A&A, 710, A67 (2026)
https://doi.org/10.1051/0004-6361/202557894

Astronomy & Astrophysics

Beyond single tracers: Convolutional neural network-based inference of galaxy mass profiles from combined gas and stellar kinematics

J. Expósito-Márquez[1,2,⋆], A. Di Cintio[2,1], C. Brook[2,1], J. Sarrato-Alós[1,2], and A. V. Macciò[3,4,5]

[1] Instituto de Astrofísica de Canarias (IAC), Calle Via Láctea s/n, 38205 La Laguna, Tenerife, Spain
[2] Universidad de La Laguna, Avda. Astrofísico Fco. Sánchez s/n, 38206 La Laguna, Tenerife, Spain
[3] New York University Abu Dhabi, PO Box 129188, Abu Dhabi, United Arab Emirates
[4] Center for Astro, Particle and Planetary Physics, New York University, Abu Dhabi, United Arab Emirates
[5] Max Planck Institute für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

Received 29 October 2025 / Accepted 17 February 2026

ABSTRACT

Aims. We investigated whether combining gas and stellar kinematic maps provides measurable advantages in recovering galaxy mass profiles compared to using single-component maps alone. We used deep learning models to leverage this joint information.
Methods. We developed a probabilistic convolutional neural network (CNN) framework trained and tested on mock galaxy kinematic maps from multiple cosmological simulation suites. Our model was trained on gas-only, stars-only, and combined gas and stellar velocity maps, thus allowing the direct comparison of performance across tracers. To assess robustness, we included simulations with differing feedback models and galaxy properties.
Results. Combining gas and stellar maps reduces the dispersion in the inferred mass profiles by up to a factor of ∼1.5 compared to models using either tracer independently. The CNN architecture effectively captures complementary information from the two components. However, we find limitations in generalising between simulation suites, with reduced performance when applying models trained on one suite to galaxies from another.

Key words. dark matter

1. Introduction

The Lambda cold dark matter (ΛCDM) model provides a self-consistent framework for predicting galaxy formation from the initial density fluctuations observed in the cosmic microwave background (CMB), primarily through the influence of dark matter (DM), whose nature remains unknown. It successfully accounts for key properties of galaxies, including their abundance, clustering, morphologies, and evolution (e.g. Vogelsberger et al. 2014; Schaye et al. 2015). However, observations on submegaparsec scales present challenges to the model, raising the question of whether these discrepancies arise from baryonic physics or non-standard DM properties, or if they require a revision of the standard cosmological paradigm (Bullock & Boylan-Kolchin 2017).

Understanding how mass is distributed within galaxies is crucial for testing the CDM paradigm. Analysis of the rotation velocity of gas in low surface brightness galaxies, for example, allows us to derive and fit their underlying DM distribution (e.g. Moore 1994; Gentile et al. 2004; de Blok et al. 2008; Lelli et al. 2016; Katz et al. 2017). On the other hand, in pressure-supported galaxies which are devoid of gas, such as the dwarf spheroidal galaxies (dSphs) found within the Local Group, the kinematic information on which dynamical modelling relies 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 modelling (e.g. van der Marel 1994; Kleyna et al. 2001; Battaglia et al. 2008; Read et al. 2019; Collins et al. 2021; Bañares-Hernández et al. 2026), Schwarzschild modelling (e.g. Schwarzschild 1979; Cappellari et al. 2006; van den Bosch & de Zeeuw 2010; Breddels et al. 2013; Breddels & Helmi 2013), or distribution-function modelling techniques (e.g. Amorisco & Evans 2012; Binney & Vasiliev 2023; Croce et al. 2023; Arroyo-Polonio et al. 2025).

Despite their strengths, these dynamical models face certain limitations. In Jeans modelling, uncertainties in the stellar velocity anisotropy lead to a well-known degeneracy with the underlying mass profile (Binney & Mamon 1982), although significant progress has been made in the last years to break this degeneracy by introducing information about the shape of the velocity distribution function Read et al. (2021). Schwarzschild modelling, while more flexible, is sensitive to the quality and completeness of the available data (Kowalczyk et al. 2017). The three approaches are also affected by projection effects, which complicate the reconstruction of the galaxy's three-dimensional structure, and the models typically require significant computational resources and careful treatment of system-specific properties.

In contrast with dynamical modelling, several authors (e.g. Walker et al. 2009; Wolf et al. 2010; Amorisco & Evans 2012; Campbell et al. 2017; Errani et al. 2018) have developed simple formulae to estimate the dynamical mass of galaxies enclosed within specific radii, where they have found velocity anisotropy and/or other factors to introduce minimum uncertainty. These

⋆ Corresponding author: julenexp@iac.es

estimators rely on the line-of-sight velocity dispersion of the stellar component of the galaxies and on the half-light radius, and they aim to obtain an unbiased mass estimation that holds for a broad range of galaxy properties and that is minimally affected by projection limitations.

More recently, the combination of machine learning techniques and hydrodynamical simulations of galaxies has proven to be an effective tool for analysing complex data and uncovering underlying patterns. These methods have been applied to various astrophysical problems, such as dynamical mass estimation and density profile determination using line-of-sight data. Ho et al. (2019) applied a convolutional neural network (CNN) model based on probability distribution functions of positions and velocities of galaxies to estimate the dynamical masses of galaxy clusters. In addition, Nguyen et al. (2023) developed a graph neural network capable of accurately recovering the DM density profiles of mock spherical dwarf galaxies in dynamical equilibrium. Expósito-Márquez et al. (2023) employed 2D probability distribution functions of projected stellar positions and kinematics as input to a CNN to estimate the inner slopes (150 pc from the centre) of DM density profiles in dwarf galaxies. This approach resulted in a model that was able to differentiate between cusps and cores in cosmological hydrodynamical simulations and, when applied to several Milky Way dSphs, was able to recover an inner slope similar to that in previous studies (Brook & Di Cintio 2015; Read et al. 2019). Sarrato-Alós et al. (2025) used a similar approach to develop a CNN capable of recovering the full dynamical mass profiles of dispersion-supported galaxies with great success, although encountering a generalisation problem when cross-testing the model within different simulation models. de los Rios et al. (2023) and de los Rios et al. (2025) used a similar methodology to analyse mock photometry and interferometry images from cosmological hydrodynamical simulations, successfully inferring their dynamical mass profiles.

Given the success of such studies, it is natural to ask whether the uncertainty and degeneracies in the inference of DM distributions could be further reduced in galaxies that contain sufficient gas to allow for rotation curve analysis by simultaneously incorporating the stellar kinematics and the dynamical information provided by the gas component. This approach was successfully tested on a study of the isolated dwarf irregular galaxy WLM by Leung et al. (2021), which presented significant improvement in constraining the mass distribution with multi-tracer models.

For this article we investigated the potential advantages of using deep learning methods to integrate both gas and stellar velocity fields, a capability that allows a more complete exploitation of available kinematic information. We extended a probabilistic convolutional neural network, previously developed for stellar kinematic data alone (Expósito-Márquez et al. 2023; Sarrato-Alós et al. 2025), to incorporate inputs from realistically constructed mock HI observation maps as well. The model was trained and tested on a large suite of high-resolution cosmological hydrodynamical simulations, with systematic comparisons made between models trained on gas-only, stars-only, and combined kinematic maps. We also evaluated the generalisation ability of the network across different simulation suites in order to assess the model's robustness to variations in galaxy formation physics. We find that our model further reduces the scatter in the estimation of dynamical masses compared to the results presented in Sarrato-Alós et al. (2025), which already demonstrated improved performance over classical mass estimators.

This paper is structured as follows. In Sect. 2, we describe the galaxy simulations used to train and evaluate the CNN, from the NIHAO (Wang et al. 2015) and AURIGA (Grand et al. 2017) projects. Sect. 3 outlines the CNN architectures developed for stellar-only, gas-only, and combined input data, detailing both the network structure and the input/output configurations. In Sect. 4 we present the main results of our study. Sect. 4.1 focuses on a comprehensive analysis of model performance using the NIHAO galaxy sample, examining potential biases across different regions of parameter space and comparing the performance of models with varying input types. In Sect. 4.2, we evaluate the model's generalisation capability by training and testing across different simulation suites, and we discuss the implications for applying the model to real observational data. We summarise our conclusions in Sect. 5.

2. Simulation dataset

We use a comprehensive dataset of realistic galaxy formation simulations to test the ability of a CNN to recover mass profiles. From these simulations, we extract stellar and gas data to serve as input to the network, and compute the corresponding enclosed mass profiles as target outputs. To ensure physical realism, we used cosmological hydrodynamical simulations rather than idealised ones, prioritising a more representative dataset over a more easily generated but less realistic alternative.

Our goal is to select a dataset that spans the parameter space of galaxy properties as broadly as possible, while still providing a sufficiently large number of galaxies for robust training and evaluation. Simulations from the NIHAO project (Wang et al. 2015) satisfy these criteria. Even so, as demonstrated by Sarrato-Alós et al. (2025), a model trained and tested on a single simulation model could lead to strong biases, so to assess the generalizability of our model, we additionally incorporate data from the AURIGA dataset (Grand et al. 2017).

2.1. The NIHAO project

The NIHAO project consists of a series of cosmological hydrodynamical zoom-in simulations run with the parameters of Planck Collaboration XIII (2016): $H_0 = 100\,h$ km s^{-1} Mpc^{-1} with $h = 0.671$, $\Omega_m = 0.3175$, $\Omega_\Lambda = 0.6824$, $\Omega_b = 0.049$, and $\sigma_8 = 0.8344$.

The galaxy formation model incorporates ultraviolet heating, ionisation, and metal-line cooling (Shen et al. 2010). Star formation and feedback follow the prescription adopted in the Making Galaxies In a Cosmological Context (MaGICC) simulations (Stinson et al. 2013), which successfully reproduce galaxy scaling relations over a wide mass range (Brook et al. 2012). Star formation occurs in regions exceeding a density threshold of $n_{th} > 10.3$ cm^{-3}, assuming a Chabrier (2003) initial mass function. Stellar energy feedback into the interstellar medium (ISM) is implemented through a combination of blast-wave supernova feedback (Stinson et al. 2006) and early stellar feedback from massive stars. The adopted particle masses and force softenings resolve the mass profile to below 1% of the virial radius, ensuring that galaxy half-light radii are well resolved.

2.2. The AURIGA project

The cosmological parameters used in the simulations from the AURIGA project also come from Planck Collaboration XIII (2016). The AURIGA galaxy formation model includes magnetohydrodynamics, primordial, and metal-line cooling with self-shielding, stellar feedback, and thermal feedback from black holes in both radio and quasar accretion modes. Star formation is implemented according to the Kennicutt–Schmidt law

(Schmidt 1959), adopting a Chabrier (2003) initial mass function. The model resolves star-forming regions and feedback processes at high spatial resolution, enabling the formation of thin and thick disc components consistent with those observed in the Milky Way (Grand et al. 2017). The resulting galaxies reproduce realistic rotation curves, star formation histories, and structural properties, and follow key empirical relations, including the stellar-to-halo mass relation and the Tully–Fisher relation.

2.3. Galaxy selection

Our dataset consists of 6158 zoom-in cosmological simulations from the NIHAO project and 956 from the AURIGA project. In addition to the fiducial NIHAO simulations, we included variants featuring different star formation density thresholds $n_{\rm th}$ (Dutton et al. 2020), simulations incorporating black hole physics (Blank et al. 2019), and high-resolution re-runs of six NIHAO halos (Buck et al. 2019).

We selected all galaxies, satellites included, with stellar masses in the range of $M_* = 10^{5.5}$–10^{11} $M_\odot$ containing at least 100 stellar and gas particles and a high-resolution particle mass fraction exceeding 95% with respect to low-resolution particles at the edges of the box. To further enhance the dataset quality, we visually inspected all selected galaxies and manually excluded ongoing mergers and severely disrupted systems. We refer to this dataset as the full set.

We also constructed a limited set with the further constraints of cold gas masses of $M_{\rm cold} > 10^{7.5}$ $M_\odot$ and $M_{\rm cold}/M_* > 0.1$, where we used the mass of HI as a proxy for the cold gas, thus reducing the number of galaxies to 5001 from NIHAO and 907 from AURIGA. This set ensures that the galaxies have a relevant amount of cold gas. We refer to this dataset as the cold gas set.

3. Neural network

In this work, we tackle the problem of inferring the dark matter content underlying the kinematic, dynamical, and photometric properties of galaxies. This constitutes a complex pattern recognition task, well-suited to modern deep learning approaches. In particular, convolutional neural networks (CNNs) are an effective tool for capturing and modelling the structured information present in such multi-dimensional datasets. We used the Python package TENSORFLOW (Abadi et al. 2015) to construct a CNN based on the one used in Expósito-Márquez et al. (2023) and Sarrato-Alós et al. (2025).

3.1. Input data

To construct the inputs of our network, we projected all our galaxies on the plane of the sky in 12 different orientations with increasing inclination. For each of these projections we constructed two maps with star information (Sect. 3.1.1) and three maps with cold gas information (Sect. 3.1.2).

3.1.1. Star data

Typically, the number of stars for which spectroscopic data are available for LG dwarf galaxies is of 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, we divided each simulated galaxy's complete sample of stars into several subsets, each made of randomly selected stars, and we use just one of those subsets for each projection of our dataset. 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 100 stars.

These projected stars are defined by their position $(x_{\rm proj}, y_{\rm proj})$ and their line-of-sight velocity $v_{\rm LOS}$; we used this information to construct the images with star information that function as inputs of our CNN. The generated images 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. We constructed two maps:

- A PDF sampled at 64x64 points with the distribution of stars in {x,y} phase space, between -0.5 $R_{\rm hl}$ and 0.5 $R_{\rm hl}$ in each coordinate, in the reference system where $(x, y) = (0, 0)$ is the centre of the galaxy. $R_{\rm hl}$ is the projected half-stellar count radius of the galaxy, which we use as a proxy for the half-light radius.
- A PDF sampled at 64×64 points with the distribution of stars in $\{\hat{R}_{\rm proj}, \hat{v}_{\rm LOS}\}$ phase space, where $\hat{R}_{\rm proj} = \sqrt{x^2 + y^2}/{\rm R_{hl}}$ is the radial position normalised by the half-light radius and $\hat{v}_{\rm LOS} = v_{\rm LOS}/P_{98\%}$ is the line-of-sight velocity normalised by the 98% 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.

We note that the adopted spatial limits effectively exclude stellar data beyond these regions. During the testing phase, multiple boundary conditions and normalisation strategies were evaluated. The limits used in this study correspond to those that yielded the best overall performance on the current dataset. A plausible explanation is that including such data in the input PDFs may dilute the more informative signal from stars located nearer to the galaxy centre, where the gravitational potential is most tightly constrained. In Fig. 1 we show an example of the star maps for an edge-on galaxy from NIHAO.

3.1.2. Gas data

We generated mock HI observations using the MARTINI code, described in Oman et al. (2019). This tool enables the production of synthetic, spatially resolved HI line data cubes directly from hydrodynamical simulation snapshots. It offers a comprehensive suite of features, including spectral modelling and the incorporation of observational effects such as noise contamination and beam convolution.

For the same galaxy projections used to create the star maps as explained in Sect. 3.1.1, we used MARTINI to create a spectral data cube of 64 channels with a spectral resolution of 5 km/s. The input for MARTINI consists of the position and average velocities of the gas particles, their HI component mass, their temperature, and their softening length, along with the distance to the observer, fixed at 1 Mpc as a typical distance for LG galaxies (Walker 2003). We set the space covered by the HI line data cubes from the centre of the galaxy as double the radius where the HI superficial density drops below 1 $M_\odot$ pc^{-2}, $R_{\Sigma_{\rm HI} < 1\ M_\odot {\rm pc}^{-2}}$.

To reduce dimensionality while preserving essential information, we extracted the first three statistical moments of the HI emission:

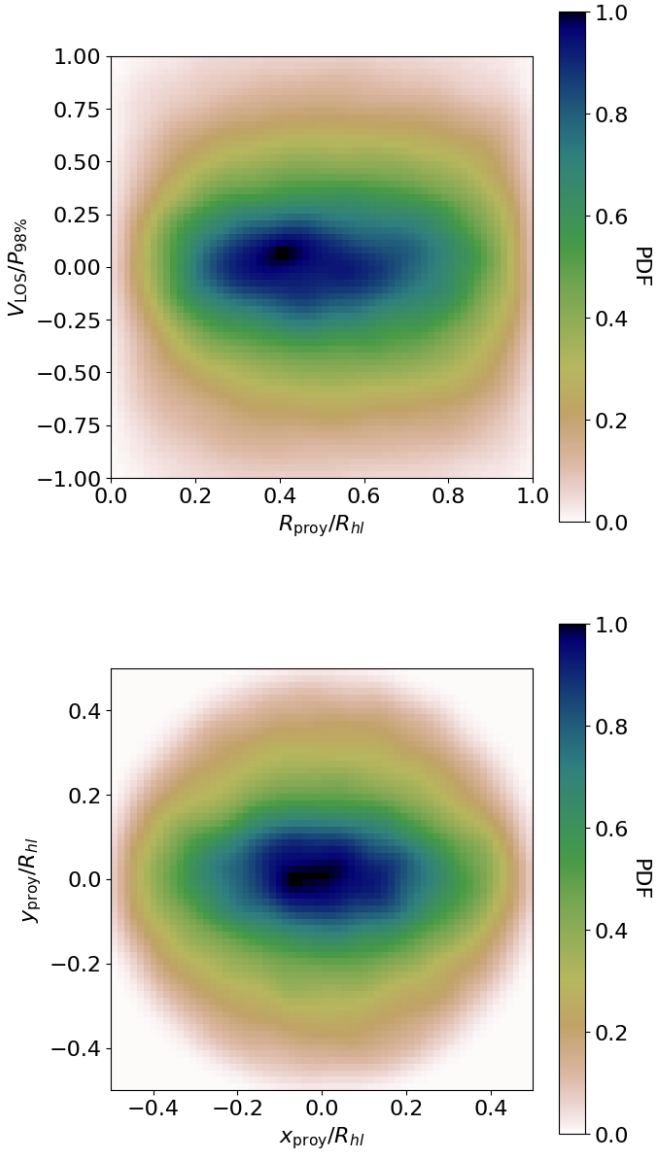

Fig. 1. Star information input maps for a single edge-on galaxy from the NIHAO sample of $M_* = 10^{10.68}\ M_\odot$. Top: PDF in the $\{\hat{R}_{proj}, \hat{v}_{LOS}\}$ phase space. Bottom: PDF in the {x,y} phase space. Both obtained following the procedure described in Sect. 3.1.1.

- a 64×64 grid with a side length of 4 $R_{\Sigma_{HI} < 1\ M_\odot pc^{-2}}$, with the line-of-sight integrated intensity;
- a 64×64 grid with the average line-of-sight velocity covering the same physical space;
- a 64×64 grid with the line-of-sight velocity dispersion covering the same physical space.

To ensure a sufficient signal-to-noise ratio for the first and second moments, we applied a mask to exclude regions with a column density below $10^{19.5}$ atoms cm^{-2}, as recommended by Oman et al. (2019). In Fig. 2, we show an example of the gas maps for different inclinations of a NIHAO galaxy.

3.2. Architecture

The architecture of our neural network is shown in Fig. 3. The CNN is organised into five parallel branches, each corresponding to one of the input map types introduced in Sects. 3.1.1 and 3.1.2. This design allows the network to learn feature representations that are specific to the stellar and gas components before they are combined. Within each branch, the input is processed by two convolutional blocks, each consisting of a convolutional layer followed by max-pooling, which extracts spatial features and compresses the information into progressively higher-level representations. To reduce overfitting and improve the ability to generalise across galaxies with different morphologies and inclinations, dropout layers are applied after these blocks. The outputs of the five branches are then flattened into one-dimensional vectors and concatenated into a single feature representation. This combined vector is passed through a sequence of fully connected layers that capture correlations across the stellar and gas channels while gradually reducing dimensionality. The final output of the network consists of N values corresponding to the predicted enclosed mass profile. In this study we used $N = 10$, with points distributed between 0.6 R_{hl} and 2.4 R_{hl}. Dropout layers were also included between selected dense layers to improve training stability.

To provide a scale to the normalised input data, we expanded this representation with a set of global galaxy parameters, concatenated to the output of the first dropout layer after the concatenation. These include the projected half-light radius R_{hl} and the line-of-sight velocity dispersion σ_v. For the gas component, we added both structural and dynamical properties: the radius at which the HI surface density falls below 1 $M_\odot$ pc^{-2} ($R_{\Sigma_{HI}<1M_\odot pc^{-2}}$), the rotational velocity of the cold gas at this radius (v_{cold}), and the mean cold gas velocity dispersion within the same region (σ_{cold}).

A central feature of this design is its modularity. By selectively activating input branches and global parameters, the model can be trained in several configurations: using only stellar data, only gas data, or both combined. This flexibility is crucial for our broader goal of testing how different tracers contribute to the accuracy and robustness of dynamical mass inferences.

To produce reliable uncertainty estimates, we extended the CNN with a normalising flow. The 32-dimensional feature vector from the penultimate dense layer serves as input to a normalising flow model (Papamakarios et al. 2021), implemented with the ltu-ili Python package (Learning the Universe Implicit Likelihood Inference; Ho et al. 2024). We adopted a masked autoregressive flow (MAF) (Papamakarios et al. 2018), which learns a set of invertible transformations conditioned on the CNN features, mapping a simple Gaussian distribution into the posterior over enclosed masses. The resulting model outputs an N-dimensional joint probability density function (PDF) that captures correlations between the mass estimates at different radii. This probabilistic approach provides a principled way to assess uncertainties, ensuring that the predictions are not only accurate, but also interpretable. In the context of our study, this capability is essential for identifying systematic biases, evaluating generalisation across simulation suites, and ultimately guiding the application of the model to real observational data.

4. Results

4.1. NIHAO results

We defined three different models for our network based on the number of channels and the given information.

- Gas model: the inputs are the three gas channels described in Sect. 3.2, the radius where the HI superficial density drops below 1 $M_\odot$ pc^{-2} ($R_{\Sigma_{HI} < 1\ M_\odot pc^{-2}}$), the velocity of the gas at that radius, and the mean velocity dispersion of the gas inside that radius.

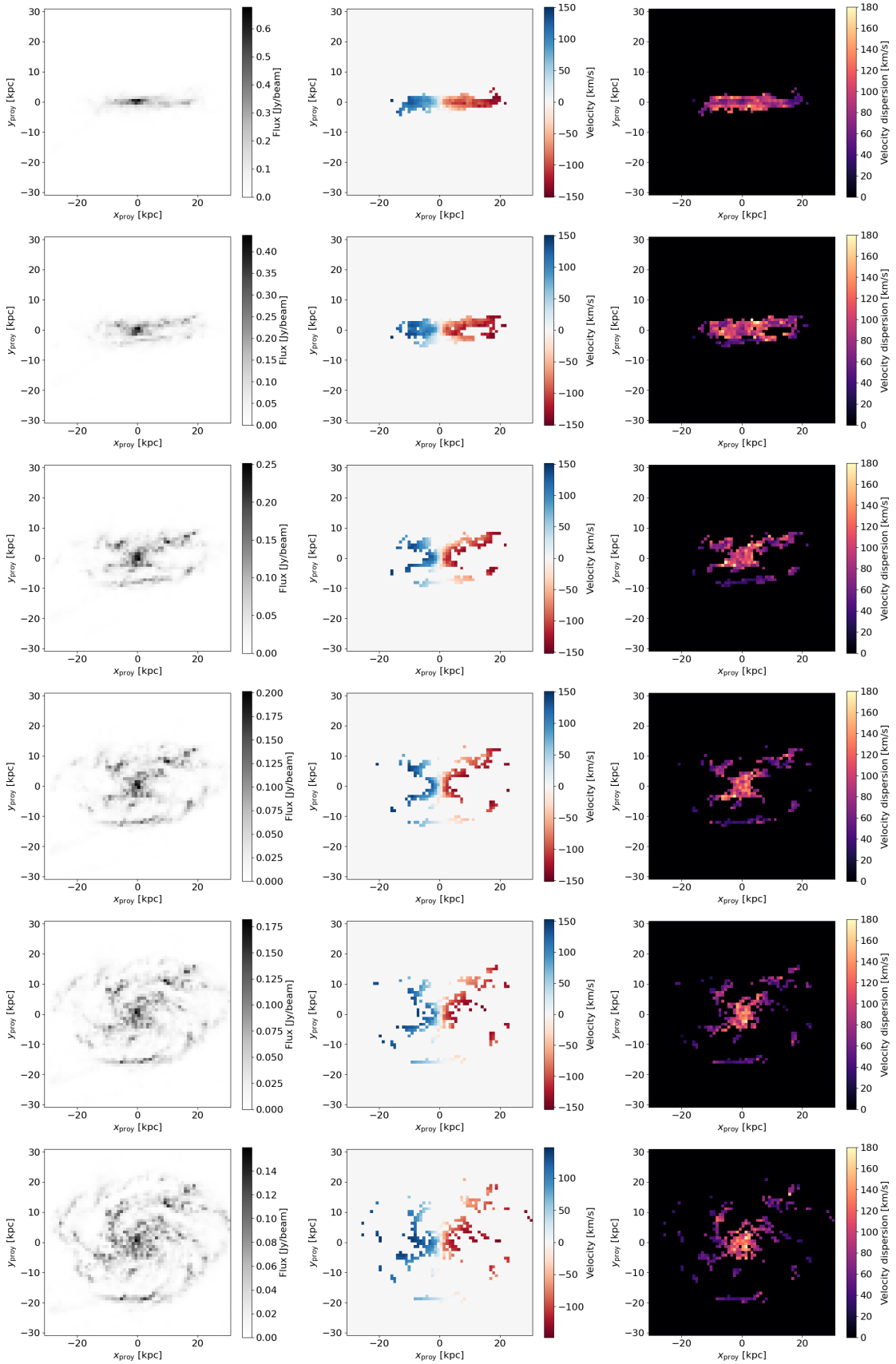

Fig. 2. Gas information input maps for a single galaxy from the NIHAO sample of $M_* = 10^{10.68}$ $M_\odot$ at increasing inclinations from edge-on to 60 degrees. From left to right: HI intensity, average line-of-sight velocity, and velocity dispersion maps, obtained following the procedure described in Sect. 3.1.2.

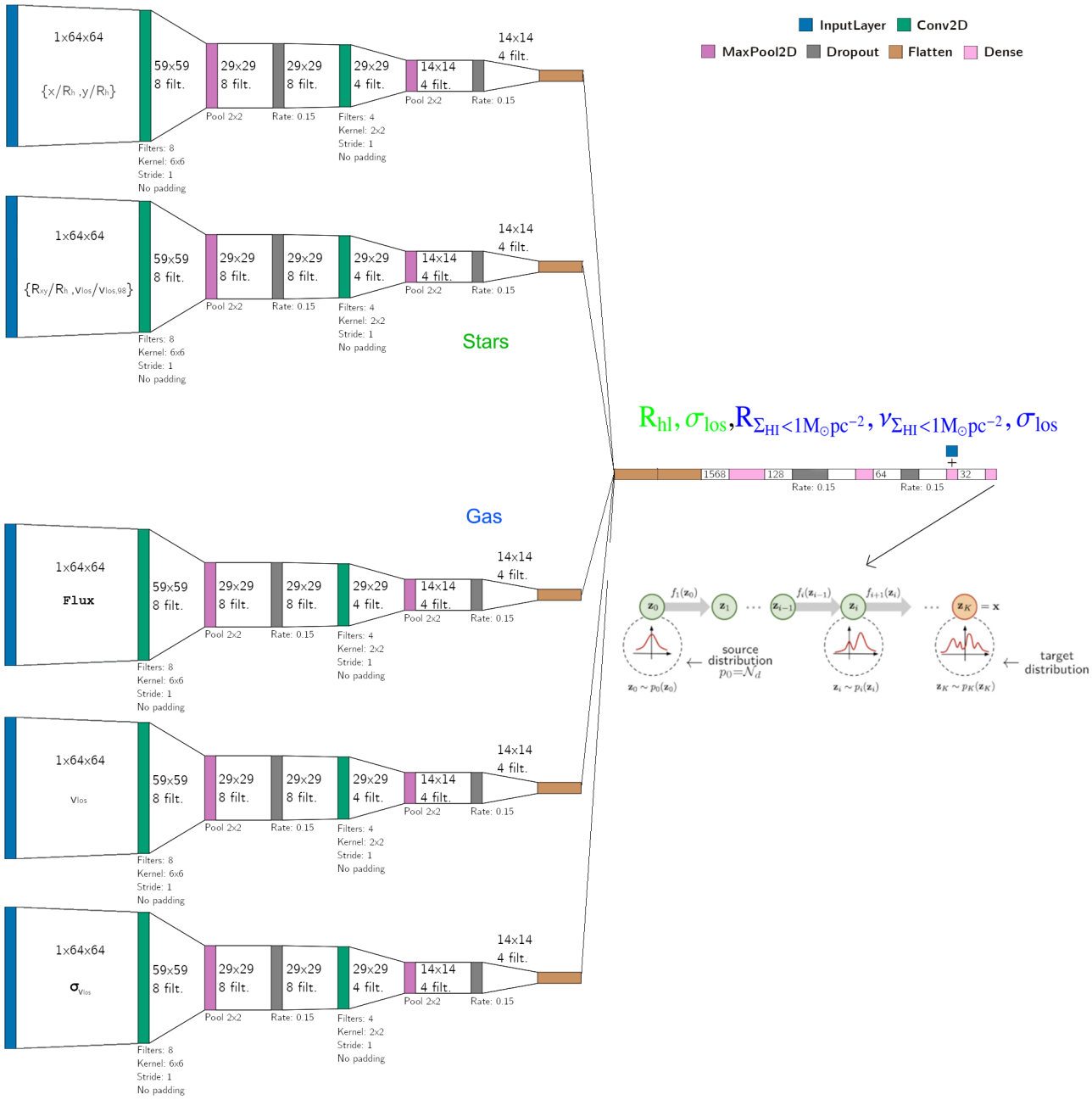

Fig. 3. Scheme of our CNN architecture. Five parallel branches (two for star information, three for gas information) pass through several convolutional and pooling layers independently, further reducing the dimensionality; they then merge for a final sequence of dense and pooling layers, producing an N parameter output. After training the CNN, the 32 neurons of the penultimate layer are used as inputs to train a normalising flow model. The flow model learns a series of transformations to an N-dimensional Gaussian PDF, which are conditioned on the inputs, and outputs a posterior N-dimensional joint PDF. The final output represents the value estimated for the dynamical mass of the galaxy enclosed within N different radii.

– Star model: the inputs are the two star channels described in Sect. 3.2, the projected half-light radius, and the mean velocity dispersion of the stars.
– StarGas model: the inputs are all of the above, i.e. the five channels described in Sect. 3.2, the value of $R_{\Sigma_{\rm HI}<1\ M_\odot {\rm pc}^{-2}}$, the velocity of the gas at that radius and the mean velocity dispersion of the gas inside that radius, the projected half-light radius, and the mean velocity dispersion of the stars.

In Fig. 4, we compare the results for the full set ($M_* = 10^{5.5}$–10^{11} $M_\odot$ containing at least 100 stellar and gas particles and a high-resolution particle mass fraction exceeding 95%) and the cold gas set (same restrictions, plus cold gas masses of $M_{\rm cold} > 10^{7.5}$ $M_\odot$ and $M_{\rm cold}/M_* > 0.1$), using both the Gas model and the StarGas model.

Only 18.8% of the galaxies from the full set do not meet the cold gas set criteria, but their presence in the dataset increases the dispersion on the enclosed mass prediction ∼1.7 times throughout the complete profile with respect to training only on the cold gas set. On the other hand, the StarGas model undergoes only a slight increase in homogeneous dispersion over the whole range of distances when using the full set instead of the cold gas set, which shows that the full model is relatively stable in the

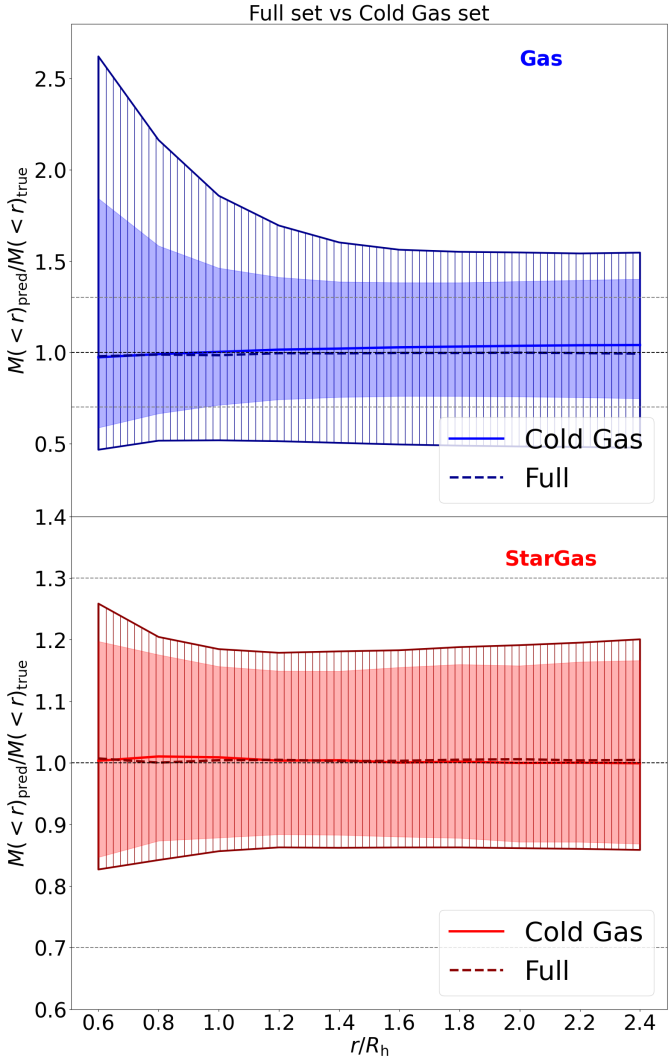

Fig. 4. Ratio of the mass predicted by the neural network models to the real mass enclosed within different radii of the galaxies in the test sets of NIHAO galaxy projections. Top: results for the Gas model. Bottom: results for the StarGas model.

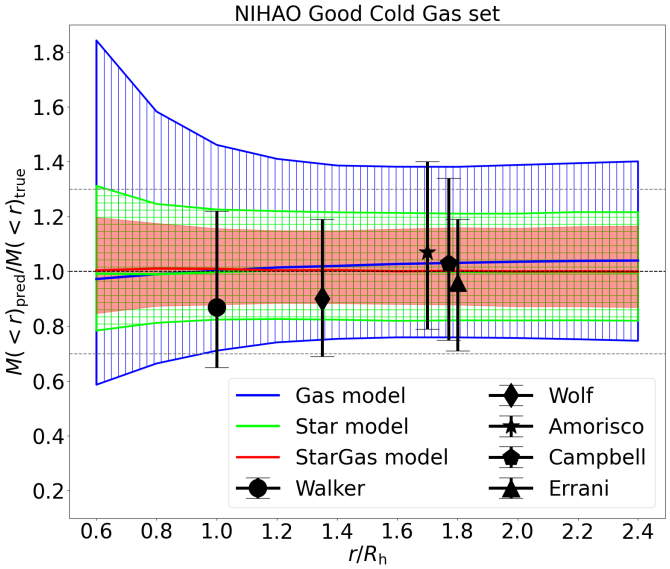

Fig. 5. Ratio of the mass predicted by the neural network models to the real mass enclosed within different radii of the galaxies in the test sets of NIHAO galaxy projections. The coloured lines show the median ratio using the mass estimated by the CNN for the training set, while the shaded regions indicate the 1σ errorbars. Blue: Results when using the three gas channels described in Sect. 3.2, the radius where the cold gas superficial density falls below 1 $M_\odot$ pc^{-2}, and the mean velocity dispersion inside that radius. Green: Results when using the two star channels described in Sect. 3.2, the projected half-light radius, and the mean velocity dispersion of the stars. Red: Results when using all of the above. The predicted-to-true enclosed mass ratios resulting from applying the literature mass estimators to NIHAO galaxies are shown as black symbols with 1σ errorbars.

Table 1. Mean value and 1σ scatter of the predicted-to-true enclosed mass ratio $M(<r)_{\rm pred}/M(<r)_{\rm true}$ for the Star, Gas, and StarGas models.

	$M(<r)_{\rm pred}/M(<r)_{\rm true}$		
NN model	0.6 $R_{\rm hl}$	$R_{\rm hl}$	2.4 $R_{\rm hl}$
Star	$0.99^{+0.32}_{-0.21}$	$1.00^{+0.23}_{-0.17}$	$1.00^{+0.22}_{-0.18}$
Gas	$0.97^{+0.87}_{-0.39}$	$1.00^{+0.46}_{-0.29}$	$1.03^{+0.36}_{-0.29}$
StarGas	$1.00^{+0.19}_{-0.16}$	$1.00^{+0.14}_{-0.13}$	$1.00^{+0.17}_{-0.13}$

presence of galaxies without gas information in the dataset and is able to make use of the stellar information provided in those cases even in a global training.

As we are interested in studying the performance of the model when making use of a dual source of information, we use the cold gas set throughout the rest of the analysis.

In Fig. 5, we show the accuracy of the model trained on NIHAO galaxies at all radii for the testing set and using the three models. The numerical results at 0.6 $R_{\rm hl}$, $R_{\rm hl}$, and 2.4 $R_{\rm hl}$ are shown in Table 1. The bias in the recovery of the mass at different radii is negligible in all three cases, with a maximum error for any of the radii of 0.03, although the Gas model presents a tendency to underestimate enclosed mass at smaller radii and overestimate it at larger radii. On the other hand, in comparison to the Star model, the 1σ uncertainty is reduced by a factor of ∼1.5 throughout the profile when using gas information along with the star information, with a particularly significant improvement in the innermost part of the galaxy. The combination of gas and star information allows the model to improve its prediction at all studied radii.

The Gas model, by itself, performs significantly worse than the Star and StarGas models at all radii, with dispersions of ∼2.2 times those of the StarGas model in the outer part of the galaxy and above ∼3.5 in the inner side, making the model unsuitable to be used just with cold gas information from the galaxies. These results point to a possible decoupling between gas morphology and gravitational potential in NIHAO dwarf galaxies, likely driven by strong feedback-related effects.

We also compared the results with several mass estimators (Walker et al. 2009; Wolf et al. 2010; Amorisco & Evans 2012; Campbell et al. 2017; Errani et al. 2018). The StarGas model shows less error in the mass recovery for all cases while at the same time maintaining a negligible bias.

We next tested whether the model, which was trained over such a wide mass range, could be applied equally well to low- and high-mass galaxies. In Fig. 6 we show the accuracy of the StarGas model for a testing set of only dwarf galaxies ($M_* < 10^{9.5}$ $M_\odot$) and high-mass galaxies ($M_* > 10^{9.5}$ $M_\odot$). The

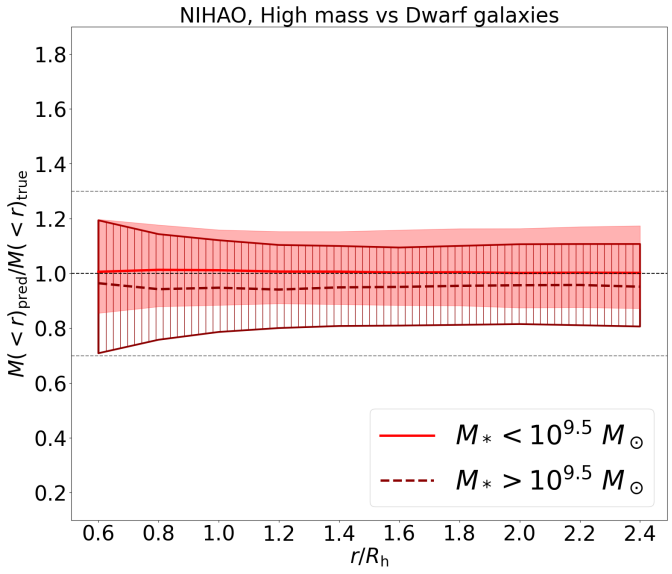

Fig. 6. Ratio of the mass predicted by the neural network models to the real mass enclosed within different radii of the galaxies in the test sets of NIHAO galaxy projections. Shown are the results for the StarGas model applied to two subsets of galaxies with $M_* < 10^{9.5}\ M_\odot$ and $M_* > 10^{9.5}\ M_\odot$.

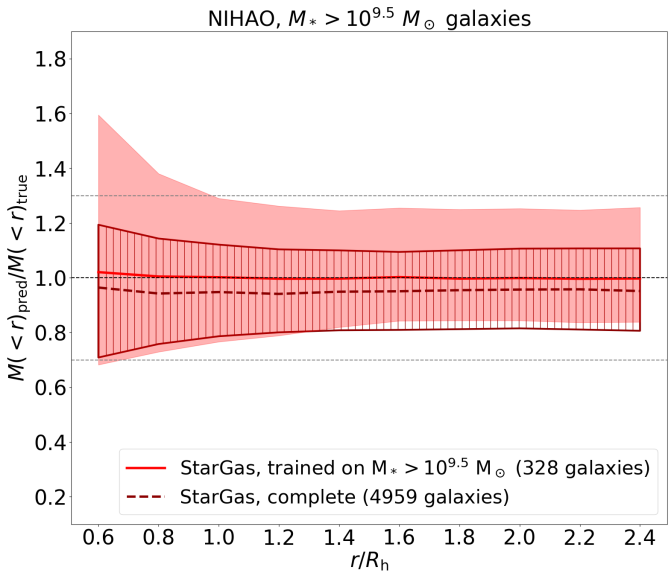

Fig. 7. Ratio of the mass predicted by the StarGas model to the real mass enclosed within different radii of the galaxies in the test sets of NIHAO galaxy projections. Shown are the results applied to a subset of galaxies with $M_* > 10^{9.5}\ M_\odot$ for the main StarGas model and for a modified StarGas model trained only with a subset of 328 high-mass galaxies with $M_* > 10^{9.5}\ M_\odot$.

StarGas model shows a clear bias towards underestimating the masses of high-mass galaxies, up to predicting less than ∼7% of the total mass in the average prediction of the entire high-mass subset.

This bias is most likely a direct consequence of an over-representation of dwarf galaxies in the training set; from the total of 4959 galaxies in the training dataset, only 328 have a stellar mass greater than $10^{9.5}\ M_\odot$ due to the large number of satellites in the dataset. This may have led to an improvement in the optimisation of the complete training by tending to reduce the overall estimated mass, even if that increased the error in the massive galaxy part of the set.

The effect of training a model only with high-mass galaxies can be seen in Fig. 7. When training only on high-mass galaxies, the bias towards underestimating their mass is eliminated. On the other hand, the dispersion increases by a factor of 3 in comparison to the dispersion presented in the model trained with all the galaxies. This shows that a greater number of galaxies, even if they are dwarf galaxies, gives relevant information to the model in regard to inferring mass profiles for high-mass galaxies, even if it also generates a bias. The way of dealing with this effect will be explored in future work.

We also investigated the performance of the models on galaxies with differing kinematic support, specifically comparing rotation-supported and dispersion-supported systems. Dispersion-supported galaxies were identified using the criterion $\kappa_{co} < 0.5$. The parameter κ_{co}, introduced in Correa et al. (2017), quantifies the fraction of stellar kinetic energy invested in ordered rotation and is defined as

$$\kappa_{co} = \frac{1}{K_s} \sum_{i}^{r<30\ \mathrm{kpc}\ ;\ j_{z,i}>0} \frac{m_i}{2} \left(\frac{j_{\parallel,i}}{R_{\perp,i}} \right)^2, \quad (1)$$

where K_s is the total kinetic energy of the stellar component, and the sum is taken over all star particles within 30 kpc of the galaxy's centre that have positive specific angular momentum $j_\parallel$ aligned with the total stellar angular momentum vector, m_i is the mass of the i-th stellar particle, and $R_{\perp,i}$ denotes its distance from the galaxy's stellar rotation axis.

We show the results in Fig. 8. The Star model works significantly worse with rotation-supported galaxies than with dispersion-supported galaxies, and the StarGas model also shows a clear increase in error in mass recovery for rotation-supported galaxies. A similar bias to that encountered with high-mass galaxies is present with the Star model, but it is mostly eliminated on the StarGas model. The Gas model shows much greater dispersion, but no bias, and a better performance on rotation-supported galaxies than on dispersion-supported ones, as expected. It is therefore reasonable to conclude that the information provided by gas is allowing us to correct the bias found when using only stellar information, while reducing dispersion even though it is much greater when using gas alone.

We finally studied the effect of inclination in the model predictions. In Fig. 8 we compare the performance of the three models on a subset of the testing dataset consisting exclusively of edge-on galaxies and galaxies inclined at 60° with respect to the line of sight.

The Star model exhibits a systematic bias with respect to galaxy inclination. For edge-on galaxies, the enclosed mass is consistently overestimated across all radii, with a mean ratio of $M(<r)_{pred}/M(<r)_{true} = 1.07$. In contrast, for galaxies inclined at 60°, the model underestimates the mass, with $M(<r)_{pred}/M(<r)_{true} = 0.89$ at small radii and 0.96 at larger radii. These trends indicate that the Star model lacks the capacity to infer galaxy inclination and appropriately correct its mass predictions. Instead, it interprets the inclination-induced suppression of projected stellar velocities as a signature of lower mass. During training, the model compensates for this effect by adjusting its predictions to minimise the overall error across the full dataset, where galaxies with all inclinations are present. As a result, inclination-dependent biases persist in the model's output. The Star model is thus strongly subject to projection effects, as happens with dynamical modelling.

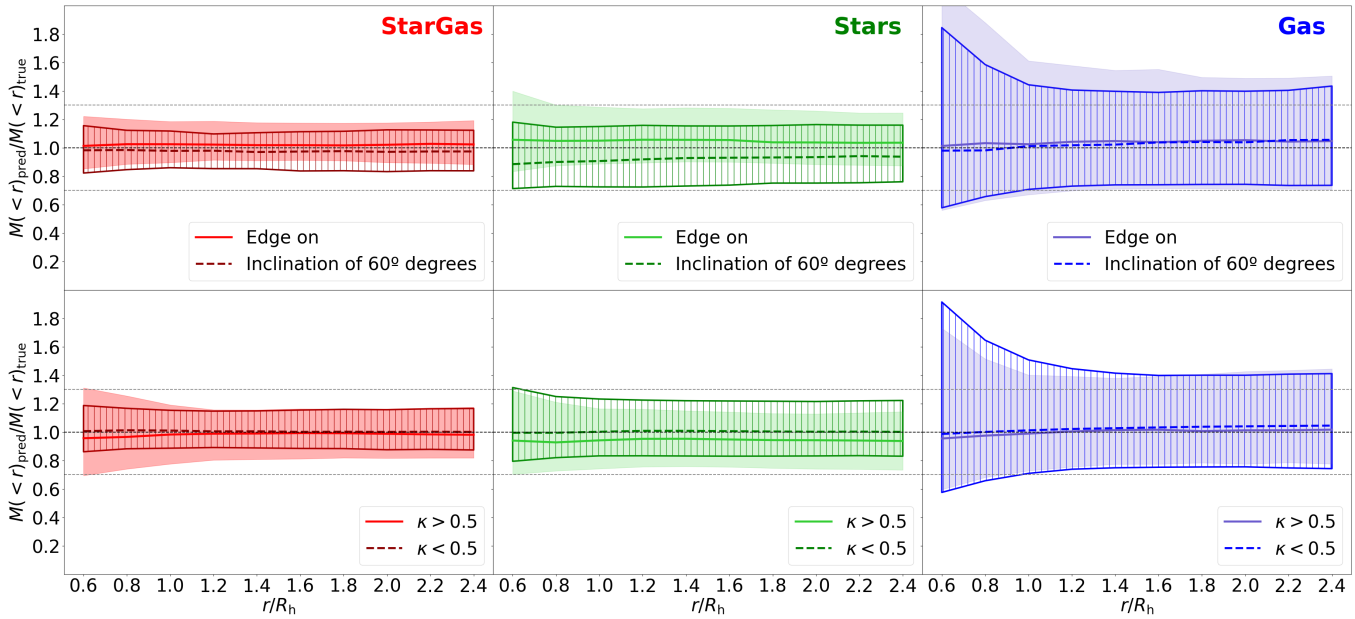

Fig. 8. Ratio of the mass predicted by the neural network models to the real mass enclosed within different radii of the galaxies in the test sets of NIHAO galaxy projections. Shown are the results for two subsets consisting of rotation-supported galaxies with $\kappa > 0.5$ and of dispersion-supported galaxies with $\kappa < 0.5$, and for two subsets consisting of edge-on galaxies and galaxies at an inclination of 60° degrees.

The Gas model, instead, does not exhibit a significant inclination-dependent bias; however, it shows increased scatter in its predictions for edge-on galaxies. This may be due to a stronger degeneracy between edge-on disc galaxies and spheroidal systems observed at various inclinations. In contrast, non-spheroidal galaxies at intermediate inclinations may be more easily distinguishable from spheroidal ones, enabling the gas-related input channels to better infer the underlying gravitational potential.

The StarGas model achieves significantly lower dispersion than either the Star or Gas models and displays only a minor bias, likely inherited from the stellar input channels. Nevertheless, the mean residuals of the predicted-to-true enclosed mass ratio remain below 0.05 at all radii for both inclination subsets. This again highlights the model's ability to partially correct the individual limitations of using stellar or gas information alone by leveraging their combined input.

4.2. Cross-testing between different simulation datasets

Training and testing a neural network on a single suite of hydrodynamical simulations may lead to overfitting to the specific physical and numerical characteristics of that simulation set. While the model may perform well within that domain, its reliability on observational data, or even other simulations, can be compromised. Ideally, the network should learn general features in the input maps that are predictive of the mass profile, rather than adapting to the peculiarities of a particular simulation. A robust way to assess this generalisation capability is through cross-simulation testing: training the model on one simulation suite and evaluating its performance on a different one, then comparing the results to those obtained when testing within the same suite.

We conducted a cross-testing analysis between the NIHAO and AURIGA simulation suites. A similar comparison was previously performed in Sarrato-Alós et al. (2025) using a model equivalent to our Star model, which revealed a clear limitation in generalisation across simulations. For this work, we extended that analysis by examining how the inclusion of gas information influences the model's ability to generalise between different simulation suites.

Table 2. Mean value and 1σ scatter of predicted-to-true enclosed mass ratio $M(< r)_{\rm pred}/M(< r)_{\rm true}$ for various training and testing combinations across NIHAO and AURIGA simulation suites using the StarGas model.

$M(< r)_{\rm pred}/M(< r)_{\rm true}$, StarGas model				
Trained	Tested	0.6 $R_{\rm hl}$	$R_{\rm hl}$	2.4 $R_{\rm hl}$
NIHAO	NIHAO	$0.99^{+0.24}_{-0.16}$	$1.00^{+0.16}_{-0.13}$	$1.01^{+0.17}_{-0.14}$
	AURIGA	$0.94^{+0.37}_{-0.22}$	$0.95^{+0.32}_{-0.20}$	$0.96^{+0.26}_{-0.17}$
AURIGA	NIHAO	$0.93^{+1.07}_{-0.32}$	$0.97^{+0.57}_{-0.25}$	$1.03^{+0.36}_{-0.40}$
	AURIGA	$1.01^{+0.26}_{-0.22}$	$1.01^{+0.21}_{-0.20}$	$1.01^{+0.14}_{-0.15}$
BOTH	NIHAO	$1.00^{+0.23}_{-0.14}$	$1.02^{+0.18}_{-0.16}$	$1.00^{+0.22}_{-0.12}$
	AURIGA	$0.98^{+0.28}_{-0.22}$	$0.98^{+0.24}_{-0.20}$	$0.99^{+0.20}_{-0.16}$

Figure 9 and Table 2 present the results of the cross-simulation analysis for the StarGas model. When the model is trained on both NIHAO and AURIGA simultaneously, its performance on each individual suite is comparable to that of models trained exclusively on the respective suite. In both cases, we observe a slight reduction in performance, which likely reflects a reduced emphasis on simulation-specific features and an improvement in the model's generalisation ability.

In contrast, when testing NIHAO galaxies with a model trained solely on AURIGA, the performance degrades severely, particularly at small radii. This failure may stem from the fact that approximately 90% of the NIHAO sample consists of dwarf satellite galaxies, whose inner mass profiles are strongly influenced by NIHAO's feedback prescription, known to differ significantly from that used in AURIGA. As a result, the

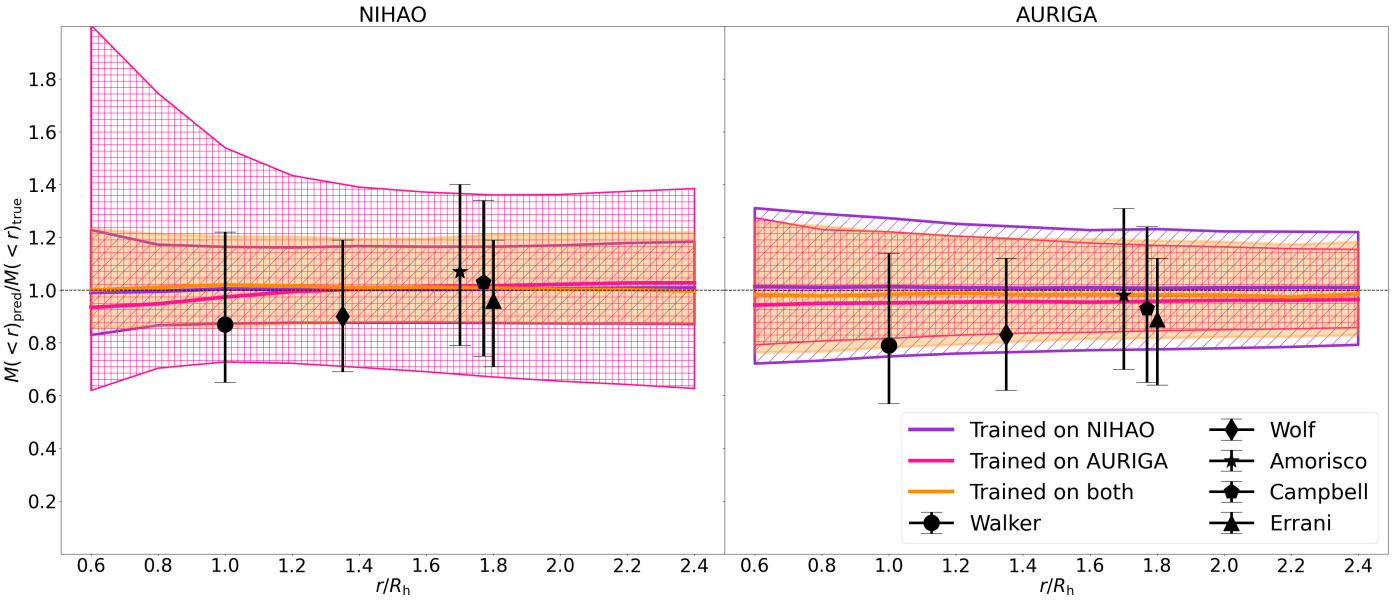

Fig. 9. Ratio of the mass predicted by the neural network models to the real mass enclosed within different radii of the galaxies in the test sets of NIHAO and AURIGA galaxy projections. The coloured lines show the median ratio using the mass estimated by the CNN for the training set, while the shaded regions indicate the 1σ errorbars. Purple: Results of the StarGas model trained on NIHAO galaxies. Pink: Results of the StarGas model trained on AURIGA galaxies. Orange: Results of the StarGas model trained on both simulation suites. The predicted-to-true enclosed mass ratios resulting from applying literature mass estimators to both NIHAO (left) and AURIGA (right) galaxies are shown as black symbols with 1σ errorbars.

AURIGA trained model is likely unable to reproduce structural features such as the cores found in many NIHAO dwarfs, leading to large dispersion and systematic overestimation of the enclosed mass in the inner regions.

On the other hand, the model trained only on NIHAO performs consistently when applied to AURIGA galaxies. While the dispersion increases slightly relative to the AURIGA-trained model, the results remain stable, with only a mild bias toward underestimating the enclosed mass.

5. Conclusions

In this work, we developed and tested a probabilistic deep learning framework to infer the enclosed dynamical mass profiles of galaxies by combining stellar and gas kinematic information from realistic cosmological hydrodynamical simulations. Following previous work done by Expósito-Márquez et al. (2023) and Sarrato-Alós et al. (2025), our model was built around a multiple channel convolutional neural network (CNN) with a normalising flow for uncertainty quantification that uses projected stellar and gas 2D maps to predict the mass distribution across different radii.

By training on a diverse and physically motivated dataset drawn from the NIHAO simulation suites, we systematically explored the strengths and limitations of the approach. We used both mock HI observations using the MARTINI code (Oman et al. 2019) and kinematic information from the star particles of the simulations. We compared variants of the model that use only stellar data, only gas data, or a combination of both, and we studied the effects of galaxy inclination on performance, as well as the biases of the model when applied to specific regions of the galaxies parameter space. We also trained and tested the model on the AURIGA simulation suites and we performed a cross-simulation testing of the model to study its generalisation capabilities.

The main results of our study are the following:

– The StarGas model, which integrates both stellar and gas input maps, consistently outperforms single-tracer models and standard literature mass estimators (Walker et al. 2009; Wolf et al. 2010; Amorisco & Evans 2012; Campbell et al. 2017; Errani et al. 2018) when trained and tested on NIHAO galaxies. It reduces the scatter in mass profile predictions to a standard deviation of ∼30% across a range of radii and yields unbiased residuals at all scales (Fig. 5);
– The Star model is affected by strong biases related to galaxy inclination, significantly overestimating the mass in edge-on systems and underestimating it at intermediate inclinations. These projection effects are not effectively disentangled by the stellar channels alone. The Gas model avoids systematic inclination biases, but shows greater dispersion, especially for edge-on systems. The StarGas model partially mitigates the inclination-related biases of the Star model and the increased uncertainty of the Gas model, demonstrating that combining both tracers provides complementary constraints on the gravitational potential (Fig. 8);
– Cross-testing reveals that models trained on a single simulation suite are vulnerable to generalisation issues due to differing feedback implementations and numerical effects. The StarGas model trained on AURIGA performs poorly when tested on NIHAO galaxies, with the gas implementation not outperforming previous results with a star-only simpler model (Sarrato-Alós et al. 2025), particularly in the inner regions. The bad performance is likely due to its inability to model cored mass profiles in dwarf galaxies since AURIGA galaxies do not form cores. On the other hand, the model trained on NIHAO performs significantly better when tested on AURIGA galaxies. Training on a mixed dataset from multiple suites improves the model's generalisation capacity, with only moderate increases in scatter. The model trained on both NIHAO and AURIGA achieves stable

performance across simulation types. These results indicate that the StarGas model has better robustness to simulation-specific systematics (Fig. 9) than the star-only model from Sarrato-Alós et al. (2025).

Overall, our results demonstrate the power of combining multiple kinematic tracers, such as HI and stellar maps, in a unified deep learning framework to infer galaxy mass profiles in a data-driven and probabilistic manner. Future work will focus on extending this model to observational datasets and further exploring the model's interpretability and sensitivity to physical galaxy properties.

Acknowledgements. CB is supported by Projects PID2021-122603NB-C22 and PID2024-156100NB-C22 financed by MICIU/AEI /10.13039/501100011033 / FEDER, EU. ADC acknowledges fundings from the Agencia Estatal de Investigación, under the 2023 call "Ayudas para Incentivar la Consolidación Investigadora", grant number CNS2023-144669, proyecto "TINY", and the 2024 call "Proyectos de Generación de Conocimiento", grant number PID2024-160009NA-I00, proyecto "INGENIO". This research is also co-funded by the European Union (Widening Participation, ExGal-Twin, GA 101158446). The authors wish to acknowledge the contribution of the IAC High-Performance Computing support team and hardware facilities to the results of this research. The freely available softwares *pynbody* (Pontzen et al. 2013) and *MARTINI* (Oman et al. 2019) have been used for part of this analysis. The authors thank the AURIGA projects' PIs for making their data publicly available at `https://wwwmpa.mpa-garching.mpg.de/auriga/data.html`.

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from `tensorflow.org`
Amorisco, N. C., & Evans, N. W. 2012, MNRAS, 419, 184
Arroyo-Polonio, J. M., Pascale, R., Battaglia, G., et al. 2025, A&A, 699, A347
Bañares-Hernández, A., Read, J. I., & Júlio, M. P. 2026, A&A, 705, A212
Battaglia, G., Helmi, A., Tolstoy, E., et al. 2008, ApJ, 681, L13
Binney, J., & Mamon, G. A. 1982, MNRAS, 200, 361
Binney, J., & Vasiliev, E. 2023, MNRAS, 520, 1832
Blank, M., Macciò, A. V., Dutton, A. A., & Obreja, A. 2019, MNRAS, 487, 5476
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
Brook, C. B., Stinson, G., Gibson, B. K., Wadsley, J., & Quinn, T. 2012, MNRAS, 424, 1275
Buck, T., Obreja, A., Macciò, A. V., et al. 2019, MNRAS, 491, 3461
Bullock, J. S., & Boylan-Kolchin, M. 2017, ARA&A, 55, 343
Campbell, D. J. R., Frenk, C. S., Jenkins, A., et al. 2017, MNRAS, 469, 2335
Cappellari, M., Bacon, R., Bureau, M., et al. 2006, MNRAS, 366, 1126
Chabrier, G. 2003, PASP, 115, 763
Collins, M. L. M., Read, J. I., Ibata, R. A., et al. 2021, MNRAS, 505, 5686
Correa, C. A., Schaye, J., Clauwens, B., et al. 2017, MNRAS, 472, L45
Croce, A., Pascale, R., Giunchi, E., et al. 2023, A&A, 682
de Blok, W. J. G., Walter, F., Brinks, E., et al. 2008, AJ, 136, 2648
de los Rios, M., Petač, M., Zaldivar, B., et al. 2023, MNRAS, 525, 6015
de los Rios, M., Gioia, S., Iocco, F., & Trotta, R. 2025, Dark Matter profiles of "in silico" galaxies: deep learning inference
Dutton, A. A., Buck, T., Macciò, A. V., et al. 2020, MNRAS, 499, 2648
Errani, R., Peñarrubia, J., & Walker, M. G. 2018, MNRAS, 481, 5073
Expósito-Márquez, J., Brook, C. B., Huertas-Company, M., et al. 2023, MNRAS, 519, 4384
Gentile, G., Salucci, P., Klein, U., Vergani, D., & Kalberla, P. 2004, MNRAS, 351, 903
Grand, R. J. J., Gómez, F. A., Marinacci, F., et al. 2017, MNRAS, 467, 179
Ho, M., Rau, M. M., Ntampaka, M., et al. 2019, ApJ, 887, 25
Ho, M., Bartlett, D. J., Chartier, N., et al. 2024, LtU-ILI: An All-in-One Framework for Implicit Inference in Astrophysics and Cosmology
Katz, H., Lelli, F., McGaugh, S. S., et al. 2017, MNRAS, 466, 1648
Kleyna, J. T., Wilkinson, M. I., Evans, N. W., & Gilmore, G. 2001, ApJ, 563, L115
Kowalczyk, K., Łokas, E. L., & Valluri, M. 2017, MNRAS, 470, 3959
Lelli, F., McGaugh, S. S., & Schombert, J. M. 2016, AJ, 152, 157
Leung, G. Y. C., Leaman, R., Battaglia, G., et al. 2021, MNRAS, 500, 410
Moore, B. 1994, Nature, 370, 629
Nguyen, T., Mishra-Sharma, S., Williams, R., & Necib, L. 2023, Phys. Rev. D, 107
Oman, K. A., Marasco, A., Navarro, J. F., et al. 2019, MNRAS, 482, 821
Papamakarios, G., Pavlakou, T., & Murray, I. 2018, Masked Autoregressive Flow for Density Estimation
Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. 2021, J. Mach. Learn. Res., 22, 1
Planck Collaboration XIII. 2016, A&A, 594, A13
Pontzen, A., Roškar, R., Stinson, G., & Woods, R. 2013, pynbody: N4-body/SPH analysis for python, Astrophysics Source Code Library
Read, J. I., Walker, M. G., & Steger, P. 2019, MNRAS, 484, 1401
Read, J. I., Mamon, G. A., Vasiliev, E., et al. 2021, MNRAS, 501, 978
Sarrato-Alós, J., Brook, C., Di Cintio, A., et al. 2025, A&A, 703, A140
Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, MNRAS, 446, 521
Schmidt, M. 1959, ApJ, 129, 243
Schwarzschild, M. 1979, ApJ, 232, 236
Shen, S., Wadsley, J., & Stinson, G. 2010, MNRAS, 407, 1581
Stinson, G., Seth, A., Katz, N., et al. 2006, MNRAS, 373, 1074
Stinson, G. S., Brook, C., Macciò, A. V., et al. 2013, MNRAS, 428, 129
van den Bosch, R. C. E., & de Zeeuw, P. T. 2010, MNRAS, 401, 1770
van der Marel, R. P. 1994, MNRAS, 270, 271
Vogelsberger, M., Genel, S., Springel, V., et al. 2014, MNRAS, 444, 1518
Walker, A. R. 2003, Distances to Local Group Galaxies, eds. D. Alloin, & W. Gieren (Berlin, Heidelberg: Springer Berlin Heidelberg), 265
Walker, M. G., Mateo, M., Olszewski, E. W., et al. 2009, ApJ, 704, 1274
Wang, L., Dutton, A. A., Stinson, G. S., et al. 2015, MNRAS, 454, 83
Wolf, J., Martinez, G. D., Bullock, J. S., et al. 2010, MNRAS, 406, 1220