

Trial by FIRE: Probing the dark matter density profile of dwarf galaxies with GraphNPE

Tri Nguyen,^{1,2,3,4,5*} Justin Read,⁶ Lina Necib,^{4,5} Siddharth Mishra-Sharma^{3,5†},

Claude-André Faucher-Giguère^{1,2}, and Andrew Wetzel⁹

¹Center for Interdisciplinary Exploration and Research in Astrophysics, Northwestern University, 1800 Sherman Ave, Evanston, IL 60201

²NSF-Simons AI Institute for the Sky, 172 E. Chestnut St., Chicago, IL 60611, USA

³Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁴Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁵NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Cambridge, MA 02139, USA

⁶Department of Physics, University of Surrey, Guildford, GU2 7XH, UK

⁷Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁸Department of Physics, Harvard University, Cambridge, MA 02138, USA

⁹Department of Physics & Astronomy, University of California, Davis, CA 95616, USA

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The Dark Matter (DM) distribution in dwarf galaxies provides crucial insights into both structure formation and the particle nature of DM. GraphNPE (Graph Neural Posterior Estimator), first introduced in Nguyen et al. (2023), is a novel simulation-based inference framework that combines graph neural networks and normalizing flows to infer the DM density profile from line-of-sight stellar velocities. Here, we apply GraphNPE to satellite dwarf galaxies in the FIRE-2 Latte simulation suite of Milky Way-mass halos, testing it against both Cold and Self-Interacting DM scenarios. Our method demonstrates superior precision compared to conventional Jeans-based approaches, recovering DM density profiles to within the 95% confidence level even in systems with as few as 30 tracers. Moreover, we present the first evaluation of mass modeling methods in constraining two key parameters from realistic simulations: the peak circular velocity, V_{\max} , and the peak virial mass, M_{200m}^{peak} . Using only line-of-sight velocities, GraphNPE can reliably recover both V_{\max} and M_{200m}^{peak} within our quoted uncertainties, including those experiencing tidal effects ($\gtrsim 63\%$ of systems are recovered with our 68% confidence intervals and $\gtrsim 92\%$ within our 95% confidence intervals). The method achieves 10 – 20% accuracy in V_{\max} recovery, while M_{200m}^{peak} is recovered to 0.1 – 0.4 dex accuracy. This work establishes GraphNPE as a robust tool for inferring DM density profiles in dwarf galaxies, offering promising avenues for constraining DM models. The framework’s potential extends beyond this study, as it can be adapted to non-spherical and disequilibrium models, showcasing the broader utility of simulation-based inference and graph-based learning in astrophysics.

Key words: dark matter – galaxies: dwarf – galaxies: structure – stars: kinematics and dynamics

1 INTRODUCTION

Dark matter (DM) makes up about 85% of matter in the Universe, playing an integral role in the formation and evolution of galaxies and clusters (White & Rees 1978; Navarro et al. 1997; Subramanian et al. 2000; Dekel et al. 2003; Springel et al. 2005). Despite its significance, the particle nature of DM remains among the greatest outstanding questions in modern astrophysics and cosmology. While the standard Lambda Cold Dark Matter (Λ CDM) paradigm has successfully explained the large-scale structure of the Universe, such as the fluctuations in the cosmic microwave background (CMB) (e.g. Bennett et al. 2013; Planck Collaboration et al. 2020) and the distribution of galaxy clusters (e.g. Tinker et al. 2008; Rozo et al. 2010), dis-

crepancies arise at the small scales. Observations of dwarf galaxies in the Local Group have uncovered small-scale challenges, including the core-cusp problem (Flores & Primack 1994; Moore 1994; Navarro et al. 1996; Read & Gilmore 2005; Spekkens et al. 2005; Oh et al. 2011; Pontzen & Governato 2014; Oh et al. 2015; Read et al. 2019; De Leo et al. 2024), missing satellites problem (Klypin et al. 1999; Moore et al. 1999; Read & Erkal 2019), too-big-to-fail (Boylan-Kolchin et al. 2011), satellite plane (Kunkel & Demers 1976; Lynden-Bell 1976), and diversity of shapes problem (Oman et al. 2015; Creasey et al. 2017; Zavala et al. 2019; Hayashi et al. 2020a) (for a review, refer to Bullock & Boylan-Kolchin 2017). These challenges can be partially addressed by a more complete understanding of baryonic processes (Governato et al. 2010; Pontzen & Governato 2012; Fitts et al. 2017; Garrison-Kimmel et al. 2017; Kim et al. 2018; Simpson et al. 2018; Garrison-Kimmel et al. 2019; Buck et al. 2019; Sales et al. 2022), environmental effects such as

* E-mail: trivnguyen@northwestern.edu

† Currently at Anthropic; worked performed while at MIT/IAIFI.

tidal interaction with the Milky Way and other galaxies (Boylan-Kolchin et al. 2011; Mayer et al. 2006; Read et al. 2006), and survey selection effects (Tollerud et al. 2008; Drlica-Wagner et al. 2020). However, open questions still remain (Pawlowski et al. 2015; Santos-Santos et al. 2020), motivating alternative theories of DM such as self-interaction (Spergel & Steinhardt 2000).

Characterized by their high mass-to-light ratio, small size, and proximity to the Galaxy, the Milky Way’s satellite galaxies present an exciting frontier for exploring and testing the particle nature of DM. Key observables, including the satellite’s inner DM density profile, their mass function, luminosity function, and phase-space distribution, can offer stringent constraints on DM properties, such as self-interaction cross-section (Rocha et al. 2013; Kaplinghat et al. 2016; Tulin & Yu 2018; Read et al. 2018; Correa 2021; Toguz et al. 2022; Júlio et al. 2023), DM dissipation (Kaplan et al. 2010; Cyr-Racine & Sigurdson 2013; Shen et al. 2021, 2024; Gemmell et al. 2024; Roy et al. 2024), DM-baryon interaction (Nadler et al. 2019), and mass of warm and fuzzy DM (Bode et al. 2001; Kennedy et al. 2014; Lovell et al. 2014; Bose et al. 2016; Alvey et al. 2021; Hayashi et al. 2021; Dalal & Kravtsov 2022; Tan et al. 2024).

The inner density profile of DM in a galaxy is affected by two sources: the DM particle properties and interactions, and the baryonic processes. In the absence of baryons, cold DM models are expected to exhibit a Navarro-Frenk-White (NFW) profile, where the inner profile is r^{-1} (Navarro et al. 1997). Self-interacting DM (SIDM) models, however, largely predict a cored profile, where $r \propto r^0 \propto \text{constant}$ in the inner galaxy¹ (Tulin & Yu 2018). Dissipation can instead lead to stronger cusps, for example Shen et al. (2021) found cusps that scale as $r^{-1.5}$.

Baryonic processes, such as feedback, can also significantly influence the density profiles (e.g. Di Cintio et al. 2014a,b; Read et al. 2019; De Leo et al. 2024). Feedback can lead to a diverse range of cored and cuspy profiles, depending on where a galaxy is in its cycle of gas inflows and outflows (El-Badry et al. 2016, 2017; Muni et al. 2025). However, if the gas is stripped, such as in Milky Way satellites, the galaxy can be “frozen into” a cored state. This effect is particularly relevant for a specific mass range, including classical dwarfs and Milky Way-mass galaxies (e.g. Lazar et al. 2020). Consequently, dwarf galaxies, particularly the ultra-faint, are the ideal laboratory for investigating the inner profile affected solely by the particle nature of DM (e.g. Zoutendijk et al. 2021b,a).

Apart from serving as a direct probe of the particle nature of DM through its inner slope of the DM density profile, dwarf galaxies are an excellent target for indirect detection of DM. They are relatively close, making the annihilation/decay rates of DM into standard model particles quite high and more easily detectable, and generally devoid of astrophysical backgrounds that might mimic DM signals (e.g., the potential population of millisecond pulsars near the Galactic Center muddling the waters as to whether DM is the source of the Galactic center excess (Cirelli et al. 2011; Ackermann et al. 2011; Geringer-Sameth & Koushiappas 2011; Mazziotta et al. 2012; Ackermann et al. 2014; Martinez 2015; Ackermann et al. 2015; Geringer-Sameth et al. 2015a,b; Anderson et al. 2015; Leane & Slatyer 2019; Ajello et al. 2024)).

Given the critical information that is encoded in the density profile of DM, one needs to work through recovering such profiles, particularly in dwarf galaxies where the effect of baryons is minimal.

¹ In some cases, there are scenarios where gravothermal collapse occurs, leading to a strong cusp in the center of galaxies with SIDM (e.g. Balberg et al. 2002; Koda & Shapiro 2011; Nishikawa et al. 2020).

To tackle such problem, one could recover the underlying gravitational potential of the galaxy through the dynamics of its stars. Traditionally, this has been done using the Jeans dynamical modeling techniques (Jeans 1915; Bonnivard et al. 2015), which relate the velocity dispersion of tracer stars within these galaxies to their DM-dominated gravitational potential. Jeans methods are straightforward to implement. Most models assume spherical symmetry, but tests on non-spherical mocks show that for less than $\sim 10,000$ tracer stars, this is reasonable approximation (Read & Steger 2017; Genina et al. 2020). Some models have explored axisymmetric fits to nearby dwarfs (Hayashi et al. 2020b, 2023), while even highly tidally disrupting dwarfs have been shown to be successfully modeled, providing contaminating tidal debris can be correctly distinguished from bound stars (Read et al. 2018; De Leo et al. 2024).

A long-standing challenge for Jeans methods, applied to just line-of-sight velocity data, has been *mass-anisotropy degeneracy* (Merrifield & Kent 1990; Wilkinson et al. 2002; Łokas & Mamon 2003; de Lorenzi et al. 2009; Read & Steger 2017; El-Badry et al. 2017; Genina et al. 2020). This refers to a degeneracy between the enclosed mass, that we want to infer, $M(< r)$, and the velocity anisotropy parameter $\beta(r)$ ($\beta = 0, 1, -\infty$ corresponds to an isotropic, radial anisotropy, and tangential anisotropy velocity dispersion, respectively; Binney & Tremaine 2008). However, this degeneracy can be broken in a number of ways. Firstly, where available, proper motion data can be used (Strigari et al. 2007b; Read & Steger 2017; Strigari et al. 2018; Massari et al. 2020; Vitali et al. 2024). Secondly, we can use higher order moments of the velocity distribution, for example through “Virial Shape Parameters” (Merrifield & Kent 1990; Łokas 2009; Richardson & Fairbairn 2014; Read & Steger 2017; Mamon et al. 2013). Thirdly, we can simultaneously fit multiple tracer populations moving within the same gravitational potential (e.g. Walker & Peñarrubia 2011; Amorisco & Evans 2012; Zhu et al. 2016; Read & Steger 2017). And finally, we can move away from Jeans modeling to Schwarzschild (e.g. Breddels & Helmi 2013, 2014; Kowalczyk et al. 2019) or distribution function modeling (Wilkinson et al. 2002; Pascale et al. 2018; Read et al. 2021). These latter methods simultaneously fit all moments of the velocity distribution function, $f(v)$, by assuming some functional form for f (or a range of functional forms).

Distribution function methods maximize the information content in the data, while providing a natural way to incorporate survey selection functions and binary star contamination (e.g. Spencer et al. 2018; Kervick et al. 2022). However, previously proposed methods require models that are expensive to compute, limiting the parameter space that can be explored, especially as we move beyond spherical models. Traditional distribution function methods also provide no natural route to moving beyond the assumption of pseudo-dynamical equilibrium.

In this paper, we further improve and test a new method, GRAPHNPE, a simulation-based inference (SBI) framework that can massively accelerate distribution function modeling. First introduced by Nguyen et al. 2023 (hereafter N23), GRAPHNPE employs normalizing flows (Papamakarios & Murray 2016; Papamakarios et al. 2019) and graph neural networks (Scarselli et al. 2009; Kipf & Welling 2016; Gilmer et al. 2017; Battaglia et al. 2018) to parametrically model the density profiles. SBI uses complex simulations to implicitly model the likelihood, thus directly incorporating these simulations into the inference process (e.g. Cranmer et al. 2020). In this work, we train GRAPHNPE on idealized simulations of dwarf galaxies that are spherical and in equilibrium. However, we emphasize that the framework can be readily adapted to analyze non-spherical

or disequilibrium mock datasets, which we plan to explore in future studies.

Mass modeling methods have been extensively tested on mock data, including realistic triaxial mocks (Read & Steger 2017), tidally stripped mocks (Read et al. 2018; De Leo et al. 2024), and cosmologically realistic satellites (Genina et al. 2020). Genina et al. (2020) applied the GRAVSPHERE, a higher-order moment Jeans modeling code (Read & Steger 2017), to satellites from the APOSTLE simulations (Sawala et al. 2016), examining the impact of tidal stripping on the recovery of mass profiles. Their focus was on recovering central DM densities, cusp slopes, and half-stellar radius mass $M_{\star,1/2}$.

In this work, we test GRAPHNPE on similarly realistic dwarf galaxies data in Milky Way-like environments, using the Latte suite (Wetzel et al. 2016) from the Feedback In Realistic Environments² (FIRE-2; Hopkins et al. 2018) simulations. We will use the cold DM suite of Wetzel et al. (2016, 2023) as well as the self-interacting DM suite from Sameie et al. (2021); Vargya et al. (2022); Arora et al. (2024). In doing so, we aim to assess the robustness of GRAPHNPE against phenomena such as tidal interactions and validate their reliability in a realistic astrophysical environment, for both cored and cuspy profiles. Moving forward, the ultimate goal is to apply GRAPHNPE to real observational datasets of dwarf galaxies in the Local Group.

We also investigate the recovery of the peak circular velocity of halos, V_{\max} , and their peak virial mass, M_{200m} —both of key interest for constraining cosmological models (e.g. Burkert 1995; Bullock et al. 2001; van den Bosch et al. 2003; Rodriguez-Puebla et al. 2016; Read & Erkal 2019; Kim & Peter 2021). This is the first study to evaluate how well mass modeling methods can constrain V_{\max} and M_{200m} for simulated dwarf galaxies in realistic environments. This test is particularly timely given that two state-of-the-art mass modeling methods in the literature, CJAM (Jeans Anisotropic Multi-Gaussian Expansion; Watkins et al. 2013) and GRAVSPHERE, yield quite different inferences when applied to the same data (Zoutendijk et al. 2021a).

This paper is structured as follows. Section 2 provides an overview of the Latte suite of FIRE-2 simulations and the procedure for constructing the dwarf galaxy dataset. Section 3 details the forward modeling approach and the machine learning architecture of GRAPHNPE. Section 4 evaluates GRAPHNPE against two Jeans-based methods on a subset of FIRE-2 dwarf galaxies: a simple Jeans model using an unbinned Gaussian likelihood (Section 4.1) and GRAVSPHERE (Section 4.2). Section 5 presents the GRAPHNPE result on the full FIRE-2 dataset, under varying degrees of tidal effects. Section 6 provides a brief comparison of the performance between the CDM and SIDM samples. Section 7 discusses differences between the simulations and real observations (Section 7.1) and compares GRAPHNPE with other mass modeling methods (Section 7.2). Finally, Section 8 concludes the paper.

2 FIRE-2 SUITES OF SIMULATION

2.1 FIRE-2 Simulations

In this paper, we aim to test the validity of GRAPHNPE on a more realistic scenarios than previously done in N23, with the end goal of applying GRAPHNPE to the Milky Way’s satellites. To that end, we use the satellite population of the Milky Way-mass simulated galaxies as part of the Latte suite (Wetzel et al. 2016). The Latte suite is a set of zoom-in simulations of Milky Way-mass galaxies, in

which the hydrodynamics is based on the FIRE-2 physics (Hopkins et al. 2018). In order to ensure robustness of the results in a varied density profiles, we will use the CDM galaxies (Wetzel et al. 2016, 2023), along with SIDM galaxies presented in (Sameie et al. 2021; Vargya et al. 2022; Arora et al. 2024). We discuss each of these properties in turn.

FIRE-2 physics: The simulations are run using the GIZMO³ code base (Hopkins 2014, 2015), which utilizes a mesh-free, finite-mass (MFM) hydrodynamic solver, along with a version of the Tree-PM gravity solver from GADGET-3 (Springel 2005). MFM is a mesh-free Lagrangian finite-volume Godunov method that provides adaptive spatial resolution while maintaining conservation of mass, energy, momentum, and angular momentum. It integrates the advantages of both smoothed particle hydrodynamics (SPH) and traditional grid-based methods, enabling adaptive resolution and the accurate capture of fluid dynamics in highly complex astrophysical phenomena. The FIRE-2 physics model (Hopkins et al. 2018) additionally includes star formation (Hopkins et al. 2013), stellar feedback from radiation pressure, supernovae Ia and II, stellar winds, photoelectric heating, and photoionization, along with a detailed treatment of cooling, heating, UV background (Faucher-Giguère et al. 2009), metal mixing in the interstellar medium (Su et al. 2017; Escala et al. 2018) and enrichment from stars and supernovae (Ma et al. 2016).

Latte simulations: The Latte suite, first introduced in Wetzel et al. (2016), consists of isolated Milky Way-mass halos with $M_{200m} = 1 - 2 \times 10^{12} M_{\odot}$ at $z = 0$ (Sanderson et al. 2018; Wetzel et al. 2023). These halos are selected from a periodic box with a length of 85.5 Mpc as part of the AGORA project (Kim et al. 2014). The simulations start at $z = 99$, with initial conditions generated using the Multi-Scale Initial Conditions code (MUSIC; Hahn & Abel 2011).

Cosmology: The simulations adopt a standard Λ CDM cosmology with parameters consistent with the Planck 2018 results (Planck Collaboration et al. 2020). Specifically, the hosts are run with AGORA cosmology ($\Omega_{\Lambda} = 0.728$, $\Omega_m = 0.272$, $\Omega_b = 0.0455$, $\sigma_8 = 0.807$, $n_s = 0.961$, $h = 0.702$). Star particles have an initial mass resolution of $7070 M_{\odot}$ and a spatial resolution of 4 pc. DM particles have a mass resolution of $35,000 M_{\odot}$ and a spatial resolution of 40 pc (Wetzel et al. 2023).

Dark Matter Models: To make sure that we span a variety of inner profiles, we include in our analysis simulations with Cold Dark Matter (CDM) as well as Self-Interacting Dark Matter (SIDM) (see e.g. Tulin & Yu 2018, for a review). We analyze the fiducial CDM simulations from the FIRE-2 public data release (Wetzel et al. 2023) and the SIDM simulations presented in Sameie et al. (2021); Vargya et al. (2022); Arora et al. (2024).⁴ For the SIDM simulations, we use two hosts, m12f and m12m, each simulated with two constant DM self-interaction cross-sections of $\sigma/m = 1$ and $10 \text{ cm}^2/\text{g}$. DM self-interaction is implemented using the Monte Carlo scattering algorithm in Rocha et al. (2013); Peter et al. (2013), which determines the probability of interaction using a spline kernel with adaptive smoothing length (Monaghan & Lattanzio 1985). All scattering events are modeled as elastic and isotropic in the center of mass frame.

As discussed in Sameie et al. (2021); Vargya et al. (2022), the SIDM simulations are run with a modified version of the FIRE-2 physics model. Specifically, these simulations ignore the thermal-

² <https://fire.northwestern.edu/>

³ <https://bitbucket.org/phopkins/gizmo-public>

⁴ Prior SIDM simulations using the FIRE-2 galaxy formation model include simulations of $10^{10} M_{\odot}$ halos (Robles et al. 2017; Fitts et al. 2019) and dissipative SIDM scenarios (Shen et al. 2021, 2024); however, these are not included in this study.

simulation	M_{200m} [M_\odot]	R_{200m} [kpc]	σ/m [cm 2 /g]	N_{gal}
m12c_CDM	1.35×10^{12}	351	0	23
m12b_CDM	1.43×10^{12}	358	0	16
m12f_CDM	1.71×10^{12}	380	0	24
m12i_CDM	1.18×10^{12}	336	0	12
m12m_CDM	1.58×10^{12}	371	0	36
m12f_SIDM1	1.40×10^{12}	352	1	30
m12m_SIDM1	1.24×10^{12}	337	1	42
m12f_SIDM10	1.35×10^{12}	346	10	19
m12m_SIDM10	1.20×10^{12}	333	10	34

Table 1. Simulation specifications of the FIRE-2 simulations used in this work. Values for virial mass and radius are taken from Wetzel et al. (2023).

to-kinetic energy conversion in the unresolved Sedov-Taylor shock expansion phase from mass loss in massive stars. This leads to reduced star formation rates and lower stellar masses in the simulated galaxies, ultimately also increasing the number of satellites, as shown in Table 1. In general, this is advantageous for our study, as it provides another robustness test for our model; however, it complicates direct comparisons between CDM and SIDM scenarios (Section 6).

2.2 Test samples of dwarf galaxies

Given that our goal is to apply GRAPHNPE on Milky Way satellite galaxies, we will focus our studies to the satellites of the Milky Way-mass simulated galaxies. In FIRE-2, halos and subhalos are identified using a modified version of the ROCKSTAR⁵ six-dimensional halo finder (Behroozi et al. 2013), which accounts for multi-mass and multi-species particles. For each subhalo, we use the DM particles assigned by ROCKSTAR and identify its associated star particles. We use a similar procedure to Necib et al. (2019) to assign member star particles: we require that the star particles lie within the current r_{200m} (the radius within which the average density is 200 times the mean matter density of the Universe) of the subhalo, and have velocities within 3σ of the subhalo’s stellar velocity dispersion

We select subhalos based on their radial distances from the hosts and their halo-to-stellar mass ratios. Specifically, we include galaxies located within $3R_{200m}$ of the hosts at $z = 0$. While our primary goal is to apply GRAPHNPE to the Milky Way’s satellites, extending the selection to more distant dwarf galaxies allows us to assess GRAPHNPE’s performance across a broader subhalo population and in environments affected by tidal effects (Section 5).

In addition, we select subhalos with halo-to-stellar mass ratios satisfying $M_{\text{halo}}/M_\star > 10^2$ and with the number of star particles between $N_\star \in [20, 5000]$. This roughly corresponds to a stellar mass range of approximately $M_\star \in [10^5, 10^7] M_\odot$ and a halo mass range of $M_{\text{halo}} \in [10^7, 10^{10}] M_\odot$.⁶ The number of stars are chosen to encompass observations of classical and faint dwarfs. For instance, the highest number of resolved stars observed for Milky Way dwarfs, Fornax and Sculptor, are 2483 and 1365 stars, respectively (Walker et al. 2009a).

Next, we center each subhalo, and compute the positions and velocities of star and DM particles of each subhalo relative to the subhalo’s DM barycenter. We identify the center of each subhalo using the shrinking sphere method (Power et al. 2003) on the DM

particles only. Starting with an initial guess for the center, we iteratively compute the DM barycenter of all particles within a specified radius. At each step, the radius is reduced by 10%, and the center is updated to the newly calculated barycenter. The iterations proceed until the center converges to within 0.1% of the previous iteration. In rare cases, when a subhalo is located too close to its host or another subhalo, the computed center may still be offset. These cases are manually checked. Additionally, the stellar and DM centers are not always perfectly aligned, potentially complicating the analysis. For simplicity, we assume these centers coincide.

Finally, in this study, we do not account for observational effects. This includes measurement uncertainties, survey selection biases, and sources of contamination, such as misidentified member stars and binaries. These effects will be explored in future work.

3 METHODOLOGY

GRAPHNPE (Graph Neural Posterior Estimator) employs simulation-based inference (SBI; see Cranmer et al. 2020 for a review) to overcome limitations in the likelihood construction and sampling process in Jeans modeling. In SBI, the likelihood function is implicitly encoded within the simulations, allowing for more realistic physical processes to be incorporated than analytic models. Neural posterior estimation (NPE), in particular, leverages neural networks to maximize the information extracted from simulated data and map it directly to posterior distributions (Cranmer & Louppe 2016; Papamakarios & Murray 2016).

The GRAPHNPE model consists of a graph neural network (GNN; Scarselli et al. 2009) for feature extraction and a conditional normalizing flow (Jimenez Rezende & Mohamed 2015; Papamakarios et al. 2019) for density estimation. Stellar kinematic data, which consists of line-of-sight velocities and two projected coordinates, are represented as graphs. During the forward pass, the GNN compresses the graph representation into summary features, which are then used as context by the flow to model the posterior distribution.

Both the GRAPHNPE model and training simulations are largely the same with those introduced in N23. Below, we summarize key assumptions and a few notable changes.

3.1 Training Simulation & Forward Model

Similar to N23, we generate training samples of dwarf galaxies using a Monte Carlo simulation based on analytic models. The simulation procedure is as follows:

For each simulated galaxy, we first adopt some parametric phase-space distribution function $f(\vec{x}, \vec{v}; \theta)$ for its tracer stars, where \vec{x}, \vec{v} are the positions and velocities, respectively. The parameter set θ defines the DM density, velocity anisotropy, and tracer density profiles. We can then generate the population of N tracer stars by independently sampling each star from $f(\vec{x}, \vec{v}; \theta)$.

The likelihood of each simulated galaxy, given the parameter set θ , is determined by the probability of observing the set of positions and velocities, $\{\vec{x}_i, \vec{v}_i\}$, for each of the N tracer stars, where $i = 1, \dots, N$:

$$\mathcal{L}_{\text{NPE}} = \prod_{i=1}^N \mathcal{T}[f](\{\vec{x}_i, \vec{v}_i\}; \theta), \quad (1)$$

where \mathcal{T} represents some transformations on f to incorporate any observational effect. In our forward model, \mathcal{T} simply projects the 6-D positions and velocities into a 2D projected coordinates and line-of-sight velocities along a random axis. Any additional observational

⁵ <https://bitbucket.org/awetzel/rockstar-galaxies>

⁶ Peak halo mass range of $M_{\text{halo}}^{\text{peak}} \in [10^8, 10^{10}] M_\odot$.

effects, such as measurement uncertainties, selection biases, etc., can be folded directly into \mathcal{T} , which we will explore in future work.

In the context of NPE, the likelihood informs the relationship between the observed data $\{\vec{x}_i, \vec{v}_i\}$ and the parameter θ . While the model does not explicitly learn \mathcal{L} , it implicitly incorporates this information during training by approximating the posterior distribution $p(\theta | \{\vec{x}_i, \vec{v}_i\})$.

We assume the galaxies to be spherically symmetric and in dynamical equilibrium (i.e., the distribution function $f(\vec{x}, \vec{v}; \theta)$ does not explicitly depend on time). These are common assumptions in traditional Jeans analysis, and as previously discussed, might not be robust against realistic Milky Way dwarfs. Ideally, one can construct a more realistic family of distribution functions from hydrodynamic simulations, such as FIRE-2. However, such simulations are computationally demanding and currently impractical for generating the large training sets required for our framework. Moreover, as we will demonstrate in Section 7.1, even high-resolution zoom-in simulations struggle to capture the observed population of dwarf galaxies, particularly the ultra-faints.

In addition, training on hydrodynamic simulations could make the model sensitive to specific features of the simulations, such as baryonic prescriptions or even specific code implementations. These include differences in subgrid physics models, numerical solvers, resolution, or the treatment of star formation, feedback processes, and gas cooling. Variations in these implementation choices across simulation codes can lead to divergent results, even when simulating similar physical systems. As a result, models trained on one set of simulations might overfit to these implementation-specific details, reducing their capacity to generalize to other simulations or real observational data.

For these reasons, we aim to focus only on investigating whether the additional kinematic information leveraged by GRAPHNPE can help improve the robustness of the mass modeling, while keeping the simplified assumptions in Jeans analysis.

As in N23, we adopt the same parametric function for the DM density profile and tracer density distribution, and update the velocity anisotropy profile. Namely, we assume the DM density profiles follow a generalized Navarro-Frenk-White profile (Navarro et al. 1997):

$$\rho_{\text{dm}}^{\text{gNFW}}(r) = \rho_0 \left(\frac{r}{r_{\text{dm}}} \right)^{-\gamma} \left(1 + \frac{r}{r_{\text{dm}}} \right)^{-(3-\gamma)}, \quad (2)$$

where ρ_0 is the density normalization, r_{dm} is the DM scale radius, and γ is the inner slope.

The tracer density distribution $\nu(r)$ follows the 3-D Plummer profile (Plummer 1911),

$$\nu(r) = \frac{3L}{4\pi r_{\star}^3} \left(1 + \frac{r^2}{r_{\star}^2} \right)^{-5/2}, \quad (3)$$

where L is the total luminosity and r_{\star} is the tracer scale radius. The tracer mass density $\Sigma_{\star}(R)$ is the projection of $\nu(r)$ along the line-of-sight,

$$\Sigma_{\star}(R) = \frac{L}{\pi r_{\star}^2} \left(1 + \frac{R^2}{r_{\star}^2} \right)^{-2}, \quad (4)$$

where R is the projected radius. We assume the stellar contribution to the total mass to be negligible and ignore the luminosity term L , which describes the normalization of Eq. 3 and Eq. 4.

The velocity anisotropy profile $\beta(r)$ is defined as

$$\beta(r) \equiv 1 - \frac{\sigma_t^2}{\sigma_r^2}, \quad (5)$$

where σ_r and σ_t are the radial and tangential velocity dispersions. The velocity anisotropy ranges from $-\infty$ to 1, where $\beta = 1, 0, -\infty$ corresponds to a radial, isotropic, and tangential velocity profile, respectively. We assume a functional form for $\beta(r)$ following the Osipkov-Merritt (OM) profile (Osipkov 1979; Merritt 1985),

$$\beta^{\text{OM}}(r) = \frac{\beta_0 + (r/r_a)^2}{1 + (r/r_a)^2}, \quad (6)$$

where β_0 is the normalization and r_a is the anisotropy scale radius. The scale radius r_a thus determines the transition from β_0 at small radii to a radially-biased orbits at larger radii. In contrast to the N23 model, which always set β_0 to 0, we let β_0 be a free parameter within the range $[-0.5, 1]$.

In total, the current model has three DM parameters (ρ_0, r_s, γ) and three stellar parameters (β_0, r_a, r_{\star}). We expand the prior range for these parameters from N23 and summarize them in Table 2. Additionally, we select only galaxies with 3-D velocity dispersions within $0.1 - 50$ km/s for the final training set, which should sufficiently include the population of classical and ultra-faint dwarfs (Pace 2024).⁷ Among Milky Way dwarfs, the highest and lowest observed line-of-sight velocity dispersions—excluding the Small and Large Magellanic Clouds (SMC and LMC)—is 12.1 ± 0.2 km/s for Fornax (Walker et al. 2009a; Górski et al. 2011; Muñoz et al. 2018; Wang et al. 2019) and $1.2^{+0.9}_{-0.6}$ km/s for Tucana V (Hansen et al. 2024).

To construct and sample distribution functions, we use the AGAMA (Action-based Galaxy Modeling Architecture; Vasiliev 2019) library. AGAMA is a C++ framework that provides tools for distribution function modeling, potential solving, and orbit integration, with a particular focus on action-based methods, making it well-suited for equilibrium dynamical models.

For each galaxy, we first sample the tracer counts from a Poisson distribution with a mean of 100, then sample that many tracers from the distribution function independently. We then apply five random projections by selecting a random line-of-sight axis and projecting the galaxy onto the 2D plane perpendicular to this axis. We extract only the line-of-sight velocities v_{los} and the projected coordinates (x, y) for each galaxy. We do not consider proper motions here, although they can readily include in future work and can provide additional constraints.

3.2 Machine learning framework

We adopt the same machine learning setup as in N23, with minor updates to the network architecture (see Appendix B). The full schematic of our method is available in Figure S1 of N23. Each galaxy is modeled as an undirected graph consisting of a set of nodes and edges connecting them. Each node represents an individual tracer star, with node features given by its projected radius from the center of the galaxy, i.e., $r = \sqrt{x^2 + y^2}$, where x and y are the 2D projected coordinates, and its line-of-sight velocity v_{los} . The edges are constructed by connecting each star to its k -nearest neighbors, with $k = 20$.

Representing galaxies as graphs is advantageous because it allows for a *permutation-invariant* neural network architecture, such as a GNN. This approach enables the model to efficiently handle varying numbers of tracer stars, which is particularly important for realistic

⁷ Assuming an isotropic velocity distribution, the 3-D velocity dispersion $\sigma_{3\text{-D}}$ is related to the line-of-sight velocity dispersion σ_{los} via $\langle \sigma_{3\text{-D}}^2 \rangle = 3 \langle \sigma_{\text{los}}^2 \rangle$. For $\sigma_{3\text{-D}}$ in the range of $(0.1 - 50)$ km/s, σ_{los} is approximately within $(0.06 - 28.9)$ km/s.

Parameter	Name	Equation	Prior
$\log_{10}(\rho_0/(\mathrm{M}_\odot \mathrm{kpc}^{-3}))$	DM density normalization	gNFW (Eq. 2)	[3, 10]
$\log_{10}(r_{\mathrm{dm}}/\mathrm{kpc})$	DM scale radius	gNFW (Eq. 2)	[-2, 2]
γ	DM inner slope	gNFW (Eq. 2)	[-1, 2]
r_\star/r_{dm}	Stellar scale radius	Plummer (Eq. 4)	[0.2, 1]
r_a/r_\star	Velocity anisotropy scale radius	OM (Eq. 6)	[0.1, 10]
β_0	Velocity anisotropy normalization	OM (Eq. 6)	[-0.5, 1]

Table 2. Prior ranges of the training simulations before the velocity dispersion cut.

applications, where observed dwarf galaxies have a broad range of tracer counts.

As mentioned, the GRAPHNPE model consists of a GNN to extract summary features from the input graph representation, and a conditional flow to model the posterior distributions from these summary features. The architecture remains largely the same with the N23 model, with a notable change in the graph convolution layers in the GNN. Instead of the Chebyshev convolutional layers (ChebConv), the GNN now employs graph attention layers (GATConv, Veličković et al. 2017). These layers use the self-attention mechanism (Vaswani et al. 2017) to learn the edge weighting between connected nodes based on the node features. This allows the model to learn the most relevant connections directly from the training data, rather than relying on some predetermined weights or features (e.g. Euclidean distance between nodes, as in Villanueva-Domingo & Villaescusa-Navarro 2022; de Santi et al. 2023; Cuesta-Lazaro & Mishra-Sharma 2023). We found that although both ChebConv and GATConv have similar performance on the validation dataset (as also observed in N23), GATConv are less prone to overfit and more robust against tidal effects (Section 5).

During training, the GNN and normalizing flows are optimized simultaneously using the maximum likelihood objective of the flows (Equation B1). For additional training details, we refer readers to Appendix B. We use 5×10^6 and 5×10^5 galaxies for the training and validation set, respectively. The increase in the number of training samples compared to N23 reflects the wider range of the prior distributions and the inclusion of β_0 . The training converges after approximately 22 hours on a NVIDIA Tesla V100. Once trained, the model can sample the posterior almost instantaneously, *regardless of the number of tracers*.

4 COMPARISON WITH JEANS-BASED METHODS

4.1 Simple Jeans modeling

We compare the performance of GRAPHNPE and Jeans modeling on the FIRE dwarf galaxies. While Jeans methods are generally considered fast, they can be significantly more time-consuming than GRAPHNPE when modeling the full mock dataset. As mentioned, once trained, GRAPHNPE can draw posterior samples almost instantaneously, whereas Jeans modeling may take several hours to an entire day, especially for galaxies with a large tracer count. For this reason, we randomly select four galaxies from each FIRE-2 simulation, resulting in a total of 36 galaxies.

For each galaxy, we apply the fitting procedure outlined in Strigari et al. (2008), which is commonly employed in the literature (e.g. Geringer-Sameth et al. 2015b; Chang & Necib 2021). Briefly, we first fit the tracer mass density profile $\Sigma_\star(R)$ for r_\star , and then perform a joint fit for $\rho_{\mathrm{dm}}(r)$ and $\beta(r)$ by modeling the line-of-sight velocity

distribution as an unbinned Gaussian likelihood,

$$\mathcal{L}_{\mathrm{Jeans}} = \prod_{i=1}^N \frac{(2\pi)^{-1/2}}{\sqrt{\sigma_{\mathrm{los}}^2(R_i) + \Delta_i^2}} \exp \left[-\frac{1}{2} \left(\frac{(\vec{v}_i - \langle v \rangle)^2}{\sigma_{\mathrm{los}}^2(R_i) + \Delta_i^2} \right) \right], \quad (7)$$

where R_i is the projected radius, \vec{v}_i and Δ_i are the line-of-sight velocity and its measurement uncertainty of star i , and $\langle v \rangle$ is the mean velocity of the tracer population. The line-of-sight velocity dispersion profile $\sigma_{\mathrm{los}}(r)$ is related to $\rho_{\mathrm{dm}}(r)$ and $\beta(r)$ via Jeans equations (see Appendix A). Since we do not account for measurement uncertainty, Δ_i is set to zero for all stars. To ensure consistency with GRAPHNPE, the prior distribution is identical for the DM and anisotropy parameters.

For the remainder of this work, we refer to this model as “Simple Jeans” to distinguish it from more advanced Jeans-based methods. These include, for example, GRAVSPHERE (Read & Steger 2017, discussed further in Section 4.2), CJAM (Jeans Anisotropic Multi-Gaussian Expansions; Emsellem et al. 1994; Cappellari 2008; Watkins et al. 2013; Cappellari 2015), and MAMPOSSt (Mamon et al. 2013).

For each galaxy, we also calculate the true density and anisotropy profile from the DM and star particles. We divide the particles into non-overlapping radial bins containing equal number of particles, and calculate the mean and standard deviation of the relevant quantities within each bin. For $\rho_{\mathrm{dm}}(r)$, we use bins of 500 DM particles, which should sufficiently resolve the density profiles. For $\beta(r)$, we use bins of 50 star particles, though this number may be reduced slightly when necessary to guarantee at least two bins.

4.1.1 DM density and velocity anisotropy profiles

Figure 1 and Figure 2 show the recovered DM density profiles $\rho_{\mathrm{dm}}(r)$ and velocity anisotropy profiles $\beta(r)$ for a subset of the selected galaxies, with the remaining profiles provided in Appendix C1. Each panel shows a single dwarf galaxy, labeled with its ROCKSTAR halo ID at $z = 0$ and a corresponding letter (e.g., A, B, C, D) for readability, and compares the profiles derived using GRAPHNPE (blue) and Jeans modeling (orange). The posterior median profiles are shown as solid lines, with shaded bands indicating the 68% and 95% confidence intervals, while black circles with error bars represent the true profiles.⁸ The dashed vertical lines indicate $(0.25, 1, 4) r_{1/2}$, where $r_{1/2}$ is the 3-D half-stellar mass radius.

Both GRAPHNPE and Jeans modeling can recover the DM density to within the 95% confidence intervals. However, the velocity anisotropy profile is generally not well-constrained, with significant uncertainties persisting in most cases. The OM profile provides a reasonably good fit for these satellites, with the exception of Halo A, which may also be affected by the low number of tracer stars.

⁸ Note that error bars are not visible for the true density profiles.

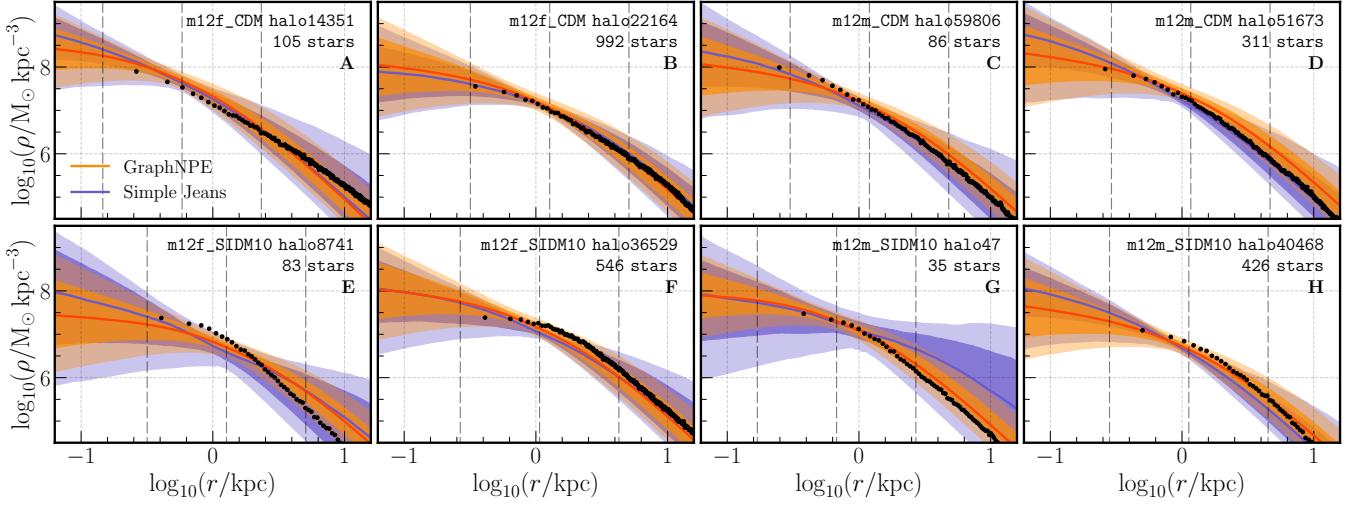


Figure 1. Comparison between the inferred DM density profiles $\rho_{\text{dm}}(r)$ from GRAPHNPE (orange) and Jeans modeling (blue) for a selection of FIRE-2 galaxies. Each panel shows the inferred and true profiles for an individual galaxy. The solid line and shaded bands denote the median, 68%, and 95% confidence intervals of the inferred density profiles. The true density profiles, calculated directly from the DM particle data, is represented by black circles with error bars. The vertical dashed lines indicate $0.25, 1.0, 4.0 r_{1/2}$, where $r_{1/2}$ is the 3-D half-stellar mass radius.

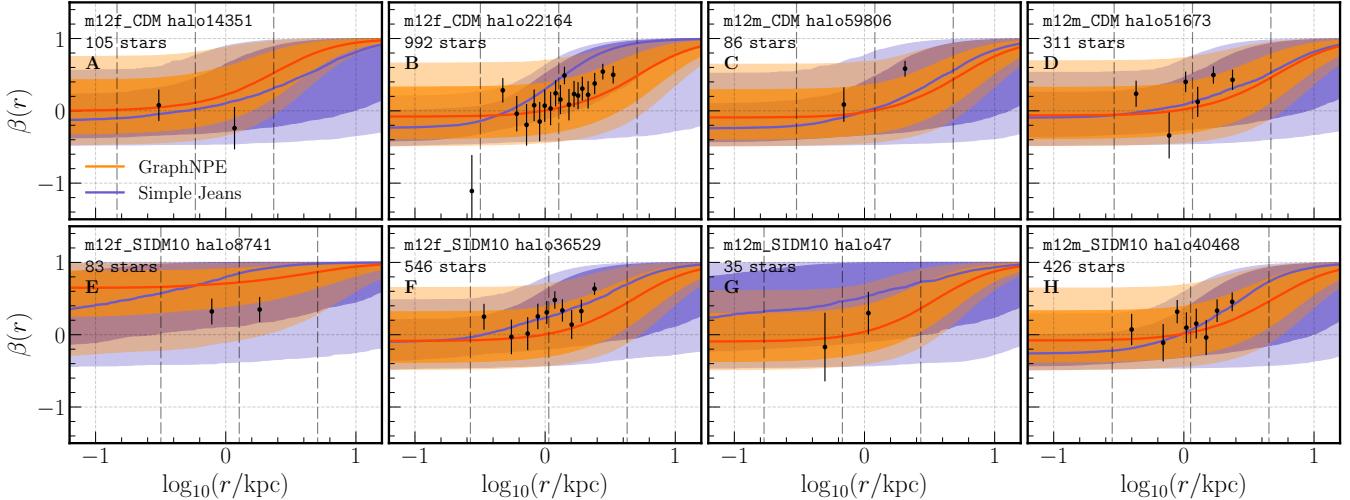


Figure 2. Comparison between the inferred velocity anisotropy profiles $\beta(r) = 1 - \sigma_t^2 / \sigma_r^2$, where σ_t and σ_r are the tangential and radial velocity dispersions, from GRAPHNPE (orange) and Jeans modeling (blue) for a selection of FIRE-2 galaxies. Values of $\beta = 0, 1 - \infty$ corresponds to isotropic, radially-biased, and tangentially-biased orbits. Panels are the same as in Figure 1.

In general, GRAPHNPE provides tighter constraints on the profiles compared to Jeans modeling. This is particularly evident when the tracer count is low, as seen in Halo E and Halo G. Conversely, when the tracer star count is high, the performance of GRAPHNPE and Jeans modeling becomes more similar. In the case of Halo B, the posterior distributions of the recovered density profiles $\rho_{\text{dm}}(r)$ show nearly perfect overlap, with Jeans modeling offering slightly tighter constraints. However, GRAPHNPE outperforms Jeans modeling in constraining the anisotropy profile $\beta(r)$. We observe similar performance trend across all profiles, as detailed in Appendix C1.

The trend in the performance differences between GRAPHNPE and Jeans is consistent with findings in N23. The improvement at low tracer counts likely results from GRAPHNPE's ability to account for

the full distribution function $f(\vec{x}, \vec{v})$. By training on Monte Carlo simulations with the likelihood described in Equation 1, which explicitly incorporates the full $f(\vec{x}, \vec{v})$, GRAPHNPE effectively performs amortized inference on this likelihood. In contrast, the Jeans likelihood in Equation 7 depends explicitly only on the velocity dispersion, which is the second-order moment of $f(\vec{x}, \vec{v})$. This also explains the similarity in the recovered density profiles at high tracer counts. As the tracer count increases, the line-of-sight velocity dispersion profiles can be better constrained to larger radii. This helps break the degeneracy between the density profile $\rho_{\text{dm}}(r)$ and the anisotropy profile $\beta(r)$, allowing Jeans modeling to recover more accurate density profiles that closely align with those from GRAPHNPE (e.g. Read & Steger 2017; Read et al. 2021; Chang & Necib 2021).

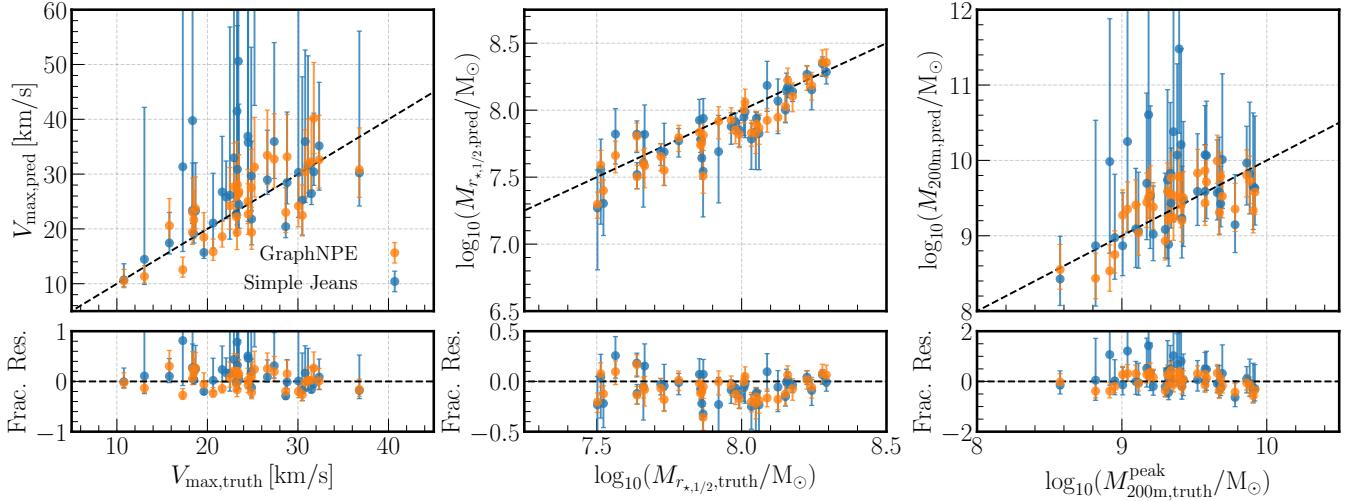


Figure 3. Comparison between the inferred mass and structural parameters from GRAPHNPE (orange) and Jeans modeling (blue). From left to right, the columns show the peak circular velocity V_{\max} , the half-stellar radius mass $M_{r_{\star,1/2}}$, and the virial mass M_{200m} . V_{\max} and M_{200m} are calculated by extrapolating the DM density profiles to r_{\max} and r_{200m} , respectively. The top and bottom rows show the median and fractional residual of the predicted values against the true values, respectively, with the error bars denoting the 64th percentile ranges. The black dashed line represents the one-to-one correlation. Since the virial mass M_{200m} might not be well-defined for tidally stripped halos, the recovered M_{200m} values are compared to the peak virial mass achieved throughout the halos' evolutionary history.

4.1.2 Mass and structural parameters

We perform a detailed comparison between the performance of GRAPHNPE and Jeans modeling by recovering key mass and structural parameters for all selected galaxies. Specifically, for each galaxy, we calculate:

(i) **The peak circular velocity:** $V_{\max} \equiv V_{\text{circ}}(r_{\max})$, where r_{\max} is the radius at which the circular velocity, $V_{\text{circ}}(r) = \sqrt{GM(r)/r}$, reaches its maximum value. V_{\max} provides a robust mass estimator that is relatively insensitive to the halo boundary (e.g., Klypin et al. 1999; Peñarrubia et al. 2008).

(ii) **The half-stellar radius total mass:** $M_{r_{\star,1/2}} \equiv M(r_{\star,1/2})$, where $r_{\star,1/2}$ is the 3-D radius enclosed half of the stellar mass. Previous studies have shown that the mass profile is maximally constrained near $r_{\star,1/2}$ for a single tracer population with only line-of-sight velocity measurements, since $M_{r_{\star,1/2}}$ is approximately proportional to the velocity dispersion $\langle \sigma_{\text{los}}^2 \rangle$ (e.g. Walker et al. 2009b; Wolf et al. 2010; Campbell et al. 2017; Errani et al. 2018).

(iii) **The virial mass:** $M_{200m} \equiv M(r_{200m})$, where r_{200m} is the radius at which the mean halo density is 200 times the *mean matter density* of the Universe.

To calculate the predicted values of V_{\max} and M_{200m} , we extrapolate the DM density profiles to r_{\max} and r_{200m} , respectively. Since many Milky Way dwarf galaxies, as well as those in our test samples, have likely experienced some degree of tidal stripping, the outer regions of their DM halos have already been significantly stripped, and the present-day M_{200m} is not well-defined. Therefore, we compare our mass estimates to the peak virial mass of the halos throughout their evolutionary history, M_{200m}^{peak} .

We emphasize that this is the first study to evaluate the ability of mass modeling methods to recover V_{\max} and M_{200m}^{peak} in mock galaxies within realistic environments. Genina et al. (2020), which applied an older version of GRAVSPHERE on dwarf galaxies in the APOSTLE simulations, is the most similar to this work. Their study examined

	V_{\max} (km/s)		
	AE ¹	SE ²	NLL ³
Simple Jeans	7.73 ± 2.07	210.01 ± 138.03	4.42 ± 0.17
GRAPHNPE	3.60 ± 0.38	18.15 ± 3.02	3.51 ± 0.31
	$\log_{10}(M_{r_{\star,1/2}} / M_{\odot})$		
	AE	SE	NLL
Simple Jeans	0.11 ± 0.01	0.02 ± 0.00	-0.46 ± 0.09
GRAPHNPE	0.11 ± 0.01	0.02 ± 0.00	-0.59 ± 0.08
	$\log_{10}(M_{200m} / M_{\odot})$		
	AE	SE	NLL
Simple Jeans	0.41 ± 0.08	0.38 ± 0.14	1.07 ± 0.07
GRAPHNPE	0.24 ± 0.02	0.08 ± 0.01	0.33 ± 0.03

¹ Absolute Error(AE). Lower is better.

² Squared Error (SE). Lower is better.

³ Negative Log-Likelihood (NLL). Lower is better. To estimate the NLL of each parameter, the posterior was first fitted using Gaussian Kernel Density Estimate (KDE) and evaluated at the target point.

Table 3. Performance of GRAPHNPE and Jeans modeling for the maximum circular velocity V_{\max} , the half-light mass $M_{r_{\star,1/2}}$, and the virial mass M_{200m} over the selected sample of galaxies for multiple metrics.

how well GRAVSPHERE can constrain $M_{r_{\star,1/2}}$ and compared this with other mass estimators (see their Figures 3 and 4) and explored the impact of tidal stripping on mass profile recovery (their Figure 9). However, they did not investigate the recovery of V_{\max} or M_{200m}^{peak} , which we analyze for the first time in this context. The effects of tidal stripping on these quantities are further explored in Section 5.

Figure 3 displays the predicted and true values for V_{\max} , $M_{r_{\star,1/2}}$, and M_{200m}^{peak} (left to right columns) as derived from GRAPHNPE (orange) and Jeans modeling (blue). The predicted values correspond to the median of the posterior, with error bars representing the 68% confidence interval for each parameter. The fractional residual (bottom panel) is defined as $\Delta V_{\max} = (V_{\max,\text{pred}} - V_{\max,\text{truth}})/V_{\max,\text{truth}}$ for V_{\max} and $\Delta M = \log(M_{\text{pred}}/M_{\text{truth}})$ for $M_{r_{\star,1/2}}$ and M_{200m}^{peak} .

To further quantify the performance of each method, we compute standard performance metrics, including the absolute error (AE), squared error (SE), and negative log-likelihood (NLL). Table 3 presents the mean and the standard error for each metric across all selected galaxies. The AE and SE metrics are computed between the true values and predicted values (which we take to be the median of the posteriors) with the SE metric more sensitive to outliers. The NLL metric, on the other hand, takes into account the overall shape of the posterior. If the posterior were perfectly Gaussian, the NLL would be reduced to a chi-square statistic, with an additional normalization term that accounts for the standard deviation. However, since the posteriors of V_{\max} and M_{200m} exhibit long tails, as hinted from the asymmetric error bars in Figure 3, the NLL provides a more robust assessment. To compute the NLL, we first fit the posterior using Gaussian Kernel Density Estimation⁹ (KDE) and then evaluate it at the true value.

Both GRAPHNPE and Jeans modeling recover V_{\max} and M_{200m} within the 95% confidence intervals. However, it is evident from Figure 3 and Table 3 that GRAPHNPE provides more accurate and tighter constraints compared to Jeans modeling. Specifically, Jeans tends to overestimate V_{\max} and M_{200m}^{peak} , while also producing significantly wider 68% confidence intervals. As shown in Table 3, Jeans performs worse according to all three performance metrics: AE, SE, and NLL.

On the other hand, both GRAPHNPE and Simple Jeans achieve similarly accurate performance in recovering $M_{r_{*,1/2}}$. This is expected, as $M_{r_{*,1/2}}$ strongly correlates with the line-of-sight velocity dispersion, making it less susceptible to the mass-anisotropy degeneracy and uncertainties in modeling $\beta(r)$. This relationship has been established by various mass estimators (Walker et al. 2009b; Wolf et al. 2010; Campbell et al. 2017; Errani et al. 2018). Similarly, mass modeling methods have demonstrated comparable accuracy in estimating $M_{r_{*,1/2}}$, even in realistic galactic environments (Breddels & Helmi 2013; Genina et al. 2020), and additionally shown that $M_{r_{*,1/2}}$ is less sensitive to assumptions on DM profile (Breddels & Helmi 2013). From Table 3, we observe that the AE and SE metrics for GRAPHNPE and Simple Jeans are similar, with GRAPHNPE achieving a marginally better NLL; however, the performance differences remain well within one standard deviation for both methods.

4.2 Higher-moment methods

Having demonstrated that GRAPHNPE produces tighter and more accurate constraints on the dark matter density and velocity anisotropy profiles compared to Simple Jeans, we now turn to a comparison with a more advanced Jeans-based method. Specifically, we evaluate the performance of GRAPHNPE against GRAVSPHERE, a non-parametric mass modeling approach that incorporates higher-order velocity moments (Read & Steger 2017). In addition to the velocity dispersion, GRAVSPHERE computes the “Virial Shape Parameters” (VSPs; Merrifield & Kent 1990; Richardson & Fairbairn 2014):

$$v_{s1} = \frac{2}{5} \int_0^\infty GM(5 - 2\beta)v\sigma_r^2 r dr = \int_0^\infty \Sigma_\star \langle v_{\text{los}}^4 \rangle R dR \quad (8)$$

and

$$v_{s2} = \frac{4}{35} \int_0^\infty GM(7 - 6\beta)v\sigma_r^2 r^3 dr = \int_0^\infty \Sigma_\star \langle v_{\text{los}}^4 \rangle R^3 dR. \quad (9)$$

Here r and R denote the 3-D and projected radius, respectively. VSPs are particularly advantageous as they depend only on $\langle v_{\text{los}}^4 \rangle$.

⁹ We use a KDE bandwidth of 1 for V_{\max} and 0.1 for $M_{r_{*,1/2}}$ and M_{200m} .

and $\beta(r)$ (where the anisotropy term arises from the integration over the projected radius), rather than directly involving its fourth-order moment counterpart (see e.g. Merrifield & Kent 1990; Richardson & Fairbairn 2014). This imposes two additional constraints on the velocity anisotropy, helping to break the mass-anisotropy degeneracy.

GRAVSPHERE has been described and extensively tested in Read & Steger (2017); Read et al. (2018); Genina et al. (2020); Collins et al. (2021); De Leo et al. (2024). Additionally, Read et al. (2021) provides a comprehensive comparison between an earlier version of GRAVSPHERE and other mass modeling techniques. In this paper, we use the public version of GRAVSPHERE as described in Collins et al. (2021) which bins the data using the BINULATOR method.¹⁰ Below, we highlight the key considerations for this comparison:

GRAVSPHERE assumes the following form for the velocity anisotropy (Baes & van Hese 2007),

$$\beta(r) = \beta_0 + (\beta_\infty - \beta_0) \frac{1}{1 + (r/r_a)^\eta}, \quad (10)$$

where β_0 and β_∞ are the asymptotic anisotropies at small and large radii, r_a is the anisotropy transition radius, and η controls the sharpness of the transition. The OM anisotropy described in Equation 6 is a special case of this parameterization, where $\beta_\infty = 1$ and $\eta = 2$. In GRAVSPHERE, the priors are defined on the symmetrized version of $\beta(r)$, defined as

$$\tilde{\beta}(r) = \frac{\sigma_r^2 - \sigma_t^2}{\sigma_r^2 + \sigma_t^2} = \frac{\beta}{2 - \beta}. \quad (11)$$

This reparameterization ensures that $\tilde{\beta}$ remains finite, as it is constrained to $[-1, 1]$, whereas β can diverge to $-\infty$. The priors of GRAVSPHERE are uniform over $\tilde{\beta}_0 \in [-0.5, 1]$, $\tilde{\beta}_\infty \in [-0.5, 1]$, $\eta \in [0, 3]$, with a fixed $r_a = 1$ kpc. For comparison, the priors of GRAPHNPE in Table 2 roughly translate to a range of $\tilde{\beta}_0 \in [-0.2, 1]$ and $r_a \in [2 \times 10^{-3}, 10^3]$ kpc, with $\tilde{\beta}_\infty = 1$ and $\eta = 2$ fixed by the definition of the OM profile.

Additionally, GRAVSPHERE uses the CORENFWTIDES model for the DM density profile (Collins et al. 2021) and a “three Plummer” model for the tracer density profile (Rojas-Niño et al. 2016; Read & Steger 2017). For a more detailed description of the model, we refer readers to Section 4.1 in Collins et al. (2021). Here, we summarize the asymptotic behavior of the CORENFWTIDES profile. The DM density follows $\rho_{\text{dm}}^{\text{CNFWt}} \propto r^{-\gamma}$, where γ transitions from γ_{in} at small radii to γ_{out} at large radii. Both the inner and outer slopes, γ_{in} and γ_{out} , are free parameters, with prior distributions $[0, 1]$ and $[3.01, 5]$ respectively.¹¹ In comparison, the gNFW profile in GRAPHNPE follows $\rho_{\text{dm}}^{\text{gNFW}} \propto r^{-\gamma}$ at small radii, where γ can vary between $[-1, 2]$ (see Table 2), and always transition to $\rho_{\text{dm}}^{\text{gNFW}} \propto r^{-3}$ at large radii.

Lastly, unlike GRAPHNPE, GRAVSPHERE partitions the data into radial bins using the BINULATOR code, described in Section 4.1.1 of Collins et al. (2021). BINULATOR automatically adjusts the bins to contain an equal number of stars and fits a generalized Gaussian probability to each bin to robustly estimate its mean, variance, and kurtosis. This improves upon the binning routines of the previous version of GRAVSPHERE and enhances performance in the low tracer count limit. In this work, GRAVSPHERE uses 10 stars per radial bin for the light profile and 20 stars per radial bin for the line-of-sight velocity dispersion profile. The stars are assumed to have negligible velocity uncertainties. The model simultaneously fits the mass,

¹⁰ Code available at <https://github.com/justinread/gravsphere>.

¹¹ The parameters γ_{in} and γ_{out} are equivalent to n and δ in $\rho_{\text{dm}}^{\text{CNFWt}}$. We use this notation for consistency with the gNFW profile.

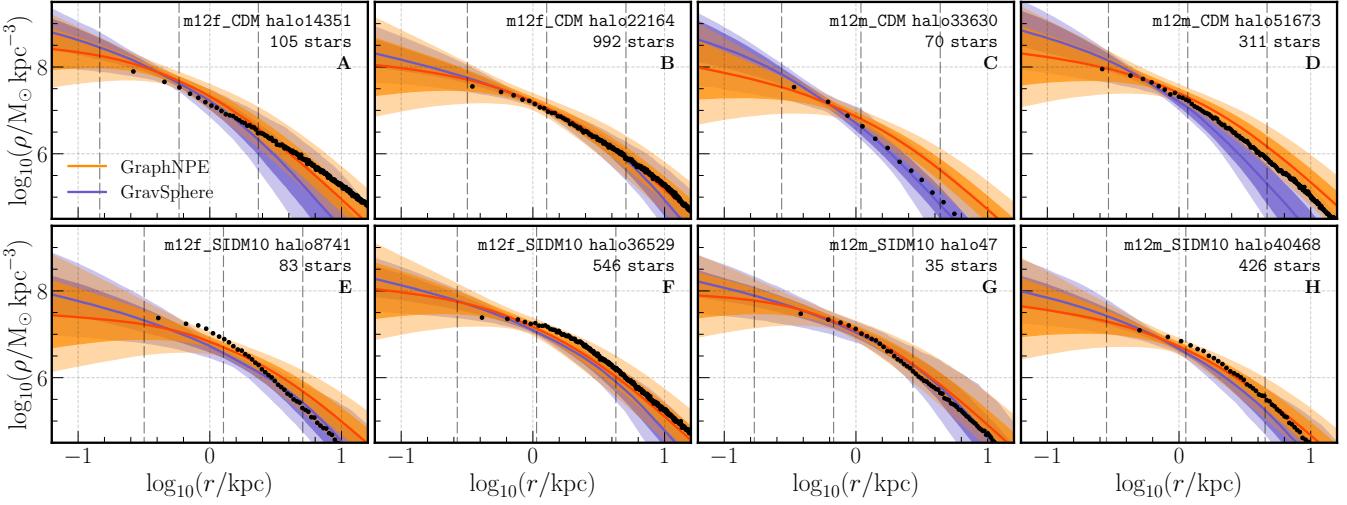


Figure 4. Comparison between the inferred DM density profiles $\rho_{\text{dm}}(r)$ from GRAPHNPE (orange) and GRAVSPHERE (blue) for a selection of FIRE-2 galaxies. Panels are the same as in Figure 1.

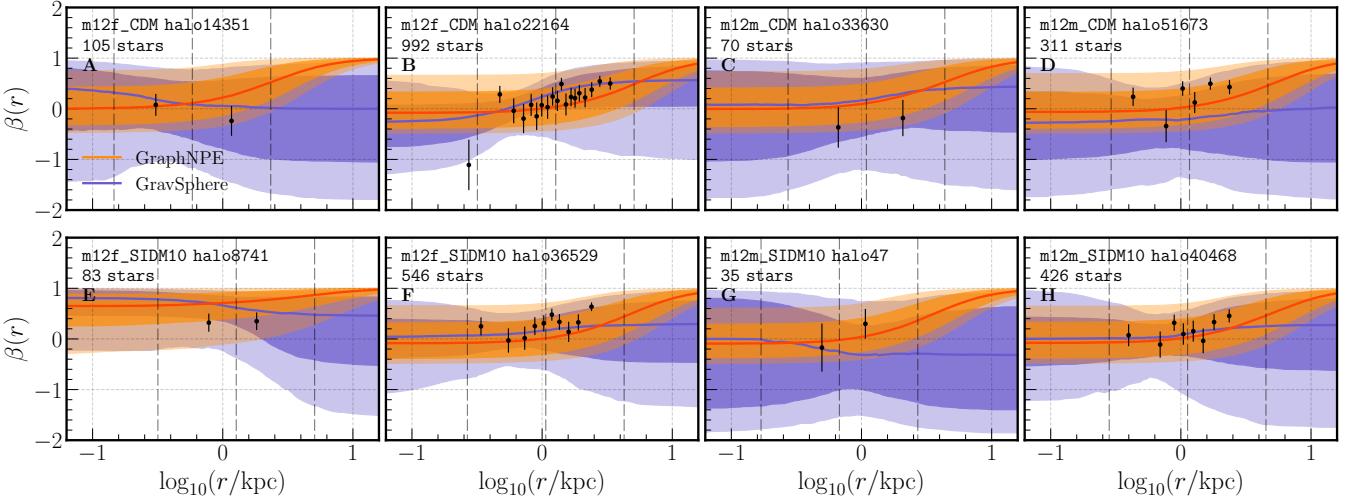


Figure 5. Comparison between the inferred velocity anisotropy profiles $\beta(r)$ from GRAPHNPE (orange) and GRAVSPHERE (blue) for a selection of FIRE-2 galaxies. Panels are the same as in Figure 1.

velocity anisotropy, and tracer density profiles by minimizing the chi-square statistic between the observable quantities—namely Σ_\star and σ_{los} within each bin, and the VSPs.

Due to differences in modeling choices and prior distributions, it is difficult to directly compare between GRAPHNPE and GRAVSPHERE. Thus, unlike in Section 4.1, here we do not seek to establish which method can provide a better constraint on the density profile. Instead, our goal is to establish a baseline performance for GRAPHNPE and to highlight its key strengths and limitations relative to advanced Jeans-based methods like GRAVSPHERE. To this end, we apply GRAVSPHERE to the same sample of FIRE-2 dwarf galaxies presented in Section 4.1. Due to the higher computational cost of GRAVSPHERE as compared to Simple Jeans, we limit our analysis to four simulations: m12f, m12m, m12f_SIDM10, and m12m_SIDM10, comprising a total of 16 dwarf galaxies.

4.2.1 Result

Figure 4 shows the inferred DM density profiles $\rho_{\text{dm}}(r)$ for a selection of FIRE-2 galaxies (with the rest shown in Appendix C1). Each panel shows an individual galaxy and follows the same format as Figure 1: the median, 68%, and 95% confidence intervals of the inferred $\rho_{\text{dm}}(r)$ from GRAPHNPE and GRAVSPHERE are displayed in orange and blue, respectively, while the true profile is shown as black circles with error bars. Note that the galaxy samples are selected to highlight the differences between GRAPHNPE and GRAVSPHERE, and thus do not necessarily overlap with the sample in Figure 1.

In general, the performance between GRAVSPHERE and GRAPHNPE are comparable. Both methods effectively constrain the density profiles within their respective 95% confidence intervals. Interestingly, we do not observe the significant performance degradation, as seen with Simple Jeans in Section 4.1. It may seem surprising that GRAVSPHERE performs so well even when there are few data points,

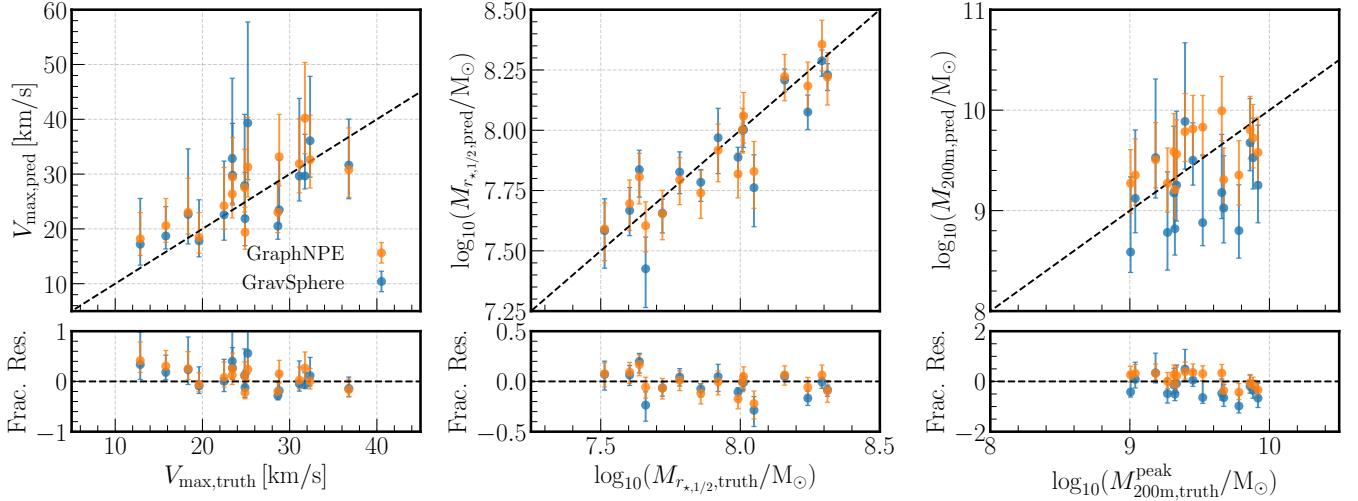


Figure 6. Comparison between the inferred mass and structural parameters from GRAPHNPE (orange) and GRAVSPHERE (blue). From left to right, the columns show the peak circular velocity V_{\max} , the enclosed mass $M_{r_{*,1/2}}$, and the virial mass M_{200m} . Panels are the same as in Figure 3.

	V_{\max} (km/s)		
	AE	SE	NLL
GRAVSPHERE	4.71 ± 0.87	33.60 ± 12.65	3.59 ± 0.06
GRAPHNPE	4.15 ± 0.58	22.23 ± 4.74	3.33 ± 0.09
$\log_{10}(M_{r_{*,1/2}} / M_{\odot})$			
	AE	SE	NLL
GRAVSPHERE	0.09 ± 0.02	0.02 ± 0.01	-0.78 ± 0.24
GRAPHNPE	0.08 ± 0.02	0.01 ± 0.00	-0.90 ± 0.14
$\log_{10}(M_{200m} / M_{\odot})$			
	AE	SE	NLL
GRAVSPHERE	0.41 ± 0.06	0.23 ± 0.06	0.79 ± 0.08
GRAPHNPE	0.27 ± 0.03	0.09 ± 0.01	0.36 ± 0.03

Table 4. Performance of GRAPHNPE and GRAVSPHERE for the maximum circular velocity V_{\max} , the half-light mass $M_{r_{*,1/2}}$, and the virial mass M_{200m} over the selected sample of galaxies for multiple metrics.

as earlier versions of the code certainly struggled in this regime (e.g. Read et al. 2021). However, here we use the improved version of GRAVSPHERE introduced in Collins et al. (2021) that uses the BINULATOR to bin the data. This was designed to work well even for very few stellar velocity data points, making GRAVSPHERE suitable also for modeling ultra-faint dwarfs.

The inner profile appears to be more tightly constrained by GRAVSPHERE. However, we note that this is likely due to differences in the prior distribution on the inner slope γ , with $\gamma_{\text{in}}^{\text{GS}} \in [0, 1]$ and $\gamma^{\text{gNFW}} \in [-1, 2]$. The upper-95% confidence intervals of the density profiles between the two methods tend to be in agreement, though the profiles from GRAVSPHERE are slightly steeper in some cases. For the lower-95%, GRAPHNPE allows the density to flatten more significantly and even decline as r decreases, thus favoring more core-like structure. The decline in the central density is unphysical; however, by allowing γ to go negative, GRAPHNPE avoids running into the prior edge for profiles with pronounced cores.

We find that GRAVSPHERE predicts steeper outer density profiles than GRAPHNPE, likely also due to differences in the prior distribution of γ . Specifically, GRAVSPHERE allows the outer slope $\gamma_{\text{out}}^{\text{GS}}$ to vary between $[3.01, 5]$, whereas GRAPHNPE assumes a fixed outer

behavior $\rho_{\text{dm}}^{\text{gNFW}} \propto r^{-3}$. As a result, GRAVSPHERE struggles with shallower outer profiles (e.g. Halo A and Halo D), while GRAPHNPE has difficulty with steeper ones (e.g. Halo C and Halo E).

Figure 5 presents the velocity anisotropy profiles $\beta(r)$. Within the $(0.25, 4) r_{*,1/2}$ region, where kinematic tracers are well-sampled, the inferred anisotropy profiles from both GRAVSPHERE and GRAPHNPE are well constrained. Outside this range, the weaker constraints from GRAVSPHERE arise from its more flexible prior and the lack of tracer data, whereas both GRAPHNPE and Simple Jeans assumes the OM profile, which imposes stricter assumptions on $\beta(r)$ but may be less accurate in some cases. For example, this is evident in Halo A and Halo E, where GRAVSPHERE provides a better fit compared to the OM profile.

As in Section 4.1.2, we compute the same mass and structural parameters, i.e. V_{\max} , $M_{r_{*,1/2}}$, and M_{200m}^{peak} , for each galaxy, with the results shown in Figure 6. Compared to Simple Jeans, the performance of GRAVSPHERE closely aligns with that of GRAPHNPE. Notably, despite differences in modeling choices and prior distributions, the predicted values of V_{\max} , $M_{r_{*,1/2}}$, and M_{200m}^{peak} from GRAVSPHERE and GRAPHNPE are remarkably consistent, with their 68% confidence intervals largely overlapping. However, we note that GRAVSPHERE consistently underestimates M_{200m}^{peak} , which is somewhat expected given its preference of steeper outer slopes (Figure 4). Lastly, we evaluate the AE, SE, and NLL performance metrics and present the mean and standard error in Table 4. The performance metrics are comparable for GRAPHNPE and GRAVSPHERE across all mass and structural parameters, with GRAPHNPE slightly outperforming GRAVSPHERE.

5 TIDAL EFFECTS

We now evaluate GRAPHNPE’s performance on the full FIRE-2 mock dataset. In this section, we assess the robustness of GRAPHNPE across varying degree of tidal effects, which has been extensively studied in dwarf galaxies in the Local Group (e.g. Read et al. 2006; Battaglia et al. 2015; Pace et al. 2022; De Leo et al. 2024) and cosmological simulations (e.g. Kazantzidis et al. 2011; Tomozeiu et al. 2016; Frings et al. 2017; Fattahi et al. 2018; Shipp et al. 2023, 2024). Here, we do

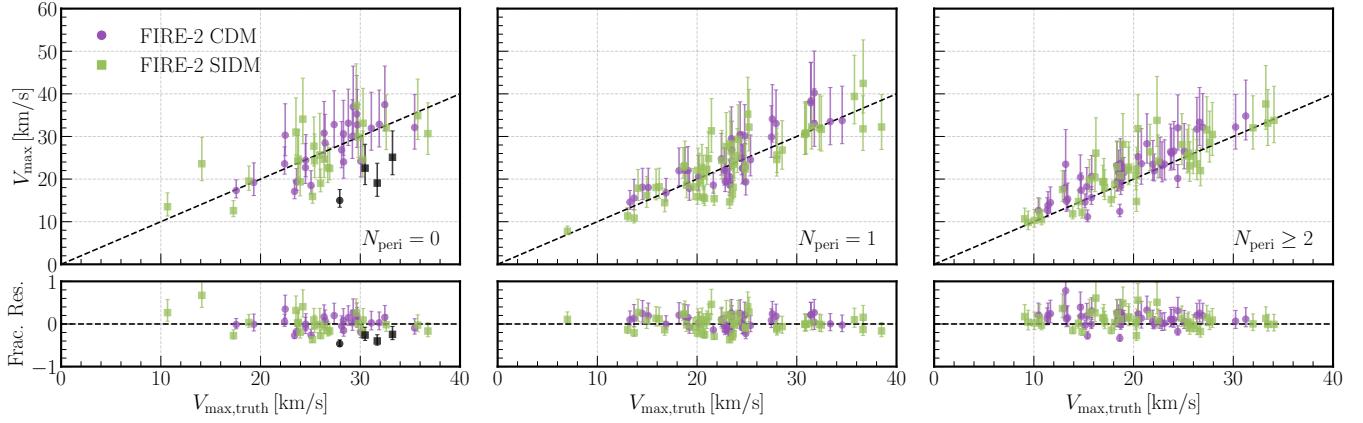


Figure 7. Recovery of the peak circular velocity V_{\max} of dwarf galaxies in FIRE-2 dataset. The top row shows the predicted V_{\max} (the median of the posteriors) versus the true V_{\max} , while the bottom row displays the residuals. Error bars represent the 68% confidence intervals. CDM galaxies are shown as purple circles, while SIDM galaxies are represented by green squares; outliers for both are highlighted in black. The black dashed line represents the one-to-one correlation. Galaxies are grouped by the number of pericentric passages N_{peri} , with each column corresponding to a different grouping.

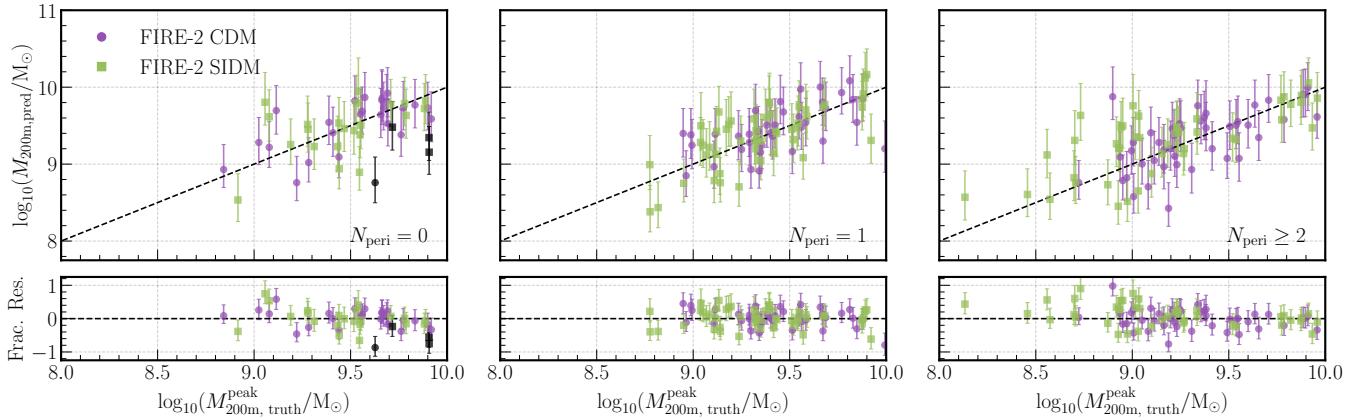


Figure 8. Recovery of the peak virial mass M_{200m} of dwarf galaxies in the FIRE-2 dataset. Panels are the same as Figure 7.

not focus on satellites that are massively disrupted, but rather on those where tidal signatures, such as tidal tails or velocity gradients, are not clearly detectable. In FIRE-2, Shipp et al. (2023) shows that this is common, with more than half of satellites exhibiting significant tidal features that remain undetected when analyzed using mock Dark Energy Survey (DES; Flaugher et al. 2015; Abbott et al. 2018, 2021) observations.

5.1 Quantifying tidal effects with orbital parameters

To quantify the degree of tidal effects, we track the orbital history of each galaxy over the simulation snapshots and compute the following parameters: (1) the number of pericentric passages N_{peri} , (2) the last pericentric distance r_{peri} , (3) the time since the last pericentric passage t_{peri} , and (4) the current distance to host r_{curr} . Satellites that are more tidally disrupted are expected to have lower pericentric distances r_{peri} and to have completed more orbits around their host galaxies (higher N_{peri}). The latter condition can result from either a high apocentric distance or an early infall time; however, we do not examine these factors separately. The time since the last pericentric passage t_{peri} provides insight into how long the satellite has

had to relax or recover from its most recent tidal interaction, which could influence the persistence of tidal features. Lastly, the current distance r_{curr} indirectly measures the ongoing tidal forces and also a directly observable quantity. For a detailed exploration of the orbital dynamics and histories of satellites in Milky Way-mass halos within FIRE-2 simulations, we refer readers to Santistevan et al. (2023).

To account for variations in the mass and size of the host halos, r_{peri} and r_{curr} are expressed in units of the virial radius, R_{200m} , of the host halo at the time of the pericentric passage or the current epoch, respectively.

It is important to note that the above orbital parameters do not directly probe tidal effects. For instance, the density within the half-light radius of a satellite relative to its host can influence its susceptibility to tidal effects, with more compact satellites being more resilient. As a result, tidal effects are commonly quantified using density-based metrics (e.g., Pace et al. 2022; Shipp et al. 2023) or mass loss indicators, such as the ratio of current to peak values of V_{\max} , M_{200m} , and M_{\star} (e.g., Barber et al. 2015; Fattahi et al. 2018). However, while density metrics and mass loss indicators provide a more direct measure of tidal effects, they are not readily available in observations, making them difficult to apply consistently. Orbital

parameters, though more indirect, are more easily inferred from observations.

5.2 Example: result with N_{peri}

We assess the performance of GRAPHNPE as a function of the orbital parameters discussed above, with a primary focus on V_{max} and M_{200m}^{peak} .

Figures 7 and 8 present the predicted versus true values (top rows) and residuals (bottom rows) of V_{max} and M_{200m}^{peak} , respectively, across three different bins of the number of pericentric passages $N_{\text{peri}} = 0, 1$, and ≥ 2 (left to right columns). Similar figures for other orbital parameters, including r_{peri} , t_{peri} , and r_{curr} , are provided in Appendix C2. As before, the predicted values and their error bars are the median and 68% confidence intervals of the posterior distributions. Additionally, we distinguish CDM and SIDM galaxies with purple circles and green squares, respectively, with outliers denoted in black.

Figures 7 and 8 clearly illustrate the impact of tides on the predictions of the present-day V_{max} and the peak M_{200m}^{peak} . As the number of pericentric passages increase, tidal effects accumulate, and GRAPHNPE tends to overestimate the V_{max} and M_{200m}^{peak} . Despite this, we find a good alignment between the predicted and true values in all three bins. In particular, we find that GRAPHNPE can recover V_{max} and M_{200m}^{peak} within the 68% confidence interval for approximately 64% and 71% of galaxies in the full samples, respectively. Although not shown in the figures, GRAPHNPE captures approximately 93 % and 95% of galaxies within the 95% confidence interval for V_{max} and M_{200m}^{peak} , respectively. Interestingly, this is similar to Genina et al. (2020), which showed that GRAVSPHERE recover the enclosed mass distributions to within 68% and 95% confidence for 60% and 90% of their samples, respectively.

To further investigate the effects of tides, we calculate the fraction of galaxies within the 68% and 95% confidence intervals, which we now refer to as the “coverage fractions”, across different N_{peri} bins. Table 5 presents the mean and associated error of the coverage fractions for V_{max} and M_{200m}^{peak} for three N_{peri} bins. The errors are estimated using bootstrapping methods, where we resample the dataset with replacement multiple times to generate a distribution of coverage fractions, using the standard deviation of this distribution as the uncertainty estimate.

As N_{peri} increases, the coverage fraction decreases, which is expected due to GRAPHNPE’s tendency to overestimate V_{max} and M_{200m}^{peak} for more tidally disrupted galaxies. The recovery is overall quite strong: even for galaxies with multiple pericentric passages, we recover their true V_{max} and M_{200m}^{peak} values within the 68% confidence interval for $57.2 \pm 5.1\%$ and $71.2 \pm 5.0\%$ of the galaxies, and within the 95% confidence interval for $89.6 \pm 3.4\%$ and $93.0 \pm 2.8\%$, respectively. We note that the first bin has a lower coverage fraction, which we attribute to a few outlier data points, highlighted as black data points in Figures 7 and 8. These outliers will be discussed in more detail below; here, we note that removing them increases the 68% and 95% coverage fractions in the first N_{peri} bin to 69.6% and 92.0% for V_{max} and to 71.6% and 91.9% for M_{200m}^{peak} .

As noted, a few outlier galaxies in the $N_{\text{peri}} = 0$ bin exhibit significantly underestimated V_{max} and M_{200m}^{peak} values by GRAPHNPE. These galaxies, highlighted as black data points in Figures 7 and 8, are located at the high-mass end, with true values of $V_{\text{max}} \gtrsim 28$ km/s and $\log_{10} M_{200m}^{\text{peak}} \gtrsim 9.5$, making them less likely to be strongly affected by tidal effects. Upon closer examination, we find

that these galaxies are all nearing their first pericentric passage and are likely experiencing significant tidal shocks due to their proximity to the host galaxy. This example underscores a key limitation of N_{peri} : while it effectively tracks cumulative tidal effects, it does not capture instances where a galaxy is undergoing strong, transient tidal interactions. Therefore, to gain a more comprehensive understanding of the model’s performance and the complexity of tidal interactions, it is crucial to consider additional orbital parameters, as discussed in Section 5.3.

GRAPHNPE’s tendency to overestimate V_{max} and M_{200m}^{peak} can be explained as follows. As the satellite passes through its pericenter, the energy injection from tidal shock can “puff up” the stellar profiles, extending the stellar velocity dispersion profiles to larger radii while simultaneously lowering the overall velocity dispersion (e.g., Read et al. 2006; Peñarrubia et al. 2008). Additionally, Read et al. (2006) have shown that the tidal tails can lead to a rise in the projected velocity dispersion profiles beyond the tidal stripping radius, making the satellites appear dynamically hotter in projection. In the context of Jeans modeling, this increased velocity dispersion at large radii can result in an overestimation of the enclosed mass in these regions. Although GRAPHNPE is trained on the full distribution function, as defined in Equation 1, the velocity dispersion is still expected to play a dominant role in determining the mass profiles.

The presence of unbound stars can artificially inflate the velocity dispersion, further leading to an overestimation of the mass. The member star assignment procedure described in Section 3 mitigates this issue but is unlikely to completely eliminate it. While assignments based on the halo potential, such as using V_{max} (Samuel et al. 2020), can return a more pristine sample of member stars, we opt to use our procedure since such information is not directly accessible in observational data.

In a recent work, Chiang et al. (2024) demonstrates that tidal stripping tends to isotropize the anisotropy profile. Since GRAPHNPE assumes the OM anisotropy profile, which inherently tends towards radially-biased orbits at large radii, it can overestimate β at these radii for disrupted galaxies. This, in turn, leads to an overestimation in the 3-D velocity dispersion and, consequently, the mass.

In the case of V_{max} , because we infer present-day values, highly disrupted galaxies often have their outer halos stripped, causing an overestimation of their current V_{max} .

To briefly summarize, we emphasize that, despite these challenges, GRAPHNPE’s overall performance remains strong. In particular, GRAPHNPE recovers V_{max} and M_{200m}^{peak} within the 68% and 95% confidence intervals for a large fraction of galaxies. This accuracy matches that of the previous version of GRAVSPHERE in Genina et al. (2020) for enclosed mass distributions, while here we specifically demonstrate similar reliability for V_{max} and M_{200m}^{peak} . The strong performance on M_{200m}^{peak} suggests that (1) present-day stellar kinematics encode sufficient information to infer the peak halo mass and (2) GRAPHNPE is capable of effectively extracting this information. The former is somewhat unsurprising, as past studies have shown that M_{200m}^{peak} can be recovered if the tidal radius is sufficient far from the stellar half-stellar mass radius of satellites (e.g. Read et al. 2006; Errani et al. 2018; Read et al. 2018; Read & Erkal 2019). This result is further highlighted by the fact that the simple Jeans model in Section 4.1 appears to overestimate the peak mass more consistently, as shown from the samples in Figure 3.

	V_{\max}	M_{200m}^{peak}		
	68% conf.	95% conf.	68% conf.	95% conf.
$N_{\text{peri}} = 0$	$64.0 \pm 6.8\%$	$90.7 \pm 3.9\%$	$67.7 \pm 6.4\%$	$90.7 \pm 4.0\%$
$N_{\text{peri}} = 1$	$69.4 \pm 4.8\%$	$96.6 \pm 1.9\%$	$71.6 \pm 4.8\%$	$98.9 \pm 1.1\%$
$N_{\text{peri}} \geq 2$	$57.2 \pm 5.1\%$	$89.6 \pm 3.4\%$	$71.2 \pm 5.0\%$	$93.0 \pm 2.8\%$
$r_{\text{peri}}/R_{\text{host},200m}^{\text{peri}} \geq 0.5$	$70.2 \pm 5.7\%$	$93.7 \pm 3.1\%$	$75.1 \pm 5.6\%$	$98.4 \pm 1.6\%$
$0.2 \leq r_{\text{peri}}/R_{\text{host},200m}^{\text{peri}} < 0.5$	$63.5 \pm 5.4\%$	$94.8 \pm 2.6\%$	$72.9 \pm 4.9\%$	$96.2 \pm 2.2\%$
$r_{\text{peri}}/R_{\text{host},200m}^{\text{peri}} < 0.2$	$49.5 \pm 8.3\%$	$88.1 \pm 5.4\%$	$61.4 \pm 8.4\%$	$91.0 \pm 5.0\%$
$r_{\text{curr}}/R_{\text{host},200m} \geq 1$	$67.3 \pm 5.2\%$	$92.2 \pm 2.9\%$	$68.2 \pm 5.1\%$	$95.4 \pm 2.2\%$
$0.5 \leq r_{\text{curr}}/R_{\text{host},200m} < 1$	$67.2 \pm 5.2\%$	$93.2 \pm 2.8\%$	$71.6 \pm 4.9\%$	$96.4 \pm 2.0\%$
$r_{\text{curr}}/R_{\text{host},200m} < 0.5$	$54.9 \pm 6.9\%$	$94.3 \pm 3.2\%$	$75.4 \pm 5.9\%$	$94.3 \pm 3.1\%$
$t_{\text{peri}}/\text{Gyr} > 5$	$69.5 \pm 6.7\%$	$96.0 \pm 2.9\%$	$67.3 \pm 6.7\%$	$100.0 \pm 0.0\%$
$2 \leq t_{\text{peri}}/\text{Gyr} < 5$	$61.2 \pm 6.2\%$	$93.3 \pm 3.3\%$	$77.7 \pm 5.4\%$	$91.4 \pm 3.8\%$
$t_{\text{peri}}/\text{Gyr} < 2$	$61.0 \pm 6.0\%$	$91.1 \pm 3.5\%$	$68.8 \pm 5.5\%$	$97.1 \pm 2.1\%$

Table 5. Summary of coverage fractions for the 68% and 95% confidence intervals for different bins of orbital parameters. The mean and associated errors are reported for each case.

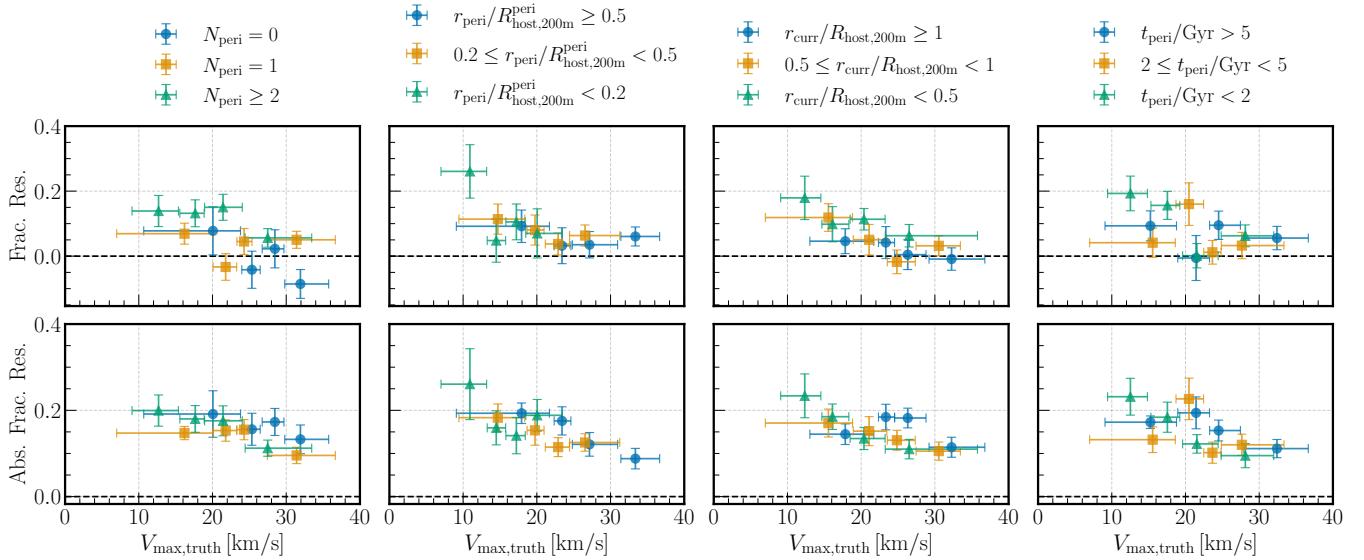


Figure 9. The net (top rows) and absolute (bottom rows) fractional residuals of V_{\max} across different bins of the orbital parameters. From left to right, the panels show bins of the number of pericentric passages N_{peri} , the last pericentric distance r_{peri} in units of the host R_{200m} at the time, the current distance r_{curr} in units of the host current R_{200m} , and the time since last pericenter t_{peri} .

5.3 Result with all orbital parameters

We now examine the performance of GRAPHNPE for the remaining orbital parameters: namely, the last pericentric distance r_{peri} , the current distance r_{curr} , and the time since last pericenter t_{peri} . As before, Table 5 presents the 68% and 95% coverage fractions, now calculated across three separate bins for each of the above parameters, resulting in a total of nine bins.

The coverage fractions generally align with their respective confidence intervals. In the case of V_{\max} , the trends in r_{peri} , r_{curr} , and t_{peri} are consistent with that of N_{peri} : more tidally disrupted bins tend to have lower coverage due to GRAPHNPE’s bias towards higher V_{\max} values. On the other hand, the trends for M_{200m}^{peak} are much less pronounced. With the exception of r_{peri} , the coverage fractions remain roughly the same across all bins of orbital parameters.

Interestingly, GRAPHNPE appears to be conservative in estimating M_{200m}^{peak} , with overly large confidence intervals that cover more sam-

ples than expected. In some cases (e.g., $t_{\text{peri}} > 5$ Gyr, $N_{\text{peri}} = 1$), GRAPHNPE can recover M_{200m}^{peak} for close to 100% of the populations. The broad confidence intervals might arise from extrapolating the DM density profiles when computing M_{200m}^{peak} , which introduce additional uncertainties. This issue could be mitigated with a more accurate model for the profile tails or a more extended tracer population.

Among the four orbital parameters, the pericentric distance r_{peri} shows the clearest trend, with the coverage fraction dropping significantly in the most disrupted bin, specifically $r_{\text{peri}} < 0.2R_{\text{host},200m}^{\text{peri}}$. This is consistent with findings from prior works (e.g. Shipp et al. 2023, 2024; Montero-Dorta et al. 2024). Nevertheless, even under these conditions, we still recover V_{\max} within the 68% confidence interval for about 50% of the sample and within the 95% confidence interval for about 90%. For M_{200m}^{peak} , approximately 60% of the sample

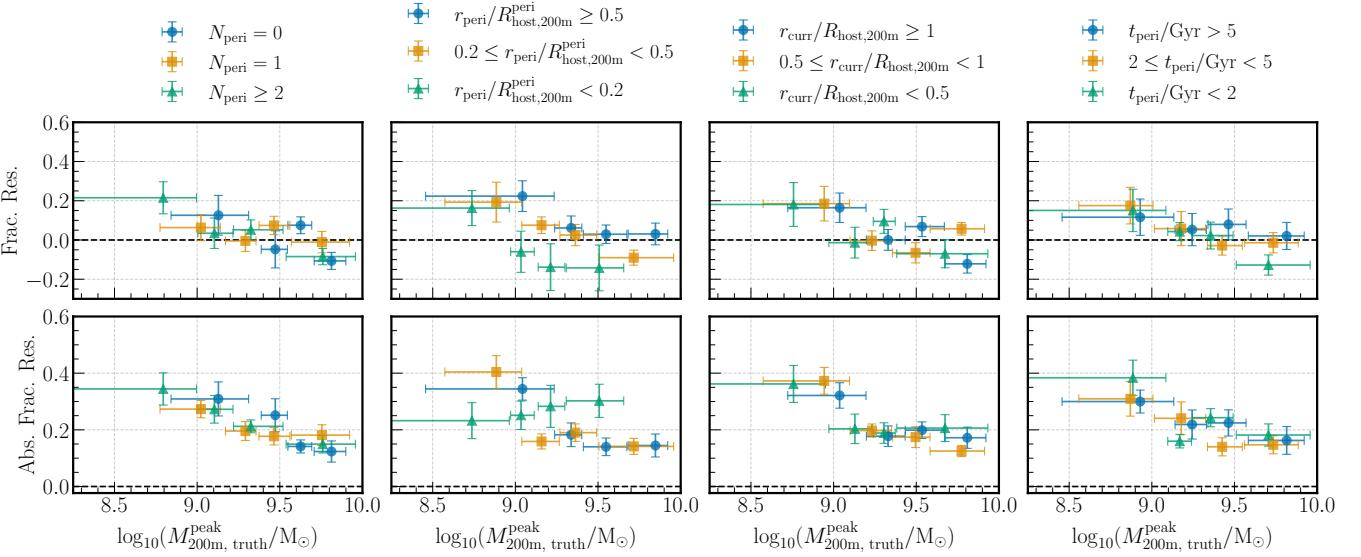


Figure 10. The net (top rows) and absolute (bottom rows) residuals of M_{200m}^{peak} across different bins of the orbital parameters. Panels are the same as Figure 9.

falls within the 68% confidence interval, while 90% falls within the 95% confidence interval.

To better quantify the accuracy of GraphNPE on predicting V_{\max} and M_{200m}^{peak} , we compute their mean residuals across different bins of the four orbital parameters. We first note an important consideration. It is evident from Figures 7 and 8 that there are distinct populations of galaxies within the three N_{peri} bins. Specifically, more massive galaxies—both in their peak and present-day masses—are more likely to occupy lower N_{peri} bin. Galaxies with high N_{peri} likely fell into their host halos at earlier times, resulting in generally lower values of M_{200m}^{peak} , as they had less time to accrete mass before being incorporated into their host. Additionally, these galaxies are expected to have experienced more stripping, leading to lower values of present-day V_{\max} . It is therefore crucial to disentangle tidal effects from those of galaxy populations. More massive galaxies tend to host more stars, and the tighter constraints by GraphNPE observed in the lower N_{peri} bin could arise either from the absence of significant tidal effects or from the larger number of stars. Therefore, for each bin of the tidal parameter, we further subdivide the galaxies into three equal-sized bins based on V_{\max} and M_{200m}^{peak} . Each sub-bin contains at least 10 galaxies to ensure statistical robustness in our analysis.

Figures 9 and 10 present the fractional residuals for V_{\max} and M_{200m}^{peak} , respectively, across different bins of N_{peri} , r_{peri} , r_{curr} , and t_{peri} (left to right columns). The top row shows the net (signed) residuals, while the bottom row shows the absolute (unsigned) residuals. In each panel, the color and marker indicate the bins of the corresponding tidal parameter, with blue circles, yellow squares, and green triangles representing bins in order from least to most likely to be tidally disrupted. The y-position of each data point represents the mean residuals of galaxies in the bin, with the vertical error bars showing the standard errors in estimating the mean. The x-position corresponds to the mean of the true V_{\max} or M_{200m}^{peak} values for all galaxies in the bin, while the horizontal error bars indicate the range of true V_{\max} or M_{200m}^{peak} values within the bin. We do not include the outliers discussed previously in the calculation.

The fractional residual of V_{\max} exhibits similar trends across the

panels of Figure 9. Within the same tidal bin, GraphNPE overestimates V_{\max} for galaxies at low V_{\max} , as expected since these galaxies are more susceptible to tides. As V_{\max} increases, both the net residual and absolute residual steadily decrease. Beyond $V_{\max} \gtrsim 20 \text{ km/s}$, the net residual approaches zero, while the absolute residual flattens out at approximately 10–20%. When comparing between different tidal bins, we find that galaxies in the more disrupted bins tend to have lower values of V_{\max} , consistent with the findings in Figure 7. At similar V_{\max} values, the mean residuals tend to be larger in more disrupted bins, although there is significant scatter in both V_{\max} and residuals within each bin. For V_{\max} in the range 10–20 km/s, the absolute residuals are typically around 20%, peaking at approximately $25 \pm 10\%$ for galaxies with $V_{\max} \approx 11 \text{ km/s}$ in the $r_{\text{peri}} < 0.2 R_{\text{host},200m}$ bin. At $V_{\max} \gtrsim 20 \text{ km/s}$, both the net and absolute residuals flatten across all tidal bins, as also previously noted, resulting in no significant performance differences.

We observe a similar trend in the residual of M_{200m}^{peak} in Figure 10 for bins of N_{peri} , r_{peri} , and r_{curr} . Specifically, GraphNPE overestimates the true values at low M_{200m}^{peak} , with both the net and absolute residuals decreasing in magnitude as M_{200m}^{peak} increases. At higher masses, the residuals flatten across all tidal bins, similar to the behavior seen for V_{\max} . This transition occurs around $M_{200m}^{\text{peak}} \sim 10^9 M_{\odot}$, consistent with the expectation for halos with $V_{\max} \sim 20 \text{ km/s}$, based on $M_{200m} - V_{\max}$ relation (see e.g. Bullock et al. 2001; Rodríguez-Puebla et al. 2016). Quantitatively, the absolute residual decreases from ~ 0.4 dex at the low-mass end to ~ 0.2 dex at higher M_{200m}^{peak} . In the case of t_{peri} , although we find a clear correlation between the true M_{200m}^{peak} and the residuals, there are no significant performance differences across the t_{peri} bins, indicating that time since pericenter does not strongly impact GraphNPE’s ability to recover M_{200m}^{peak} .

6 COMPARISON BETWEEN CDM AND SIDM

We briefly compare the performance of GraphNPE between the FIRE-2 CDM and SIDM dwarf galaxies. Prior works have shown that

	V_{\max}		M_{200m}	
	68% conf.	95% conf.	68% conf.	95% conf.
CDM	$64.5 \pm 6.4\%$	$93.3 \pm 3.3\%$	$71.9 \pm 7.6\%$	$97.2 \pm 2.7\%$
SIDM1	$62.4 \pm 5.9\%$	$88.3 \pm 4.0\%$	$65.4 \pm 5.6\%$	$89.8 \pm 3.7\%$
SIDM10	$71.0 \pm 6.6\%$	$94.3 \pm 3.2\%$	$71.0 \pm 6.5\%$	$97.9 \pm 2.0\%$

Table 6. Summary of coverage fractions for the 68% and 95% confidence intervals for the CDM and SIDM samples. The mean and associated errors are reported for each case.

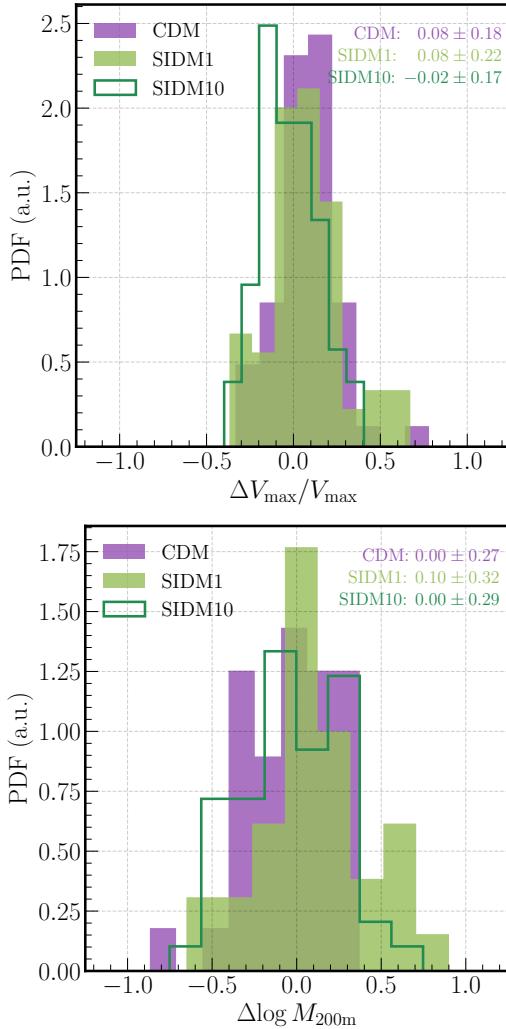


Figure 11. Normalized distributions of fractional residuals in V_{\max} (top) and M_{200m}^{peak} (bottom) for dwarf galaxies in the FIRE-2 CDM (purple), SIDM1 (green), and SIDM10 (dark green) simulations. The mean and standard deviation of the distribution for each simulation are indicated in the legend. The CDM, SIDM1, and SIDM10 simulations have 50, 72, and 53 mock galaxies, respectively.

SIDM subhalos are more prone to tidal disruption effects than their CDM counterparts, due to their lower concentration and evaporation driven by ram pressure stripping (Spergel & Steinhardt 2000; Tulin & Yu 2018; Robles et al. 2017; Vargya et al. 2022). Here, we only consider the global trend across the entire population, irrespective of the orbital parameters discussed in Section 5.

To minimize the environmental impacts of the host galaxies, we only consider $m12f_{\text{CDM}}$ and $m12m_{\text{CDM}}$ from the CDM simulations, comparing them against their SIDM counterparts. We also compare

the SIDM simulations with different cross-sections: $\sigma/m = 1 \text{ cm}^2/\text{g}$ and $\sigma/m = 10 \text{ cm}^2/\text{g}$, which we simply denote as SIDM1 and SIDM10, respectively. The CDM, SIDM1, and SIDM10 simulations consist of 50, 72, and 53 mock galaxies, respectively.

We remind readers of an important caveat: the SIDM simulations are run with a modified version of the FIRE-2 physics model that ignore the thermal-to-kinetic energy conversion during shock expansion from massive star mass loss (Section 2). As a result, the SIDM galaxies exhibit lower star formation rates and stellar masses, with a higher number of surviving galaxies due to increased survival probabilities. While we do not expect this modification to affect performance on individual CDM vs. SIDM galaxies, it introduces a systematic difference (independent of SIDM physics) in the galaxy populations.

Figure 11 shows the normalized distribution of fractional residuals for V_{\max} (top) and M_{200m} (bottom), with CDM and SIDM simulations represented in purple and green, respectively. The mean and standard deviation of each distribution are indicated in the figure.

In the top panel, the mean residuals suggest that GRAPHNPE tends to overestimate V_{\max} for CDM and SIDM1, consistent with the trends discussed in Section 5. This is further evidenced by the extended tails of the CDM and SIDM1 distributions. The model performs best on the SIDM10 samples, where the V_{\max} residual distribution is more centered around zero. The residual distributions of M_{200m} in the bottom panel are more comparable between the simulations. The SIDM1 distribution peaks the highest at zero, but also exhibits some outliers toward positive residuals, contributing to a larger overall spread.

Table 6 presents the 68% and 95% coverage fractions for both V_{\max} and M_{200m} across the CDM and SIDM samples. The reported means and associated errors are derived from 1000 bootstrap resamples. The model shows comparable performance on the CDM and SIDM10 simulations. In contrast, SIDM1 exhibits lower overall coverage, likely driven by the long tail of high residuals shown in Figure 6.

To conclude, despite these differences, the distributions remain broadly comparable. However, given the limited sample size, the large spread in the distributions, and the caveats mentioned above, it is difficult to draw definitive conclusions. For both V_{\max} and M_{200m} , the mean residuals between the simulations remain consistent within 1σ with each other. We leave a more detailed analysis to future work.

7 DISCUSSION

7.1 FIRE-2 Test Samples vs. Observational Samples

Our ultimate goal is applying GRAPHNPE to real observational datasets of dwarf galaxies in the Local Group. Therefore, it is crucial not only to evaluate the performance of GRAPHNPE on more realistic simulations, such as FIRE-2, but to also determine how well the dwarf galaxies in our test datasets match observations. Here, we compare the properties of FIRE-2 dwarf galaxies with those of the Milky Way

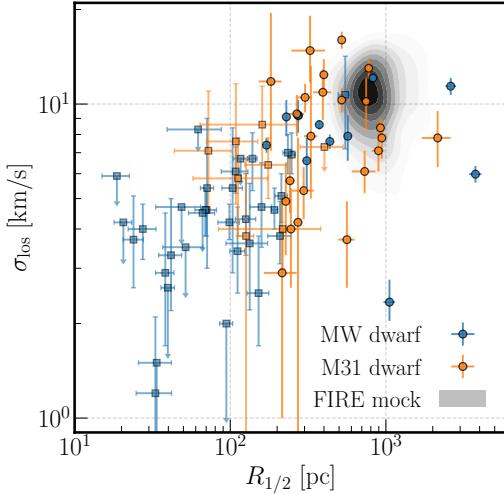


Figure 12. The line-of-sight velocity dispersion σ_{los} versus the projected half-stellar mass radius $R_{\star,1/2}$ for Milky Way dwarfs (blue, excluding SMC and LMC), M31 dwarfs (orange), and dwarfs in our FIRE-2 mock dataset (gray contours). The contours are shown at every 10% intervals. Upper limits are denoted by downward arrows. Ultra-faint dwarfs ($M_V > -7.7$) are denoted as squares.

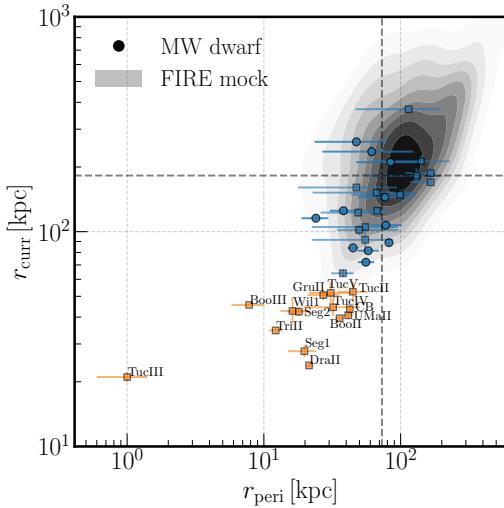


Figure 13. The pericentric distance r_{peri} versus current distance r_{curr} for Milky Way dwarfs (blue and orange circles) and our FIRE-2 mock dataset (gray contours). The contours are shown at every 10% intervals. Ultra-faint dwarfs ($M_V > -7.7$) are denoted as squares.

to identify key discrepancies and what constitutes a reliable dwarf target for our mass modeling method in future analysis.

We emphasize that here we do not seek to compare the FIRE-2 physics model directly with observations. Rather, we aim to assess how the ranges of properties in existing observational samples compare to those of the FIRE-2 dwarf galaxies included in our test samples, and highlight which real dwarf galaxies are considered “out-of-distribution”. This is important, as it allows us to identify the regimes where the results in Section 5 may not fully generalize, providing critical context for interpreting the results of this study.

Previous works have shown that satellites in FIRE-2 simulations broadly match the observed satellite population in the Local Group. Wetzel et al. (2016) and Garrison-Kimmel et al. (2019) demonstrated

that the simulated satellite stellar mass functions and line-of-sight velocity dispersions are consistent with those of the Milky Way and M31, down to approximately $M_\star \sim 10^5 M_\odot$ and $\sigma_{\text{los}} \sim 5 \text{ km/s}$. This agreement excludes ultra-faint dwarfs, as FIRE-2 does not yet resolve this regime due to limitations in mass and spatial resolutions.

Figure 12 shows the line-of-sight velocity dispersion σ_{los} versus the projected half-stellar mass radius $R_{\star,1/2}$. Data for Milky Way (blue) and M31 (orange) dwarfs are from the Local Volume catalog presented in Pace (2024). The gray contours represent the distribution of σ_{los} and $R_{1/2}$ of dwarf galaxies from our FIRE-2 mock catalog. Ultra-faint dwarfs, defined as $M_V > -7.7$ (Simon 2019), are marked as squares to highlight that they are not expected to be resolved in FIRE-2.

As expected, the dwarf galaxies in our mock samples predominantly occupy regions of the parameter space corresponding to higher σ_{los} compared to the Milky Way and M31 samples. As discussed above, this discrepancy arises because FIRE-2 simulations do not resolve ultra-faint dwarfs. Additionally, the FIRE-2 dwarfs generally exhibit larger half-stellar mass radii $R_{\star,1/2}$ compared to the observed Milky Way and M31 samples, consistent with the findings of Shen et al. (2024). Of the 43 Milky Way dwarfs shown, only three are within this 90% contour of the simulation, namely Bootes III, Fornax, and Sextans.

Despite differences between the FIRE-2 mock samples and observations, it is important to note that the majority of the observed data samples remain well within the prior distribution. From Table 2, the Plummer scale radius can range from $r_\star \in [20, 20000] \text{ pc}$,¹² while the velocity dispersion can range from $\sigma_{3-D} \in [0.1, 50] \text{ km/s}$ (or $\sigma_{\text{los}} \in [0.06, 28.9] \text{ km/s}$, assuming isotropy). The three galaxies with the smaller half-stellar mass radii are Triangulum II, Segue 1, and Willman 1 with values of $20.56^{+2.74}_{-2.61} \text{ pc}$, $23.90^{+3.74}_{-3.40} \text{ pc}$, and $27.43^{+5.89}_{-5.35} \text{ pc}$, respectively (see Pace 2024 and references therein). While these technically fall within the prior range, Triangulum II is very close to the lower boundary. This proximity may artificially truncate the posterior distribution, preventing a proper exploration of smaller values and potentially biasing parameter inferences. To better accommodate such observations, we plan to extend the prior limits in future analysis.

More importantly, as shown in N23, the performance of GraphNPE is largely scale-free. In our idealized simulations, we observed no significant differences in performance across samples with varying DM and stellar profiles (see Figure 1 of N23). This robustness likely extends to more realistic simulations, such as FIRE. That said, environmental effects—particularly tidal effects—could introduce scale-dependent differences in performance, as less massive and less compact galaxies are likely to be more disrupted. Comparing the FIRE-2 populations to observed dwarf galaxies is not straightforward. The FIRE-2 mock samples are generally more massive yet less compact than observed galaxies, leading to competing effects that may partially offset each other. Additionally, Shipp et al. (2023) showed that FIRE-2 satellites may be more prone to tidal effects than observations, with many satellites disrupting with relatively high densities.

Therefore, we now compare the orbital properties of our mock galaxies with Milky Way dwarfs to highlight out-of-distribution real dwarf galaxies. A prior work from Samuel et al. (2020) has shown that the radial distribution of FIRE-2 satellites are consistent with

¹² For a Plummer profile, the projected half-stellar mass radius is equal to the scale radius, i.e., $R_{\star,1/2} = r_\star$.

Local Group and Milky Way-analog in the SAGA survey (Geha et al. 2017; Mao et al. 2021, 2024) for $M_\star \gtrsim 10^5 M_\odot$.

Figure 13 shows the pericentric distance r_{peri} versus the current distance r_{curr} of the Milky Way dwarfs (circles) and the FIRE-2 simulations (gray contours, shown at every 10% interval. Blue and orange circles represent dwarfs within and outside of the 90% contour, respectively. Similarly, ultra-faint dwarfs ($M_V > -7.7$) are marked as squares and not expected to be resolved by FIRE-2. The vertical and horizontal dashed lines correspond to 0.2 and 0.5 times R_{200m} of the Milky Way, assumed to be 365 ± 76 kpc (Deason et al. 2020). Orbital properties are taken from Pace et al. (2022), which uses mixture models to determine the systemic proper motions of the Milky Way satellites and subsequently fits their orbits, accounting for both the Milky Way and LMC potentials. About half of the Milky Way satellites are within the 90% contour of the simulations, which typically occupy the high r_{peri} and r_{curr} regions. This difference highlights a numerical limitation: simulated satellites too close to their hosts are often disrupted and become much more challenging to track with ROCKSTAR (Behroozi et al. 2013; Mansfield et al. 2024).

There are 13 satellites within the 90% contours that have a mean $r_{\text{peri}} < 0.2R_{200m}$ and $r_{\text{curr}} < 0.5R_{200m}$ of the Milky Way. These satellites belong to the more disrupted tidal bins in Section 5, so we expect GRAPHNPE to perform worse for these cases.

The 14 satellites outside of the 90% contour are annotated in Figure 13. Note that all of these satellites are necessarily disrupted—some may have large densities and thus are more resilient to tidal effects. However, for those that are disrupted, the degree of tidal effects may not be fully captured in the FIRE-2 simulations due to numerical limitations. As such, the results presented in Section 5 may require extrapolation to account for these effects.

Additionally, we list satellites identified as potentially disrupted in Pace et al. (2022) and other studies. Tucana III and Antlia III has been identified as being tidally disrupted based on priors studies (see Drlica-Wagner et al. 2015; Shipp et al. 2018; Li et al. 2018 for Tucana III and Ji et al. 2021; Vivas et al. 2022 for Antlia III). Pace et al. (2022) also identified additional satellites potentially undergoing tidal disruption due to their low pericentric distances and low densities relative to the Milky Way, including Bootes I, Bootes III, Crater II, Grus II, Segue 2, and Tucana IV. In a forthcoming work, we will provide a more detailed discussion of the tidal effects on these satellites and their implications for GRAPHNPE. For now, we direct readers to the discussion in Pace et al. (2022) and the references therein.

7.2 Comparison with distribution function fitting

As briefly discussed in Section 1, our method is a form of massively accelerated distribution function (DF) fitting (Wilkinson et al. 2002; Pascale et al. 2018; Read et al. 2021). Traditional DF methods rely on the assumption of steady-state dynamics, where the phase-space distribution is expressed as a function of conserved integrals of motion $I(\{\vec{x}_i, \vec{v}_i\})$. Under this assumption, the likelihood is given by,

$$\mathcal{L}_{\text{DF}} = \prod_{i=1}^N f(I(\{\vec{x}_i, \vec{v}_i\}); \theta). \quad (12)$$

The integral of motion $I(\{\vec{x}_i, \vec{v}_i\})$ is chosen to be the energy-angular momentum $\{E, L\}$, or the action-angle variable J , as in Read et al. (2021). The likelihood of a given phase-space coordinate, $\{\vec{x}_i, \vec{v}_i\}$, is calculated by first computing I and evaluating the DF at $f(I)$. Conversely, to sample $\{\vec{x}_i, \vec{v}_i\}$ from $f(I)$, AGAMA can generate samples

in the space of I and then transform them back to phase-space coordinates using an improved version of the torus mapping approach in Binney & McMillan (2016).

We expect the performance of GRAPHNPE to be comparable to the AGAMA DF method Read et al. (2021). However, we note a few key advantages:

First, as with other NPE methods, GRAPHNPE is amortized, i.e. once trained, the model can be applied to any observation without re-training. In contrast, likelihood-based methods, including DF methods, must be rerun for each new dataset. This significantly limits the parameter space that can be explored with these models, particularly when extending to non-spherical cases. For systems with 1000 tracers, the AGAMA DF method in Read et al. (2021) takes approximately $O(10^3)$ minutes per run, comparable to the total training time of GRAPHNPE. However, once trained, GRAPHNPE performs inference in milliseconds, regardless of the tracer counts, making it far more efficient for large-scale applications. While computational cost may not be a major concern for real observations, given that only a few dwarf galaxies have been observed (Pace 2024), it becomes critical when testing and exploring the modeling parameter space on mock simulations. For instance, the detailed evaluations of mass modeling and tidal effects in Section 5 would be impractical with DF methods on the full FIRE-2 mock dataset, due to their high computational cost.

Second, to evaluate the likelihood, traditional DF methods must map an observed phase-space coordinate $\{\vec{x}_i, \vec{v}_i\}$ to the integrals of motion I . For datasets with incomplete phase-space information (e.g., missing proper motions or line-of-sight distances), these methods require marginalization over the missing coordinates through Monte Carlo sampling and evaluation of complex multidimensional integrals. This process is not only computationally intensive, but can also introduce numerical noise. In contrast, the NPE likelihood in Equation 1 is never explicitly evaluated, but instead learned implicitly through the training process. Consequently, generating the training set requires only sampling from $f(I)$, mapping $I \rightarrow \{\vec{x}_i, \vec{v}_i\}$, and discarding the corresponding components of $\{\vec{x}_i, \vec{v}_i\}$ that are missing in the data. This approach thus entirely circumvents the need for complex marginalization integrals required in traditional DF methods.

Finally, it is unclear how traditional DF methods can be extended to model non-equilibrium dynamics. While GRAPHNPE in this study also assumes pseudo-dynamical equilibrium—via the use of orbital integrals of motion for generating the training set—this assumption is not intrinsic to the method. By construction, GRAPHNPE can incorporate more general dynamical states simply by expanding the training data to include systems undergoing tidal disruption, mergers, or other out-of-equilibrium processes. This flexibility makes simulation-based approaches such as GRAPHNPE a promising path forward for extending mass modeling beyond equilibrium assumptions (e.g. Ural et al. 2015; Widmark & Johnston 2025), an avenue we leave for future work.

8 CONCLUSION

Dwarf galaxies and their mass density profiles hold the key to understanding structure formation and the particle nature of DM. Accurate mass modeling of these galaxies is essential for constraining DM properties and distinguishing between different theoretical scenarios, such as CDM and SIDM. As new spectroscopic data from surveys such as MUSE-Faint (Zoutendijk et al. 2020, 2021b,a; Júlio et al. 2023) and DELVE (DECam Local Volume Exploration sur-

vey; Drlica-Wagner et al. 2022; Tan et al. 2025) continue to push the limits of kinematic measurements to the ultra-faint regime, it is increasingly important to develop robust modeling techniques capable of extracting meaningful constraints even with low stellar tracer counts. Traditional Jeans-based methods often struggle in this regime due to sparse kinematic data, making it critical to adopt approaches that maximize the information extracted from limited observations.

In this work, we assess the performance of GraphNPE, an SBI framework first introduced in Nguyen et al. (2023), on simulated dwarf galaxies in Milky Way-like galactic environments. The framework employs GNNs and normalizing flows to infer DM density profiles from line-of-sight stellar velocities. We train GraphNPE on a Monte Carlo simulation of dwarf galaxies, where each system is generated by sampling from a spherical equilibrium DF model. Unlike traditional Jeans-based approaches, GraphNPE can leverage the full phase-space distribution of tracers, incorporating higher-order velocity moments and spatial correlations to maximize the information extracted from stellar kinematics.

We apply our framework to 96 dwarf galaxies from the five CDM simulations of the Latte (Wetzel et al. 2016, 2023) suite of FIRE-2 simulations (Hopkins et al. 2018), along with 61 satellites from the SIDM simulations with a self-interaction cross-section of $\sigma/m = 1 \text{ cm}^2/\text{g}$ and 46 from those with $\sigma/m = 10 \text{ cm}^2/\text{g}$ (Sameie et al. 2021; Vargya et al. 2022; Arora et al. 2024).

In Sections 4.1 and 4.2, we compared GraphNPE to two Jeans-based methods: a Simple Jeans model using an unbinned Gaussian likelihood (Strigari et al. 2008), and GRAVSPHERE, a higher-order moment method incorporating virial shape parameters (Read & Steger 2017). We randomly selected 36 and 16 FIRE-2 galaxies for these comparisons, respectively. For each galaxy, we assess the inferred DM density $\rho(r)$ and velocity anisotropy profiles $\beta(r)$, as well as three mass and structural parameters: the peak circular velocity V_{\max} , the half-stellar radius mass $M_{r_{*,1/2}}$, and the peak virial mass M_{200m}^{peak} . We summarize our findings below:

- Compared to Simple Jeans, we find that GraphNPE produces tighter and more accurate constraints on the DM density and anisotropy profiles. This is particularly evident when tracer counts are low, as GraphNPE continues to provide robust constraints, whereas Simple Jeans struggles to reliably recover the density and anisotropy profiles. When tracer counts are high, however, the performances of GraphNPE and Simple Jeans become more comparable, as the additional line-of-sight velocities can help constrain the anisotropy and mitigate the mass-anisotropy degeneracy.
- When we compare GraphNPE to the GRAVSPHERE higher-order Jeans model, we find that the performance is overall similar, with discrepancies primarily arising from differences in modeling choices and prior distributions. Notably, GRAVSPHERE tends to favor steeper inner and outer density profiles due to its prior distribution.
- To assess the accuracy of different methods in recovering V_{\max} , $M_{r_{*,1/2}}$, and M_{200m} , we compute the absolute errors and square errors between the predicted and true values of these parameters, as well as the negative log-likelihood of the predicted posterior distributions. We summarize these metrics in Table 3 for GraphNPE versus Jeans and Table 4 for GraphNPE versus GRAVSPHERE. Overall, GraphNPE significantly outperforms the Jeans model across all metrics. While GraphNPE also achieves slightly better recovery of mass and structural parameters compared to GRAVSPHERE, the improvements are relatively modest. We discuss key advantages of GraphNPE against traditional modeling methods such as GRAVSPHERE and DF fitting in Section 7.2.

In Section 5, we analyzed how well GraphNPE can recover V_{\max} and M_{200m}^{peak} across different levels of tidal effects n in the full FIRE-2 mock dataset. Our focus is not on satellites that are heavily disrupted, but rather on those where tidal signatures, such as tidal tails or velocity gradients, are subtle or difficult to detect, a scenario found to be common in Shipp et al. (2023). To quantify tidal effects, we consider four orbital parameters: the number of pericentric passages N_{peri} , the last pericentric distance r_{peri} , the current distance from the host r_{curr} , and the time since the last pericentric passage t_{peri} . We present results on the impact of tidal effects on the recovery of V_{\max} and M_{200m}^{peak} , with Section 5.2 providing a detailed example using three N_{peri} bins and Section 5.3 extending the analysis to the remaining orbital parameters. Our conclusion is as follows:

- Overall, the recovery is strong for both V_{\max} and M_{200m}^{peak} . We observe that GraphNPE systematically overestimates both parameters for the more tidally disrupted bins. Despite this, for the full FIRE-2 dataset, GraphNPE recovers V_{\max} within the 68% and 95% confidence intervals for 64% and 93% of galaxies, respectively, while for M_{200m}^{peak} , the corresponding fractions are 71% and 95%.
- The overestimation is primarily observed in lower-mass galaxies ($V_{\max} < 20 \text{ km/s}$ and $M_{200m}^{\text{peak}} < 10^9 M_{\odot}$) that have undergone significant tidal disruption. In contrast, GraphNPE remains robust for more massive galaxies, even when they experience strong tidal effects. The mean absolute residuals in V_{\max} increase from 10% in the least disrupted galaxies to 20% in the most disrupted ones, while for M_{200m}^{peak} , they rise from 0.2 to 0.4 dex.
- For V_{\max} , the coverage fraction, defined as the proportion of true values that fall within the expected confidence interval, shows a clear decreasing trend in the more disrupted bins for all orbital parameters. In contrast, the trends for M_{200m}^{peak} are much less pronounced (see Table 5).
- Of the four orbital parameters, r_{peri} shows the clearest trends, with the 68% coverage drops significantly for the most disrupted bin ($r_{\text{peri}} < 0.2 R_{\text{host},200m}^{\text{peri}}$), consistent with past findings (e.g. Shipp et al. 2023, 2024; Montero-Dorta et al. 2024). However, even in this case, GraphNPE recovers V_{\max} within the 68% and 95% confidence interval for approximately 50% and 90% of the samples, respectively. Similarly, for M_{200m}^{peak} , around 60% of the sample falls within the 68% confidence interval, while just under 90% falls within the 95% confidence interval.
- We find that GraphNPE tends to be slightly conservative when predicting M_{200m}^{peak} , as indicated by coverage fractions that exceed their expected confidence intervals, typically by a few percents. For example, across the entire population, the coverage fraction is 71% at the 68% confidence level (see Table 5). This weaker constraint likely comes from the additional uncertainties introduced when extrapolating the DM density profiles to r_{200m} . These uncertainties could be mitigated by employing more accurate and flexible tail models for the DM density profiles or by extending the tracer populations to larger radii.

Lastly, we briefly compare the performance of GraphNPE across the CDM and SIDM simulations with different cross-sections by examining their residuals. Overall, the model performs similarly for CDM and the high-interaction SIDM scenario ($\sigma/m = 10 \text{ cm}^2/\text{g}$), while showing lower coverage for the low-interaction SIDM case ($\sigma/m = 1 \text{ cm}^2/\text{g}$). Despite some differences in distributions, the mean residuals for both V_{\max} and M_{200m} remain consistent within 1σ across simulations. These findings demonstrate the robustness

of GRAPHNPE while highlighting the need for further testing with larger datasets to draw more definitive conclusions.

It is important to note that GRAPHNPE is trained on idealized dynamical models assuming spherical equilibrium, rather than on galaxies in complex hydrodynamical environments like those in FIRE. A common challenge for machine learning methods is their ability to generalize to out-of-distribution data, where deviations from the training distribution can degrade performance. Despite this, GRAPHNPE robustly recovers the density and velocity anisotropy profiles, as well as key mass and structural parameters, even when applied to galaxies in realistic hydrodynamical simulations.

We further highlight the novelty and importance of evaluating how well mass modeling methods can recover V_{\max} and M_{200m}^{peak} of dwarf galaxy simulations under realistic galactic environments. A previous study by [Genina et al. \(2020\)](#) conducted a comprehensive test of GRAVSPHERE on APOSTLE dwarf galaxies, demonstrating that its recovery of $M_{r,\star,1/2}$ is comparable to that of mass estimators while also examining the impact of tidal effects. Our work additionally examines V_{\max} and M_{200m}^{peak} , both of which are crucial for constraining cosmological models (e.g. [Burkert 1995](#); [Bullock et al. 2001](#); [van den Bosch et al. 2003](#); [Rodríguez-Puebla et al. 2016](#); [Read & Erkal 2019](#); [Kim & Peter 2021](#)). This evaluation is particularly timely, as discrepancies between leading mass modeling approaches, such as CJAM ([Watkins et al. 2013](#)) and GRAVSPHERE, suggest that different methods can yield significantly different mass estimates when applied to the same dataset ([Zoutendijk et al. 2021a](#)), underscoring the need for rigorous validation.

In ongoing work, we aim to apply GRAPHNPE to observed dwarf galaxies in the Milky Way and the Local Group. We discuss key differences between galaxies in our mock FIRE-2 dataset and real observations in Section 7. Before applying GRAPHNPE to observational data, two key considerations need to be addressed: incorporating measurement uncertainties and masking contamination from background and binary stars. Both can be readily handled by integrating them into the training or inference process (see e.g. [Wang et al. 2023](#)).

In this work, we establish GRAPHNPE as a robust and efficient DF-based mapping method for inferring DM density profiles in dwarf galaxies, providing a promising avenue for constraining DM models. By leveraging machine learning, GRAPHNPE enables rapid, amortized inference, making it particularly well-suited for large-scale applications such as systematically exploring the mass modeling parameter space and investigating the effects of tidal disruption across mock simulations.

While this study applies GRAPHNPE to spherical equilibrium models, its framework is not inherently restricted to such cases, unlike traditional DF methods. GRAPHNPE can be extended to out-of-equilibrium systems (e.g. [Ural et al. 2015](#); [Widmark & Johnston 2025](#)) by expanding the training set to include galaxies undergoing tidal disruption, mergers, and other dynamical perturbations, which we will pursue in future work. This will allow GRAPHNPE to recover mass profiles in systems where equilibrium assumptions break down, further broadening its applicability to realistic astrophysical environments. More broadly, our work highlight the power of simulation-based inference and graph-based learning in astrophysics, paving the way for more adaptable and data-driven approaches to mass modeling in the era of increasingly precise kinematic surveys.

ACKNOWLEDGMENTS

We thank Jenna Samuel, Nondh Panithanpaisal, Nora Shipp, Stephanie O’Neil, Arpit Arora, Xiaowei Ou, Tjitske Starkenburg for helpful discussions.

TN, SM, and LN is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, <http://iaifi.org/>). TN is also supported by a CIERA Postdoctoral Fellowship. LN is also supported by the Sloan Fellowship, the NSF CAREER award 2337864, NSF award 2307788. JIR would like to acknowledge support from STFC grants ST/Y002865/1 and ST/Y002857/1. CAFG was supported by NSF through grants AST-2108230 and AST-2307327; by NASA through grants 21-ATP21-0036 and 23-ATP23-0008; and by STScI through grant JWST-AR-03252.001-A. AW received support from NSF, via CAREER award AST-2045928 and grant AST-2107772. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics of U.S. Department of Energy under grant Contract Number DE-SC0012567.

This work used Bridges-2 ([Brown et al. 2021](#)) at Pittsburgh Supercomputing Center through allocation phy210068p from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Other numerical calculations were run on the Northwestern computer cluster Quest, the Caltech computer cluster Wheeler, Frontera allocation FTA-Hopkins/AST20016 supported by the NSF and TACC, XSEDE/ACCESS allocations ACI-1548562, TGAST140023, and TG-AST140064 also supported by the NSF and TACC, and NASA HEC allocations SMD-16-7561, SMD-17-1204, and SMD-16-7592. The computations in this work were, in part, run at facilities supported by the Scientific Computing Core at the Flatiron Institute, a division of the Simons Foundation. The data used in this work were, in part, hosted on equipment supported by the Scientific Computing Core at the Flatiron Institute, a division of the Simons Foundation.

SOFTWARE

This research makes use of the following packages: AGAMA ([Vasiliev 2019](#)), GRAVSPHERE ([Read et al. 2018](#); [Collins et al. 2021](#)), IPYTHON ([Perez & Granger 2007](#)), JUPYTER ([Kluyver et al. 2016](#)), MATPLOTLIB ([Hunter 2007](#)), NUMPY ([Harris et al. 2020](#)), PYTORCH ([Paszke et al. 2019](#)), PYTORCH GEOMETRIC ([Fey & Lenssen 2019](#)), PYTORCH LIGHTNING ([Falcon et al. 2020](#)), SciPY ([Virtanen et al. 2020](#)), ZUKO ([Rozet et al. 2024](#))

DATA AVAILABILITY

The CDM FIRE-2 simulations are publicly available through the FIRE project website (<http://fire.northwestern.edu/data/>). The SIDM simulation data used in this study are not publicly available but can be provided upon reasonable request to the corresponding author. Additional simulation data, including the specific analysis outputs and derived data products generated for this study, are available upon reasonable request to the corresponding author. The trained model weights are available at upon reasonable request to the corresponding author.

REFERENCES

- Abbott T. M. C., et al., 2018, *ApJS*, **239**, 18
 Abbott T. M. C., et al., 2021, *ApJS*, **255**, 20
 Ackermann M., et al., 2011, *Phys. Rev. Lett.*, **107**, 241302
 Ackermann M., et al., 2014, *Phys. Rev. D*, **89**, 042001
 Ackermann M., et al., 2015, *Phys. Rev. Lett.*, **115**, 231301
 Ajello M., McDaniel A., Karwin C., di Mauro M., Drlica-Wagner A., Fermi-Lat Collaboration 2024, in American Astronomical Society Meeting Abstracts. p. 315.03
 Alvey J., et al., 2021, *MNRAS*, **501**, 1188
 Amorisco N. C., Evans N. W., 2012, *MNRAS*, **419**, 184
 Anderson B., Chiang J., Cohen-Tanugi J., Conrad J., Drlica-Wagner A., Llena Garde M., Zimmer S., 2015, arXiv e-prints, p. arXiv:1502.03081
 Arora A., et al., 2024, *ApJ*, **974**, 223
 Ashton G., et al., 2022, *Nature Reviews Methods Primers*, **2**, 39
 Baes M., van Hese E., 2007, *A&A*, **471**, 419
 Balberg S., Shapiro S. L., Inagaki S., 2002, *ApJ*, **568**, 475
 Barber C., Starkenburg E., Navarro J. F., McConnachie A. W., 2015, *MNRAS*, **447**, 1112
 Battaglia G., Sollima A., Nipoti C., 2015, *MNRAS*, **454**, 2401
 Battaglia P. W., et al., 2018, arXiv e-prints, p. arXiv:1806.01261
 Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, *ApJ*, **762**, 109
 Bennett C. L., et al., 2013, *ApJS*, **208**, 20
 Binney J., 1980, *MNRAS*, **190**, 873
 Binney J., McMillan P. J., 2016, *MNRAS*, **456**, 1982
 Binney J., Tremaine S., 2008, Galactic Dynamics: Second Edition
 Bode P., Ostriker J. P., Turok N., 2001, *ApJ*, **556**, 93
 Bonnivard V., Combet C., Maurin D., Walker M. G., 2015, *MNRAS*, **446**, 3002
 Bose S., Hellwing W. A., Frenk C. S., Jenkins A., Lovell M. R., Helly J. C., Li B., 2016, *MNRAS*, **455**, 318
 Boylan-Kolchin M., Bullock J. S., Kaplinghat M., 2011, *MNRAS*, **415**, L40
 Breddels M. A., Helmi A., 2013, *A&A*, **558**, A35
 Breddels M. A., Helmi A., 2014, *ApJ*, **791**, L3
 Brown S. T., Buitrago P., Hanna E., Sanielevici S., Scibek R., Nystrom N. A., 2021, in Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions. PEARC '21. Association for Computing Machinery, New York, NY, USA, doi:10.1145/3437359.3465593, <https://doi.org/10.1145/3437359.3465593>
 Buck T., Macciò A. V., Dutton A. A., Obreja A., Frings J., 2019, *MNRAS*, **483**, 1314
 Bullock J. S., Boylan-Kolchin M., 2017, *ARA&A*, **55**, 343
 Bullock J. S., Kolatt T. S., Sigad Y., Somerville R. S., Kravtsov A. V., Klypin A. A., Primack J. R., Dekel A., 2001, *MNRAS*, **321**, 559
 Burkert A., 1995, *ApJ*, **447**, L25
 Campbell D. J. R., et al., 2017, *MNRAS*, **469**, 2335
 Cappellari M., 2008, *MNRAS*, **390**, 71
 Cappellari M., 2015, arXiv e-prints, p. arXiv:1504.05533
 Chang L. J., Necib L., 2021, *MNRAS*, **507**, 4715
 Charbonnier A., et al., 2011, *MNRAS*, **418**, 1526
 Chiang B. T., van den Bosch F. C., Schive H.-Y., 2024, arXiv e-prints, p. arXiv:2411.03192
 Cirelli M., et al., 2011, *J. Cosmology Astropart. Phys.*, **2011**, 051
 Collins M. L. M., et al., 2021, *MNRAS*, **505**, 5686
 Correa C. A., 2021, *MNRAS*, **503**, 920
 Cranmer K., Louppe G., 2016, *J. Brief Ideas*
 Cranmer K., Brehmer J., Louppe G., 2020, *Proceedings of the National Academy of Science*, **117**, 30055
 Creasey P., Sameie O., Sales L. V., Yu H.-B., Vogelsberger M., Zavala J., 2017, *MNRAS*, **468**, 2283
 Cuesta-Lazaro C., Mishra-Sharma S., 2023, arXiv e-prints, p. arXiv:2311.17141
 Cyr-Racine F.-Y., Sigurdson K., 2013, *Phys. Rev. D*, **87**, 103515
 Dalal N., Kravtsov A., 2022, arXiv e-prints, p. arXiv:2203.05750
 De Leo M., Read J. I., Noël N. E. D., Erkal D., Massana P., Carrera R., 2024, *MNRAS*, **535**, 1015
 Deason A. J., Fattah A., Frenk C. S., Grand R. J. J., Oman K. A., Garrison-Kimmel S., Simpson C. M., Navarro J. F., 2020, *MNRAS*, **496**, 3929
 Dekel A., Arad I., Devor J., Birnboim Y., 2003, *ApJ*, **588**, 680
 Di Cintio A., Brook C. B., Macciò A. V., Stinson G. S., Knebe A., Dutton A. A., Wadsley J., 2014a, *MNRAS*, **437**, 415
 Di Cintio A., Brook C. B., Dutton A. A., Macciò A. V., Stinson G. S., Knebe A., 2014b, *MNRAS*, **441**, 2986
 Drlica-Wagner A., et al., 2015, *ApJ*, **813**, 109
 Drlica-Wagner A., et al., 2020, *ApJ*, **893**, 47
 Drlica-Wagner A., et al., 2022, *ApJS*, **261**, 38
 Durkan C., Bekasov A., Murray I., Papamakarios G., 2019, arXiv e-prints, p. arXiv:1906.04032
 El-Badry K., Wetzel A., Geha M., Hopkins P. F., Kereš D., Chan T. K., Faucher-Giguère C.-A., 2016, *ApJ*, **820**, 131
 El-Badry K., Wetzel A. R., Geha M., Quataert E., Hopkins P. F., Kereš D., Chan T. K., Faucher-Giguère C.-A., 2017, *ApJ*, **835**, 193
 Emsellem E., Monnet G., Bacon R., 1994, *A&A*, **285**, 723
 Errani R., Peñarrubia J., Walker M. G., 2018, *MNRAS*, **481**, 5073
 Escala I., et al., 2018, *MNRAS*, **474**, 2194
 Falcon W., et al., 2020, PyTorchLightning/pytorch-lightning: 0.7.6 release, doi:10.5281/zenodo.3828935, <https://doi.org/10.5281/zenodo.3828935>
 Fattah A., Navarro J. F., Frenk C. S., Oman K. A., Sawala T., Schaller M., 2018, *MNRAS*, **476**, 3816
 Faucher-Giguère C.-A., Lidz A., Zaldarriaga M., Hernquist L., 2009, *ApJ*, **703**, 1416
 Fey M., Lenssen J. E., 2019, arXiv e-prints, p. arXiv:1903.02428
 Fitts A., et al., 2017, *MNRAS*, **471**, 3547
 Fitts A., et al., 2019, *MNRAS*, **490**, 962
 Flaugher B., et al., 2015, *AJ*, **150**, 150
 Flores R. A., Primack J. R., 1994, *ApJ*, **427**, L1
 Frings J., Macciò A., Buck T., Penzo C., Dutton A., Blank M., Obreja A., 2017, *MNRAS*, **472**, 3378
 Garrison-Kimmel S., et al., 2017, *MNRAS*, **471**, 1709
 Garrison-Kimmel S., et al., 2019, *MNRAS*, **487**, 1380
 Geha M., et al., 2017, *ApJ*, **847**, 4
 Gemmell C., Roy S., Shen X., Curtin D., Lisanti M., Murray N., Hopkins P. F., 2024, *ApJ*, **967**, 21
 Genina A., et al., 2020, *MNRAS*, **498**, 144
 Geringer-Sameth A., Koushiappas S. M., 2011, *Phys. Rev. Lett.*, **107**, 241303
 Geringer-Sameth A., Koushiappas S. M., Walker M. G., 2015a, *Phys. Rev. D*, **91**, 083535
 Geringer-Sameth A., Koushiappas S. M., Walker M., 2015b, *ApJ*, **801**, 74
 Germain M., Gregor K., Murray I., Larochelle H., 2015, arXiv e-prints, p. arXiv:1502.03509
 Gilmer J., Schoenholz S. S., Riley P. F., Vinyals O., Dahl G. E., 2017, arXiv e-prints, p. arXiv:1704.01212
 Górski M., Pietrzyński G., Gieren W., 2011, *AJ*, **141**, 194
 Governato F., et al., 2010, *Nature*, **463**, 203
 Hahn O., Abel T., 2011, *MNRAS*, **415**, 2101
 Hansen T. T., Simon J. D., Li T. S., Sharkey D., Ji A. P., Thompson I. B., Reggiani H. M., Galarza J. Y., 2024, *ApJ*, **968**, 21
 Harris C. R., et al., 2020, *Nature*, **585**, 357
 Hayashi K., Chiba M., Ishiyama T., 2020a, *ApJ*, **904**, 45
 Hayashi K., Chiba M., Ishiyama T., 2020b, *ApJ*, **904**, 45
 Hayashi K., Ferreira E. G. M., Chan H. Y. J., 2021, *ApJ*, **912**, L3
 Hayashi K., Hirai Y., Chiba M., Ishiyama T., 2023, *ApJ*, **953**, 185
 He K., Zhang X., Ren S., Sun J., 2015, arXiv e-prints, p. arXiv:1502.01852
 Hendrycks D., Gimpel K., 2016, arXiv e-prints, p. arXiv:1606.08415
 Hopkins P. F., 2014, GIZMO: Multi-method magneto-hydrodynamics+gravity code, Astrophysics Source Code Library, record ascl:1410.003
 Hopkins P. F., 2015, *MNRAS*, **450**, 53
 Hopkins P. F., Narayanan D., Murray N., 2013, *MNRAS*, **432**, 2647
 Hopkins P. F., et al., 2018, *MNRAS*, **480**, 800
 Hunter J. D., 2007, *Computing in Science and Engineering*, **9**, 90
 Jeans J. H., 1915, *MNRAS*, **76**, 70
 Ji A. P., et al., 2021, *ApJ*, **921**, 32
 Jimenez Rezende D., Mohamed S., 2015, arXiv e-prints, p. arXiv:1505.05770

- Jones G. L., Qin Q., 2022, *Annual Review of Statistics and Its Application*, **9**, 557
- Júlio M. P., et al., 2023, *A&A*, **678**, A38
- Kaplan D. E., Krnjaic G. Z., Rehmann K. R., Wells C. M., 2010, *J. Cosmology Astropart. Phys.*, **2010**, 021
- Kaplinghat M., Tulin S., Yu H.-B., 2016, *Phys. Rev. Lett.*, **116**, 041302
- Kazantzidis S., Łokas E. L., Callegari S., Mayer L., Moustakas L. A., 2011, *ApJ*, **726**, 98
- Kennedy R., Frenk C., Cole S., Benson A., 2014, *MNRAS*, **442**, 2487
- Kerrick C., Walker M. G., Peñarrubia J., Koposov S. E., 2022, *ApJ*, **929**, 77
- Kim S. Y., Peter A. H. G., 2021, *arXiv e-prints*, p. arXiv:2106.09050
- Kim J.-h., et al., 2014, *ApJS*, **210**, 14
- Kim S. Y., Peter A. H. G., Hargis J. R., 2018, *Phys. Rev. Lett.*, **121**, 211302
- Kingma D. P., Ba J., 2014, arXiv preprint arXiv:1412.6980
- Kipf T. N., Welling M., 2016, *arXiv e-prints*, p. arXiv:1609.02907
- Kluyver T., et al., 2016, in , IOS Press. pp 87–90, doi:10.3233/978-1-61499-649-1-87
- Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, *ApJ*, **522**, 82
- Koda J., Shapiro P. R., 2011, *MNRAS*, **415**, 1125
- Kowalczyk K., del Pino A., Łokas E. L., Valluri M., 2019, *MNRAS*, **482**, 5241
- Kunkel W. E., Demers S., 1976, in *The Galaxy and the Local Group*. p. 241
- Lazar A., et al., 2020, *MNRAS*, **497**, 2393
- Leane R. K., Slatyer T. R., 2019, *arXiv e-prints*, p. arXiv:1904.08430
- Li T. S., et al., 2018, *ApJ*, **866**, 22
- Łokas E. L., 2009, *MNRAS*, **394**, L102
- Łokas E. L., Mamon G. A., 2003, *MNRAS*, **343**, 401
- Loshchilov I., Hutter F., 2016, *arXiv e-prints*, p. arXiv:1608.03983
- Loshchilov I., Hutter F., 2019, in *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bkg6RiCqY7>
- Lovell M. R., Frenk C. S., Eke V. R., Jenkins A., Gao L., Theuns T., 2014, *MNRAS*, **439**, 300
- Lynden-Bell D., 1976, *MNRAS*, **174**, 695
- Ma X., Hopkins P. F., Faucher-Giguère C.-A., Zolman N., Muratov A. L., Kereš D., Quataert E., 2016, *MNRAS*, **456**, 2140
- Mamon G. A., Biviano A., Boué G., 2013, *MNRAS*, **429**, 3079
- Mansfield P., Darragh-Ford E., Wang Y., Nadler E. O., Diemer B., Wechsler R. H., 2024, *ApJ*, **970**, 178
- Mao Y.-Y., Geha M., Wechsler R. H., Weiner B., Tollerud E. J., Nadler E. O., Kallivayalil N., 2021, *ApJ*, **907**, 85
- Mao Y.-Y., et al., 2024, *ApJ*, **976**, 117
- Martinez G. D., 2015, *MNRAS*, **451**, 2524
- Massari D., Helmi A., Mucciarelli A., Sales L. V., Spina L., Tolstoy E., 2020, *A&A*, **633**, A36
- Mayer L., Mastropietro C., Wadsley J., Stadel J., Moore B., 2006, *MNRAS*, **369**, 1021
- Mazziotta M. N., Loparco F., de Palma F., Giglietto N., 2012, *Astroparticle Physics*, **37**, 26
- Merrifield M. R., Kent S. M., 1990, *ApJ*, **99**, 1548
- Merritt D., 1985, *ApJ*, **90**, 1027
- Monaghan J. J., Lattanzio J. C., 1985, *A&A*, **149**, 135
- Montero-Dorta A. D., Rodriguez F., Artale M. C., Smith R., Chaves-Montero J., 2024, *MNRAS*, **527**, 5868
- Moore B., 1994, *Nature*, **370**, 629
- Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, *ApJ*, **524**, L19
- Muñoz R. R., Côté P., Santana F. A., Geha M., Simon J. D., Oyarzún G. A., Stetson P. B., Djorgovski S. G., 2018, *ApJ*, **860**, 66
- Muni C., Pontzen A., Read J. I., Agertz O., Rey M. P., Taylor E., Kim S. Y., Gray E. I., 2025, *MNRAS*, **536**, 314
- Nadler E. O., Gluscevic V., Boddy K. K., Wechsler R. H., 2019, *ApJ*, **878**, L32
- Navarro J. F., Eke V. R., Frenk C. S., 1996, *MNRAS*, **283**, L72
- Navarro J. F., Frenk C. S., White S. D. M., 1997, *ApJ*, **490**, 493
- Necib L., Lisanti M., Garrison-Kimmel S., Wetzel A., Sanderson R., Hopkins P. F., Faucher-Giguère C.-A., Kereš D., 2019, *ApJ*, **883**, 27
- Nguyen T., Mishra-Sharma S., Williams R., Necib L., 2023, *Phys. Rev. D*, **107**, 043015
- Nishikawa H., Boddy K. K., Kaplinghat M., 2020, *Phys. Rev. D*, **101**, 063009
- Oh S.-H., de Blok W. J. G., Brinks E., Walter F., Kennicutt Robert C. J., 2011, *AJ*, **141**, 193
- Oh S.-H., et al., 2015, *AJ*, **149**, 180
- Oman K. A., et al., 2015, *MNRAS*, **452**, 3650
- Osipkov L. P., 1979, *Pisma v Astronomicheskii Zhurnal*, **5**, 77
- Pace A. B., 2024, *arXiv e-prints*, p. arXiv:2411.07424
- Pace A. B., Erkal D., Li T. S., 2022, *ApJ*, **940**, 136
- Papamakarios G., Murray I., 2016, *arXiv e-prints*, p. arXiv:1605.06376
- Papamakarios G., Pavlakou T., Murray I., 2017, *arXiv e-prints*, p. arXiv:1705.07057
- Papamakarios G., Nalisnick E., Jimenez Rezende D., Mohamed S., Lakshminarayanan B., 2019, *arXiv e-prints*, p. arXiv:1912.02762
- Pascale R., Posti L., Nipoti C., Binney J., 2018, *MNRAS*, **480**, 927
- Paszke A., et al., 2019, *arXiv e-prints*, p. arXiv:1912.01703
- Pawlowski M. S., Famaey B., Merritt D., Kroupa P., 2015, *ApJ*, **815**, 19
- Peñarrubia J., Navarro J. F., McConnachie A. W., 2008, *ApJ*, **673**, 226
- Perez F., Granger B. E., 2007, *Computing in Science and Engineering*, **9**, 21
- Peter A. H. G., Rocha M., Bullock J. S., Kaplinghat M., 2013, *MNRAS*, **430**, 105
- Planck Collaboration et al., 2020, *A&A*, **641**, A6
- Plummer H. C., 1911, *MNRAS*, **71**, 460
- Pontzen A., Governato F., 2012, *MNRAS*, **421**, 3464
- Pontzen A., Governato F., 2014, *Nature*, **506**, 171
- Power C., Navarro J. F., Jenkins A., Frenk C. S., White S. D. M., Springel V., Stadel J., Quinn T., 2003, *MNRAS*, **338**, 14
- Read J. I., Erkal D., 2019, *MNRAS*, **487**, 5799
- Read J. I., Gilmore G., 2005, *MNRAS*, **356**, 107
- Read J. I., Steger P., 2017, *MNRAS*, **471**, 4541
- Read J. I., Wilkinson M. I., Evans N. W., Gilmore G., Kleyne J. T., 2006, *MNRAS*, **367**, 387
- Read J. I., Walker M. G., Steger P., 2018, *MNRAS*, **481**, 860
- Read J. I., Walker M. G., Steger P., 2019, *MNRAS*, **484**, 1401
- Read J. I., et al., 2021, *MNRAS*, **501**, 978
- Richardson T., Fairbairn M., 2014, *MNRAS*, **441**, 1584
- Robles V. H., et al., 2017, *MNRAS*, **472**, 2945
- Rocha M., Peter A. H. G., Bullock J. S., Kaplinghat M., Garrison-Kimmel S., Ofiorbe J., Moustakas L. A., 2013, *MNRAS*, **430**, 81
- Rodríguez-Puebla A., Behroozi P., Primack J., Klypin A., Lee C., Hellinger D., 2016, *MNRAS*, **462**, 893
- Rojas-Niño A., Read J. I., Aguilar L., Delorme M., 2016, *MNRAS*, **459**, 3349
- Roy S., Shen X., Barron J., Lisanti M., Curtin D., Murray N., Hopkins P. F., 2024, *arXiv e-prints*, p. arXiv:2408.15317
- Rozet F., Divo F., Schnake S., 2024, *probabilists/zuko*: Zuko 1.1.0, doi:10.5281/zenodo.7625672
- Rozo E., et al., 2010, *ApJ*, **708**, 645
- Sales L. V., Wetzel A., Fattahi A., 2022, *Nature Astronomy*, **6**, 897
- Sameie O., et al., 2021, *MNRAS*, **507**, 720
- Samuel J., et al., 2020, *MNRAS*, **491**, 1471
- Sanderson R. E., et al., 2018, *ApJ*, **869**, 12
- Santistevan I. B., Wetzel A., Tollerud E., Sanderson R. E., Samuel J., 2023, *MNRAS*, **518**, 1427
- Santos-Santos I. M. E., et al., 2020, *MNRAS*, **495**, 58
- Sawala T., et al., 2016, *MNRAS*, **457**, 1931
- Scarselli F., Gori M., Tsoi A. C., Hagenbuchner M., Monfardini G., 2009, *IEEE Transactions on Neural Networks*, **20**, 61
- Shen X., Hopkins P. F., Necib L., Jiang F., Boylan-Kolchin M., Wetzel A., 2021, *MNRAS*, **506**, 4421
- Shen X., Hopkins P. F., Necib L., Jiang F., Boylan-Kolchin M., Wetzel A., 2024, *ApJ*, **966**, 131
- Shipp N., et al., 2018, *ApJ*, **862**, 114
- Shipp N., et al., 2023, *ApJ*, **949**, 44
- Shipp N., et al., 2024, *arXiv e-prints*, p. arXiv:2410.09143
- Simon J. D., 2019, *ARA&A*, **57**, 375
- Simpson C. M., Grand R. J. J., Gómez F. A., Marinacci F., Pakmor R., Springel V., Campbell D. J. R., Frenk C. S., 2018, *MNRAS*, **478**, 548
- Skilling J., 2004, in Fischer R., Preuss R., Toussaint U. V., eds, *American Institute of Physics Conference Series* Vol. 735, Bayesian Inference and

- Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. pp 395–405, doi:10.1063/1.1835238
- Speagle J. S., 2019, arXiv e-prints, p. arXiv:1909.12313
- Spekkens K., Giovanelli R., Haynes M. P., 2005, AJ, 129, 2119
- Spencer M. E., Mateo M., Olszewski E. W., Walker M. G., McConnachie A. W., Kirby E. N., 2018, AJ, 156, 257
- Spergel D. N., Steinhardt P. J., 2000, Phys. Rev. Lett., 84, 3760
- Springel V., 2005, MNRAS, 364, 1105
- Springel V., et al., 2005, Nature, 435, 629
- Strigari L. E., Koushiappas S. M., Bullock J. S., Kaplinghat M., 2007a, Phys. Rev. D, 75, 083526
- Strigari L. E., Bullock J. S., Kaplinghat M., 2007b, ApJ, 657, L1
- Strigari L. E., Koushiappas S. M., Bullock J. S., Kaplinghat M., Simon J. D., Geha M., Willman B., 2008, ApJ, 678, 614
- Strigari L. E., Frenk C. S., White S. D. M., 2018, ApJ, 860, 56
- Su K.-Y., Hopkins P. F., Hayward C. C., Faucher-Giguère C.-A., Kereš D., Ma X., Robles V. H., 2017, MNRAS, 471, 144
- Subramanian K., Cen R., Ostriker J. P., 2000, ApJ, 538, 528
- Tan C. Y., Dekker A., Drlica-Wagner A., 2024, arXiv e-prints, p. arXiv:2409.18917
- Tan C. Y., et al., 2025, ApJ, 979, 176
- Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, ApJ, 688, 709
- Toguz F., Kawata D., Seabroke G., Read J. I., 2022, MNRAS, 511, 1757
- Tollerud E. J., Bullock J. S., Strigari L. E., Willman B., 2008, ApJ, 688, 277
- Tomozeiu M., Mayer L., Quinn T., 2016, ApJ, 818, 193
- Tulin S., Yu H.-B., 2018, Phys. Rep., 730, 1
- Ural U., Wilkinson M. I., Read J. I., Walker M. G., 2015, Nature Communications, 6, 7599
- Vargya D., Sanderson R., Sameie O., Boylan-Kolchin M., Hopkins P. F., Wetzel A., Graus A., 2022, MNRAS, 516, 2389
- Vasiliev E., 2019, MNRAS, 482, 1525
- Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., 2017, arXiv e-prints, p. arXiv:1706.03762
- Veličković P., Cucurull G., Casanova A., Romero A., Liò P., Bengio Y., 2017, arXiv e-prints, p. arXiv:1710.10903
- Villanueva-Domingo P., Villaescusa-Navarro F., 2022, ApJ, 937, 115
- Virtanen P., et al., 2020, Nature Methods,
- Vitral E., et al., 2024, ApJ, 970, 1
- Vivas A. K., Martínez-Vázquez C. E., Walker A. R., Belokurov V., Li T. S., Erkal D., 2022, ApJ, 926, 78
- Walker M. G., Peñarrubia J., 2011, ApJ, 742, 20
- Walker M. G., Mateo M., Olszewski E. W., 2009a, AJ, 137, 3100
- Walker M. G., Mateo M., Olszewski E. W., Peñarrubia J., Evans N. W., Gilmore G., 2009b, ApJ, 704, 1274
- Wang M. Y., et al., 2019, ApJ, 881, 118
- Wang B., Leja J., Villar V. A., Speagle J. S., 2023, ApJ, 952, L10
- Watkins L. L., van de Ven G., den Brok M., van den Bosch R. C. E., 2013, MNRAS, 436, 2598
- Wetzel A. R., Hopkins P. F., Kim J.-h., Faucher-Giguère C.-A., Kereš D., Quataert E., 2016, ApJ, 827, L23
- Wetzel A., et al., 2023, ApJS, 265, 44
- White S. D. M., Rees M. J., 1978, MNRAS, 183, 341
- Widmark A., Johnston K. V., 2025, arXiv e-prints, p. arXiv:2501.13148
- Wilkinson M. I., Kleyna J., Evans N. W., Gilmore G., 2002, MNRAS, 330, 778
- Wolf J., Martinez G. D., Bullock J. S., Kaplinghat M., Geha M., Muñoz R. R., Simon J. D., Avedo F. F., 2010, MNRAS, 406, 1220
- Zavala J., Lovell M. R., Vogelsberger M., Burger J. D., 2019, Phys. Rev. D, 100, 063007
- Zhu L., van de Ven G., Watkins L. L., Posti L., 2016, MNRAS, 463, 1117
- Zoutendijk S. L., et al., 2020, A&A, 635, A107
- Zoutendijk S. L., et al., 2021a, arXiv e-prints, p. arXiv:2112.09374
- Zoutendijk S. L., Brinchmann J., Bouché N. F., den Brok M., Krajanović D., Kuijken K., Maseda M. V., Schaye J., 2021b, A&A, 651, A80
- de Lorenzi F., et al., 2009, MNRAS, 395, 76
- de Santi N. S. M., et al., 2023, ApJ, 952, 69

van den Bosch F. C., Mo H. J., Yang X., 2003, MNRAS, 345, 923

APPENDIX A: JEANS DYNAMICAL MODELING

Let $f(\vec{x}, \vec{v})$ be the phase-space distribution function that describes the positions and velocities of tracer stars in a self-gravitating system (i.e. the dwarf galaxy). Following the derivation from Binney (1980); Binney & Tremaine (2008), we assume the system follows the collisionless Boltzmann equations:

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \frac{\partial f}{\partial \vec{x}} - \frac{\partial \Phi}{\partial \vec{x}} \cdot \frac{\partial f}{\partial \vec{v}} = 0, \quad (\text{A1})$$

where Φ is the gravitational potential of the system. For dwarf galaxies, the potential is dominated by their DM components. We can thus ignore contributions of the tracer populations and write the potential in terms of the DM enclosed mass or density profiles.

We will now solve Eq. A1 and derive a relation between Φ and the stellar velocity dispersions. Multiplying by a velocity component v_i and integrating over all velocities, Eq. A1 becomes:

$$\frac{\partial}{\partial t} (v \langle v_j \rangle) + \frac{\partial}{\partial x_i} (v \langle v_i v_j \rangle) + v \frac{\partial \Phi}{\partial x_j} = 0, \quad (\text{A2})$$

where $v = \int d^3 \vec{v} f(\vec{x}, \vec{v})$ is the number density of the tracer stars, and $\langle \rangle$ denotes the average over the velocities. Assuming a steady-state solution, we can ignore the first term with the explicit time derivative. Then, assuming spherical symmetry and working in a spherical coordinate system centered on the dwarf galaxy (r, θ, ϕ) , we may rewrite the equation as:

$$\frac{1}{v} \left[\frac{\partial}{\partial r} (v \sigma_r^2) + \frac{2\beta(r)}{r} (v \sigma_r^2) \right] = -\frac{GM(r)}{r^2}, \quad (\text{A3})$$

where $\sigma_i = \sqrt{\langle v_i^2 \rangle - \langle v_i \rangle^2}$ denotes the velocity dispersion of the system. This is commonly known as the “spherical Jeans equations”. Here, we rewrite the gravitational potential as a function of the enclosed mass $M(r)$ using the Poisson equation, i.e., $\Phi = -GM(r)/r$. We have also introduced the velocity anisotropy term:

$$\beta(r) = 1 - \frac{\sigma_\theta^2 + \sigma_\phi^2}{2\sigma_r^2}, \quad (\text{A4})$$

which effectively captures how the distribution of stellar velocities deviates from isotropy. The velocity anisotropy β ranges from $(-\infty, 1]$, where $\beta = 1, 0, -\infty$ indicates a radial, isotropic, and tangential velocity profile, respectively.

We integrate the spherical Jeans equations (Eq. A3) to get the radial velocity dispersion profile $\sigma_r(r)$,

$$\sigma_r^2(r) = \frac{1}{v(r)g(r)} \int_r^\infty g(\tilde{r}) \frac{GM(\tilde{r})v(\tilde{r})}{\tilde{r}^2} d\tilde{r}, \quad (\text{A5})$$

where the function $g(r)$ is

$$g(r) = \exp \left(2 \int \frac{\beta(\tilde{r})}{\tilde{r}} d\tilde{r} \right). \quad (\text{A6})$$

The radial velocity dispersion profile $\sigma_r(r)$ is a function of the enclosed mass profile $M(r)$ and thus the density profile $\rho(r)$. From Eq. A5 and Eq. A6, we see that different combinations of the velocity anisotropy $\beta(r)$ and the enclosed mass profile $M(r)$ can result in the same radial velocity dispersion profile $\sigma_r(r)$. This manifests as the mass-anisotropy degeneracy, as mentioned, and represents a fundamental limitation of the spherical Jeans equations. In practice,

because only the projected radii and line-of-sight velocities are available, we project Eq. A5 using the Abel transformation. This gives the line-of-sight velocity dispersion profile $\sigma_{\text{los}}(R)$:

$$\sigma_{\text{los}}^2(R) = \frac{2}{\Sigma_{\star}(R)} \int_R^{\infty} \left(1 - \beta(r) \frac{R^2}{r^2}\right) \frac{v(r)\sigma_r^2(r)r}{\sqrt{r-R^2}} dr, \quad (\text{A7})$$

where $\Sigma_{\star}(R)$ is the surface mass density of the tracer stars at the projected radius R . It is simply the projection of $v(r)$ along the line-of-sight. Note that in Eq. A5, we again see the $\beta(r)$ degenerates with $\sigma_r(r)$ and thus $M(r)$.

In the framework of Jeans dynamical modeling, the density profiles $\rho(r)$, velocity anisotropy profiles $\beta(r)$, and the tracer mass density $\Sigma_{\star}(R)$ are assumed some functional forms with free parameters and fit simultaneously. To fit the free parameters, one typically construct the likelihood of the line-of-sight velocity dispersion profile $\sigma_{\text{los}}(R)$ and apply maximum likelihood estimation (MLE) with Bayesian sampling techniques, e.g. Markov Chain Monte Carlo (MCMC, Jones & Qin 2022; Speagle 2019), Nested Sampling (Skilling 2004; Ashton et al. 2022). The likelihood can be either binned (e.g., Strigari et al. 2007a; Charbonnier et al. 2011) or unbinned (e.g., Strigari et al. 2008). Traditionally, the Gaussian unbinned likelihood from Strigari et al. (2008) has been commonly used (e.g., Geringer-Sameth et al. 2015b):

$$\mathcal{L} = \prod_i^{N_{\text{star}}} \frac{(2\pi)^{-1/2}}{\sqrt{\sigma_{\text{los}}^2(R_i) + \Delta_i^2}} \exp \left[-\frac{1}{2} \left(\frac{(v_i - \langle v \rangle)^2}{\sigma_{\text{LOS}}^2(R_i) + \Delta_i^2} \right) \right], \quad (\text{A8})$$

where R_i is the projected radius, and v_i and Δ_i are the line-of-sight velocity and its measurement uncertainty of star i . $\langle v \rangle$ is the mean velocity of the tracer population, and σ_{los}^2 is the intrinsic line-of-sight velocity dispersion given by Eq. A7.

APPENDIX B: ADDITIONAL DETAILS ON MACHINE LEARNING ARCHITECTURE AND TRAINING

As noted, the machine learning architecture is similar to that in N23. The model consists of a GNN embedding network and a normalizing flow for density estimation.

During the forward pass, node features are first projected onto a 64-dimensional latent space using a multi-layer perceptron (MLP). The graph is then processed through 3 GATConv layers, each with a hidden size of 128, two attention heads, and a Leaky ReLU activation (He et al. 2015). Next, we average the node features and pass the result through 4 MLP layers, each with a hidden size of 128 and a Gaussian Error Linear Unit (GELU; Hendrycks & Gimpel 2016) activation. Each MLP layer is followed by batch normalization and dropout with a rate of 0.4. The resulting summary feature is a 128-dimensional vector that is invariant to the ordering of the nodes (i.e. permutation-invariant). Finally, the summary feature is used as the conditioning features for the normalizing flows, which consists of 6 Neural Spline Flows (NSF; Durkan et al. (2019)) transformations with 8 knots. Note that this is another difference from the architecture in N23, which used 4 Masked Autoregressive Transformations (MAF; Germain et al. 2015; Papamakarios et al. 2017). We find that while NSFs have similar performance on the validation dataset as MAFs, they better generalize to the FIRE dataset.

During training, we minimize the negative log-likelihood given by the flows. Let the embedding network, which includes the projection MLP layer, the 4 GAT layers, and the output 4 MLP layers, be $g_{\phi}(\mathcal{G}_i)$ where ϕ is the trainable parameters and \mathcal{G}_i is the input graph. The

optimization objective is simply:

$$\mathcal{L} = -\log p_{\varphi}(\theta | g_{\phi}(\mathcal{G}_i)), \quad (\text{B1})$$

where φ is the trainable parameters of the flows and θ is the parameters of interest, i.e. the DM and stellar parameters.

We train the feature extractor and the flows simultaneously using the AdamW (Loshchilov & Hutter 2019; Kingma & Ba 2014) gradient descent optimizer with a cosine annealing learning rate scheduler (Loshchilov & Hutter 2016). The optimizer has a peak learning rate 5×10^{-4} and weight decay coefficient 0.01. The scheduler has 40,000 warm-up steps and 80,000 decay steps. The training batch size is 64. As noted, we use 5×10^6 and 5×10^5 galaxies for the training and validation set, respectively. The training converges after approximately 50 epochs or about 22 hours on a NVIDIA Tesla V100.

APPENDIX C: ADDITIONAL RESULTS

C1 Comparison with Jeans-based methods

Figures C1 and C2 compare the inferred dark matter density profiles $\rho(r)$ for GRAPHNPE and Jeans modeling across the full set of selected FIRE galaxies in the CDM and SIDM simulations, respectively. Figures C3 and C4 show the velocity anisotropy profiles $\beta(r)$ for the same galaxies. Similarly, Figures C5 and C6 compare the inferred $\rho(r)$ profiles for GRAPHNPE and GRAVSPHERE in the CDM and SIDM simulations, respectively, while Figures C3 and C4 present the corresponding $\beta(r)$ profiles.

C2 Tidal disruption

In Section 5.2, we show the predicted versus true values for the peak circular velocity V_{max} and the peak virial mass M_{200m} across three bins of the number of pericentric passages N_{peri} . Here, we show results for three additional orbital parameters: the last pericentric distance r_{peri} (Figure C9), the current distance r_{curr} (Figure C10), and the time since last pericenter t_{peri} (Figure C11).

This paper has been typeset from a TeX/LaTeX file prepared by the author.

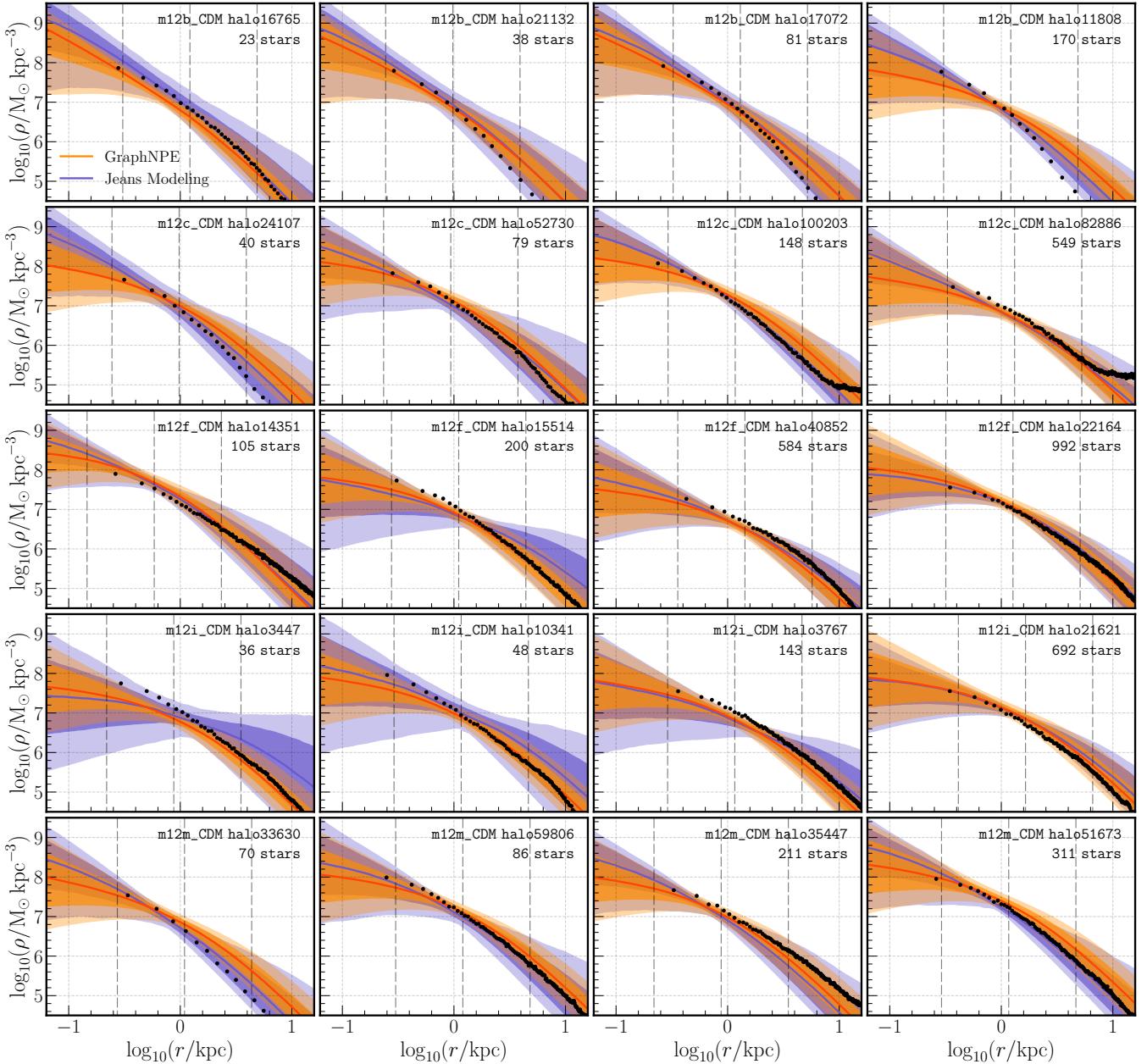


Figure C1. Comparison between the inferred DM density profiles $\rho_{\text{dm}}(r)$ from GRAPHNPE (orange) and Jeans modeling (blue) for the CDM FIRE galaxies. Panels are the same as in Figure 1.

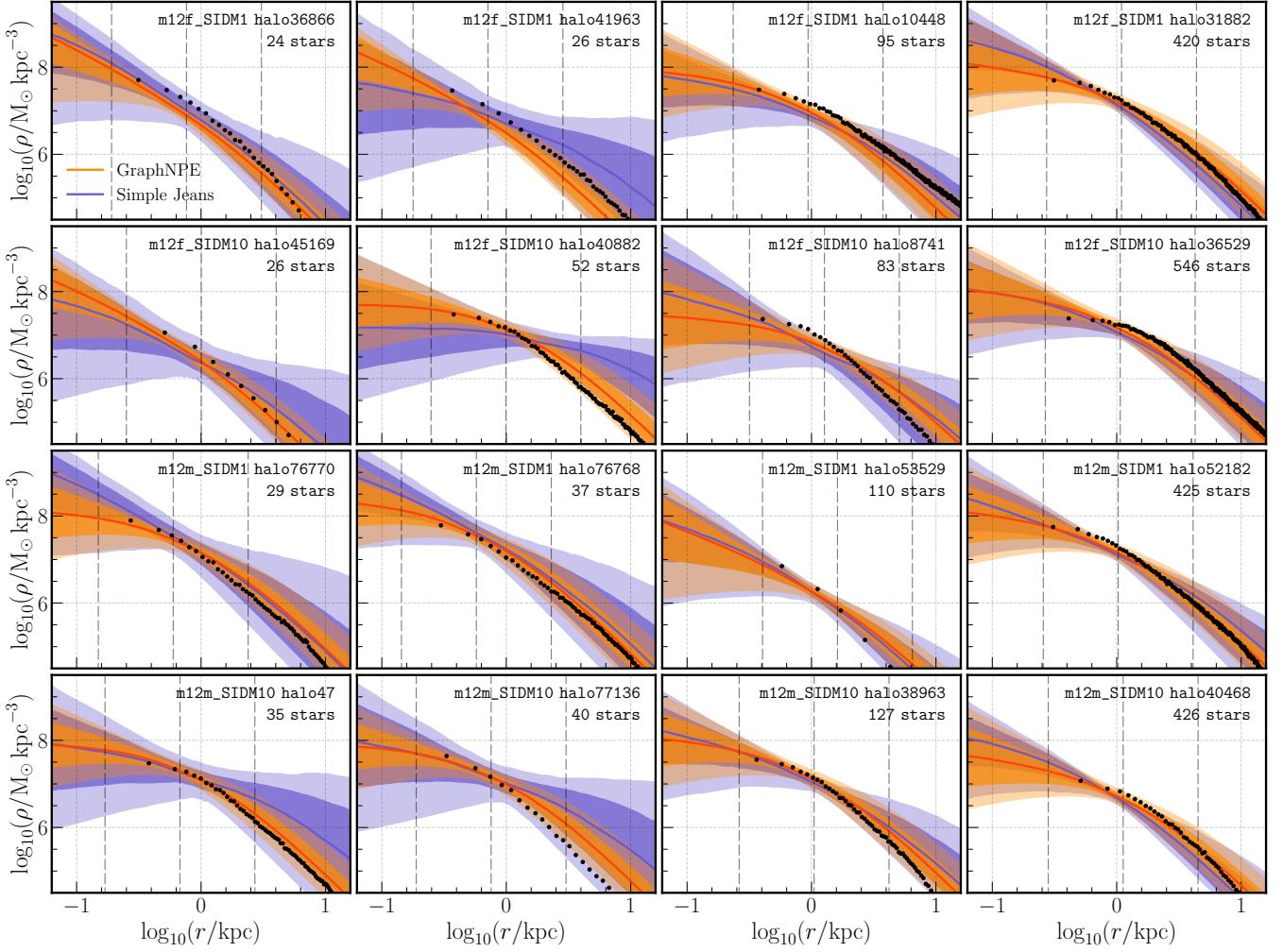


Figure C2. Comparison between the inferred DM density profiles $\rho_{\text{dm}}(r)$ from GRAPHNPE (orange) and Jeans modeling (blue) for the SIDM FIRE galaxies. Panels are the same as in Figure 1.

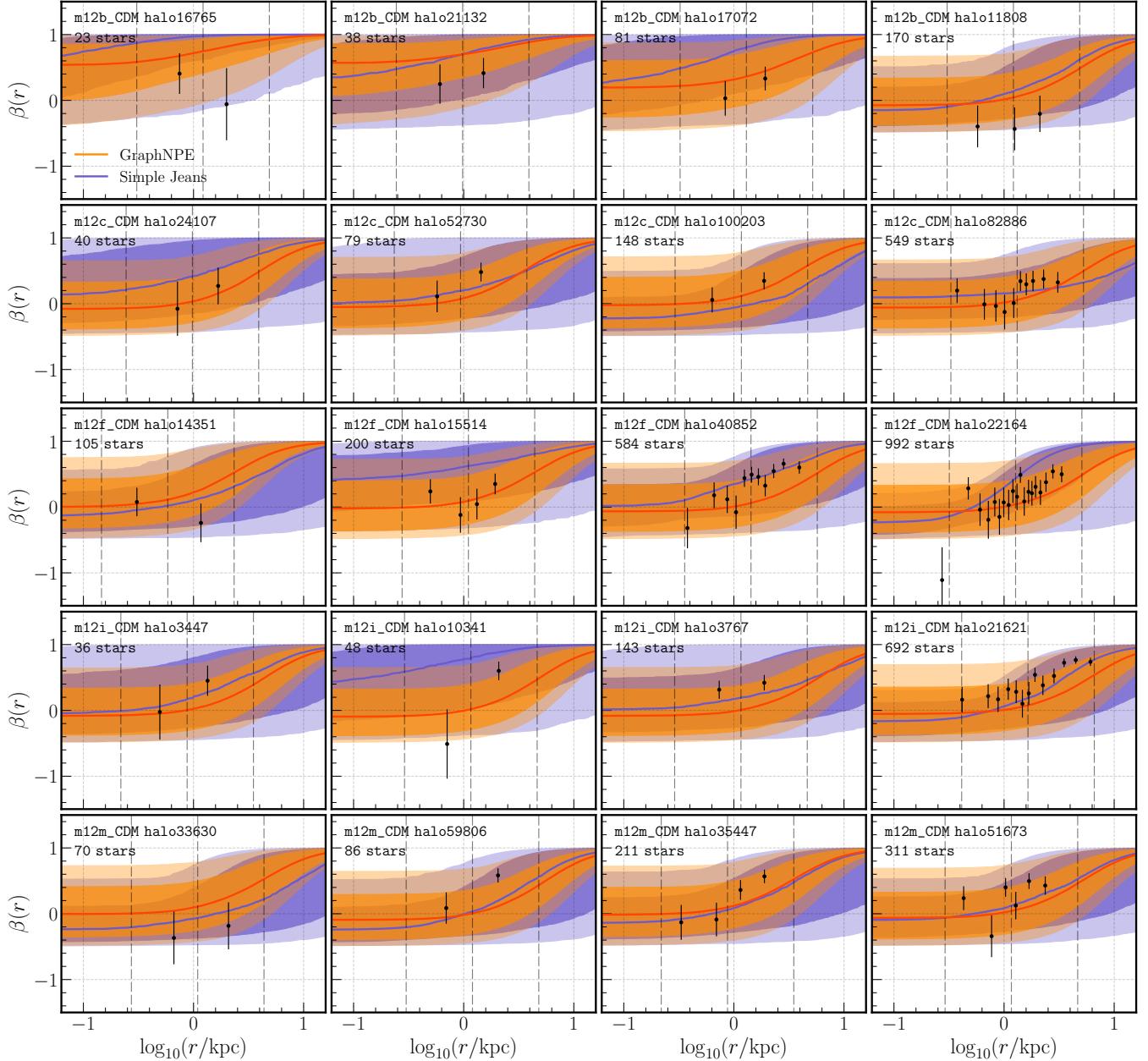


Figure C3. Comparison between the inferred velocity anisotropy profiles $\beta(r)$ from GRAPHNPE (orange) and Jeans modeling (blue) for the CDM FIRE galaxies. Panels are the same as in Figure 2.

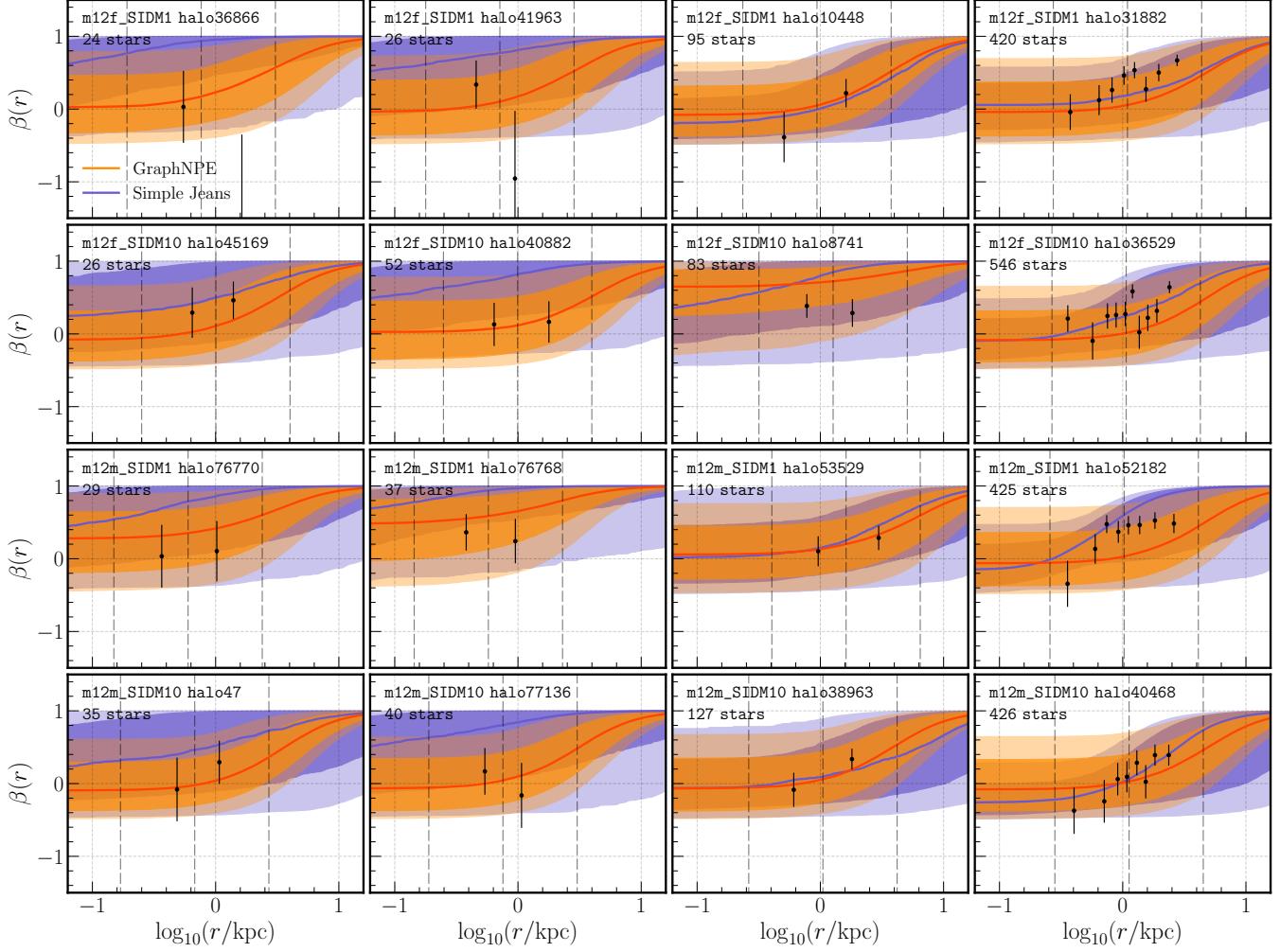


Figure C4. Comparison between the inferred DM density profiles $\rho_{\text{dm}}(r)$ from GRAPHNPE (orange) and GRAVSPHERE (blue) for the SIDM FIRE galaxies. Panels are the same as in Figure 2.

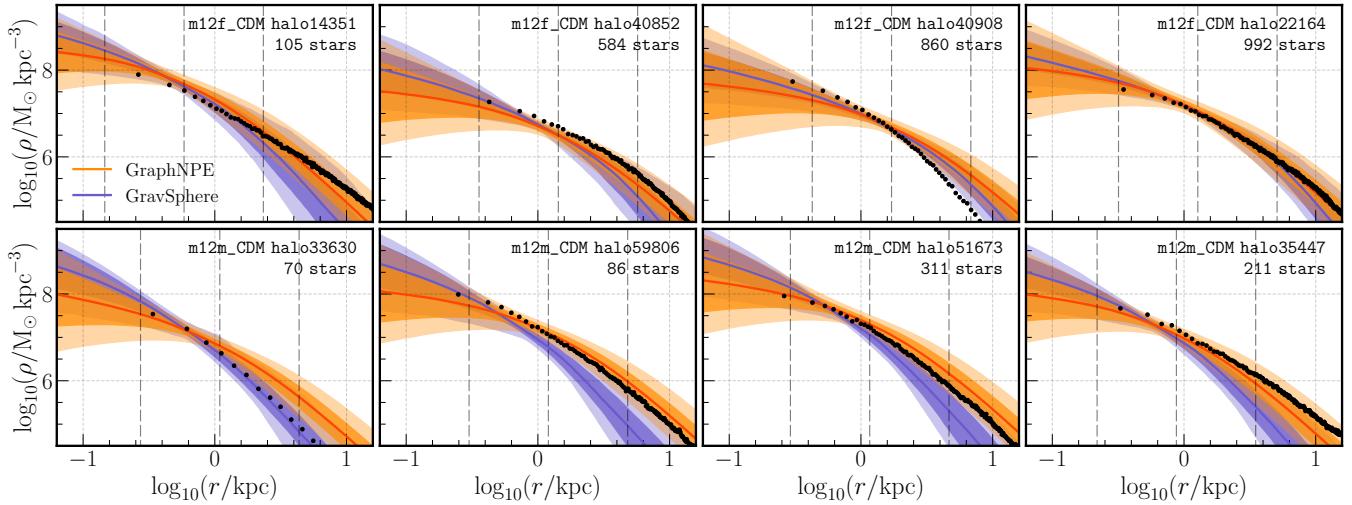


Figure C5. Comparison between the inferred DM density profiles $\rho_{\text{dm}}(r)$ from GRAPHNPE (orange) and GRAVSPHERE (blue) for the CDM FIRE galaxies. Panels are the same as in Figure 1.

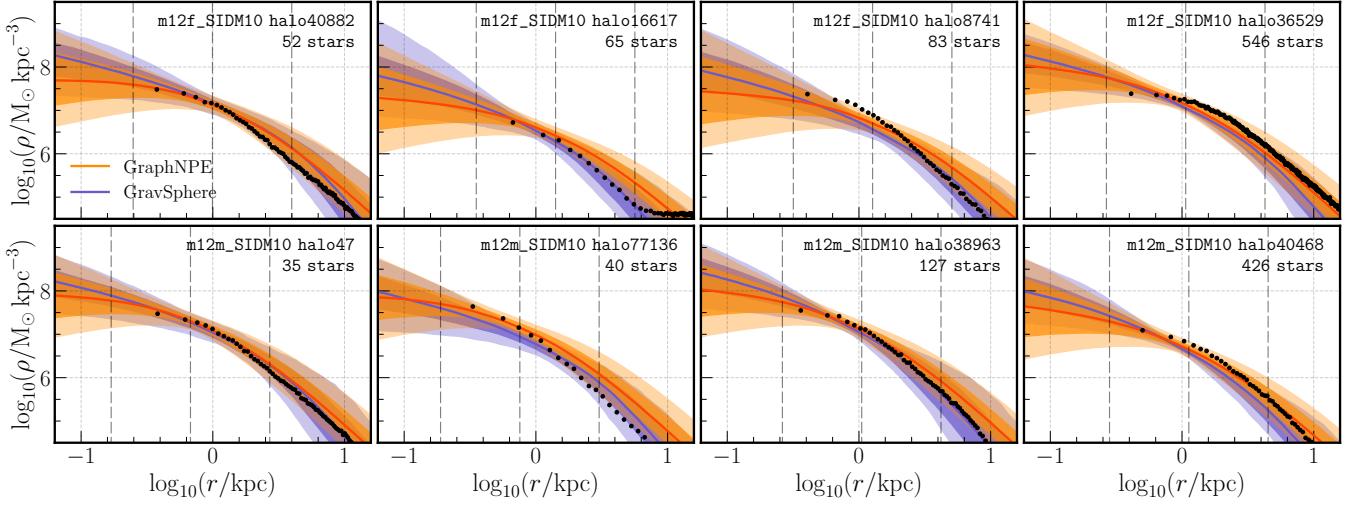


Figure C6. Comparison between the inferred DM density profiles $\rho_{\text{dm}}(r)$ from GRAPHNPE (orange) and GRAVSHERE (blue) for the SIDM FIRE galaxies. Panels are the same as in Figure 1.

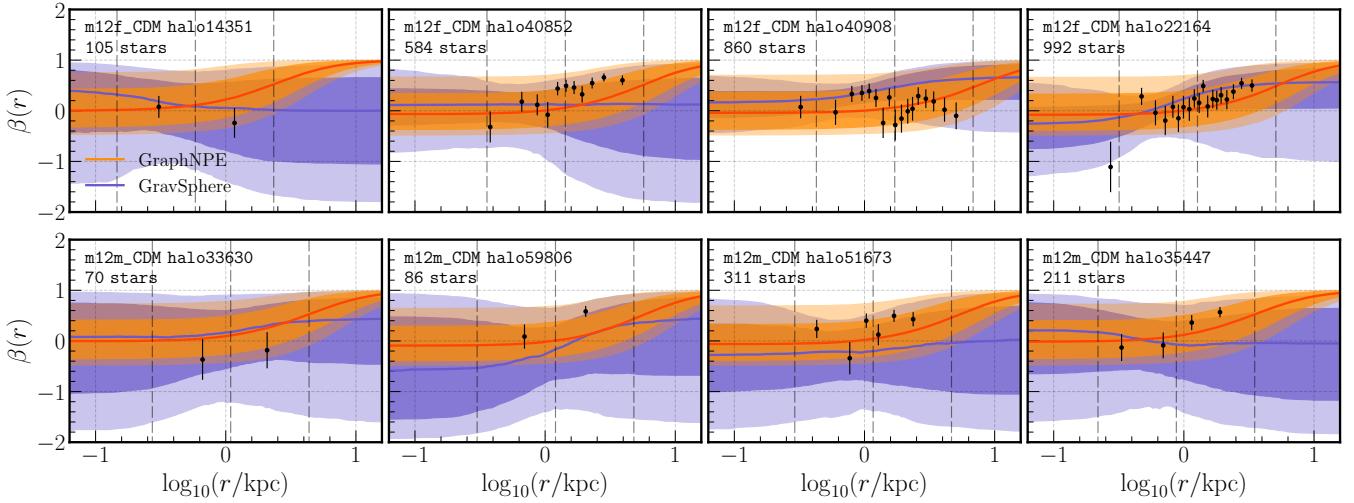


Figure C7. Comparison between the inferred velocity anisotropy profiles $\beta(r)$ from GRAPHNPE (orange) and GRAVSHERE (blue) for the CDM FIRE galaxies. Panels are the same as in Figure 2.

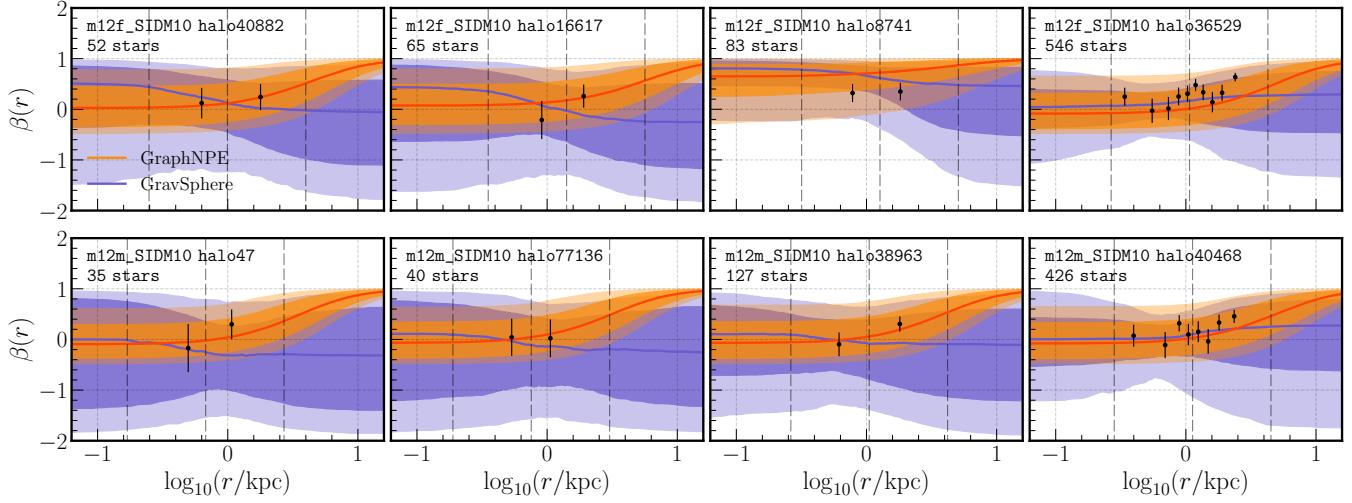


Figure C8. Comparison between the inferred velocity anisotropy profiles $\beta(r)$ from GRAPHNPE (orange) and GRAVSPHERE (blue) for the SIDM FIRE galaxies. Panels are the same as in Figure 2.

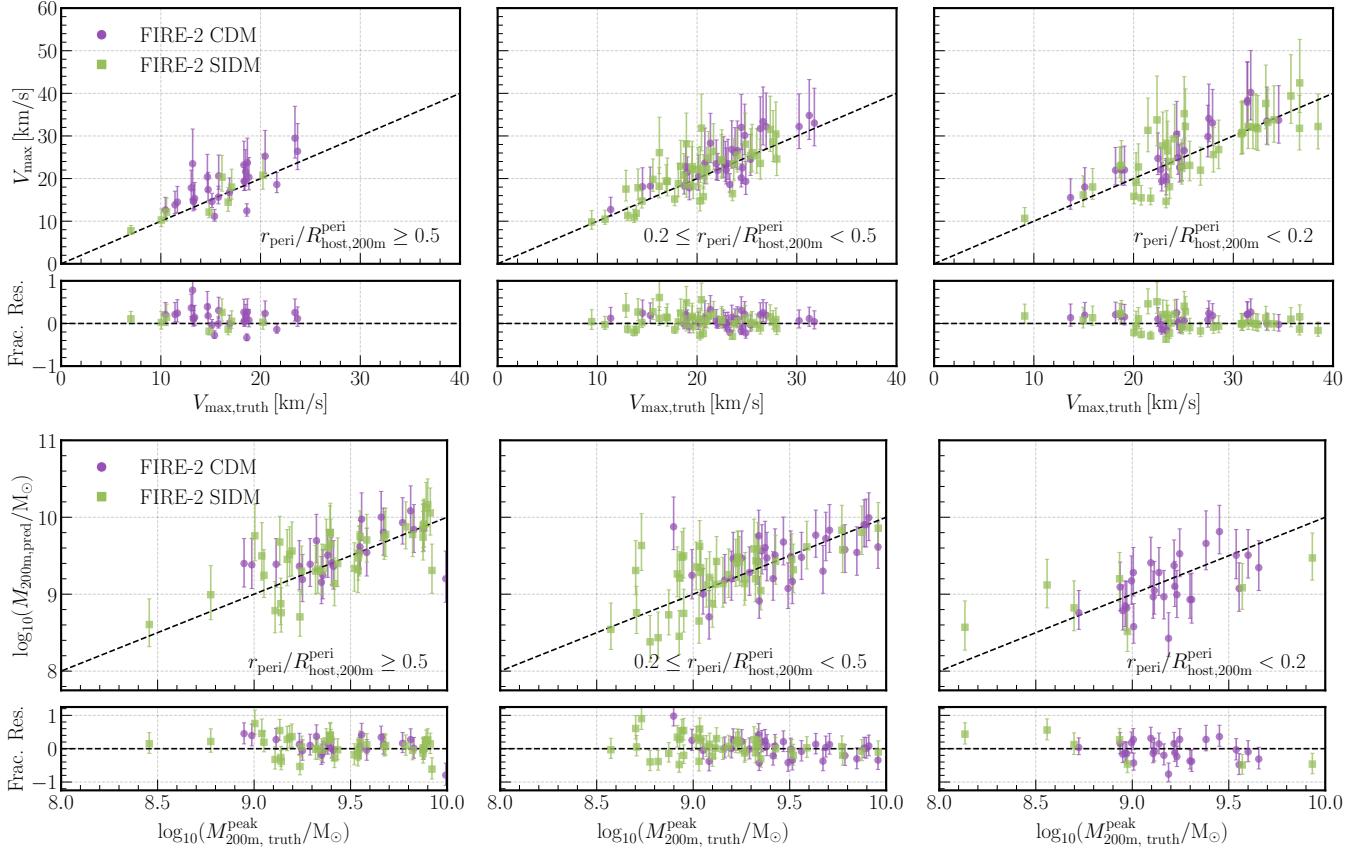


Figure C9. Recovery of the peak circular velocity V_{\max} and the peak virial mass M_{200m} of dwarf galaxies in the FIRE dataset. Galaxies are grouped by the last pericentric distances r_{peri} , in units of the host R_{200m} , with each column corresponding to a different grouping. Panels are the same as Figure 7.

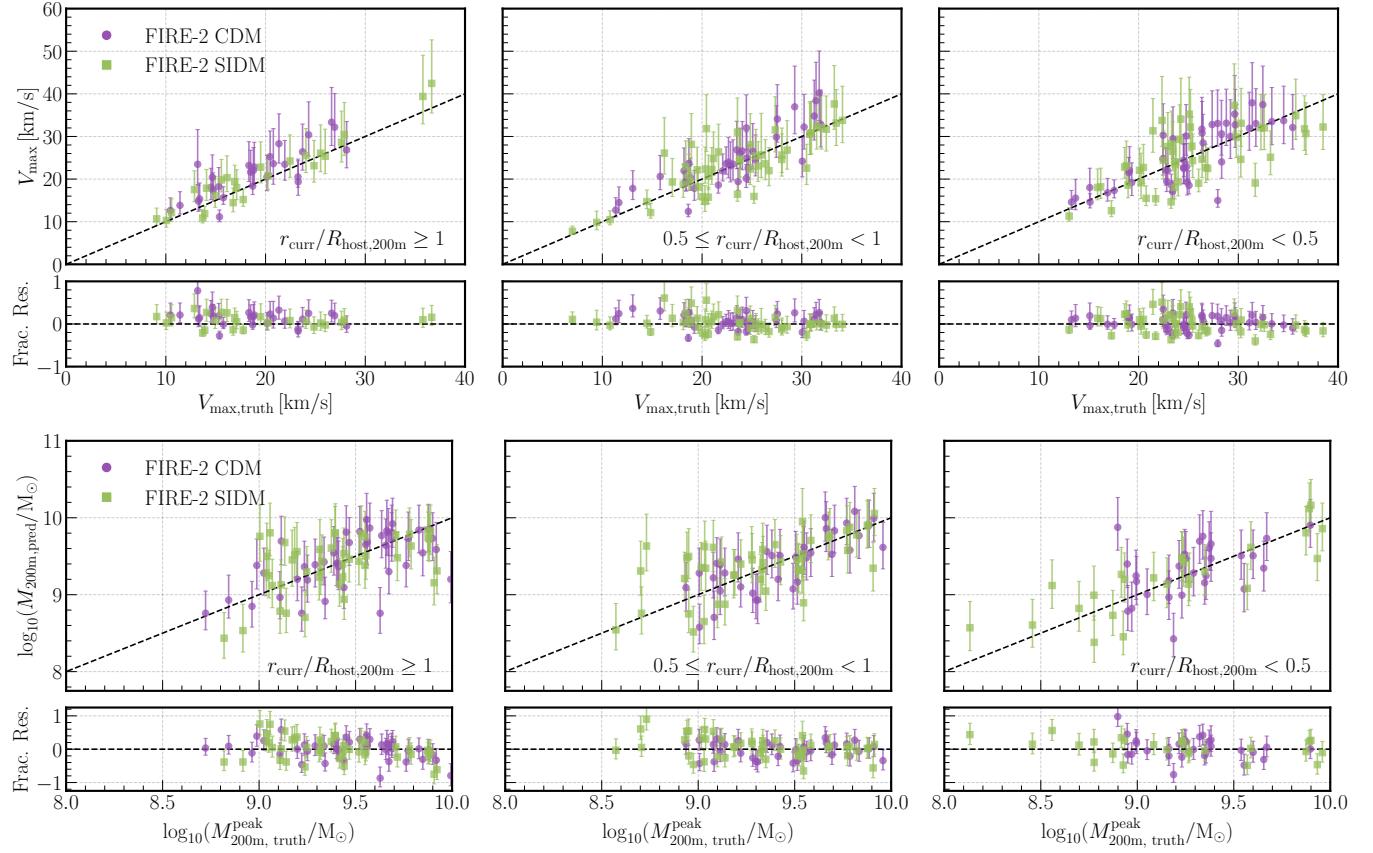


Figure C10. Recovery of the peak circular velocity V_{\max} and the peak virial mass M_{200m} of dwarf galaxies in the FIRE dataset. Galaxies are grouped by the current distance r_{curr} , in units of the host R_{200m} , with each column corresponding to a different grouping. Panels are the same as Figure 7.

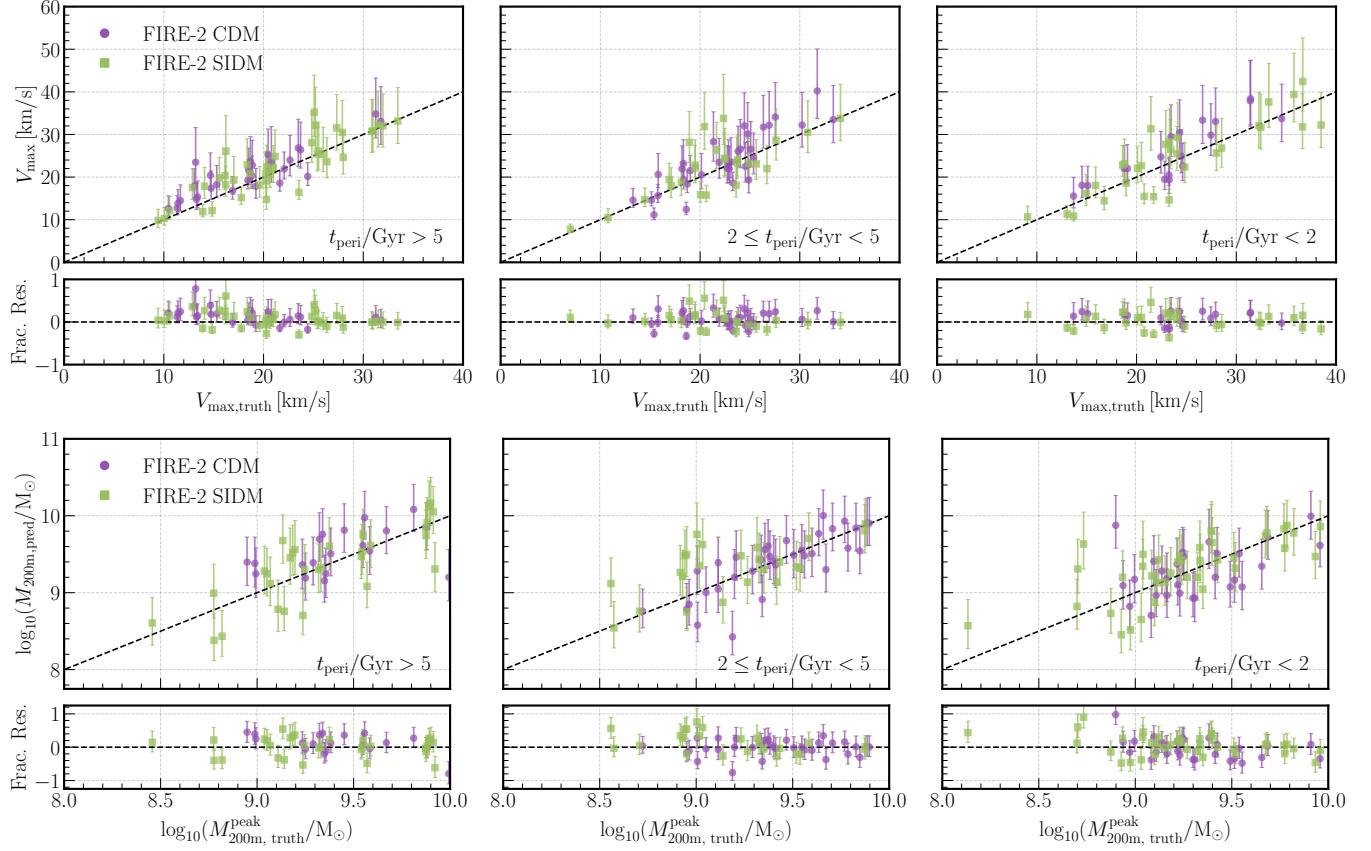


Figure C11. Recovery of the peak circular velocity V_{\max} and the peak virial mass M_{200m} of dwarf galaxies in the FIRE dataset. Galaxies are grouped by the time since last pericenter t_{peri} with each column corresponding to a different grouping. Panels are the same as Figure 7.