أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Inferring Dark Matter Density Profiles from 5D Stellar Kinematics using Simulation-Based Inference

Nawaf Bandar Alrefaie, **Ting Li**

nawaf.alrefaie@kaust.edu.sa, ting.li@astro.utoronto.ca

Abstract

Dark matter makes up most of the matter in the universe, and understanding its distribution can reveal how cosmic structures form and evolve. Because dark matter cannot be observed directly, we rely on its gravitational effects on visible matter. Dwarf galaxies are considered great laboratories for studying dark matter behaviors because they have little visible matter, which makes them dark-matter dominated. Classical dynamical methods often rely on simplifying assumptions and compress stellar observations into summary quantities, which can lose useful information. Other methods also exist, but they do not make use of the traverse motions of stars across the sky. In this project, we combine physics and deep learning using simulation-based inference, specifically Neural Posterior Estimation (NPE), to infer the posterior distribution of dark matter parameters in a dwarf galaxy from observable stellar positions and velocities on the sky. We create synthetic dwarf galaxies, each with known dark matter distribution parameters and mock observed stellar positions and velocities. We then apply a graph neural network to learn the relationship between the observed stellar data and the underlying dark matter distribution. The trained model can later infer the posterior distribution of the dark matter parameters in a dwarf galaxy, given its stars' positions and velocities. This project aims to reduce the information loss in classical dynamical methods, while extending previous methods by including stars' traverse motions across the sky

Date: August 18, 2026

Code: https://github.com/KIRAxBELMONT67/Inferring-Dark-Matter-Density-Profiles-from-5D-Stellar

Project Page: https://your-project-page.github.io

Contents

1 Introduction **2**

2 Methodology **2**

2.1 Why Simulation-Based Inference? . . . 2

2.2 Interpreting the Predictions . . . 4

3 Dataset **5**

3.1 Parameter Distributions . . . 6

3.2 Galaxy Configuration and Stellar Sampling . . . 6

3.3 Dataset Construction and Storage . . . 7

4 Model Pipeline **7**

4.1 Loading and Selecting the Stellar Features . . . 7

4.2 Line-of-Sight Velocity Noise . . . 8

4.3 Training and Validation Split . . . 8

4.4 Parameter Transformation . . . 9

4.5 Standardization . . . 9

4.6 Graph Construction . . . 10

4.7 Graph Batching and Model Input . . . 11

5 Model Evaluation **11**

5.1 Truth-versus-Predicted Test . . . 12

5.2 Dark Matter Density-Recovery Test . . . 12

5.3 Dark Matter Enclosed-Mass Recovery Test . . . 13

5.4 Corner Plots . . . 14

5.5 One-Dimensional Rank-Based Calibration Test . . . 15

5.6 TARP Joint-Calibration Test . . . 16

5.7 MIRA Score . . . 17

6 Results **19**

6.1 Comparison of the Five-Dimensional and Two-Dimensional Models . . . 19

6.2 TARP Joint Calibration . . . 19

6.3 MIRA Scores . . . 20

6.4 Key Takeaways . . . 21

7 Limitations and Future Work **21**

7.1 Alternative Neural-Network Architectures . . . 21

7.2 Comparison with Jeans Analysis . . . 22

7.3 Conversion to Observable Proper Motions . . . 22

7.4 Application to Real Dwarf Galaxies . . . 22

8 Conclusion **23**

9 Related Work **23**

References **23**

1 Introduction

Dark matter constitutes most of the matter in the Universe, but it cannot be observed directly. Its distribution must instead be inferred from its gravitational effects on visible objects. Dwarf galaxies are useful laboratories for studying dark matter because they contain relatively little baryonic matter and are often dark-matter dominated. Measurements of their dark matter density profiles can help distinguish between centrally cusped and centrally cored profiles, providing tests of dark matter models and galaxy-formation physics.

The goal of this project is to infer the parameters of a dwarf galaxy's dark matter density profile from the observed positions and motions of its stars. Previous simulation-based studies have demonstrated that graph neural networks can infer density-profile parameters from projected stellar positions and line-of-sight velocities. However, stars also have transverse motions across the sky, measured as proper motions. This project investigates whether adding two proper-motion components improves the recovery of the underlying dark matter density profile, particularly the inner-slope parameter that distinguishes cusped from cored profiles.

Several assumptions and limitations apply to this study. The model is trained using synthetic dwarf galaxies generated from specified dark matter and stellar-density models. Therefore, its predictions depend on how accurately these simulations represent real galaxies. The initial experiments use a generalized Navarro–Frenk–White profile for the dark matter distribution and a Plummer profile for the stellar distribution. The observed data consist of two projected sky-position components, one line-of-sight velocity, and two proper-motion components. Proper motions are angular velocities and must be combined with an assumed or measured distance before they can be interpreted as transverse physical velocities. Measurement uncertainties, foreground contamination, model mismatch, and a limited number of observed stars may also affect the inferred posterior distributions.

To address this problem, we use simulation-based inference, specifically Neural Posterior Estimation (NPE). Synthetic galaxies are generated with known dark matter parameters, and their stars are represented as nodes in a graph. A graph neural network learns a representation of each galaxy from its stellar observations, while a conditional density estimator learns the posterior distribution of the dark matter parameters given that representation. We compare a paper-like projected-data baseline with a five-dimensional observed-data model that additionally includes the two proper-motion components. The models are evaluated using posterior accuracy, uncertainty, coverage, truth vs predicted, dark matter's desnsity and mass recovery, and corner plots to determine whether proper-motion measurements provide a measurable improvement in dark matter density-profile inference, and Tests of Accuracy with Random Points (TARP), Mass In Random Areas (MIRA) test

2 Methodology

2.1 Why Simulation-Based Inference?

The objective of our method is to infer the posterior distribution

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{obs}}),$$

where $\boldsymbol{\theta}$ contains the unknown parameters and $\boldsymbol{x}_{\text{obs}}$ contains the observed stellar measurements. Direct Bayesian inference would require evaluating the likelihood

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}).$$

For this problem, the relationship between the galaxy parameters and the complete set of stellar observations is sufficiently complex that the likelihood is difficult to calculate analytically.

Simulation-Based Inference (SBI) avoids the need for an explicit likelihood. Instead, it learns the relationship between parameters and observations from simulated pairs

$$(\boldsymbol{\theta}_j, \boldsymbol{x}_j).$$

These pairs allow the inference model to learn which parameter combinations are consistent with different stellar observations.

We use Neural Posterior Estimation (NPE) to train a conditional density estimator

$$q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$$

that approximates the posterior:

$$q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) \approx p(\boldsymbol{\theta} \mid \boldsymbol{x}).$$

The inferred parameter vector is

$$\boldsymbol{\theta} = (\rho_0, \gamma, r_s, r_\star, r_a).$$

For a dwarf galaxy containing $N_\star$ observed stars, the conditioning observations are

$$\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^{N_\star},$$

where the five observed features of star i are

$$\boldsymbol{x}_i = \left(\mathrm{RA}_i, \mathrm{Dec}_i, v_{\mathrm{LOS},i}, \mu_{\mathrm{RA}^*,i}, \mu_{\mathrm{Dec},i}\right).$$

Therefore, $\boldsymbol{x}$ represents the complete stellar catalogue of one dwarf galaxy rather than the measurements of a single star.

Our method represents each dwarf galaxy as a graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}).$$

The node set $\mathcal{V}$ contains the observed stars, so a galaxy with $N_\star$ stars has

$$|\mathcal{V}| = N_\star$$

nodes. Each star is represented by one node, and its node-feature vector contains its five observed quantities:

$$\mathbf{X}_i = \left(\mathrm{RA}_i, \mathrm{Dec}_i, v_{\mathrm{LOS},i}, \mu_{\mathrm{RA}^*,i}, \mu_{\mathrm{Dec},i}\right).$$

Consequently, the node-feature matrix has shape

$$\mathbf{X} \in \mathbb{R}^{N_\star \times 5}.$$

The graph connections are constructed using the k-nearest-neighbour method with

$$k = 20.$$

For every star i, its 20 nearest stellar neighbours are identified using their positions on the sky. An edge is then added between node i and each of these neighbours:

$$\mathcal{E} = \left\{(i, j) : j \in \mathcal{N}_{20}(i)\right\},$$

where $\mathcal{N}_{20}(i)$ denotes the set of the 20 nearest neighbours of star i.

This graph construction allows the graph neural network to process each star together with the measurements of nearby stars. Through graph message passing, every node combines its own features with features received from its connected neighbours. After the graph-convolution layers, the node representations are combined using graph-level pooling to produce a fixed-dimensional summary of the entire dwarf galaxy:

$$\boldsymbol{h}_{\mathcal{G}} = \mathrm{Pool}\left(\{\boldsymbol{h}_i\}_{i=1}^{N_\star}\right).$$

This pooling operation allows dwarf galaxies containing different numbers of observed stars to be represented by summary vectors of the same size.

The NPE density estimator is conditioned on this graph-level summary:

$$q_\phi\left(\boldsymbol{\theta} \mid \boldsymbol{h}_{\mathcal{G}}\right) \approx p\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right).$$

The graph neural network therefore extracts information from the stellar observations, while NPE converts the resulting galaxy-level representation into a posterior distribution over the five parameters.

The graph neural network and posterior estimator are trained together by minimizing the negative conditional log-probability

$$\mathcal{L} = -\frac{1}{N_{\mathrm{sim}}} \sum_{j=1}^{N_{\mathrm{sim}}} \log q_\phi\left(\boldsymbol{\theta}_j \mid \boldsymbol{h}_{\mathcal{G}_j}\right).$$

This objective encourages the model to assign high posterior probability to the known parameter values associated with each stellar graph.

For a new dwarf galaxy, its observed stars are converted into a graph using the same $k = 20$ nearest-neighbour rule. The trained model then returns

$$q_\phi\left(\rho_0, \gamma, r_s, r_\star, r_a \mid \boldsymbol{h}_{\mathcal{G}_{\mathrm{obs}}}\right).$$

Samples from this distribution represent the joint posterior of the five inferred parameters, including their uncertainties and correlations.

2.2 Interpreting the Predictions

For an observed dwarf galaxy, the model produces posterior samples

$$\boldsymbol{\theta}^{(s)} = \left(\rho_0^{(s)}, \gamma^{(s)}, r_s^{(s)}, r_\star^{(s)}, r_a^{(s)}\right), \qquad s = 1, \ldots, N_{\mathrm{post}}.$$

Each sample is one joint parameter combination supported by the observed stellar graph. The dark matter density and enclosed-mass profiles are reconstructed separately from every posterior sample. This converts the parameter posterior into posterior distributions over the profiles as functions of radius.

2.2.1 Reconstructing the Dark Matter Density Profile

Three inferred parameters determine the dark matter density profile. The density normalization ρ_0 sets the overall amplitude of the profile. The inner density slope γ determines whether the central profile is cuspy or cored. The scale radius r_s marks the transition between the inner and outer slopes.

The inferred parameters are substituted into the generalized Navarro–Frenk–White equation:

$$\rho_{\mathrm{DM}}(r) = \rho_0 \left(\frac{r}{r_s}\right)^{-\gamma} \left(1 + \frac{r}{r_s}\right)^{\gamma-3}.$$

For posterior sample s, the reconstructed density profile is

$$\rho_{\mathrm{DM}}^{(s)}(r) = \rho_0^{(s)} \left(\frac{r}{r_s^{(s)}}\right)^{-\gamma^{(s)}} \left(1 + \frac{r}{r_s^{(s)}}\right)^{\gamma^{(s)}-3}.$$

This calculation is repeated over the selected radii for every posterior sample. At each radius, the resulting density values form a posterior distribution:

$$\left\{\rho_{\mathrm{DM}}^{(s)}(r)\right\}_{s=1}^{N_{\mathrm{post}}}.$$

The median of these values gives the median reconstructed density profile, while their percentiles provide its credible intervals.

2.2.2 Reconstructing the Dark Matter Enclosed-Mass Profile

The enclosed-mass profile is reconstructed using the same posterior samples of ρ_0, γ, and r_s. Following the gNFW treatment used by Nguyen et al., the dark matter mass enclosed within radius r is calculated by integrating the density profile:

$$M_{\mathrm{DM}}(< r) = 4\pi \int_0^r \rho_{\mathrm{DM}}(r')\, r'^2\, \mathrm{d}r'.$$

Substituting the gNFW density equation gives

$$M_{\mathrm{DM}}(< r) = 4\pi\rho_0 r_s^3 \int_0^{r/r_s} u^{2-\gamma}(1+u)^{\gamma-3}\, \mathrm{d}u,$$

where $u = r'/r_s$.

For every posterior sample, the corresponding enclosed-mass profile is

$$M_{\mathrm{DM}}^{(s)}(< r) = 4\pi \int_0^r \rho_{\mathrm{DM}}^{(s)}(r')\, r'^2\, \mathrm{d}r'.$$

The calculation is repeated for all posterior samples and radii. At each radius, the values

$$\left\{M_{\mathrm{DM}}^{(s)}(< r)\right\}_{s=1}^{N_{\mathrm{post}}}$$

form the posterior distribution of the enclosed mass. Its median and percentiles provide the reconstructed enclosed-mass profile and its credible intervals.

Both profiles are calculated using complete joint posterior samples rather than separate marginal parameter estimates. This preserves the correlations among ρ_0, γ, and r_s when their uncertainties are propagated into $\rho_{\mathrm{DM}}(r)$ and $M_{\mathrm{DM}}(< r)$.

3 Dataset

The dataset consists of simulated dwarf galaxies generated using the Action-based Galaxy Modelling Architecture (AGAMA) library. Each galaxy has a known parameter vector

$$\boldsymbol{\theta} = (\rho_0, \gamma, r_s, r_\star, r_a),$$

together with a variable number of sampled stars. A fixed random seed of 0 is used to make the dataset generation reproducible.

3.1 Parameter Distributions

For each dwarf galaxy, the dark matter and stellar-distribution parameters are sampled from the prior distributions listed in Table 1.

Parameter	Sampling distribution	Unit
ρ_0	$\log_{10}(\rho_0/(\mathrm{M}_\odot\,\mathrm{kpc}^{-3})) \sim \mathcal{U}(5,8)$	$\mathrm{M}_\odot\,\mathrm{kpc}^{-3}$
γ	$\gamma \sim \mathcal{U}(-1,2)$	NA
r_s	$\log_{10}(r_s/\mathrm{kpc}) \sim \mathcal{U}(\log_{10}(0.1), \log_{10}(5))$	kpc
$r_\star \mid r_s$	$r_\star \sim \mathcal{U}(0.2r_s, r_s)$	kpc
$r_a \mid r_\star$	$r_a \sim \mathcal{U}(0.5r_\star, 2r_\star)$	kpc
$N_\star$	$N_\star \sim \mathrm{Poisson}(100)$	NA

Table 1 Prior distributions used to generate the simulated dwarf galaxies.

The density normalization ρ_0 is sampled uniformly in logarithmic space between

$$10^5 \leq \frac{\rho_0}{\mathrm{M}_\odot\,\mathrm{kpc}^{-3}} \leq 10^8.$$

Similarly, the scale radius r_s is sampled uniformly in logarithmic space between 0.1 and 5 kpc. Thus, ρ_0 and r_s are log-uniform in their physical units rather than uniformly distributed in linear space.

The inner slope γ is sampled uniformly between -1 and 2. After r_s is sampled, the stellar scale radius $r_\star$ is sampled conditionally between $0.2r_s$ and r_s. The anisotropy radius r_a is then sampled conditionally between $0.5r_\star$ and $2r_\star$.

The number of stars is not fixed across galaxies. For each galaxy, it is sampled independently from a Poisson distribution with mean

$$\lambda = 100.$$

Consequently, the expected number of stars per galaxy is 100, while individual galaxies may contain fewer or more stars.

3.2 Galaxy Configuration and Stellar Sampling

After sampling $(\rho_0, \gamma, r_s, r_\star, r_a)$, the corresponding galaxy is specified in AGAMA using three components.

The dark matter gravitational potential is represented by an AGAMA `Spheroid` model with

$$\alpha = 1, \qquad \beta = 3, \qquad \gamma = \gamma,$$

a scale radius equal to r_s, and a density normalization equal to ρ_0.

The stellar tracer distribution is represented by a `Plummer` density profile with total normalization

$$M_\star = 1$$

and scale radius $r_\star$.

The stellar distribution function is represented by a `QuasiSpherical` model conditioned on the dark matter potential and stellar tracer density. Its central anisotropy parameter is fixed to

$$\beta_0 = 0,$$

while its anisotropy radius is set to the sampled value of r_a.

These components are combined into one AGAMA galaxy model. The model is then sampled using the galaxy-specific value of $N_\star$ to obtain the stellar positions and velocities:

$$\{\boldsymbol{s}_i\}_{i=1}^{N_\star},$$

where $\boldsymbol{s}_i$ contains the sampled phase-space information for star i.

3.3 Dataset Construction and Storage

The parameter values for all galaxies are stored in a two-dimensional array with shape

$$(N_{\text{galaxies}}, N_{\text{parameters}}) = (N_{\text{galaxies}}, 5).$$

Each row contains

$$(\rho_0, \gamma, r_s, r_\star, r_a)$$

for one dwarf galaxy. The parameter array is stored using the `float32` numerical data type.

The stellar data cannot be stored directly in one rectangular numerical array because $N_\star$ changes between galaxies. If galaxy j contains $N_{\star,j}$ stars, its stellar array has a different first dimension from galaxies with other values of $N_\star$. Padding all galaxies to the same number of stars is not used.

Instead, the stellar data are stored in a one-dimensional NumPy object array with shape

$$(N_{\text{galaxies}},).$$

Element j of this object array contains the complete stellar array for galaxy j:

$$\text{galaxy } j \longrightarrow \{\boldsymbol{s}_{j,i}\}_{i=1}^{N_{\star,j}}.$$

This representation allows every galaxy to retain its original number of stars.

The parameter sampling, AGAMA configuration, stellar sampling, and storage procedure are repeated until the required number of galaxies N_{galaxies} has been generated. The procedure can therefore create datasets of different requested sizes, including separate training and test datasets.

Finally, the parameter array and variable-length stellar object array are saved together in a compressed NumPy `.npz` archive for later preprocessing and model training.

4 Model Pipeline

The model pipeline converts the variable-sized stellar data stored for each dwarf galaxy into standardized graph objects that can be processed by GraphNPE. The main stages are feature selection, observational-noise injection, data splitting, parameter transformation, standardization, graph construction, and graph batching.

4.1 Loading and Selecting the Stellar Features

The stored dataset contains the full six-dimensional phase-space information of every sampled star:

$$\boldsymbol{s}_i = (x_i, y_i, z_i, v_{x,i}, v_{y,i}, v_{z,i}).$$

The model uses five of these six quantities. The line-of-sight position z is removed, leaving the input vector

$$\boldsymbol{x}_i = (x_i, y_i, v_{x,i}, v_{y,i}, v_{z,i}).$$

Here, (x_i, y_i) are the exact Cartesian position components relative to the centre of the galaxy and perpendicular to the adopted z-axis.

$$v_{\mathrm{LOS},i} = v_{z,i}.$$

The retained stellar measurements for galaxy j therefore form an array

$$\mathbf{X}_j \in \mathbb{R}^{N_{\star,j}\times 5},$$

where $N_{\star,j}$ is the number of stars in that galaxy. Because $N_{\star,j}$ varies between galaxies, the first dimension of $\mathbf{X}_j$ is also variable.

4.2 Line-of-Sight Velocity Noise

Gaussian observational noise is added only to the line-of-sight velocity component. For every star, the noisy line-of-sight velocity is calculated as

$$v^{\mathrm{noisy}}_{\mathrm{LOS},i} = v_{\mathrm{LOS},i} + \epsilon_i,$$

where

$$\epsilon_i \sim \mathcal{N}(0, 0.1^2).$$

Thus, the noise has a mean of zero and a standard deviation of 0.1 in the velocity units used by the dataset. The other four input features are unchanged. The resulting input vector is

$$\boldsymbol{x}^{\mathrm{noisy}}_i = (x_i, y_i, v_{x,i}, v_{y,i}, v^{\mathrm{noisy}}_{\mathrm{LOS},i}).$$

4.3 Training and Validation Split

An index array is created to identify every galaxy:

$$\boldsymbol{I} = (0, 1, \ldots, N_{\mathrm{gal}} - 1).$$

These indices are divided into training and validation subsets using `train_test_split` with

$$\texttt{train_size} = 0.8.$$

Therefore, 80% of the galaxy indices are assigned to the training subset, while the remaining 20% are assigned to the validation subset:

$$\boldsymbol{I} = \boldsymbol{I}_{\mathrm{train}} \cup \boldsymbol{I}_{\mathrm{val}}, \qquad \boldsymbol{I}_{\mathrm{train}} \cap \boldsymbol{I}_{\mathrm{val}} = \varnothing.$$

The split is performed using galaxy indices rather than individual stars. Consequently, all stars belonging to one dwarf galaxy remain in the same subset, preventing information from the same galaxy from appearing in both training and validation.

For a separately generated test dataset, the test indices are defined as

$$\boldsymbol{I}_{\mathrm{test}} = (0, 1, \ldots, N_{\mathrm{test}} - 1).$$

Original parameter	Transformed parameter
ρ_0	$\log_{10}\rho_0$
γ	γ
r_s	$\log_{10} r_s$
$r_\star$	$\log_{10} r_\star$
r_a	$\log_{10} r_a$

Table 2 Transformations applied to the target parameters before standardization.

4.4 Parameter Transformation

Each galaxy has the target parameter vector

$$\boldsymbol{\theta} = (\rho_0, \gamma, r_s, r_\star, r_a).$$

Before standardization, the positive scale parameters are transformed to base-10 logarithmic space. The transformed target vector is

$$\boldsymbol{\theta}' = \left(\log_{10}\rho_0, \gamma, \log_{10} r_s, \log_{10} r_\star, \log_{10} r_a\right).$$

The transformation applied to each component is summarized in Table 2.

The inner slope γ is not logarithmically transformed because its prior includes zero and negative values. The transformed target parameters for all galaxies are stored in an array with shape

$$(N_{\text{gal}}, 5)$$

and data type `float32`.

4.5 Standardization

Separate standard scalers are fitted to the stellar input features and the transformed target parameters. Both scalers are fitted using the training subset only. The same fitted scalers are subsequently applied to the training, validation, and test data. This prevents information from the validation or test subsets from influencing the preprocessing parameters.

4.5.1 Stellar-feature standardization

The stellar-feature scaler is fitted incrementally using all stars from the training galaxies. The incremental fitting is necessary because the galaxies contain different numbers of stars and are stored as separate arrays.

For feature d, the standardized value is

$$\widetilde{x}_{i,d} = \frac{x_{i,d} - \mu_{x,d}^{\text{train}}}{\sigma_{x,d}^{\text{train}}},$$

where $\mu_{x,d}^{\text{train}}$ and $\sigma_{x,d}^{\text{train}}$ are calculated from the stars belonging to the training galaxies.

The scaler is updated once for each training galaxy. After all training galaxies have been processed, it contains the mean and standard deviation of each of the five stellar features across the complete collection of training stars.

4.5.2 Target-parameter standardization

A second standard scaler is fitted to the transformed parameter vectors of the training galaxies:

$$\left\{\boldsymbol{\theta}'_j : j \in \boldsymbol{I}_{\text{train}}\right\}.$$

For transformed parameter d, the standardized target is

$$\widetilde{\theta}_{j,d} = \frac{\theta'_{j,d} - \mu^{\text{train}}_{\theta,d}}{\sigma^{\text{train}}_{\theta,d}}.$$

The standardized target vector remains five-dimensional:

$$\widetilde{\boldsymbol{\theta}}_j \in \mathbb{R}^5.$$

4.6 Graph Construction

A separate graph is constructed for each dwarf galaxy. The same graph-construction procedure is applied to the training, validation, and test subsets using their corresponding galaxy indices.

For galaxy j, the selected five-dimensional stellar features are first standardized using the stellar-feature scaler:

$$\widetilde{\mathbf{X}}_j = \text{Scaler}_x(\mathbf{X}_j).$$

The standardized array is converted into a `float32` PyTorch tensor and stored as the graph node-feature matrix:

$$\mathbf{X}_{\mathcal{G}_j} = \widetilde{\mathbf{X}}_j \in \mathbb{R}^{N_{\star,j}\times 5}.$$

The graph positions are extracted from the original, unstandardized projected coordinates:

$$\mathbf{P}_j = \left\{(x_i, y_i)\right\}_{i=1}^{N_{\star,j}} \in \mathbb{R}^{N_{\star,j}\times 2}.$$

Therefore, the graph edges are determined using the original projected positions, while the node features supplied to the model are standardized.

The transformed parameter vector belonging to galaxy j is standardized using the target scaler:

$$\widetilde{\boldsymbol{\theta}}_j = \text{Scaler}_y\left(\boldsymbol{\theta}'_j\right).$$

It is then converted into a `float32` target tensor with shape

$$(1, 5).$$

4.6.1 Nearest-neighbour edges

The graph connectivity is determined using a k-nearest-neighbour graph with

$$k = 20.$$

For every stellar node, the nearest nodes are identified from their projected positions (x, y). Self-connections are permitted during this step because the graph is constructed with

`loop=True`.

The initial k-nearest-neighbour graph is directed. It is converted into an undirected graph by adding the reverse of each directed edge. The final edge set therefore satisfies

$$(i, j) \in \mathcal{E} \quad \Longrightarrow \quad (j, i) \in \mathcal{E}.$$

The resulting graph object for galaxy j contains

$$\mathcal{G}_j = \left(\widetilde{\mathbf{X}}_j, \mathcal{E}_j, \widetilde{\boldsymbol{\theta}}_j, \mathbf{P}_j\right),$$

corresponding to the following PyTorch Geometric fields:

- `x`: the standardized five-dimensional node features;
- `edge_index`: the undirected k-nearest-neighbour edges;
- `y`: the standardized five-dimensional target vector;
- `pos`: the original two-dimensional projected positions.

Each completed graph is appended to a graph list. Separate graph lists are produced for the training, validation, and test subsets:

$$\mathcal{D}_{\text{train}} = \left\{\mathcal{G}_j : j \in \boldsymbol{I}_{\text{train}}\right\},$$

$$\mathcal{D}_{\text{val}} = \left\{\mathcal{G}_j : j \in \boldsymbol{I}_{\text{val}}\right\},$$

and

$$\mathcal{D}_{\text{test}} = \left\{\mathcal{G}_j : j \in \boldsymbol{I}_{\text{test}}\right\}.$$

4.7 Graph Batching and Model Input

The graph lists are passed to the PyTorch Geometric `DataLoader`. During batching, multiple dwarf-galaxy graphs are combined into one disconnected batched graph. PyTorch Geometric records which nodes belong to each galaxy, allowing graph-level pooling to produce a separate summary for every dwarf galaxy in the batch.

Each training batch therefore contains the standardized node features, graph edges, projected positions, galaxy-membership assignments, and standardized target parameters required by GraphNPE. These graph batches are then supplied to the model during training. The graph neural network converts each dwarf galaxy into a graph-level representation, and Neural Posterior Estimation uses that representation to learn the conditional posterior over

$$(\rho_0, \gamma, r_s, r_\star, r_a).$$

5 Model Evaluation

The model is evaluated using several complementary tests. These tests examine parameter-recovery accuracy, uncertainty in the reconstructed dark matter profiles, posterior correlations, marginal calibration, and joint-posterior calibration.

Before evaluation, the posterior samples are converted from their standardized representation back to the transformed parameter space using the inverse target scaler. The logarithmically transformed parameters are then converted back to their physical values. The resulting posterior samples have the form

$$\boldsymbol{\theta}^{(s)} = \left(\rho_0^{(s)}, \gamma^{(s)}, r_s^{(s)}, r_\star^{(s)}, r_a^{(s)}\right), \qquad s = 1, \ldots, N_{\text{post}}.$$

Every evaluation described below is applied separately to the two-dimensional and five-dimensional models. This allows their parameter accuracy, posterior uncertainty, and calibration to be compared using the same evaluation procedures.

5.1 Truth-versus-Predicted Test

The truth-versus-predicted test evaluates how accurately the model recovers each inferred parameter across the complete test set. For test galaxy j and parameter d, the posterior median is used as the predicted value:

$$\widehat{\theta}_{j,d} = \text{median} \left\{ \theta_{j,d}^{(s)} \right\}_{s=1}^{N_{\text{post}}} .$$

This produces one predicted value for each parameter of every test galaxy. For a test set containing N_{test} galaxies, there are therefore N_{test} posterior-median predictions for each parameter.

The galaxies are grouped into bins according to their true parameter values. Within each truth bin, the distribution of posterior-median predictions is summarized using its median and central 68% and 95% intervals. The resulting summaries are plotted against the true parameter values.

The diagonal relation

$$\widehat{\theta} = \theta_{\text{true}}$$

represents perfect parameter recovery. Systematic displacement from this diagonal indicates bias, while the widths of the 68% and 95% bands describe the variation in the posterior-median predictions across galaxies with similar true parameters. These bands summarize the distribution of point predictions within each bin; they are not posterior credible intervals for an individual galaxy.

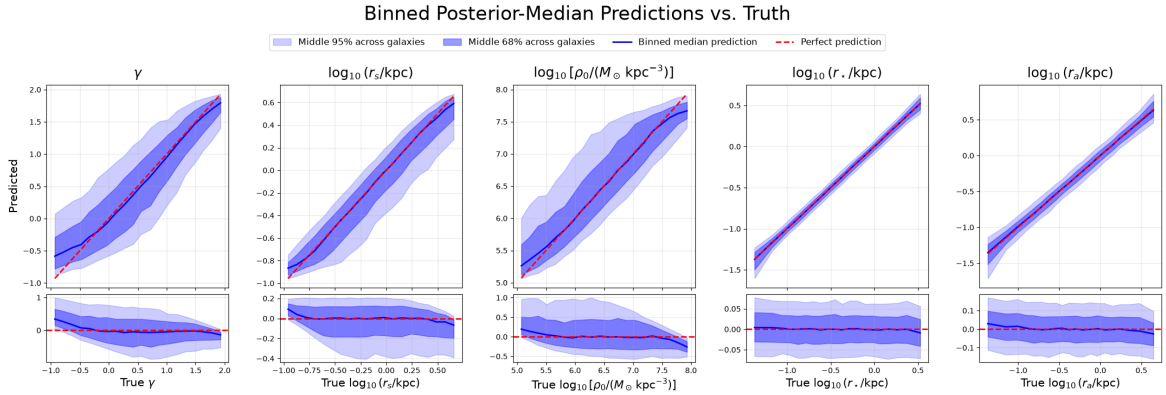


Figure 1 Truth-versus-predicted parameter recovery.

5.2 Dark Matter Density-Recovery Test

The density-recovery test evaluates whether the posterior distribution over the inferred parameters produces a dark matter density profile consistent with the true profile.

For one test galaxy, every posterior sample of

$$(\rho_0, \gamma, r_s)$$

defines one possible generalized Navarro–Frenk–White density profile:

$$\rho_{\text{DM}}^{(s)}(r) = \rho_{\text{DM}} \left(r; \rho_0^{(s)}, \gamma^{(s)}, r_s^{(s)} \right) .$$

At a fixed radius r, the gNFW density equation is evaluated using every posterior sample. This produces a posterior distribution of density values at that radius:

$$\left\{ \rho_{\text{DM}}^{(s)}(r) \right\}_{s=1}^{N_{\text{post}}} .$$

The median and central 68% and 95% credible intervals are calculated from this density posterior. The same calculation is repeated over the selected radial grid, producing the median reconstructed density profile and its pointwise credible bands:

$$\rho_{\mathrm{median}}(r), \qquad \left[\rho_{16}(r), \rho_{84}(r)\right], \qquad \left[\rho_{2.5}(r), \rho_{97.5}(r)\right].$$

The reconstructed posterior bands are plotted together with the true density profile. This comparison shows whether the model recovers the correct radial dependence and whether the posterior uncertainty contains the true profile over the radii of interest.

Because the density is calculated separately from each joint posterior sample, correlations among ρ_0, γ, and r_s are preserved during the reconstruction.

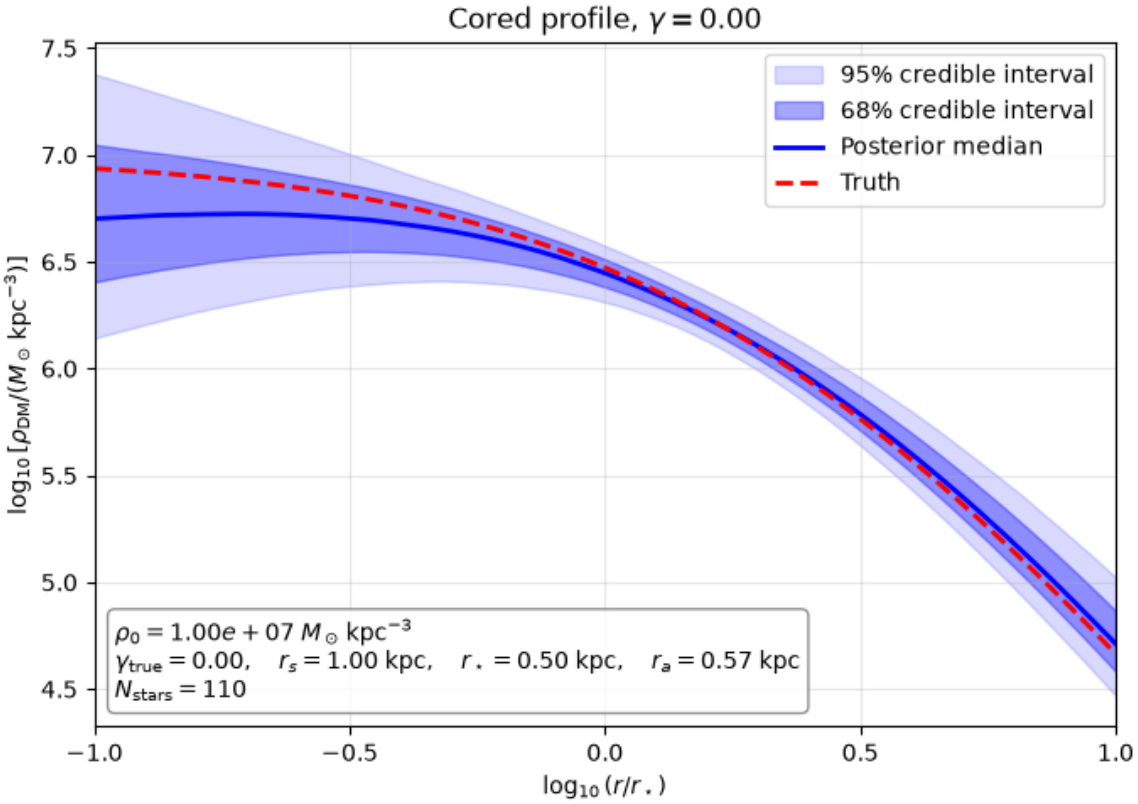


Figure 2 Dark matter density-profile recovery.

5.3 Dark Matter Enclosed-Mass Recovery Test

The enclosed-mass recovery test follows the same posterior-propagation procedure as the density-recovery test. For each posterior sample, the dark matter mass enclosed within radius r is calculated from the corresponding gNFW density profile:

$$M_{\mathrm{DM}}^{(s)}(< r) = 4\pi \int_0^r \rho_{\mathrm{DM}}^{(s)}(r')\, r'^2\, \mathrm{d}r'.$$

At each fixed radius, evaluating this equation for every posterior sample produces the enclosed-mass posterior

$$\left\{M_{\mathrm{DM}}^{(s)}(< r)\right\}_{s=1}^{N_{\mathrm{post}}}.$$

This posterior is summarized using its median and central 68% and 95% credible intervals. Repeating the procedure over the selected radial grid produces

$$M_{\mathrm{median}}(<r), \qquad [M_{16}(<r), M_{84}(<r)], \qquad [M_{2.5}(<r), M_{97.5}(<r)].$$

The median and credible bands are plotted as functions of radius and compared with the true enclosed-mass profile. This test evaluates whether the posterior provides an accurate and appropriately uncertain reconstruction of the cumulative dark matter distribution.

As in the density-recovery test, the calculation is performed using complete joint posterior samples. The correlations among ρ_0, γ, and r_s are therefore propagated into the enclosed-mass posterior.

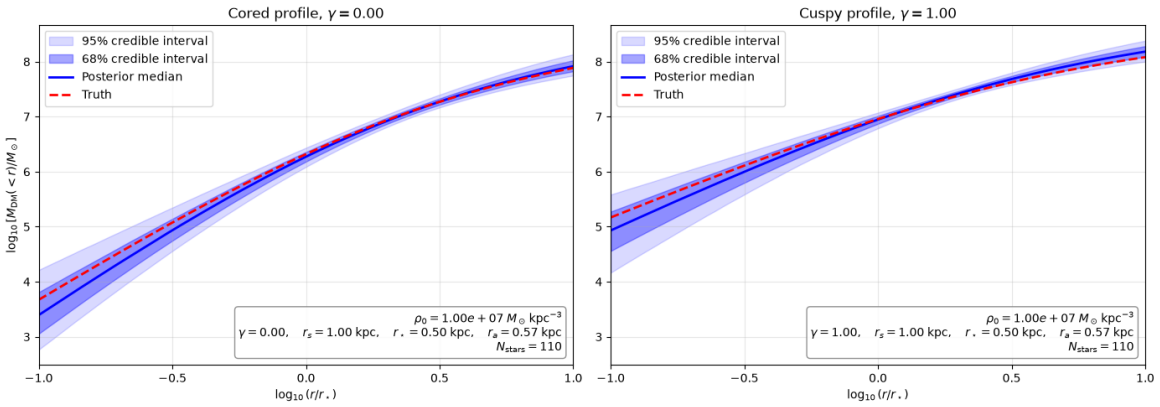


Figure 3 Dark matter enclosed-mass-profile recovery.

5.4 Corner Plots

Corner plots are used to inspect the marginal distributions and joint correlations of the inferred dark matter parameters:

$$(\log_{10}\rho_0, \gamma, \log_{10} r_s).$$

For a selected test galaxy, the corner plot is constructed using the posterior samples produced by the model. The diagonal panels show the one-dimensional marginal posterior of each parameter. The posterior median and central 68% and 95% credible intervals are calculated from the samples of the corresponding parameter.

The off-diagonal panels show the joint posterior distributions for each pair of parameters. Because the two coordinates of every plotted point are taken from the same joint posterior sample, these panels preserve and display the correlations learned by the model. The inner and outer contours summarize the central joint posterior regions.

The true value of each parameter is added to the corresponding diagonal panel. In the off-diagonal panels, the intersection of two true parameter values marks the true joint parameter location. This makes it possible to determine whether the true parameters lie within regions of high posterior support and to identify degeneracies among ρ_0, γ, and r_s.

Corner plots are generated for the same test galaxies using both the two-dimensional and five-dimensional models. Their comparison shows whether the additional input features reduce marginal uncertainty or weaken degeneracies among the inferred dark matter parameters.

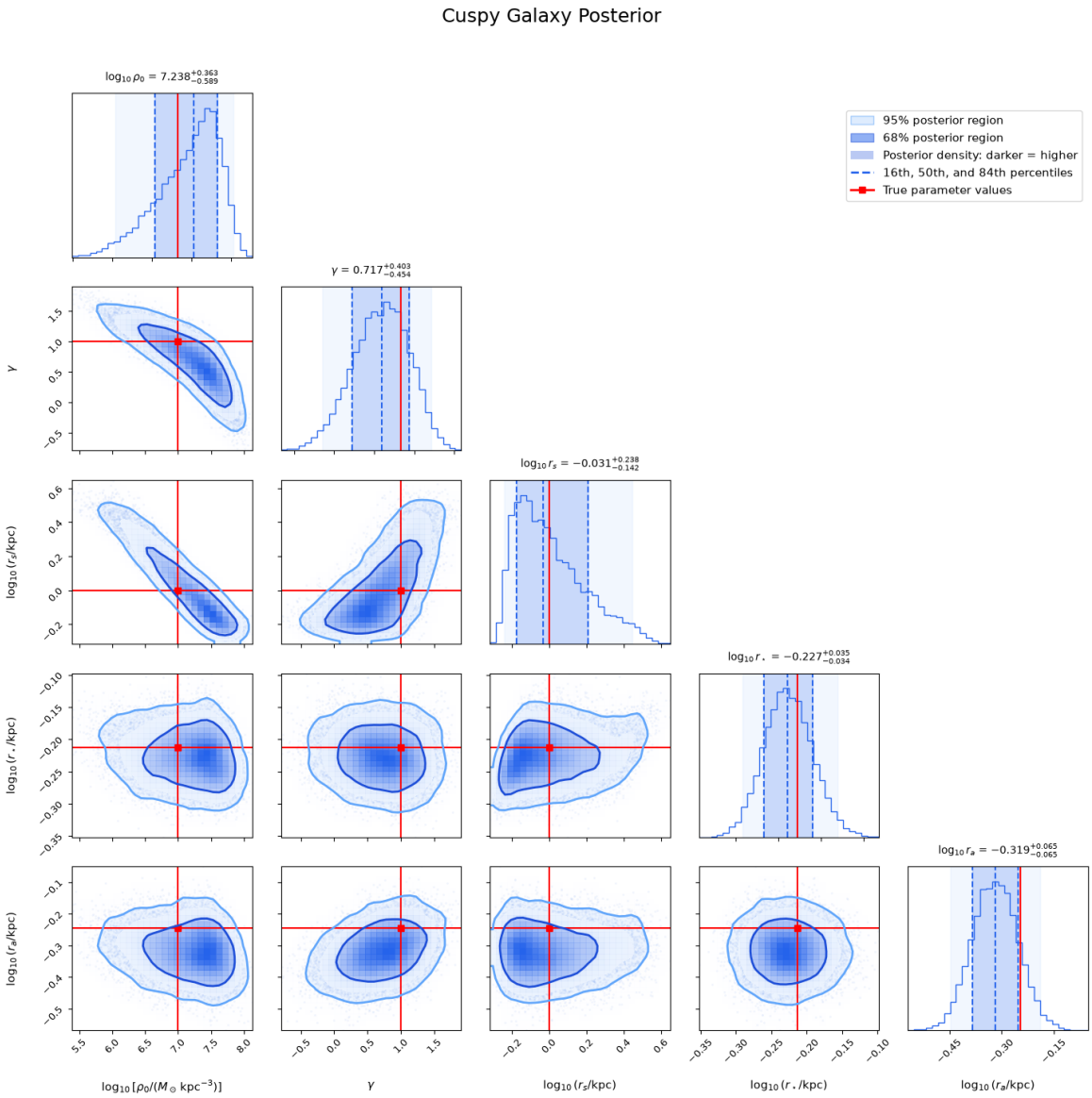


Figure 4 Corner plot of the inferred dark matter parameters.

5.5 One-Dimensional Rank-Based Calibration Test

The one-dimensional rank-based test evaluates the marginal calibration of each inferred parameter across the test galaxies.

For test galaxy j and parameter d, the normalized rank of the true value is calculated as

$$r_{j,d} = \frac{1}{N_{\text{post}}} \sum_{s=1}^{N_{\text{post}}} \mathbb{I}\left[\theta_{j,d}^{(s)} < \theta_{j,d}^{\text{true}}\right],$$

where $\mathbb{I}[\cdot]$ is the indicator function. The rank represents the fraction of posterior samples below the true parameter value.

If the marginal posterior is calibrated, the true values should appear uniformly throughout their corresponding posterior distributions. Therefore, the normalized ranks collected across the test galaxies should follow a uniform distribution:

$$r_{j,d} \sim \mathcal{U}(0,1).$$

Calibration is visualized by comparing the empirical cumulative distribution of the ranks with the ideal diagonal. At a percentile level q, a calibrated model should have approximately a fraction q of true values below the corresponding posterior percentile.

Systematic deviations from the diagonal indicate that some posterior percentiles contain the true values more or less frequently than expected. This may reveal parameter bias, posterior intervals that are too narrow, or posterior intervals that are too wide.

The test is performed independently for

$$\rho_0, \gamma, r_s, r_\star, \text{ and } r_a.$$

It therefore evaluates the calibration of each marginal posterior but does not directly test whether the complete five-dimensional joint posterior is calibrated.

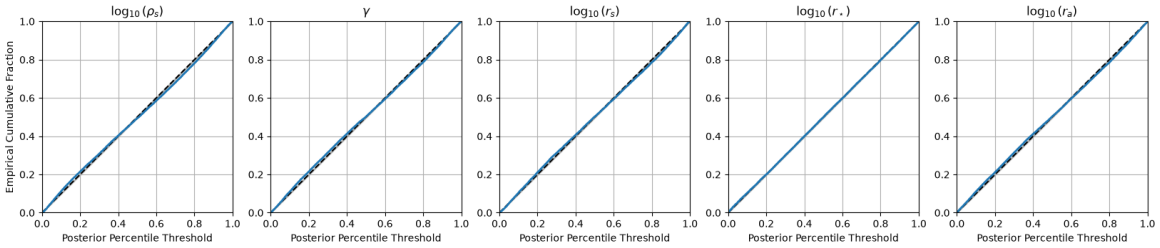


Figure 5 One-dimensional rank-based calibration test.

5.6 TARP Joint-Calibration Test

Tests of Accuracy with Random Points (TARP) are used to evaluate calibration in the complete joint parameter space. Unlike the one-dimensional rank test, TARP considers all components of

$$\boldsymbol{\theta} = (\rho_0, \gamma, r_s, r_\star, r_a)$$

simultaneously. It can therefore detect errors in the joint posterior structure, including incorrect correlations that may not appear in the marginal rank tests.

Before calculating distances, the parameter dimensions are placed on comparable numerical scales. For each test galaxy, a reference point is selected in the joint parameter space. The distance between this reference point and the true parameter vector defines a region around the reference point.

The distance between the same reference point and each joint posterior sample is then calculated. A posterior sample is counted as being inside the region when its distance from the reference point satisfies the corresponding TARP-region condition. The fraction of posterior samples inside the region provides the posterior mass associated with that region.

This procedure is repeated across the test galaxies and reference points to construct an empirical expected-coverage curve. For a calibrated joint posterior, the empirical coverage curve should follow the ideal diagonal:

$$\text{empirical coverage} = \text{expected coverage}.$$

A curve that deviates systematically from the diagonal indicates joint-posterior miscalibration. TARP complements the one-dimensional rank tests by evaluating whether the complete parameter combinations, rather than only their individual components, have the expected coverage.

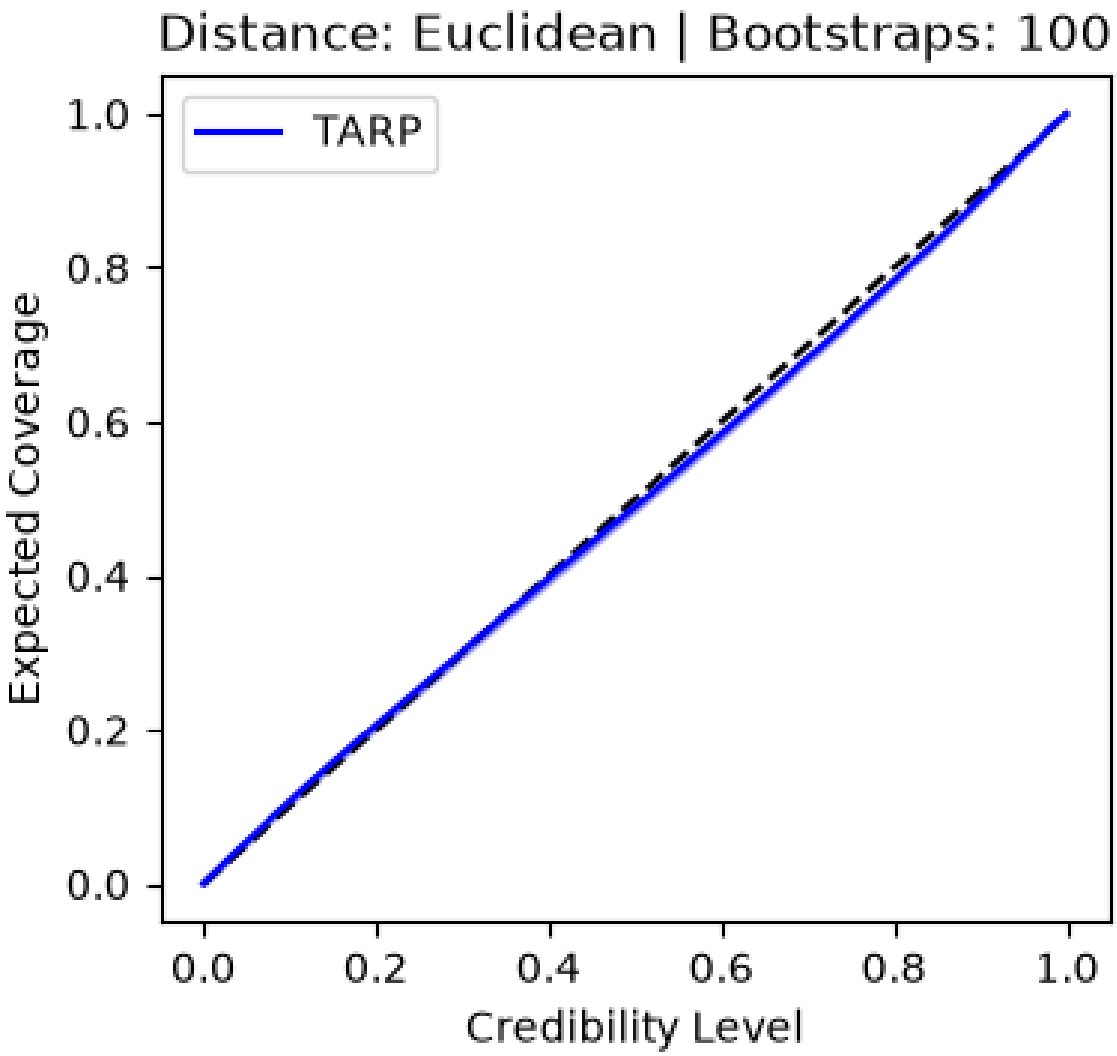


Figure 6 TARP joint-posterior calibration test.

5.7 MIRA Score

MIRA, or Mass in Random Areas, is a sample-based score for evaluating the accuracy of a conditional distribution. In this application, the candidate conditional distribution is the posterior predicted by the model:

$$q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}).$$

MIRA requires known parameter–observation pairs from the test set and samples from the posterior predicted for each observation. It does not require an explicit evaluation of the posterior density.

For each test galaxy, MIRA constructs regions in the joint parameter space and compares two quantities: whether the true parameter vector lies inside a region and how much posterior-sample probability mass is assigned to that region. These comparisons are aggregated across regions and test galaxies to produce a scalar score.

When the candidate posterior matches the true conditional distribution, the MIRA score has a

theoretical asymptotic reference value of

$$\mu_{\mathrm{MIRA}} = \frac{2}{3}.$$

The score should be interpreted by comparing its distance from this reference value with its estimated uncertainty. A larger numerical score is not automatically better because different forms of posterior miscalibration can move the score above or below the reference value.

In this work, MIRA is calculated using

$$N_{\mathrm{MIRA}} = 1{,}000$$

joint posterior samples for each test galaxy. The calculation uses the complete five-dimensional parameter vector

$$(\rho_0, \gamma, r_s, r_\star, r_a).$$

Uncertainty in the aggregate MIRA score is estimated by resampling the test galaxies, allowing the measured score to be compared with its theoretical reference.

The MIRA test is applied separately to the two-dimensional and five-dimensional models using the same test parameter–galaxy pairs and the same number of posterior samples. The model whose MIRA score is more consistent with the theoretical reference, after accounting for uncertainty, provides the more accurate approximation to the conditional posterior according to this diagnostic.

MIRA is used together with the rank-based and TARP tests rather than as a replacement for them. The one-dimensional rank tests evaluate marginal calibration, TARP evaluates joint coverage, and MIRA provides an additional scalar measure of conditional-posterior accuracy and a direct method for comparing the two models.

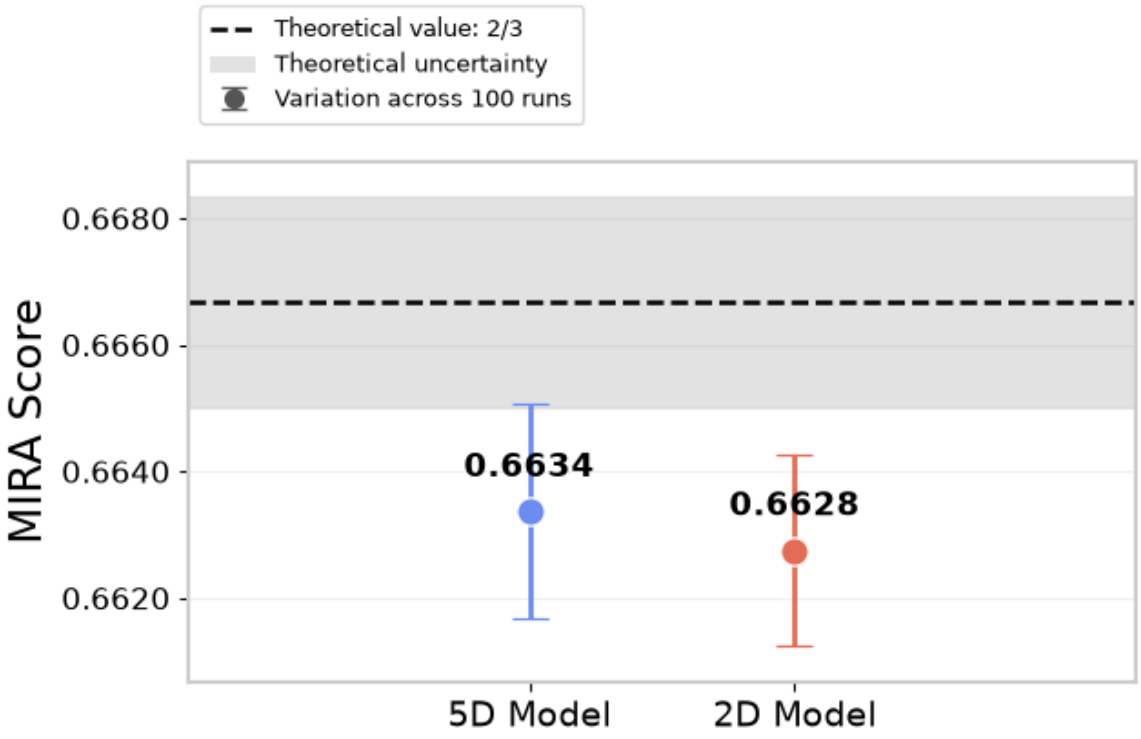


Figure 7 MIRA scores for the two-dimensional and five-dimensional models.

6 Results

The current five-dimensional model uses the simulated physical inputs

$$(x, y, v_{\mathrm{LOS}}, v_x, v_y),$$

rather than the directly observable sky coordinates and proper motions

$$(\mathrm{RA}, \mathrm{Dec}, v_{\mathrm{LOS}}, \mu_{\mathrm{RA}^*}, \mu_{\mathrm{Dec}}).$$

Therefore, the present results evaluate the benefit of including additional simulated velocity components. They do not yet represent an application to angular proper-motion measurements.

6.1 Comparison of the Five-Dimensional and Two-Dimensional Models

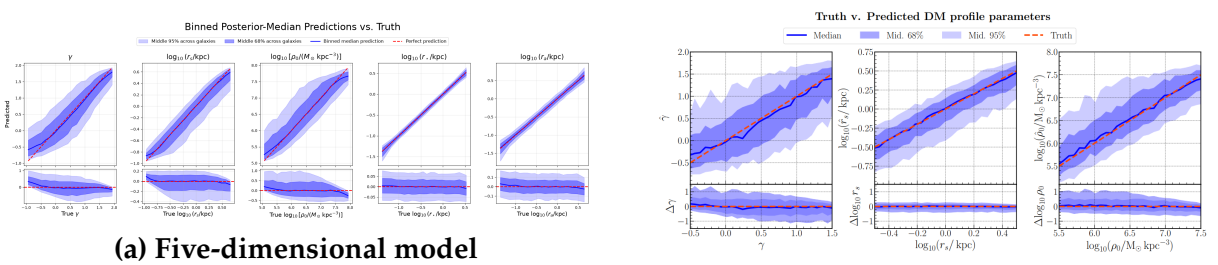
Across 20,000 test galaxies, the posterior medians predicted by the five-dimensional model closely track the true dark matter parameters. The five-dimensional model also achieved a lower test loss than the two-dimensional baseline, which uses

$$(R, v_{\mathrm{LOS}}).$$

The test losses were

$$\mathcal{L}^{\mathrm{5D}}_{\mathrm{test}} = -3.78, \qquad \mathcal{L}^{\mathrm{2D}}_{\mathrm{test}} = -0.14.$$

Figure 8 compares the truth-versus-predicted results of the five-dimensional model with those of the two-dimensional baseline in the paper.



(a) Five-dimensional model

(b) Two-dimensional baseline

Figure 8 Truth-versus-predicted parameter recovery for the five-dimensional model and the two-dimensional baseline in the paper.

6.2 TARP Joint Calibration

The TARP expected-coverage curve for the five-dimensional model remains close to the ideal-calibration diagonal. This indicates that the predicted five-dimensional joint posteriors are well calibrated over the test galaxies.

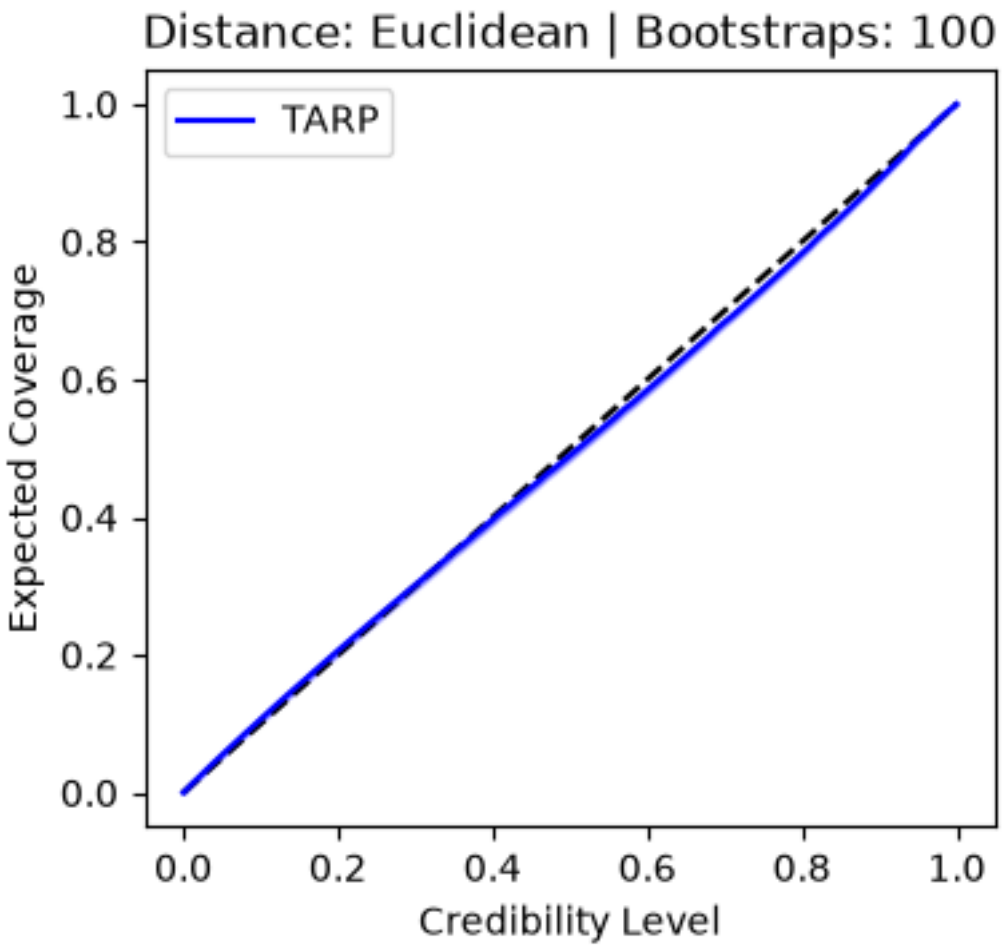


Figure 9 TARP joint-calibration results for the five-dimensional model.

6.3 MIRA Scores

The MIRA test was calculated using 1,000 joint posterior samples per test galaxy. The resulting scores were

$$\mu^{5D}_{\mathrm{MIRA}} = 0.6634, \qquad \mu^{2D}_{\mathrm{MIRA}} = 0.6628.$$

The two scores are nearly identical and are both close to the theoretical reference value

$$\mu_{\mathrm{MIRA}} = \frac{2}{3}$$

for an accurate conditional posterior. Therefore, according to the MIRA diagnostic, the two models have similar conditional-posterior accuracy despite the lower test loss achieved by the five-dimensional model.

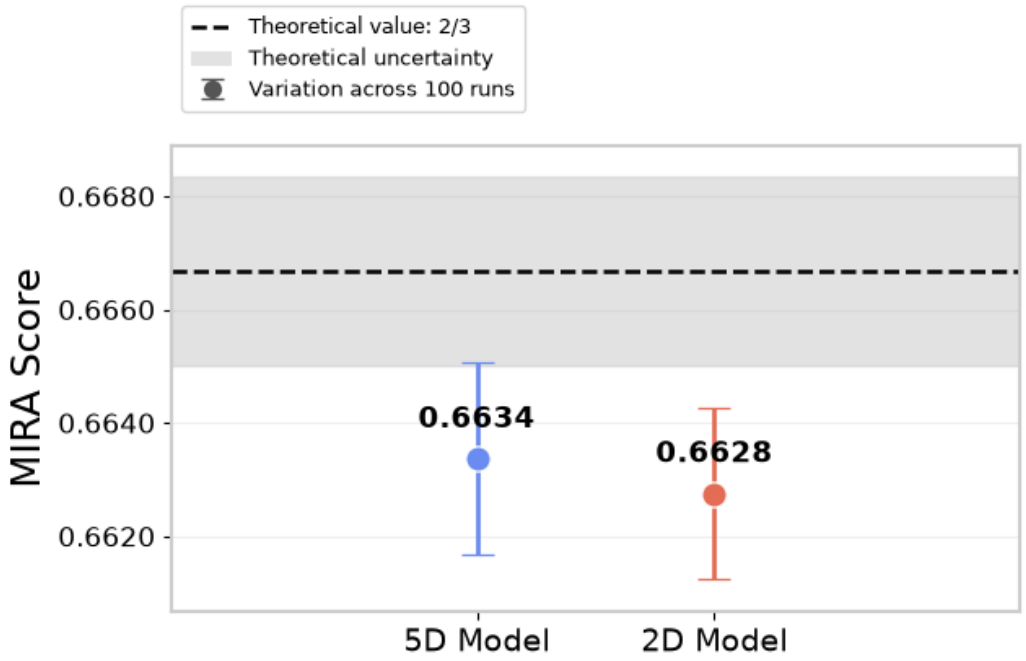


Figure 10 Comparison of the MIRA scores obtained by the five-dimensional and two-dimensional models.

6.4 Key Takeaways

- The five-dimensional model achieved a lower test loss than the two-dimensional baseline, while the two models obtained similar MIRA scores.
- On the simulated test galaxies, the five-dimensional model closely recovered the true dark matter parameters and produced well-calibrated joint posterior distributions.

7 Limitations and Future Work

The current results demonstrate the ability of GraphNPE to infer dark matter parameters from simulated stellar data. However, the present analysis is limited to idealized simulations, physical velocity inputs, and one principal neural-network architecture. Several extensions are therefore required before the method can be applied reliably to observational data.

7.1 Alternative Neural-Network Architectures

Future work will compare GraphNPE with alternative neural-network architectures, including multilayer perceptrons (MLPs) and Deep Sets.

An MLP would provide a simple non-graph baseline but requires each galaxy to be represented by a fixed-size input. This could be achieved using summary statistics, padding, or a fixed number of selected stars. A Deep Sets model can directly process variable-sized stellar catalogues using permutation-invariant operations, without constructing graph edges.

The MLP, Deep Sets, and graph neural network models should be trained using the same simulated galaxies, parameter priors, preprocessing, and posterior density estimator. Their posterior accuracy, uncertainty, calibration, and computational cost can then be compared. This comparison will determine whether the connections between neighbouring stars provide additional information beyond the individual stellar features and permutation-invariant pooling.

7.2 Comparison with Jeans Analysis

The five-dimensional inference model should also be compared with conventional Jeans analysis. Both methods should be evaluated using the same simulated test galaxies, parameter priors, stellar measurements, and measurement uncertainties.

The comparison should examine the recovery of

$$(\rho_0, \gamma, r_s, r_\star, r_a),$$

together with the reconstructed dark matter density and enclosed-mass profiles. Parameter bias, posterior uncertainty, credible-region coverage, and calibration should be evaluated for both methods.

This comparison will determine whether the five-dimensional SBI approach produces more accurate or more informative posterior constraints than Jeans analysis, particularly when transverse stellar motions are included.

7.3 Conversion to Observable Proper Motions

The current five-dimensional model uses the physical transverse velocity components

$$(v_x, v_y)$$

as inputs. These quantities are not directly equivalent to the observable angular proper motions

$$(\mu_{\mathrm{RA}^*}, \mu_{\mathrm{Dec}}).$$

Future simulations will convert the transverse velocities into proper motions using the distance to the dwarf galaxy. For a transverse velocity $v_\perp$ in km s^{-1} and distance D in kpc, the corresponding proper motion is

$$\mu = \frac{v_\perp}{4.74047 D} \quad \mathrm{mas\,yr^{-1}}.$$

The conversion must also account for the orientation of the galaxy relative to the observer so that the transverse velocities are expressed in the right-ascension and declination directions. Realistic uncertainties should then be added to the sky positions, line-of-sight velocities, proper motions, and galaxy distance.

The resulting model inputs will be

$$(\mathrm{RA}, \mathrm{Dec}, v_{\mathrm{LOS}}, \mu_{\mathrm{RA}^*}, \mu_{\mathrm{Dec}}),$$

allowing the method to be evaluated using quantities that correspond directly to astronomical observations.

7.4 Application to Real Dwarf Galaxies

After incorporating observable coordinates, proper motions, and realistic measurement uncertainties, the model should be tested on real dwarf galaxies. Draco is a suitable target because it has measured stellar positions, line-of-sight velocities, and proper-motion information.

Applying the model to Draco will require additional observational effects to be represented in the simulations, including measurement uncertainties, foreground contamination, uncertain stellar membership, incomplete spatial sampling, and the observational selection function. Differences between the simulated and real stellar populations must also be examined to reduce the risk of applying the model outside the simulation distribution.

The inferred posterior should be checked using calibration tests on realistic mock observations before interpreting the result obtained from Draco. The resulting dark matter density and enclosed-mass profiles can then be compared with constraints obtained from Jeans analysis and previous observational studies.

8 Conclusion

This work applied Graph Neural Posterior Estimation to infer

$$(\rho_0, \gamma, r_s, r_\star, r_a)$$

from simulated dwarf-galaxy stellar data. Each galaxy was represented as a graph, with stars as nodes connected to their 20 nearest neighbours. The model used simulation-based inference to predict joint parameter posteriors and reconstruct the corresponding dark matter density and enclosed-mass profiles.

Across 20,000 test galaxies, the five-dimensional model closely recovered the true dark matter parameters and achieved a lower test loss than the two-dimensional baseline (-3.78 compared with -0.14). TARP indicated that the five-dimensional posteriors were well calibrated. However, the MIRA scores were nearly identical for the five-dimensional and two-dimensional models (0.6634 and 0.6628), indicating similar posterior accuracy according to this measure.

The current model uses physical transverse velocities rather than observable proper motions. Future work will introduce realistic proper-motion measurements and uncertainties, compare alternative architectures and Jeans analysis, and test the method on real dwarf galaxies such as Draco.

9 Related Work

Chang and Necib used spherical Jeans modelling to infer dark matter density profiles from simulated stellar positions and line-of-sight velocities [1]. Although Jeans analysis is physically interpretable, it depends on assumptions such as spherical symmetry and dynamical equilibrium, and it can have difficulty distinguishing cored and cusped profiles.

Nguyen et al. introduced GraphNPE, which combines graph neural networks with simulation-based inference to estimate dark matter profile parameters from individual stellar observations [2]. Our project extends this method by adding two proper-motion components and testing whether these motions across the sky improve density-profile recovery.

References

[1] Laura J. Chang and Lina Necib. “Dark Matter Density Profiles in Dwarf Galaxies: Linking Jeans Modelling Systematics and Observation”. In: *Monthly Notices of the Royal Astronomical Society* 507 (2021), pp. 4715–4733. DOI: 10.1093/mnras/stab2440. URL: https://arxiv.org/abs/2009.00613.

[2] Tri Nguyen et al. “Uncovering Dark Matter Density Profiles in Dwarf Galaxies with Graph Neural Networks”. In: *Physical Review D* 107 (2023), p. 043015. DOI: 10.1103/PhysRevD.107.043015. URL: https://arxiv.org/abs/2208.12825.