, MNRAS, 470, 1542
Schwarzschild M., 1979, ApJ, 232, 236
Scott D., 1992, Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, Inc., New York
Sheather S. J., 2004, Stat. Sci., 19, 588
Simon J. D., Bolatto A. D., Leroy A., Blitz L., Gates E. L., 2005, ApJ, 621, 757
Spergel D. N., Steinhardt P. J., 2000, Phys. Rev. Lett., 84, 3760
Stinson G., Seth A., Katz N., Wadsley J., Governato F., Quinn T., 2006, MNRAS, 373, 1074
Tollet E. et al., 2016, MNRAS, 456, 3542
van den Bosch R. C. E., de Zeeuw P. T., 2010, MNRAS, 401, 1770
van der Marel R. P., 1994, MNRAS, 270, 271
Walker M. G., Peñarrubia J., 2011, ApJ, 742, 20
Walker M. G., Mateo M., Olszewski E. W., 2009, AJ, 137, 3100
Wang L., Dutton A. A., Stinson G. S., Macciò A. V., Penzo C., Kang X., Keller B. W., Wadsley J., 2015, MNRAS, 454, 83
Zhu L., van de Ven G., Watkins L. L., Posti L., 2016, MNRAS, 463, 1117

APPENDIX: DETAILS ON THE NEURAL NETWORK MODEL

A neural network can be formally described as a trainable and flexible approximation of a model $\mathbb{M} : d \rightarrow t$. The networks maps an input data d to a prediction $\bar{t}$ of the target t. This network is parametrized by a set of trainable weights and a set of hyperparameters. The weights are iteratively optimized during training to minimize a particular loss function, which provides a measure of how close the network prediction $\bar{t}$ is to the target t.

In this work, we use convolutional neural networks (CNNs), a class of deep neural networks (DNNs), to construct a neural network in which the input data d are the two PDFs described in Section 2.2.2, and the targets t are the inner slopes of the galaxy subsets associated with those two PDFs. We then make a mixture density convolutional

neural network (MDCNN) by embedding a mixture density layer within the CNN as the last layer.

A1 Deep neural networks

Any neural network is conformed by a set of neuron layers, defined by the following function:

$$f(\boldsymbol{x}) = g(\boldsymbol{W} \cdot \boldsymbol{x} + b), \tag{A1}$$

where $\boldsymbol{x}$ is the input of the layer, $\boldsymbol{W}$ is the weight matrix (which each element being the weight of each element of the vector $\boldsymbol{x}$), and $\boldsymbol{b}$ is a vector called the bias parameter of the layer. $g(\boldsymbol{z})$ is known as the activation function, which purpose is to break the linearity between the input and the output of the neuron.

A DNN is a neural network conformed by more than one neuron layer. The layers between the input layer (the layer that takes as inputs the input data of the neural network) and the output layer (the layer that gives as output the outputs of the neural network) are called hidden layers.

A feed-forward DNN is a DNN where the neuron layers are evaluated in sequence, passing information from layer to layer without recurrence, which means we can describe the output $\boldsymbol{h}^{(l)}$ of the lth layer as

$$\boldsymbol{h}^{(l)} = g(\boldsymbol{W}^{(l)} \cdot \boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}). \tag{A2}$$

The training of the model is done by optimizing the weight matrices $\boldsymbol{W}^{(l)}$. A model is trained on a set of input data d for which the targets t are known iteratively. In each iteration, the network performance (the similarity between the outputs $\bar{t}$ and the targets t) is evaluated using a loss function, and the weights are actualized to minimize that function by an optimization algorithm. When the loss function stops decreasing and converges to a certain value, the network is said to be optimized. The performance evaluation is done, then, on a set of independent data the model has not seen during training.

A2 Convolutional neural networks

CNNs are a particular type of DNNs especially suited for problems where spatially correlated information is crucial. The main feature of a CNN is the presence of convolutional layers, which are constructed in a way that restrict neurons in one layer to receive information only from within a small neighbourhood of the previous layer. This allows neurons to extract simple features from subsets of the previous layer, forming higher order features in subsequent layers.

A convolutional layer is designed as follows: A convolutional kernel, commonly referred to as a filter, of a given size, encoding a set of neurons, is applied to each pixel (in the case of 2D images as inputs) of the input image and its vicinity, as it scans through the whole region. A given pixel in a specific layer is only a function of the pixels in the preceding layer which are enclosed within the window defined by the kernel, known as the receptive field of the layer. This yields a feature map that encodes high values in the pixels which match the pattern encoded in the weights and biases of the corresponding neurons in the convolutional kernel, which are optimized during training (Kodi Ramanah et al. 2021).

A convolutional layer may be described as a linear operation with the discrete convolution implemented via matrix multiplication. In terms of equation (A2):

$$\boldsymbol{h}_j^{(l)} = g\left(\sum_{i \in M_j} \boldsymbol{h}^{(l-1)} \times k_{ij}^{(l)} + \boldsymbol{b}_j^{(l)}\right), \tag{A3}$$

where k is the convolutional kernel (the filter) and M_j is the receptive field of the neuron j. One convolutional layer could have multiple filters, which repeat this operation with different kernels, constructing many feature maps per layer, known as channels.

The receptive field is usually defined by the dimensions of the filter, the stride and the existence or not of padding. The application of the filter can be described as a process of sliding it over the input image of the convolutional layer. We call *stride* to the number and direction of pixels you move the filter at each step, and *padding* to the addition of empty pixels around the edges with the purpose of alleviating information loss around the edges.

Usually, a CNN is a series of pairs of convolutional layers followed by a pooling layer as a subsampling or dimensionality reduction step, a process which will reduce the initial input image to a compact representation of features. Then, that representation is reshaped as a vector, which is subsequently passed to a sequence of dense layers (LeCun, Bengio & Hinton 2015). This design allows the neural network to autonomously extract meaningful spatial features from the input image. The stack of several convolutional layers builds an internal hierarchical representation of features encoding the most relevant information from the input image. Stacking subsequent convolutional layers naturally strengthens the sensitivity of the most internal layers to features on increasingly larger scales, because the size of the receptive field becomes larger as we go deeper in the CNN.

A3 Mixture density neural networks

A MDNN is a network with layers whose outputs follow a multidimension probability distribution, called mixture density layers. This layers take as inputs n nodes, with n being the number of parameters in the desired distribution, transform their values to respect the parameter constraints of the distribution and interpret them as those parameters to construct it. When used as the last layer of the network, it allows join optimization of the features from the DNN together with a bayesian posterior backend, combining the advantages of deep feature extraction with probabilistic representation of the results.

A4 Details on the architecture

The schematic view of the architecture used in this work can be seen in Fig. 4

The convolutional sequences are constructed using pairs of convolutional and pooling layers, followed by a dropout layer. The pooling layers downsample their input along its spatial dimension using the Max Pooling method, which takes the maximum value over a certain input window for each channel. The dropout layers randomly set input units to 0 with a certain frequency called the dropout rate, and scales the rest such that the sum over all inputs is unchanged. This is done to prevent overfitting during training. The joint sequence of dense layer stars with a normalization layer that applies batch normalization to the nodes coming from the previous convolutional sequences. This normalization maintains the mean of the output close to 0 and its standard deviation close to 1. The final mixture density layer gives a probability distribution defined in the range of possible inner slopes for a given galaxy subset and transform our double channel CNN into an MDCNN. This probability distribution is understood as a posterior under the prior distribution of inner slopes with which the network has trained, which allows us to evaluate the uncertainty of the individual predictions of the model.

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