TAL TECH

DOCTORAL THESIS

Probing the Milky Way's Dark Matter Halo in the Gaia and Machine Learning Era

Sven Põder

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2025

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
93/2025

Probing the Milky Way's Dark Matter Halo in the Gaia and Machine Learning Era

SVEN PÕDER

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Science
Department of Cybernetics

NATIONAL INSTITUTE OF CHEMICAL PHYSICS AND BIOPHYSICS
Laboratory of High Energy and Computational Physics

The dissertation was accepted for the defence of the degree of Doctor of Philosophy (Applied Physics and Mathematics) on 7 November 2025

Supervisor: Dr. Joosep Pata,
High Energy and Computational Physics
National Institute of Chemical Physics and Biophysics
Tallinn, Estonia

Co-supervisor: Dr. María Benito Castaño,
Tartu Observatory
Tartu University
Tartu, Estonia

Opponents: Prof. Dr. Gabrijela Zaharijas,
University of Nova Gorica,
Nova Gorica, Slovenia

Dr. William James Pearson,
National Centre for Nuclear Research,
Warsaw, Poland

Defence of the thesis: 15 December 2025, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree elsewhere.

Sven Põder

signature

ISSN 2585-6898 (publication)
ISBN 978-9916-80-427-8 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9916-80-428-5 (PDF)
DOI https://doi.org/10.23658/taltech.93/2025

Põder, S. (2025). *Probing the MilkyWay's Dark Matter Halo in the Gaia and Machine Learning Era* [TalTech Press]. https://doi.org/10.23658/taltech.93/2025

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
93/2025

Linnutee galaktika tumeaine halo uurimine Gaia ja masinõppe ajastul

SVEN PÕDER

TAL TECH
KIRJASTUS

Contents

List of Publications

The present Ph.D. thesis is based on the following publications that are referred to in the text by Roman numbers.

I A. Bazarov, M. Benito, G. Hütsi, R. Kipper, J. Pata, and S. Põder. Sensitivity estimation for dark matter subhalos in synthetic gaia dr2 using deep learning. *Astronomy and Computing*, 41:100667, 2022

II Põder, Sven, Benito, María, Pata, Joosep, Kipper, Rain, Ramler, Heleri, Hütsi, Gert, Kolka, Indrek, and Thomas, Guillaume F. A bayesian estimation of the milky way's circular velocity curve using gaia dr3. *A&A*, 676:A134, 2023

III Põder, Sven, Pata, Joosep, Benito, María, Alonso Asensio, Isaac, and Dalla Vecchia, Claudio. Detection of stellar wakes in the milky way: A deep learning approach. *A&A*, 693:A227, 2025

Author's Contributions to the Publications

I I implemented the anomaly detection approach using an autoencoder to analyse the dark matter substructure of simulated galaxies. As I was the corresponding author, I also handled the submission process, made significant contributions to writing the manuscript and preparing the results and figures.

II My initiative was to create and optimize the data pipeline from start to finish. A large part of this was the development of a Python library (`gaia-tools`) which can transform raw Gaia observables and their measurement errors from the observational to a Galactocentric frame of reference and Cartesian or cylindrical coordinates. Ultimately, I optimized the code such that the coordinate transformation routine was integrated into a Markov Chain Monte Carlo sampling process distributed across multiple CPUs and GPUs. I was the corresponding author and handled the submission process, made significant contributions to writing the manuscript and preparing the figures in the draft.

III I was the corresponding author and handled the submission process. I contributed significantly to writing the manuscript as well as to preparing the results and figures. I developed and ran the data analysis pipeline from start to finish: running N-body simulations, deriving training datasets, implementing, tuning, and training deep learning models, and finally, evaluating models on testing datasets.

Approbation

I presented the results of the thesis at the following conferences:

1. **S. Põder**. "Detection of stellar wakes in the Milky Way: A deep learning approach", IAUS 397: Exploring the Universe with Artificial Intelligence (UniversAI): 2–6 June 2025, Athens, Greece.

2. **S. Põder**. "Data Driven Dark Matter Searches in the Milky Way", LLM retreat for PhD students and supervisors, 27–28 November 2024, Nelijärve, Estonia. *Awarded a prize*.

3. **S. Põder**. "Searching for Dark Matter Subhalos in the Milky Way using Deep Learning", 4th CERN Baltic Conference (CBC 2024), 15–17 October 2024, Tallinn, Estonia.

4. **S. Põder**. "Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning", CLUES Workshop 2024, 10–14 June 2024, Warsaw, Poland.

5. **S. Põder**. "Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning", Tuorla-Tartu meeting 2024, 6–8 May 2024, Turku, Finland.

6. **S. Põder**. "Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning", 1st European AI for Fundamental Physics Conference (EuCAIFCon), 30 April–3 May 2024, Amsterdam, Netherlands.

7. **S. Põder**. "A Bayesian Estimation of the Milky Way's Circular Velocity Curve using Gaia DR3", Kashiwa Dark Matter Symposium 2023, 5–8 December 2023, Tokyo, Japan (online).

8. **S. Põder**. "The Milky Way's dark matter halo: A Bayesian Estimation of the Milky Way's Circular Velocity Curve using Gaia DR3", 3rd CERN Baltic Conference (CBC 2023), 9–11 October 2023, Riga, Latvia.

9. **S. Põder**. "A Bayesian Estimation of the Milky Way's Circular Velocity Curve using Gaia DR3", Third EuCAPT Annual Symposium, 31 May–2 June 2023, CERN, Switzerland (online).

10. **S. Põder**. "Searching for Dark Matter Subhalos in the Milky Way using Deep Learning", Kashiwa Dark Matter Symposium 2022, 29 November–2 December 2022, Tokyo, Japan (online).

11. **S. Põder**. "Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning", 2nd CERN Baltic Conference (CBC 2022), 10–12 October 2022, Vilnius, Lithuania.

Abbreviations

ADM atomic dark matter
AE autoencoder
AHF Amiga Halo Finder
ANN artificial neural network
AOC area over the curve
API application programming interface
AUC area under the curve

CBE collisonless Boltzmann equation
CDM cold dark matter
CMB cosmic microwave background
CNN convolutional neural network
CPU central processing unit

DCT discrete cosine transform
DL deep learning
DM dark matter
DMO dark matter-only
DR data release

EDR3 Early Data Release 3
EM electromagnetic
ESA European Space Agency

FDM fuzzy dark matter
FIRE Feedback In Realistic Environments
FPR false positive rate

Gaia-CRF Gaia Celestial Reference Frame
GC Galactic Center
GPU graphics processing unit
GSP-Phot General Stellar Parametrizer from Photometry

HR Hertzsprung-Russell

ICRS International Celestial Reference System
IMF initial mass function
ISM interstellar medium

LMC Large Magellanic Cloud
LSR local standard of rest
LSST Large Synoptic Survey Telescope

MCMC Markov chain Monte Carlo
ML machine learning
MW Milky Way

NFW Navarro-Frenk-White

PTA pulsar timing array

QSO quasi-stellar object

ReLU rectified linear unit
RGB red giant branch
ROC receiver operating characteristic
RVS Radial Velocity Spectrometer

SELU scaled exponential linear unit
SHMF subhalo mass function
SIDM self-interacting dark matter
SM Standard Model of particle physics

TPR true positive rate

WDM warm dark matter

Introduction

Despite comprising roughly 85% of all matter in the cosmos, the nature and distribution of dark matter (DM) remains one of the most profound open questions in physics and astronomy. From the rotation curves of spiral galaxies to the temperature anisotropies of the cosmic microwave background, phenomena on scales spanning ten orders of magnitude insist on the presence of a non-luminous gravitating component of the Universe. While numerous direct and indirect particle detection experiments and collider searches continue to test the possible particle candidates, only cosmological and astrophysical observations have provided positive evidence for the existence of DM to date, making the Universe itself our best DM detector.

In the past 30 years, evidence gathered across cosmological scales has established a standard cold dark matter (CDM) paradigm, which predicts hierarchical structure formation: large DM halos grow through the merger and accretion of smaller subhalos. Within this framework, galaxies such as the Milky Way (MW) are expected to be embedded in massive DM halos that are composed of both a dominant smooth component and a swarm of smaller subhalos. The latter are gravitationally bound DM clumps that survive as satellites of the host halo.

Characterizing both components of galactic DM halos is essential. The properties of both the smooth halo and its substructure offer crucial indirect avenues for probing the fundamental nature of DM. Accurately mapping the large-scale distribution of DM provides constraints on the total mass and density profile of the Galaxy. Additionally, detecting DM subhalos and characterizing their abundance can discriminate between different possible DM models, including warm, fuzzy, or self-interacting DM.

In recent years, astronomy has entered a transformative "big data" era, owing to vast and increasingly precise datasets from observations. In particular, the Gaia mission has revolutionized Galactic astronomy by delivering sub-milliarcsecond astrometry and kinematics for more than a billion stars. With these data products at our disposal, the MW has become a dynamical laboratory of unprecedented fidelity, revealing previously unseen baryonic substructures and enabling detailed, kinematics-based searches for DM through the motions of luminous tracers.

The wealth of high-quality data, combined with advances in the machine learning (ML) methods, offers new and exciting opportunities to study DM in the Galaxy using data-driven techniques that leverage the statistical power of both modern observational surveys and simulations. With the above in mind, the following thesis explores both the smooth and clumpy aspects of the MW DM halo using traditional statistical and novel ML methods.

This thesis is organized as follows. Section 1 introduces the general DM problem and outlines the narrative of the standard cosmological model. In Section 2, the overall scope and objectives of the current thesis are outlined. Section 3 discusses how the MW serves as a natural laboratory for DM studies, and introduces the observational data (Gaia) and cosmological simulations used throughout this work. Section 4 presents the analysis and results from modeling the MW's disk kinematics and constraining the smooth DM halo (Publication II). Section 5 explores the use of ML methods to investigate DM substructure using MW-like and idealized N-body simulations (Publications I & III).

1 Dark matter in the ΛCDM paradigm

Evidence accumulated in observational astronomy over the past century suggests a large gap in our understanding regarding the matter composition of the world around us. Independent observations at different scales in the Universe point to the existence of invisible matter, so-called DM, which is separate from its visible baryonic counterpart. While the latter has been extensively studied and is well-described by the Standard Model of particle physics (SM), the fundamental nature of the former remains a mystery. As such, the collective effort to uncover the true nature and origin of DM is one of the largest endeavors spanning cosmology, astrophysics, and particle physics today. The following chapter aims to offer a brief overview of the current cosmological view of the Universe and the role of DM in it.

1.1 Historical context

One of the first observations hinting at unseen mass in the Universe dates back to the 1930s, when F. Zwicky analysed the velocities of galaxies in the Coma cluster [4, 5]. A comparison of the mass from visible matter and the velocity dispersion of galaxies revealed a discrepancy, which was attributed to unseen matter. Though a similar conclusion was reached in later analyses of other galaxy clusters [6, 7], at the time, the DM hypothesis had not been formulated as a fundamental problem, and the results simply remained curiosities [8].

It was in the second part of the 20th century that key pieces of evidence started to fit together, revealing a missing mass problem in the Universe. The discovery of the 21-cm (HI) emission line [9] of hydrogen made it possible to trace the abundant neutral hydrogen in the MW and other nearby galaxies. As a consequence, velocity measurements of HI lines led to the extension of rotation curves of nearby galaxies far beyond the optical disk. Rotation curves depict the rotational velocities of stars (or gas) with respect to the galactocentric distance. Since the square of the orbital velocity of a star on a circular orbit at radius R is proportional to the mass enclosed inside said radius, the resultant curve encodes information about the underlying mass distribution in the system. In the case of galaxies, the expectation was that the curve would exhibit a Keplerian drop-off of the form $v(R) \propto R^{-1/2}$, judging by the total amount of visible matter. In 1970, two important studies emerged with one measuring rotation curves of M31 [10], and the other those of M33 and NGC 300 [11]. These studies revealed that the curves remain flat up to large distances, suggesting the existence of large amounts of unseen mass in the outer parts of galaxies.

The puzzling results from rotation curve studies and galaxy cluster velocity measurements were first considered part of the same story by authors in [12] and [13] who put the observations into a larger cosmological context. The authors of [13] specifically considered the existence of a massive corona around galaxies as an explanation for the mass discrepancies observed in galaxy clusters. Furthermore, the studies pointed out that the observed mass density of galaxies is insufficient to produce a flat or closed ($\kappa \geq 0$) universe, that is, one where the dimensionless density parameter $\Omega \geq 1$. Here, $\Omega = \rho/\rho_c$ compares the observed density in the universe to its critical density[1] (ρ_c), and is related to the spatial curvature index κ. The values of the spatial curvature index describe either an open ($\kappa < 0$), flat ($\kappa = 0$), or closed ($\kappa > 0$) universe [14].

The considerations outlined above elevated the importance of the search for missing mass in the Universe. Thus, the different sources of evidence pointing to large amounts

[1] The critical density is defined as the energy density needed to produce a spatially flat universe.

of unseen matter in and around galaxies went from mere unrelated discrepancies into heralding a paradigmatic shift in our view of the Universe. It was then at the confluence of radio astronomy and extra-galactic cosmology that the stage was finally set for the DM problem to be established in the larger scientific community. [2]

1.2 Observational evidence

Currently, the evidence for DM is inferred from numerous independent phenomena at different cosmic scales. Based mainly on the review in [16], the following section offers a brief overview of the observational landscape of DM today, but is by no means exhaustive due to the diverse and rapidly evolving nature of the field.

1.2.1 Galactic rotation curves

As alluded to in the previous section, measurements of galactic rotation curves have historically played a pivotal role in establishing the DM problem. Observations consistently have shown that orbital velocities of stars in disk galaxies remain roughly constant out to larger radii, which is in contrast to the expectation from Newtonian dynamics if only visible matter is accounted for. These flat rotation curves imply the existence of a dark component to galaxies that contributes significantly to their total mass beyond the optical disk [10].

In the MW, rotation curve measurements offer a direct tracer of the underlying mass distribution, including its DM halo. Precise measurements are made possible by modern stellar surveys, most notably Gaia [17], which provide full 6D phase-space data for millions of stars in our Galaxy (see also Section 3.1). As a result, the MW's rotation curve remains an active topic of research, offering insights into both the visible and dark components of the Galaxy.

The circular velocity $v_c(R)$, which tracks the orbital speed of a star in a perfectly circular orbit at Galactocentric radius R, is related to the galactic gravitational potential under the assumption of axisymmetry through the following relation

$$\frac{\partial \Phi}{\partial R} = \frac{v_c^2}{R}, \tag{1}$$

where R is the orbital radius of the star. Assuming a spherical potential of $\Phi(R) = -GM/R$ and taking the derivative with respect to R, the circular velocity is easily derived as

$$v_c(R) = \sqrt{\frac{GM(R)}{R}}, \tag{2}$$

where $M(R)$ is the total enclosed mass. In disk galaxies like the MW, stars in the plane of the disk are on nearly circular orbits, where v_c encodes information about the axisymmetric component of the potential.

Figure 1 compiles recent measurements of the MW's rotation curve, highlighting the existence of an extended DM halo [18]. In particular, the near-flatness of the profile in the outer Galaxy cannot be explained by baryons alone. Quantitatively, rotation curve studies suggest that DM accounts for approximately 80-90% of the MW's total mass [16].

One part of this thesis presents a new inference of the circular velocity curve (also seen in Fig. 1) based on Gaia data, using a Bayesian framework that self-consistently marginalizes over systematic uncertainties. The analysis also allows estimation of the local DM

[2] The history of the DM hypothesis is a fascinating one. Although the current work does not cover it in explicit detail, the interested reader can find a thorough review in both [8] and [15].

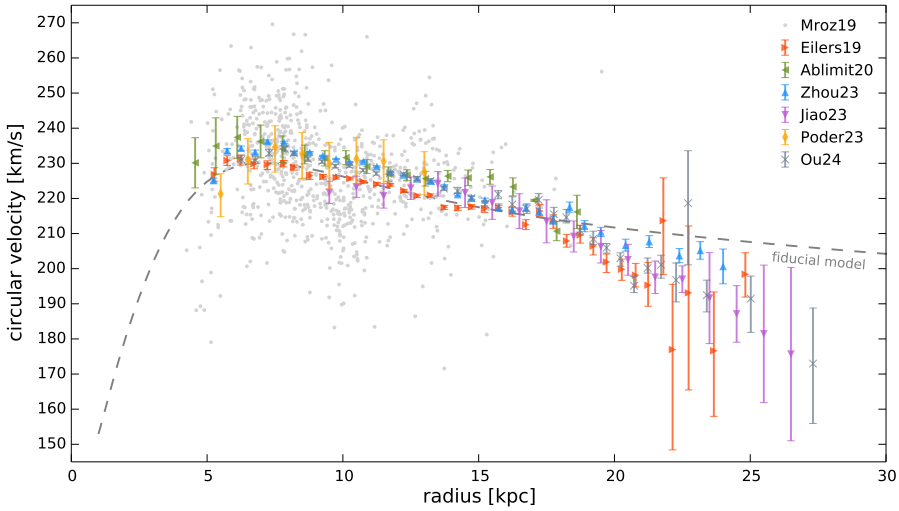


Figure 1: Compilation of circular velocity curve measurements for the MW, as presented in [18].

density and enclosed mass within the solar circle, further strengthening evidence for a dominant DM component in the Galaxy (see Section 4.2 for details).

1.2.2 Galaxy clusters

Galaxy clusters are large cosmic structures that generally contain hundreds to thousands of galaxies and are the most massive gravitationally bound structures in the Universe [14]. As was already pointed out in Section 1.1, the first clues for DM originate from studies inferring the total mass of these structures by means of the virial theorem.

In addition to galaxies, these large structures contain a considerable amount of hot intragalactic gas, which due to gravity, has been accelerated to high velocities. The thermal radiation emitted by the gas is observable with X-ray telescopes and can be used to estimate the baryon fraction $f_b = \Omega_b/\Omega_m$, assuming that galaxy clusters provide a representative sample of the Universe [19]. Studies using this method have resulted in a baryon fraction of e.g. $f_b = 0.144 \pm 0.005$ [20].

Mass measurements from observations of intragalactic gas in galaxy clusters are also complemented by gravitational lensing methods. Gravitational lensing is a phenomenon that occurs as a massive object in the foreground (e.g., a galaxy cluster) distorts the image of a highly luminous background object, such as a quasar. The amount of lensing can be used to then infer the mass of the object responsible for the distortion.

It is at the interface of these two methods that perhaps one of the most visibly remarkable signs of DM has been detected in the Bullet cluster, shown in Fig 2.

1.2.3 Large scale structure

The large-scale distribution of matter in the Universe provides powerful evidence for the existence of DM. On scales of tens to hundreds of megaparsecs, galaxies are not randomly distributed. They form a complex system of clusters, filaments, and voids, which is referred to as the cosmic web.

Comprehensive galaxy redshift surveys such as CfA2, 2dFGRS, and SDSS have mapped the distribution of galaxies across large volumes of the Universe. The statistical properties of this distribution show remarkable agreement with predictions from ΛCDM (see

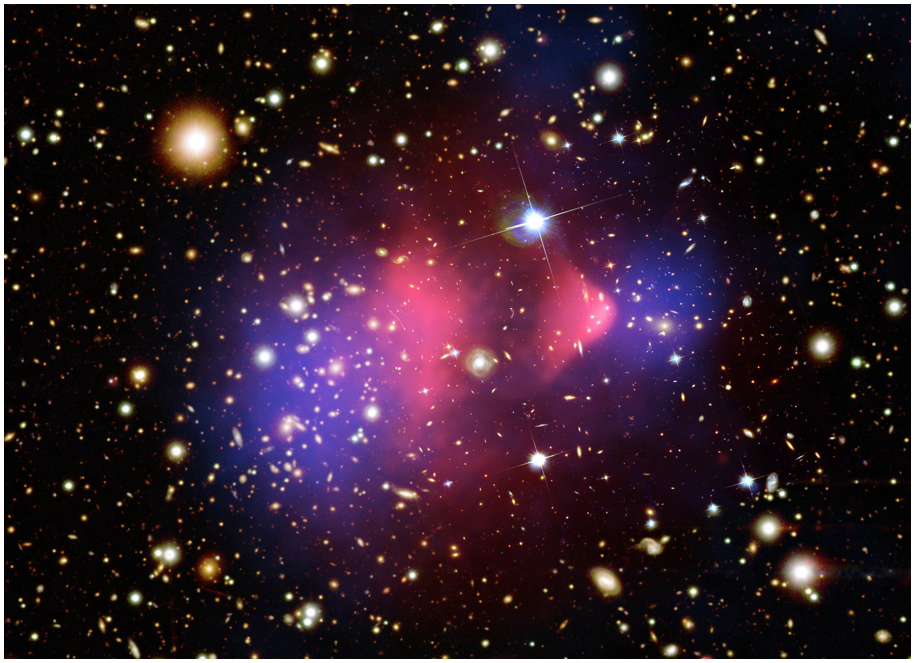


Figure 2: An image of the Bullet Cluster (1E 0657-56), which shows two galaxy clusters that have collided and passed through each other. The hot intracluster gas, observed in X-rays, is shown in pink and represents the bulk of the normal baryonic matter. The dominant mass component is mapped in blue from measurements of gravitational lensing. The image illustrates that the baryonic gas has slowed down due to EM interactions during the collision, while the DM component has passed through largely unaffected. Credit: NASA/CXC/CfA/M.Markevitch (X-ray), NASA/STScI, Magellan/U. Arizona/D. Clowe (optical and lensing map), ESO WFI (lensing map) [21].

Section 1.3) simulations. These properties are quantified with the matter power spectrum $P(k)$, whose shape and amplitude are key observables that quantify the variance of matter density fluctuations as a function of scale k.

Figure 3 shows a compilation of the measurements of $P(k)$ from different cosmological probes. These measurements constrain the matter content of the Universe and the cosmological concordance model, which implies the existence of DM. They are consistent with a Universe composed of about 85% non-baryonic DM. The agreement between observed and simulated large-scale structure strongly supports the ΛCDM framework, and therefore the existence of DM in general. In addition, combining this evidence with cosmic microwave background (CMB) anisotropy measurements has led to constraints on cosmological parameters, which point to the existence of DM in order to explain the observed clustering.

1.2.4 The cosmic microwave background

The CMB is the relic radiation left over from the Big Bang, emitted during the recombination epoch at redshift $z \approx 1100$. Before this time, the Universe was in a state of hot plasma where the photons were frequently interacting with matter. Once the Universe cooled enough for photon energies to fall below the hydrogen binding energy ($E \approx 13.6\,\mathrm{eV}$), the process $e^- + p^+ \leftrightarrow H + \gamma$ went out of equilibrium and photons decoupled from matter. No longer able to ionize neutral hydrogen atoms, the photons were finally able to propa-

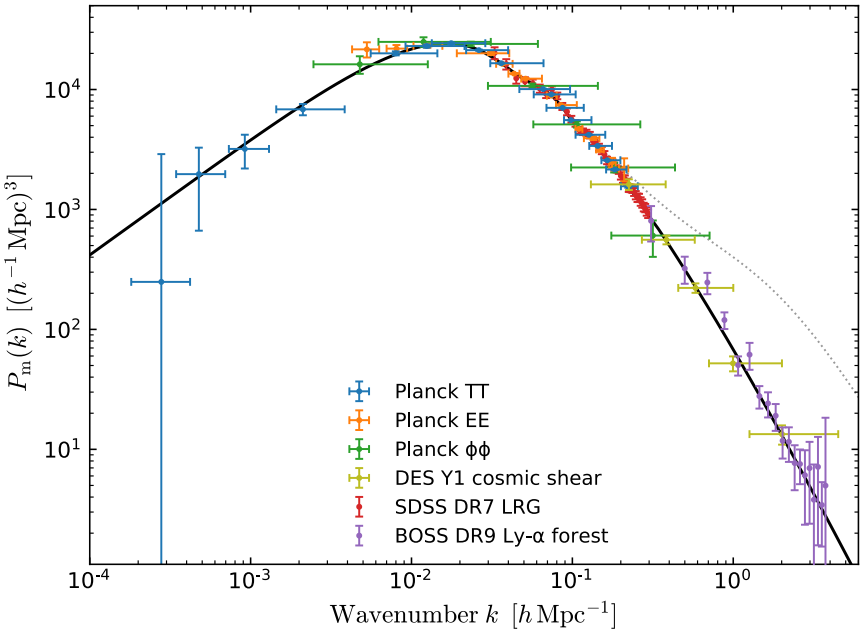


Figure 3: Matter power spectrum from various cosmological probes at redshift $z = 0$ as shown in [22].

gate freely through space and be captured in our telescopes, providing us with a snapshot of the very early Universe and its properties.

Since the CMB photons have been traversing the Universe ever since the epoch of recombination, their average temperature today is observed to be $2.725\,\mathrm{K}$ (corresponding to an energy of $6.344 \times 10^{-4}\,\mathrm{eV}$) due to redshifting from the expansion of the Universe [14]. In terms of wavelength, this energy corresponds to approximately 2 mm, placing it in the microwave range of the electromagnetic (EM) spectrum. It was first discovered in 1965 [23], during the early era of radio astronomy.

Precise CMB measurements by modern telescopes, such as the WMAP Telescope [24], or its successor Planck [22], reveal that early Universe was only nearly homogeneous with observed anisotropies present in the temperature having a typical amplitude of $\Delta T/T \approx 10^{-5}$. These fluctuations in the CMB temperature are a reflection of the primordial density perturbations in the Universe before recombination. Perturbations in the baryonic density can only start growing after decoupling from photons, meaning that the observed fluctuations in the CMB would have to be considerably larger in order to produce the structures in the Universe today [19]. The fact that the temperature perturbations are so small suggests the existence of non-baryonic DM in the early Universe, which would provide the necessary amplification to these perturbations.

1.3 The standard model of cosmology (ΛCDM)

Rapid progress in N-body cosmological simulations, coupled with a large body of observational evidence spanning disparate scales, has led to a standard model of cosmology. This theoretical framework of the Universe has been the cornerstone of cosmological studies for the past 30 years. It describes a spatially flat Universe whose energy budget is dominated by a cosmological constant (Λ) and assumes that all DM is cold (known as CDM),

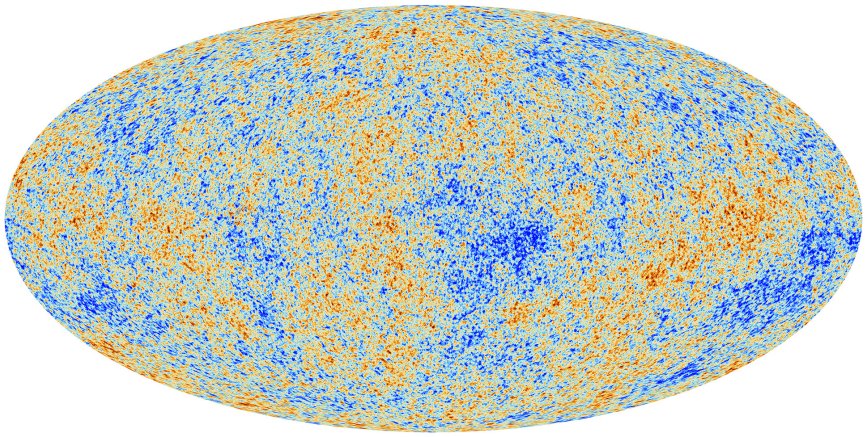


Figure 4: The cosmic microwave background. The red and blue spots reflect temperature fluctuations in the CMB, which are on the order of 10^{-5} K. Image: ©ESA and the Planck Collaboration [22].

which is why it is also called the ΛCDM model. Generally, this model is parametrized by the Hubble constant (H_0) and the dimensionless density parameters of the different constituents of the Universe. The former is a measure of the current expansion rate of the Universe, and when coupled with Hubble's law ($v = H_0 r$), gives the recessional velocity of objects in the Universe at a particular distance from the observer. The latter summarizes the total energy budget of the universe, normalized to unity via the following equation

$$\Omega_{total} = \Omega_m + \Omega_\Lambda + \Omega_k = 1, \tag{3}$$

where Ω_Λ corresponds to cosmological density of dark energy and Ω_m to the density of matter. Ω_k is related to the curvature of the Universe with a value of 0 signifying a flat universe, that is $\kappa = 0$. Precise CMB measurements by the the Planck Collaboration [25] have resulted in the following values:

$$\begin{aligned} H_0 &= 67.66 \pm 0.42\,\mathrm{km/s/Mpc}, \\ \Omega_m &= 0.3111 \pm 0.0056, \\ \Omega_{DM} h^2 &= 0.11933 \pm 0.00091, \\ \Omega_b h^2 &= 0.02242 \pm 0.00014, \\ \Omega_\Lambda &= 0.6889 \pm 0.0056, \\ \Omega_k &= 0.0007 \pm 0.0037, \end{aligned} \tag{4}$$

where the total energy density in matter is split into the density in baryons (Ω_b) and DM (Ω_{DM}), with the reduced Hubble parameter defined as $h = H_0/(100\,\mathrm{km/s/Mpc})$.

Under the ΛCDM paradigm, it is expected that all of DM in the universe is "cold". That is, it consists of collisionless, classical particles that had negligible thermal velocities early on in the Universe [16]. Furthermore, CDM particles are expected to be massive, stable over billions of years, and interact with SM particles mainly through gravity.

The most common theory is that DM originates from the very early Universe. At a time when the Universe was in a very hot and dense state, DM was in thermal equilibrium with the photon-baryon plasma through interactions of SM particles. When the Universe expanded, it cooled, causing DM to decouple from SM particles and its density to freeze

out. The leftover DM density (Ω_{DM}) is what we observe in the Universe today and is referred to as the relic or cosmological density [16].

The success of the CDM model in describing the Universe at large scales has led to the adoption of the model in the wider community. It is the default basis for cosmological simulations which model the gravitational interplay between baryons and DM [26]. In fact, much of the current understanding of galaxy formation and structure formation can be attributed to these simulations, which reveal a bottom-up scenario [27]. In this bottom-up scenario, tiny inflationary density fluctuations (with amplitudes of order 10^{-5}) provide the seeds that CDM amplifies into the first bound structures. Today's measurements of the CMB offer a snapshot of these primordial perturbations at recombination. When evolved forward under ΛCDM initial conditions, these inhomogeneities eventually reproduce the statistical clustering of galaxies and the baryon-acoustic features seen across cosmic surveys to remarkable precision.

Despite its successes, ΛCDM faces a number of small-scale challenges (e.g., missing satellites, core–cusp, "too-big-to-fail") that must be tested, in part, by probing the dark sector at sub-galactic scales [28, 27]. Therefore, to fully validate CDM as the true DM model, meaningful limits must be placed on its predictions for the minimum halo mass, internal density profiles, and substructure abundance. However, making robust sub-galactic tests of ΛCDM is hampered by the limited sensitivity of current observations at the aforementioned scales [28, 29]. To this end, theoretical predictions of small-scale dark substructure are also inconclusive due to difficulties in modeling baryonic processes relevant to galaxy formation (e.g., stellar feedback and gas dynamics) and the finite mass and spatial resolution of simulations that attempt to capture these effects [26].

1.4 Extensions to CDM

In the context of structure formation, the CDM model is only partially constrained by observations [27]. On very large scales, DM is expected to behave as CDM with general assumptions of its properties confirmed by cosmological probes on different physical scales (as is seen in Fig. 3). It is at the smaller, galactic and sub-galactic scales that the door is still open for alternative models of DM.

This is succinctly summarized by the dimensionless linear power spectrum ($\Delta^2(k)$), shown in Fig. 5, and taken from [30]. It characterizes the variance of DM density fluctuations as a function of wavenumber in log k intervals and provides an intuitive sense of how much structure is expected at different physical scales in the Universe, with smaller wavenumbers corresponding to larger scales and *vice versa*.

For $k \ll 1$, that is at galaxy cluster scales and above, DM density perturbations are in the linear regime and are described by linear perturbation theory where modes evolve independently of each other [27]. However, at $k \gg 1$ DM clustering is expected to be highly non-linear where structure formation takes place hierarchically in a bottom-up scenario (small structures collapse first). In a vanilla ΛCDM Universe, the power spectrum keeps rising well below the sub-galactic mass limit where $M \ll 10^{10}\,\mathrm{M}_{\odot}$ (shaded region in Fig. 5), forecasting the existence of DM substructures to arbitrarily low mass scales [31, 28]. At the same time, alternative DM models, such as WDM and ADM, exhibit a cutoff in their power spectra, suppressing structure formation. The corresponding physical scale k and mechanism for this cutoff is model-dependent and is ultimately a function of the particle nature of DM [27]. These alternative models therefore behave as CDM on very large scales but exhibit deviations at smaller, observationally challenging, scales. Finding evidence, which can reliably either confirm or reject DM clustering in the sub-galactic regime, is an important test of the CDM paradigm, providing valuable input in constraining possible DM

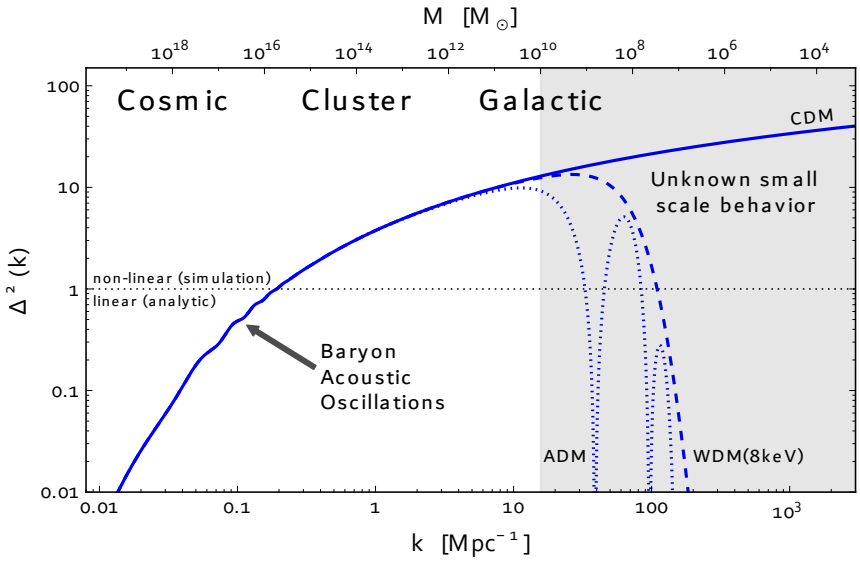


Figure 5: Dimensionless DM power spectra computed for different DM models: CDM, warm dark matter (WDM), and an example of atomic dark matter (ADM), which here represents a self-interacting dark matter (SIDM) scenario. For reference, the mass of the MW is $\approx 10^{12}\,\mathrm{M}_{\odot}$. *Adapted from [30].*

models.

2 Aims of the study

While astrophysical and cosmological observations over the past century have established the ΛCDM paradigm as the standard framework for describing the large-scale structure of the Universe, the behaviour of DM on sub-galactic scales remains poorly understood. Constraining the properties of galactic DM halos (their shape, density profile, and substructure) is crucial for testing the predictions of ΛCDM and for indirectly inferring the properties of the DM particle.

This thesis contributes to these efforts by studying the DM halo of the MW, focusing on both its smooth component as well as its substructure. At the same time, using both Bayesian inference and state-of-the-art ML methods, this work aims to leverage the increasing abundance of both observational and simulation data to develop new scalable computational tools for DM studies within our Galaxy.

The specific aims of this thesis are:

1. Determine the circular velocity curve and local DM density of the MW using Gaia DR3 data, addressing the shift from statistical to systematic limitations in Galactic dynamics through a GPU-accelerated Bayesian framework that self-consistently propagates key uncertainties related to the Sun's orbital parameters and tracer morphology (Pub. II)

2. Extend existing DM subhalo detection methodologies by developing ML-based approaches that search for DM substructure through stellar phase-space perturbations (Pub. I & III)

3. Evaluate the performance and limitations of deep learning (DL) models for detecting DM subhalos in simulated MW-like galaxies and their corresponding mock Gaia DR2 surveys (Pub. I)

4. Assess the detectability of individual DM subhalos through stellar wakes using controlled wind-tunnel N-body simulations and DL methods (Pub. III)

3 The Milky Way as a dark matter laboratory

In contemporary theories of structure formation, galaxies are thought to form within massive DM halos [27], implying that our own MW is also embedded in such a halo. Studying the distribution and substructure of DM on sub-galactic scales is a significant challenge, since the presence of DM can only be inferred from its gravitational effects on visible matter. That is, we are left to study it indirectly by analyzing the distribution and kinematics of stars. As both galactic and extragalactic astronomy are currently on the precipice of a true 'big data' era, boasting an abundance of observational data from current and expected surveys, the MW is increasingly becoming an excellent laboratory for DM studies.

In the following sections, data sources relevant to the work presented in this thesis are discussed. Specifically, a brief description of the Gaia mission and its data contents is given (relevant to Publications I & II), as well as an introduction to MW-like galaxy simulations and cosmological simulations in general (relevant to Publications I & III).

3.1 The Gaia era

Naturally, the reliability of any DM inference analysis in the Galaxy is contingent on the amount and precision of available data regarding stellar positions and velocities. A prominent herald of the new data-driven age in galactic astronomy is the European Space Agency (ESA) Gaia mission [17]. Gaia is a space-based observatory whose primary scientific goal is to map the stars in the MW in unprecedented volume and precision. Being a successor to the ESA Hipparcos mission [32] (1989 - 1993), it has already surpassed its predecessor's catalog volume by a factor of 10^4, with the total number of sources in Gaia DR3 being approximately 1.8 billion.

As of March 2025, the Gaia spacecraft has concluded its operations, with three major data releases (DRs) having been published during its active period by the Gaia Data Processing and Analysis Consortium. Though observations are no longer being carried out, the Gaia mission is expected to see yet two more releases: DR4 in 2026, and a final DR in 2030. Figure 6 shows an edge-on illustration of the MW based on Gaia data collected so far (as of January 2025). A notable feature in Fig. 6 is a warp in the galactic disk, the discovery of which was made possible by Gaia.

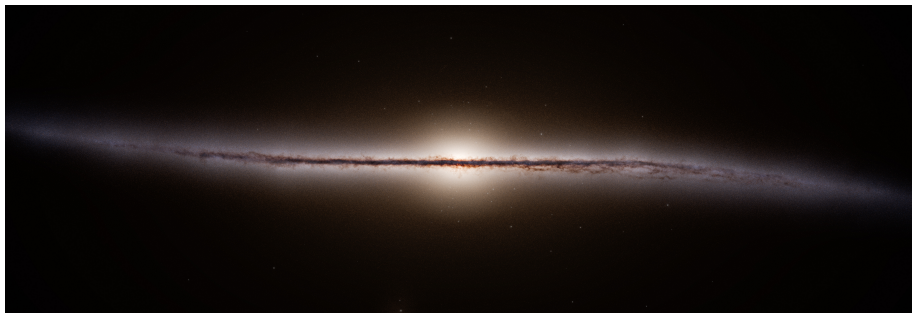


Figure 6: An edge-on depiction of the MW based on the Gaia data collected so far. Adapted from: ESA/Gaia/DPAC, Stefan Payne-Wardenaar

3.1.1 Gaia spacecraft

The Gaia space telescope was launched in 2013 from the French Guiana spaceport and started its survey in 2014 after having reached the L2 Lagrange point. Figure 7 depicts the

Gaia space telescope and its main constituents. Its payload module (second from the top) consists of three key instruments: an astrometric instrument, a photometric instrument, and the Radial Velocity Spectrometer (RVS).

The astrometric instrument measures stellar positions, parallaxes, and proper motions with unprecedented precision. Unlike Hipparcos, which relied on a predefined input catalog [17], Gaia scans the entire sky autonomously and detects minute shifts in stellar positions as it orbits the Sun at the L2 Lagrange point. Table 1 lists the median uncertainties for various astrometric parameters in Gaia DR3 as a function of G magnitude [33]. For context, the median positional precision for bright stars in the Hipparcos catalog was approximately 0.7 mas [34], whereas Gaia achieved a precision between 0.01 and 0.02 mas, an improvement by a factor of 35-70. Similar gains are seen in parallax measurements. Hipparcos had an uncertainty of ≈ 1 mas [34], while Gaia's typical parallax uncertainty at the bright end is between 0.01 and 0.02 mas.

The photometric instrument provides low-resolution photometry information across the blue (BP) and red (RP) bands, which enables the estimation of different stellar parameters such as effective temperature and surface gravity.

The RVS instrument uses the Doppler shift to measure the radial velocities v_r of stars. At launch, radial velocity standard errors were expected to achieve precision levels of $1\,\mathrm{km/s}$ for sources with $G_{RVS} \approx 11-12\,\mathrm{mag}$ and $15\,\mathrm{km/s}$ for stars at the fainter of the spectroscopic measurements ($G_{RVS} \approx 15-16\,\mathrm{mag}$) [17]. In Gaia DR3, a precision of $1.3\,\mathrm{km/s}$ at $G_{RVS} \approx 12\,\mathrm{mag}$ and $6.4\,\mathrm{km/s}$ at $G_{RVS} \approx 14\,\mathrm{mag}$ is reported, meeting or even surpassing initial expectation, specifically at the fainter end [35]. Already in Gaia DR2, the radial velocity precision for bright stars ($G_{RVS} \approx 4-8\,\mathrm{mag}$) was around $300\,\mathrm{m/s}$ [36].

The astrometric and RVS data from Gaia were fundamental to the analysis in Publication II, which relies on precise stellar position and velocity measurements to reconstruct the 3D velocity field of the stellar disk.

Table 1: Median uncertainties in Gaia EDR3 astrometric parameters by G magnitude as seen in [33].

G Magnitude	Position Uncertainty [mas]	Parallax Uncertainty [mas]	Proper Motion Uncertainty [mas/yr]
<15	0.01–0.02	0.02–0.03	0.02–0.03
17	0.05	0.07	0.07
20	0.4	0.5	0.5
21	1.0	1.3	1.4

3.1.2 Mapping the Galaxy with Gaia

In order to map the structure and dynamics in the Galaxy, it is important to determine the full six-dimensional phase-space information (positions and velocities) for individual stars.

Gaia adopted the International Celestial Reference System (ICRS) and equatorial coordinates to specify a star's position on the celestial sphere according to its right ascension (α) and declination (δ). However, to recover the 3D spatial distribution, these angular positions must be coupled with distance measurements. One possibility is to derive the distance from the parallax (ϖ) as measured by Gaia's astrometric instrument. While the inverse parallax $d = 1/\varpi$ can be reliable for nearby stars with low uncertainties, it is generally considered a noisy estimate for more distant stars [38]. This is because the previous involves a nonlinear transformation, where a small uncertainty in a star's parallax can inversely be interpreted as a large uncertainty in its heliocentric distance [39].

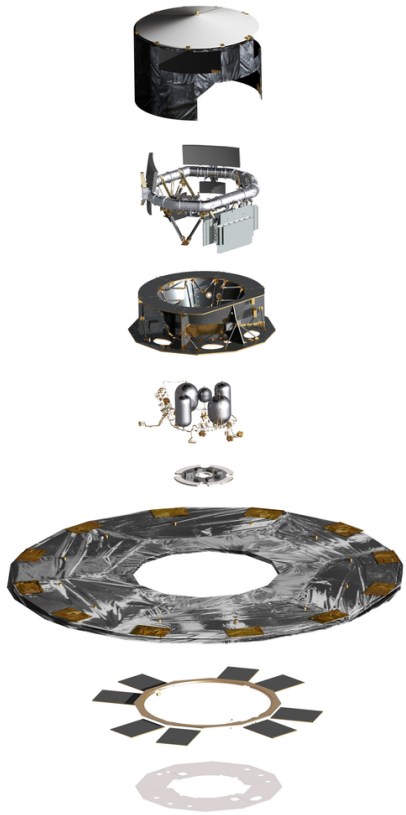


Figure 7: Exploded diagram of the Gaia spacecraft illustrating its primary components from top to bottom: thermal tent, payload module, service module, propellant systems, phased-array antenna, and deployable sunshield assembly with solar arrays. Credit: ESA/ATG medialab [37, 17]

Due to these limitations, several complementary distance catalogs have been derived from Gaia data. Examples of these catalogs include the General Stellar Parametrizer from Photometry (GSP-Phot) distances [40] and 'photogeometric' distances from [39], which use Bayesian methods to reconstruct stellar distances by leveraging both the astrometry and photometry information provided by Gaia. Nonetheless, these distances remain sensitive to the quality of parallax measurements, leading to under- or overestimation biases for distant stars. In Publication II, we studied how the adoption of either distance type affects the reconstruction of the final circular velocity curve.

The movements of stars on the celestial sphere are described by the proper motions in right ascension (μ_α) and declination (μ_δ), with units of [mas/yr]. The final kinematic observable required to reconstruct full 3D velocities is the line-of-sight or radial velocity (v_r). This is obtained by Gaia's spectroscopic instruments for only a subset of all observed stars. For instance, Gaia DR3 contains radial velocity measurements for ca. 33 million stars compared to the total of 1.8 billion observed stars.

All Gaia astrometry is reported in a reference frame which is fixed by observing extragalactic ($> 50\,\mathrm{Mpc}$) quasi-stellar objects (QSOs). These objects are located at vast distances from the Solar system, and, therefore, they are assumed to be effectively stationary on the celestial sphere. While far away, a typical QSO (e.g., a quasar) is extremely luminous, making it observable by Gaia. The latest realization of this reference frame is the Gaia Celestial Reference Frame (Gaia-CRF) 3, which is used in both Gaia Early Data Release 3 (EDR3) and DR3, and is defined based on observations of approximately 1.6 million QSO-like sources [41].

When reconstructing the 3D positions and velocities of the MW stars, it is useful to implement a change of reference from the heliocentric ICRS to a Galactocentric frame with the Galactic Center (GC) as the point of origin[3]. For illustration, Fig. 8 shows the sample used in Publication II in both galactic coordinates (as viewed from the solar location) as well as in Galactocentric Cartesian coordinates. Transforming Gaia's spherical astrometric data to Cartesian or cylindrical Galactocentric coordinates requires a sequence of rotations and translations that account for the Sun's position and velocity within the Galaxy. In the context of this thesis, transformations, along with propagation of the associated measurement uncertainties via Gaia provided covariance information, were carried out with the `gaia-tools` Python package[4].

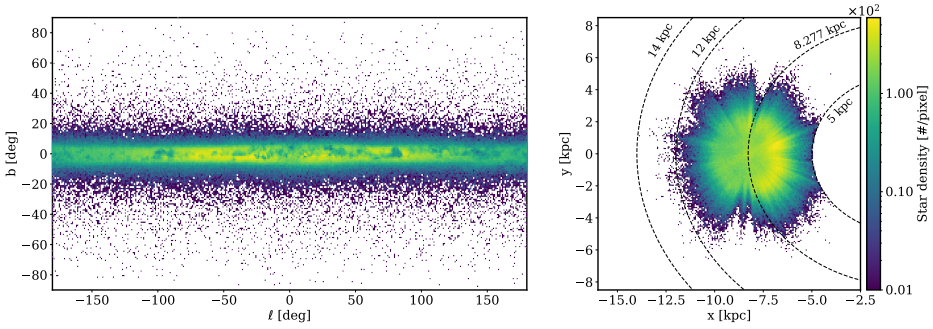


Figure 8: Sample of stars as used and depicted in Publication II. Left: stars are shown in galactic coordinates (l, b) in a heliocentric frame of reference. Right: the same sample is depicted in Galactocentric coordinates (x, y) with the location of the Sun shown on the dashed line at 8.277 kpc.

3.2 Milky Way-like galaxy simulations

Besides stellar surveys like Gaia, cosmological simulations play a crucial role in informing theories of structure and galaxy formation. Starting from ΛCDM initial conditions, these simulations model the evolution of DM, dark energy, and ordinary matter over an enormous range of physical scales. Even though the detailed nature of the first two is still unknown, their general properties, as they are understood in the context of ΛCDM, can be leveraged to reliably predict their behaviour [26]. Despite the lack of knowledge about what DM is composed of exactly and limited computational resources of the time, cosmological N-body simulations established that DM must behave as CDM on very large scales already in the 1980s [27].

[3]The Galactocentric transformation is not described in detail in the current thesis, but the interested reader can find a brief overview in Section 3.1.7 of the Gaia DR2 documentation [42].

[4]The code is hosted on GitHub and can be accessed from https://github.com/HEP-KBFI/gaia-tools

Today, the landscape of cosmological simulations is diverse, with simulation suites differing in their use of physics models, numerical methods, and initial condition implementations [43]. Advancements in computational hardware and algorithms have also resulted in substantial increases in resolution, being able to simulate trillions of particles at a time [26, 44], compared to $O(10^{3-4})$ particles possible in the 1980s [45]. Based on the type of initial conditions and simulated constituents, we can roughly divide this landscape into four major sections (see Fig. 9) [26].

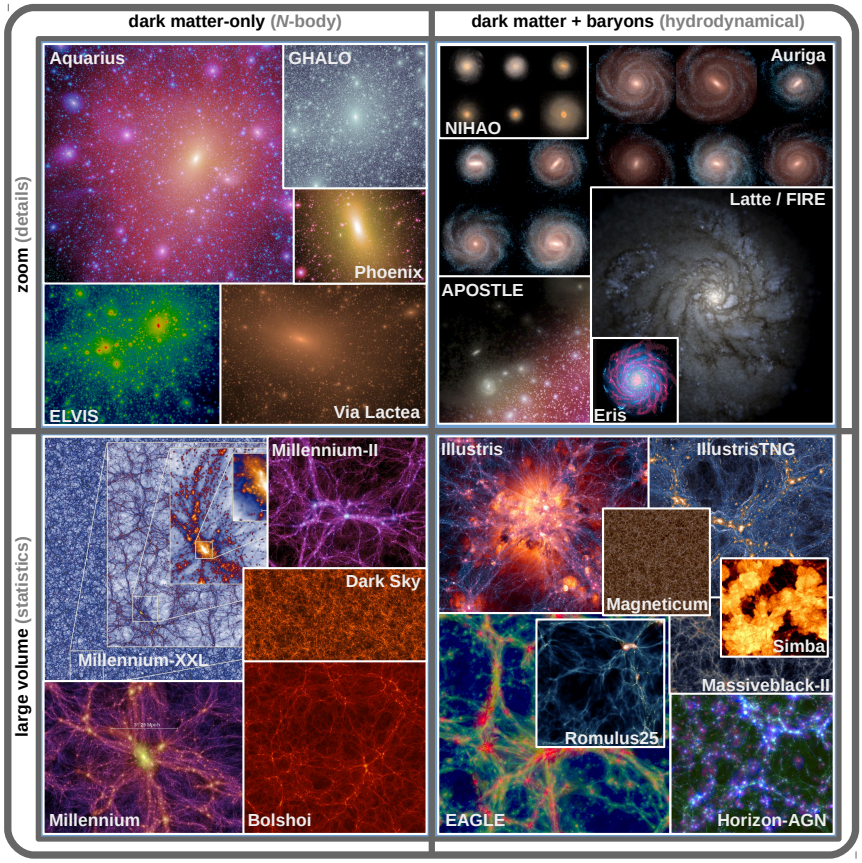


Figure 9: The landscape of contemporary cosmological simulations, as illustrated in [26], can be broadly categorized along two axes. First, by their physical constituents: dark matter-only (DMO) simulations and hydrodynamical simulations that include both DM and baryonic physics (left and right, respectively). Second, by their initial conditions: either zoom-in simulations focusing on specific regions of interest, or large periodic box simulations that model a cosmological volume of the universe at fixed resolution (top and bottom, respectively).

In terms of the initial conditions, simulations generally divide into "zoom-in" and large box (or volume) simulations. With zoom-in initial conditions [46], DM halos and the galaxy (or galaxies) therein are resolved in greater detail, resulting from an enhanced resolution of stars, gas and DM. The region of interest itself is surrounded by a low-resolution background to reduce computational cost while still preserving large-scale gravitational effects. In contrast, large box simulations employ uniformly sampled initial conditions across a much larger cosmological volume. This allows for the statistical study of galaxy formation and evolution across a wide range of scales, but typically at the expense of resolution,

meaning individual galaxies and their substructures are less well-resolved compared to zoom-in simulations.

Likewise, the particle content can be divided into DMO N-body simulations and hydrodynamical simulations, which include both DM and baryons. As the name suggests, the first have been used to study the large-scale distribution of DM, as well as the structure of DM halos [26]. The latter also includes baryonic effects by modelling astrophysical gases as inviscid ideal gases that obey the Euler equations. This treatment is accompanied by various sub-resolution approaches to model different baryonic effects such as gas cooling, stellar feedback, star formation, and so on. The hydrodynamical treatment of this gas is numerically expensive, with different numerical discretization schemes having been developed over the years. The specifics of these codes and the baryonic models are beyond the scope of this thesis, but the interested reader can find a more thorough overview in [26] and [27].

Although baryonic matter makes up roughly 5% of the energy budget of the Universe, this component is essential in reconstructing realistic properties of the visible matter in galaxies [26, 47]. Furthermore, on small scales, interactions between baryons and DM modify the inner structures of DM halos and the halo mass function [48, 49], further emphasizing the importance of understanding this interplay. As alluded to in Section 1, the interdependence between baryons and DM in sub-galactic environments is still rather poorly understood. This limited understanding is one of the key reasons why the core-cusp problem remains unresolved within the ΛCDM framework. This is a long-standing discrepancy between the steep central density profiles predicted by DMO simulations and the flatter, cored profiles observed in dwarf galaxies [28]. While hydrodynamical simulations that incorporate baryonic feedback, such as supernova-driven outflows, have shown non-trivial results in alleviating this tension [27], the effectiveness of such mechanisms appears to be highly sensitive to assumptions about star formation, feedback efficiency, and resolution. As a result, the robustness of baryonic solutions to the core-cusp problem is still actively debated [28], highlighting the relevance of modern, high-resolution simulations and novel subgrid physics models.

3.2.1 Latte galaxies

The analysis in Publication I was based on high-resolution cosmological simulations from the Feedback In Realistic Environments (FIRE) project [47, 50]. The FIRE simulations model galaxy formation in a ΛCDM cosmology while incorporating a state-of-the-art physics model to treat stellar feedback mechanisms in the interstellar medium (ISM).

We specifically utilized the Latte suite [47][5], which is a set of zoom-in hydrodynamical simulations within the FIRE-2 framework. The galaxies in the Latte suite were simulated from a set of isolated MW-mass halos at $z = 0$ with the mass required to be in the range $M_{200} = 1 - 2 \times 10^{12} M_{\odot}$. The isolation of the halos was satisfied by requiring that no neighbouring halos of similar mass were within at least $5 \times R_{200}$ or around 3 Mpc [49]. This selection of halos was made solely on the mass and isolation criteria, ignoring any information regarding their formation histories or subhalo populations.

Each simulation in the Latte suite was run using the GIZMO code in meshless finite mass mode, with mass resolutions of $m_{\rm DM} = 3.5 \times 10^4\,{\rm M}_{\odot}$ for DM particles and $m_{\rm gas} = 7.1 \times 10^3\,{\rm M}_{\odot}$ for initial gas particles. Gravitational softening lengths were set to 20 pc for DM and 4 pc for stars, while the gas softening was adaptive with a minimum of 1 pc. In Publication I, we used snapshots at $z = 0$ from three MW analogues, named `m12f`, `m12i`,

[5]The Latte suite is publicly available via the Flatiron Institute website: https://flathub.flatironinstitute.org/fire.

and m12m. The key properties of these MW-like galaxies are listed in Table 2.

Table 2: Properties of the Latte FIRE-2 galaxies m12f, m12i and m12m, which were used in this thesis. Table data is from [50].

Galaxy	$M_{\rm halo}^{\rm vir}$ $[10^{12}\,{\rm M}_\odot]$	M_* $[10^{10}\,{\rm M}_\odot]$	$m_{\rm DM}$ $[{\rm M}_\odot]$	$m_{\rm gas}$ $[{\rm M}_\odot]$
m12f	1.6	8.0	3.5×10^4	7.1×10^3
m12i	1.2	6.5	3.5×10^4	7.1×10^3
m12m	1.5	12.0	3.5×10^4	7.1×10^3

The Latte galaxies are recognized for their realism as they exhibit similar characteristics of the MW and the Local Group region in general. Notably, authors in [47] have shown that the observed dwarf galaxy population in the Galactic neighbourhood is in good agreement with those seen in the FIRE simulated analogues. Moreover, the simulations are also in agreement with different properties of the MW, including its stellar mass [50], disk morphology [51], and properties of the stellar halo [52]. For instance, in an ongoing work by Benito et al. (in preparation), we analyse the chemical bimodality of the simulated Romeo galaxy and find that its chemically selected disk components closely resemble those observed in the MW disk.

Beyond direct comparisons with current observations, the Latte galaxies have also been used to make predictions of the distribution and properties of DM subhalos in MW-mass galaxies [49, 53]. While these are important for interpreting the satellite populations of MW-like galaxies and testing the predictions of ΛCDM on small scales, the work in the current thesis utilized the simulations as a testing ground for developing and testing DM subhalo inference approaches.

3.2.2 Synthetic Gaia surveys

In addition to the MW-like galaxy simulations themselves, mock stellar surveys are useful tools to troubleshoot both new and existing DM inference methods before application to observations. Mock stellar surveys simulate stellar phase-space distributions, often tailored to closely replicate the structure, limitations, and uncertainties of specific observational surveys. For instance, by accounting for the expected observational noise from a specific instrument, one can evaluate the sensitivity and reliability of a developed inference method against realistic observational conditions. Below, largely based on the work by [54], is a description of Gaia-like mock stellar surveys utilized in Publication I.

Motivated by the advent of the Gaia mission, Sanderson et al. [54] developed the ananke framework with which data from the FIRE-2 simulated galaxies (introduced in the previous section) can be transformed into synthetic stellar phase-space surveys. Specifically, they have used the Latte suite of FIRE-2 galaxies (m12f, m12i, and m12m) to generate synthetic Gaia-like catalogs.

This was done by assuming local standards of rest (LSRs)[6] in the simulations, which can be thought of as different observational viewpoints inside the simulated galaxies that are the same distance from the GCs as the Sun is in the MW ($\approx 8.2\,{\rm kpc}$). The initial catalogs are generated by sampling stellar populations from star particles. The synthetic stars are given realistic stellar properties (mass, metallicities, absolute Gaia brightness, etc.) according to an initial mass function (IMF) and model isochrones.

[6] Formally, the local standard of rest is generally assumed to be a reference frame where stars in the Solar vicinity are moving around the GC on a circular orbit and thus individual stellar velocities are composed of both the velocity of the LSR and their peculiar velocities with respect to the LSR as: $V_\odot = V_{LSR} + V_{pec}$.

The IMF is an empirically derived law which characterizes the number of stars born in a star-forming region given a certain mass interval [14]. For stellar masses $> 1\,\mathrm{M}_{\odot}$ this is generally defined as a power-law

$$\xi(M) \propto M^{-\beta}, \tag{5}$$

where β is a dimensionless power-law index. What this relation implies is that in a given star formation region, the formation of low-mass stars is favored. In the `ananke` framework, the IMF of [55] is used, but another commonly used IMF is the Chabrier mass function [14, 56].

Isochrones are curves on the Hertzsprung-Russell (HR) diagram, which, derived from stellar evolution theory, reflect the observable properties of stars (their position on the HR diagram) of different masses but the same age and metallicity [57]. They can be used to assign realistic photometric and spectroscopic properties to synthetic stars based on their age and chemical composition.

The phase-space parameters, which are the positions x, y, z and velocities v_x, v_y, v_z are assigned by statistically sampling a phase-space distribution kernel (1D) centered on the parent particle. Although the procedure described above results in relatively realistic galaxy surveys, the smoothing lengths adopted during the 6D phase-space parameter generation can induce nonphysical effects and be troublesome when studying small-scale effects from dark subhalos.

The way in which extinction models are applied to the generated catalog represents a leap forward compared to previous mock catalogs. Dust extinction is the attenuation of stellar light as it passes through interstellar dust, primarily affecting the observed brightness and color of stars [58]. Dust grains scatter and absorb light with a shorter wavelength more than photons with longer wavelengths, and therefore, stars observed through interstellar dust appear dimmer and redder. Modeling dust extinction is crucial in creating realistic mock surveys, as it directly influences the observed stellar distributions. If not accounted for appropriately, it can significantly bias derived astrophysical parameters. In synthetic Gaia, instead of applying the empirically derived extinction map of the MW, gas evolution and distribution are tracked inside the galaxies, and dust maps are created in a more self-consistent manner. This is because the observed dust pattern of the MW represents a specific galaxy formation history that is different from the simulated galaxies and would introduce biases if haphazardly applied to the simulations. Thus, by estimating the extinction in each particular simulated galaxy, it prevents the introduction of nonphysical artifacts by preserving the correlations between dust extinction and regions of active star formation.

In addition to extinction maps, additional realism is achieved by incorporating Gaia-like observational characteristics. Observational uncertainties are convolved into the simulated true values of astrometric and kinematic properties with a realistic error model, which captures various instrumental limitations. Also, the Gaia selection function is accounted for by including only stars whose apparent brightness (accounting for extinction) falls into the range $3 < G < 21$. These imperfections certainly add a note of realism to these catalogs, which can be seen in Fig. 10.

In each of the Latte galaxies, three distinct LSRs were used to generate mock Gaia DR2 observations, therefore producing a total number of nine synthetic surveys over all three galaxies. The number of mock stars in the entirety of this dataset, after assuming magnitude limit and extinction effects, is close to 4.3×10^{10} [54]. This number represents the overall number of synthetic stars that would be observable with a Gaia-like instrument (assuming its selection function) in the specified LSRs locations.

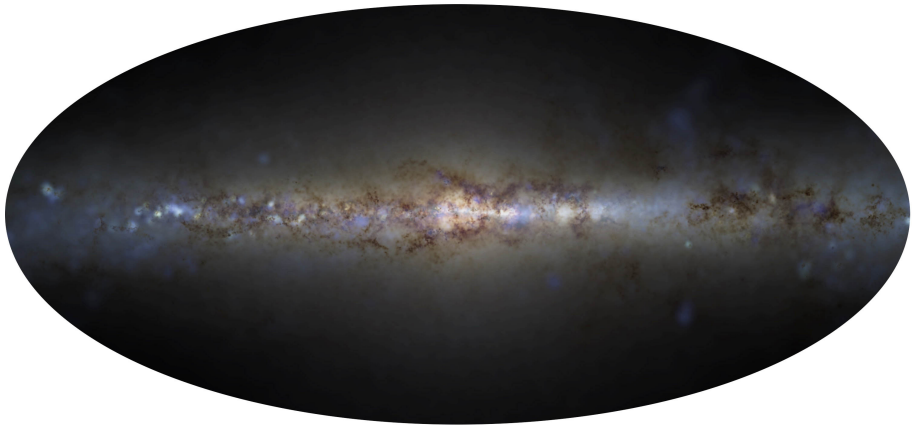


Figure 10: Aitoff projection of a synthetic survey from a particular LSR in the galaxy `m12i` *as shown in [54].*

The work in Publication I used the synthetic Gaia DR2 survey to study the detectability of dark subhalo-induced phase-space perturbations in a Gaia-like observational setting. The data used in the aforementioned study is publicly available on the FIRE website[7] along with a newer `ananke` dataset designed to mimic the contents of Gaia DR3.

[7]Data can be accessed on the FIRE website: `https://fire.northwestern.edu/ananke/`.

4 The smooth dark matter halo (Pub. II)

In contemporary structure formation theories, extended DM halos around galaxies are thought to be ubiquitous, providing the necessary gravitational anchors in which subsequent galaxy formation takes place [59]. Constraining the configuration of the MW DM halo is important for various reasons. For instance, the shape of the halo encodes information about the dynamical history of the MW [27], and is thus important for studies of Galactic archaeology. Also, since the properties and substructure of DM halos depend on the nature of DM [27], their study opens a window to probe the fundamental properties of DM.

The total DM halo of the MW is expected to consist of a dominant smooth component and a population of bound substructures known as subhalos. In this section, we focus on the study of the smooth component, the concept of which is introduced in Section 4.1. In Section 4.2, we summarize the methodology adopted in Publication II to constrain the circular velocity profile of the MW using the kinematics of red giant branch (RGB) stars. Section 4.3 summarizes the results from Publication II, where precise measurements and modeling of this velocity curve resulted in a robust estimate of the DM distribution near the Solar neighbourhood.

4.1 The smooth component

In the ΛCDM cosmological paradigm, galaxies such as the MW reside within massive DM halos, which provide the gravitational potential which governs the dynamics of visible matter (stars, gas, etc.), extending much further out than the visible stellar disk. Its precise and accurate distribution is therefore essential for understanding the dynamics and mass content of the MW [60]. Although the Galactic DM halo is also expected to exhibit substructure in the form of subhalos (see Section 5.1), the majority of the mass in DM (85 – 95% [27, 61]) is thought to be contained within a relatively smoothly distributed DM component [62]. This smooth component forms through mergers with other halos, which fall into the host galaxy, are tidally disrupted, and in time virialize within the parent halo [63, 61, 62].

Characterizing the MW DM halo is important since much of our understanding of the nature of DM and galaxy formation theories is reliant on the properties of DM halos [62]. For instance, the density profile of DM halos predicted by N-body numerical simulations in the CDM scenario, such as the Navarro-Frenk-White (NFW) profile and other generalized forms, consistently suggests steeply rising densities toward halo centers. Constraining the MW DM halo can serve as a test of ΛCDM, with deviations from the expected shape or substructure potentially pointing to extensions beyond the standard CDM framework [27].

Predictions about the CDM halo density profile often originate from DMO simulations [60]. In order to compare with observations, it is also important to factor in the effect of baryons. Cosmological simulations of galaxy formation have shown that the inclusion of baryonic processes can modify the inner structure of DM halos and alleviate some of the long-standing tensions in ΛCDM, such as the core-cusp problem. As alluded to in Section 3.2, simulating baryonic effects is easier said than done, as limitations in available resolution require the use of sub-grid models. Although empirically motivated, these models are not derived from first principles and can introduce systematic uncertainties. To complicate matters further, modifications to DM density profiles can also result from specific mechanisms of alternative DM models. For instance, in SIDM, interactions between DM particles in the presence of baryons can lead to the thermalization of the inner halo and creation of a DM core [64]. This is depicted in Fig. 11, where the density of a SIDM

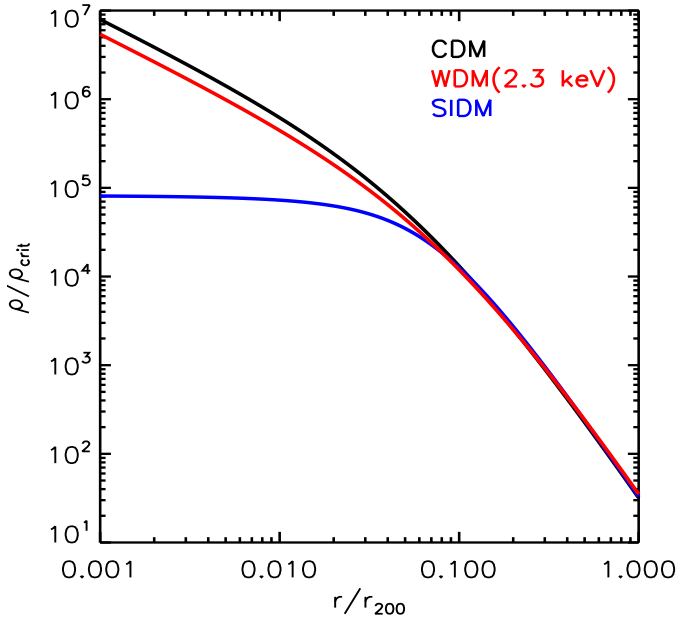


Figure 11: Spherically averaged density profiles for CDM, WDM, and SIDM as shown in [27]. Notably, CDM and WDM result in a cuspy density profile, whereas SIDM exhibits a core-like behavior at the inner radii.

starts to plateau at a certain distance from the center of the halo, whereas the density of CDM and WDM exhibits a cuspy behaviour. It is therefore not exactly clear which physical mechanisms are dominant in shaping the inner regions of galactic DM density profiles, as it can be from baryons, unknown DM physics, or a combination of both.

4.2 The circular velocity curve of the Milky Way (Pub. II)

An important tool for constraining the MW's smooth halo properties is the circular velocity curve, which reflects the total gravitational potential of the Galaxy (see Section 1). Precise measurements of this curve (either from stellar motions or other dynamical tracers) allow us to reconstruct the MW's gravitational potential and, by incorporating the contribution of baryons, gain information about the underlying DM distribution.

In Publication II, we developed a Bayesian inference approach to derive the circular velocity curve of the MW. This procedure included a careful quantification of various systematic uncertainties, such as those stemming from the Sun's Galactocentric distance and the spatial-kinematic morphology of the stellar tracers. By combining the latter with propagated measurement uncertainties, we were able to obtain a robust measurement of the MW's circular velocity profile.

4.2.1 Data and kinematic model overview

The study used a sample of ≈ 0.6 million stars on the RGB from Gaia DR3 [65]. Gaia DR3 is based on 34 months of data collection from 2014 to 2017 and contains astrometric parameters (positions, parallaxes, proper motions) for approximately 1.8 billion stars. Radial velocity measurements, which are essential for the full 6D phase-space reconstruction, are available for about 33 million stars.

RGB stars are warm tracers that are often used to constrain the dynamics of the stellar disk (e.g., [66, 67, 68]) as they have many benefits. For instance, their high luminosity allows them to be observed to large distances [59]. Also, as these are generally old stars, they are less sensitive to non-axisymmetric perturbations and can be used to study the axisymmetric component of the Galactic potential [67], which was the intent of this work. When adopting the standard assumption that the MW is in a steady state, this allows one to neglect the time derivative in the collisonless Boltzmann equation (CBE), i.e., $\partial f/\partial t = 0$ [59]. In adherence to this, we want to use equilibrium tracers, that is, stellar populations that are dynamically well mixed, which is generally not the case for young stars [69].

A practical method for connecting stellar kinematics to the underlying gravitational potential is the Jeans formalism. This formalism stems from the CBE, which describes the evolution of the phase-space distribution function in a collisionless system [59]. Integrating the CBE over velocity results in the Jeans equations. The key advantage of the Jeans formalism lies in computational speed and the fact that it does not require knowledge about the detailed shape of the distribution function, only its low-order moments [69]. The limitation to this approach is that the data must be binned in space in order to estimate those moments. In this study, we divided our stellar sample into eight radial bins in Galactocentric R with a width of 1 kpc for all bins except the last one, which was widened (2 kpc) to improve the statistical significance.

All of the stars in the final sample included their 6D phase-space coordinates: right ascension (α), declination (δ), parallax (ϖ) or other distance estimate, proper motions (μ_α and μ_δ), and radial velocity (v_r). The previous parameters, coupled with the Sun's galactocentric distance and orbital parameters, were used to transform the sample into a Galactocentric frame of reference and cylindrical coordinates. While this transformation provides a more intuitive overview of the phase-space distribution of the sample[8], it was also key in order to perform subsequent dynamical modelling with the radial Jeans equation.

The mean rotational velocity of warm stellar tracers is known to lag behind their actual circular velocity, which is a result of accumulated gravitational interactions between stars as they orbit the Galaxy [59]. In this case, the rotational velocity (v_ϕ) of a particular star can be modeled as

$$v_\phi = v_c - v_a, \tag{6}$$

where v_c is the circular velocity and v_a represents the asymmetric drift. Estimating the magnitude of the drift is an important step when inferring the axisymmetric component of the Galactic potential. The asymmetric drift inside each bin can be computed from the radial Jeans equation when combined with Eq. 6. This results in the following relation[9]

$$v_a = \frac{\sigma_R^{*2}}{v_c + \overline{v_\phi}} \left[\frac{\sigma_\phi^{*2}}{\sigma_R^{*2}} - 1 + R\left(\frac{1}{h_r} + \frac{2}{h_\sigma}\right) \right], \tag{7}$$

where v_c and $\overline{v_\phi}$ are the circular velocity and the mean rotational velocity, respectively. The terms σ_R^* and σ_ϕ^{*2} are the diagonal components from the velocity-dispersion tensor, which can be computed from the data as the variance of the Galactocentric radial and azimuthal velocity [59]. Finally, h_r and h_σ describe the spatial and kinematic morphology of

[8] After all, it is considerably easier to visualize disc dynamics from a bird's eye view of the Galaxy rather than from a viewpoint inside the disc.

[9] As it is not repeated in this thesis, see Section 3.1 in Publication II for a full explanation of this derivation.

the tracer sample. The first is the scale radius of the disc and relates to the radial number density distribution of stars. The second is the scale length of radial velocity dispersion.

4.2.2 Bayesian inference pipeline

We used the axisymmetric kinematic model to derive the circular velocities in each radial bin by using a Bayesian inference approach. That is, instead of plugging the numbers into the radial Jeans equation, we used a Markov chain Monte Carlo (MCMC) algorithm to sample the posterior of the circular velocity $v_{c,i}$ in each i-th bin. Specifically, the MCMC samples the posterior probability distribution $p(\theta|D)$ of a set of parameters θ, which in the context of this study included the circular velocities $v_{c,i}$ as well as h_r, h_σ and R_0.

The upside of this approach is that the nuisance parameters h_r, h_σ, and R_0 (hereafter denoted as $\theta_{\rm nuis}$) are now free parameters of the model. However, it is important to note that the motivation behind this inclusion was not to constrain the respective values of $\theta_{\rm nuis}$, but to propagate their uncertainties into the posterior distributions inside the bins and, by extension, into the circular velocity curve itself. This was achieved by marginalizing over $\theta_{\rm nuis}$, that is, allowing their walkers to explore the entire prior ranges during sampling, thus modifying the width of the circular velocity posterior distributions. In order to cover the range of values usually found in the literature, we opted for the following naive priors for $\theta_{\rm nuis}$:

$$\begin{aligned} R_0 &\in [7.8 - 8.5]\,\mathrm{kpc} \\ h_r &\in [2.0 - 4.0]\,\mathrm{kpc} \\ h_\sigma &\in [20.0 - 22.0]\,\mathrm{kpc}. \end{aligned} \tag{8}$$

The MCMC algorithm itself was implemented by using the Python library `emcee` [70], which is the backbone of the developed pipeline. During each iteration, the sampler produces an updated set of parameters with which to compute the likelihood function $p(D|\theta)$. This also means that during each such step, the Gaia data in each radial bin was repeatedly transformed from ICRS to the Galactocentric frame. This is because R_0, which is the Sun's Galactocentric distance, changes during each cycle as it was included in our nuisance parameters. This parameter is critical in the transformation procedure, and when it is updated, so must the positions and velocities inside the bins. In addition to this, the covariance matrix of each star also had to be propagated again according to the new parameter values. This is because the covariance information was used to recompute the weighted error of the rotational velocity inside each bin via bootstrapping.

The computational cost of all the individual routines that made up the likelihood computation at each step was severe. This led to comprehensive code optimizations, maximally utilizing any useful `NumPy` features (e.g., broadcasting) to speed up computationally expensive parts of the code. Though this resulted in non-trivial improvements, we further improvised by adopting graphics processing units (GPUs) to take care of the 'heavy lifting' in the pipeline. This was achieved by utilizing the `CuPy` library [71], which is essentially a GPU equivalent of `NumPy`. Figure 12 depicts a rudimentary scheme of the final workflow for computing the likelihood at each step of the MCMC.

In the end, a total of 6 GPUs and 12 central processing unit (CPU) cores (2 per GPU) were used to run the MCMC for a full duration of 10 000 steps. In order to avoid potential bottlenecks resulting from transferring large amounts of data to and from individual GPUs, the raw Gaia data was pre-loaded onto each device before the start of the fitting. Instead, during MCMC sampling, model parameters were transferred from the CPUs to their assigned GPU device, and after completing the Galactocentric transformation, stellar data was transferred back to the CPUs to continue the likelihood computation. Paral-

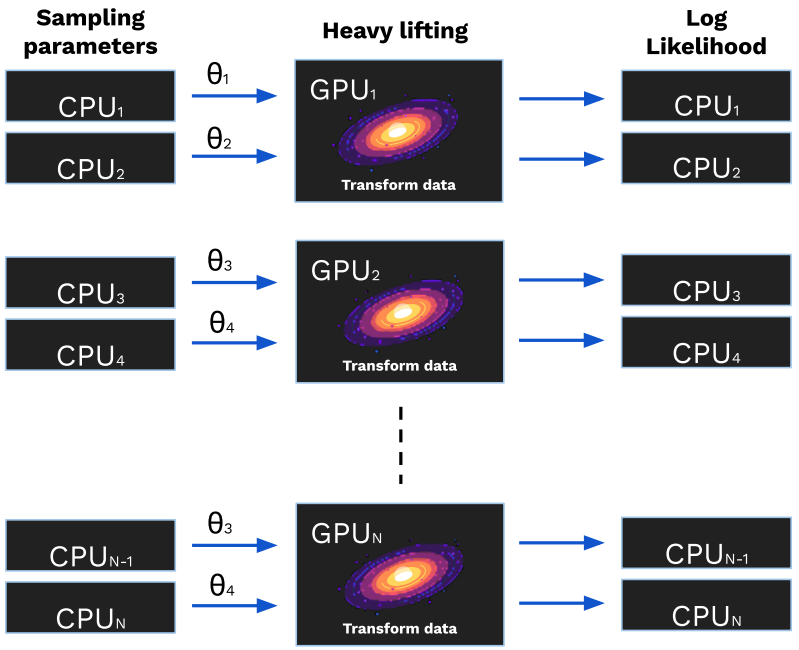


Figure 12: Illustration of a single likelihood computation step as implemented in Publication II.

lelization of the Bayesian inference pipeline across multiple CPUs and GPUs was critical to the success of this study, providing a 164-fold computational time decrease as compared to running on a single CPU.

4.3 Results and discussion

Modern surveys such as Gaia offer data with unprecedented precision and abundance. This means that as statistical uncertainties have decreased significantly, model systematics are becoming dominant sources of uncertainty and must be accounted for in a self-consistent manner.

The circular velocity obtained from Gaia DR3 as part of this study is presented in Fig. 13. Inside each bin (depicted as the dashed vertical lines in gray), the black point depicts the median of the posterior probability, with error bars denoting the 16th and 84th percentile values. The profile spans Galactocentric radii between 5 and 14 kpc. The general flatness of the rotation curve is a key observational signature indicating that the mass of the Galaxy does not consist solely of luminous matter, pointing to the existence of an extended dark component. As was mentioned in Section 1, this behavior has long been interpreted as evidence for the presence of a massive DM halo. An important caveat here is that we observe a flat slope of $0.4 \pm 0.6\,\mathrm{km s^{-1} kpc^{-1}}$ if we consider all radial bins. By removing the inner two bins, we found a slightly decreasing slope of $-1.1 \pm 0.3\,\mathrm{km s^{-1} kpc^{-1}}$. The slope of the rotation curve is therefore sensitive to the Galactocentric radii included in the analysis. More importantly, we confirmed that the shape of the rotation curve (particularly at outer radii) is also very sensitive to systematic biases in stellar distances. We saw that the slope changes significantly depending on the strictness of parallax quality cuts imposed on the sample and whether GSP-Phot or 'photogeo' distances are used.

Our Bayesian inference approach, combined with rigorous error propagation, resulted

in robust constraints on $v_c(R)$ that reflect both measurement and model uncertainties. When compared with other recent analyses, our velocity curve falls well within the range of other published measurements. This is particularly due to our large error bars from marginalizing over nuisance parameters. A key finding of this work was that uncertainty in the Galactocentric distance R_0 has a non-negligible impact on the inferred rotation curve. Because R_0 enters both the transformation to Galactocentric coordinates and the computation of the azimuthal velocity, even small shifts propagate non-trivially into the inferred circular velocity.

Finally, the computational infrastructure developed for this work enabled us to explore parameter posterior distributions efficiently despite the computational complexity of coordinate transformations and velocity uncertainty propagation. The inclusion of GPU-acceleration techniques into the MCMC pipeline was crucial in obtaining statistically converged results and ensuring our circular velocity profile is both precise and physically interpretable.

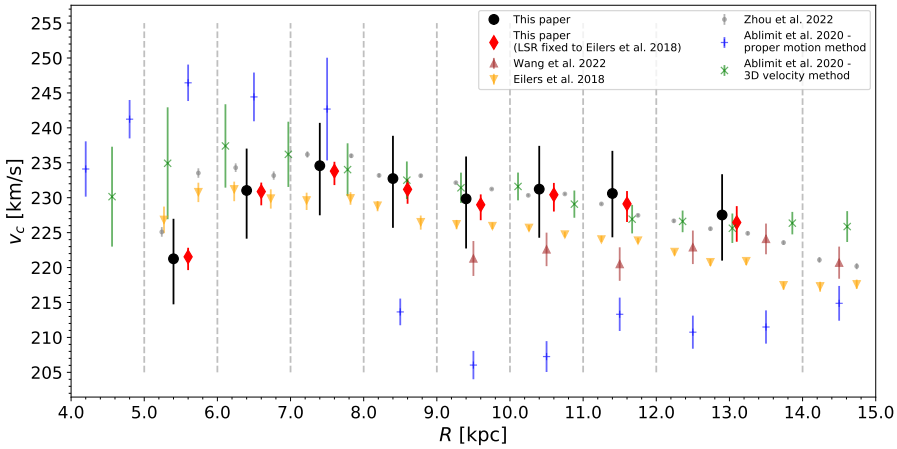


Figure 13: The circular velocity curve as obtained and shown in Publication II.

4.3.1 Dark matter density profile

Consistent with ΛCDM and rotation curve observations of other galaxies, the flatness of the MW's circular velocity curve derived in this work (Fig. 13) offers strong evidence for the presence of an extended DM halo. While the total gravitational potential is constrained by the stellar kinematics, subtracting the baryonic contributions (bulge, disk, and gas) allows us to infer the DM component (see Fig. 14).

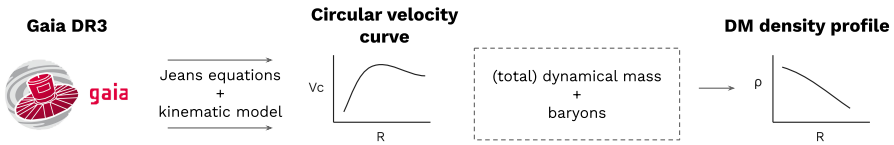


Figure 14: Illustration of the data analysis pipeline in Publication II.

The baryonic components were modeled by adopting an ensemble of density profiles, as compiled in [60], to incorporate uncertainties in the spatial distribution of stellar and

gas components. In addition, the uncertainty in the total mass within the baryonic components was accounted for by normalizing the profiles. For the stellar bulge, microlensing measurements towards the Galactic center were used, while the stellar disk was calibrated via the stellar surface density at the Solar position. The density of the smooth DM halo was modeled by a generalized NFW profile. Observed and model-predicted velocities were then compared following the Bayesian analysis of [72].

The procedure described above allowed us to propagate numerous systematic uncertainties into the final estimate of the DM mass within the inner Galaxy $R < 14\,\mathrm{kpc}$:

$$\log_{10}\left[M_{\mathrm{DM}}(R < 14\,\mathrm{kpc})/\mathrm{M}_{\odot}\right] = 11.2^{+2.0}_{-2.3}. \tag{9}$$

We were also able to estimate the local DM density $\rho_{\mathrm{DM}}(R_0)$ near the Solar radius. Historically, measuring the local DM density near the Solar system has been difficult [60, 69]. Generally, two types of these measurements exist, which are referred to as either global or local measurements [69, 19]. The first uses rotation curve (or other mass tracer) information to model the entire mass distribution of the Galaxy, given its different luminous components and the dark component. Results from global measures usually have very small statistical errors, although they are prone to significant systematic errors resulting from strong assumptions about the shape of the DM halo (including assumptions of spherical symmetry and equilibrium) [69]. In contrast, local measures, which are derived from vertical kinematics of stars near the Sun, have larger errors due to including fewer assumptions [69].

In our study, the local spherically averaged DM density at the Solar radius ($R_0 = 8.277$ kpc) was found to be

$$\rho_{\mathrm{DM}}(R_0) = \left(0.41^{+0.10}_{-0.09}\right)\,\mathrm{GeV/cm^3} = \left(0.011^{+0.003}_{-0.002}\right)\,\mathrm{M_{\odot}/pc^3}. \tag{10}$$

Figure 15 shows recent results from different global and local measures of ρ_{DM}, where the result from Publication II is depicted as the horizontal purple band. Our obtained value for ρ_{DM} is in good agreement with other recent estimates of the same type.

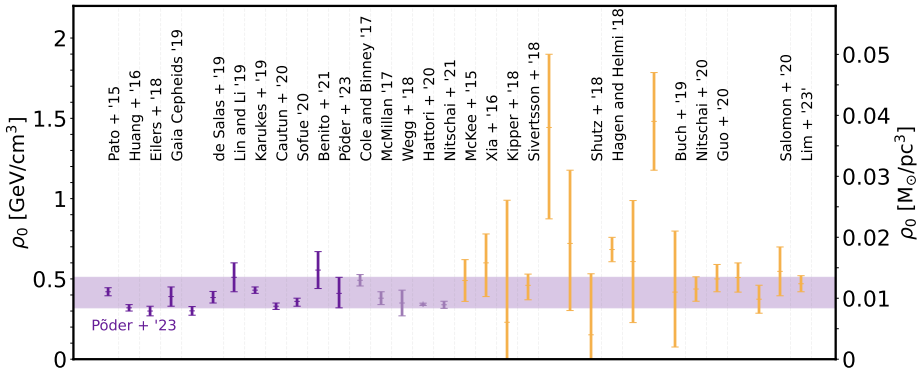


Figure 15: Estimates of the local DM density at the solar radius from a variety of studies. Purple error bars correspond to global methods, with dark purple indicating rotation curve-based analyses and lighter purple representing other global approaches. Yellow error bars show results from local methods, which typically rely on vertical kinematics of stars near the Sun. The horizontal purple band shows the result obtained in this thesis. Adapted from [73].

The different local DM density ρ_{DM} estimates encode information about the local shape of the MW's DM halo [69], which in turn has the power to inform on galaxy formation theories and probe the merger history of the Galaxy. Specifically, an account of

both local and global DM density measures can tell us whether the inner DM halo of the MW is prolate or oblate (see e.g., Fig. 1 in [69]). Furthermore, measurements of $\rho_{\rm DM}$ are critical for interpreting results from terrestrial direct detection searches. These are endeavors which aim to detect the scattering of DM from SM particles. An important equation describing this process is the expected recoil rate, which connects the interactions cross section ($\sigma_{\rm DM}$) and mass ($m_{\rm DM}$) of the DM particle and the amount of DM present in the detector ($\rho_{\rm DM}$). Since $\sigma_{\rm DM}/m_{\rm DM}$ is degenerate with respect to $\rho_{\rm DM}$, the latter has to come from an independent measure [69]. Our findings help refine this key astrophysical input by using a self-consistent dynamical modeling pipeline that includes both observational and systematic uncertainties.

4.3.2 Data products and developed software

To ensure full reproducibility, both the developed code and the input dataset used in the circular velocity curve estimation have been made publicly available. The RGB sample of nearly 0.6 million stars was uploaded to Zenodo[10]. Each entry retains its unique Gaia identifier (`source_id`), along with the star's position and proper motion in the ICRS reference frame, as provided in Gaia DR3. In addition, the dataset includes all quantities necessary for the Galactocentric transformation and subsequent uncertainty propagation. These include distance estimates, radial velocities, along with the measurement uncertainties of all astrometric and spectroscopic parameters and their correlation coefficients.

The full codebase that was used to transform the RGB sample and run the GPU-accelerated sampling via `emcee` is maintained in a GitHub repository[11]. The first version of this Python package was developed as part of the work in [74] and significantly extended throughout the work presented here. The improved implementation enabled the efficient and scalable transformation of Gaia stellar data, forming a critical component of the inference pipeline as well as a standalone result of the work itself.

4.3.3 Future outlook

This study was based on the radial Jeans equation by assuming a steady state of the MW and an axisymmetric Galactic potential. In addition to this, the disk was also expected to be symmetric with respect to the Galactic plane by omitting partial derivatives with respect to z in the Jeans equation. This simplification was not expected to change the results of the current study as our sample was chosen to be near the disk ($|z| < 0.2\,\mathrm{kpc}$), effectively minimizing any gradients in the vertical direction. In future studies, it would be interesting to improve our methodology by relaxing the assumptions in this study and also considering non-axisymmetric perturbations in different layers of the disk.

Future Gaia DRs (e.g., DR4) and the Large Synoptic Survey Telescope (LSST) will also provide additional data with which to increase the fidelity of the current study. In comparison to the data source used in this study, Gaia DR4 will contain 500 TB of data in comparison to the total data volume of 10 TB in Gaia DR3 [75]. Additional radial velocity measurements will certainly help in mapping the disk near the Solar vicinity in higher detail as well as extending the rotation curve to larger Galactocentric radii and, in the direction perpendicular to the disk, to higher $|z|$.

[10] https://zenodo.org/records/8014011
[11] https://github.com/HEP-KBFI/gaia-tools

5 The substructure of dark matter halos (Pub. I & III)

In Section 4, we saw how the circular velocity curve can be used to constrain properties related to the smooth component of the MW DM halo. In this section, we turn our focus to ML-based studies regarding the substructure of DM halos, which utilize both simulated MW analogues from the FIRE project and idealised N-body simulations performed in the computing cluster of the National Institute of Chemical Physics and Biophysics in Estonia. Sections 5.1, 5.2, and 5.3 outline the theoretical and methodological context of both studies presented in the current section. In Section 5.4, a study of the statistical imprint of subhalos on the visible galactic structure using MW-like simulations is introduced (Publication I). Following, Section 5.5 describes the formation and detection efforts of stellar wakes using wind tunnel simulations as investigated with ML models trained to recognize wake signatures in image-like datasets (Publication III). Finally, the results from both publications are summarized in Section 5.6.

5.1 Dark matter subhalos

In the ΛCDM cosmological model, structure formation takes place hierarchically in a bottom-up scenario: small DM halos form first and merge to build larger systems. While many DM halos are expected to be disrupted during this process, a significant population is predicted to survive [49] within the virial radius of the larger halo, becoming DM subhalos. These subhalos are gravitationally bound overdensities that orbit within the larger, smooth host halo (see Section 4) and represent one of the most important predictions of the CDM model [27]. Figure 16 shows an illustration of a subhalo merger tree, where smaller subhalos start to combine at higher redshift (corresponding to earlier times) and over time merge together to form a large host halo.

The absence of reliable observations of the matter power spectrum on sub-galactic scales allows several other possible models in place of CDM (see Section 1). Currently proposed alternative DM models in the literature (e.g., WDM, SIDM) behave similarly to CDM on cosmological scales but exhibit a cutoff in the power spectrum at smaller scales. The microphysical properties of any DM particle candidate model translate into predictions regarding the amount of subhalos expected at all spatial scales.

Generally, the abundance of subhalos is characterised by the subhalo mass function (SHMF), which describes the number of DM subhalos per unit mass. In CDM, the SHMF continues as a power law toward low masses, predicting an increasingly large number of subhalos [28]. However, this is not the case in alternative DM models where subhalo formation is suppressed due to cutoffs in their power spectra induced by model-specific mechanisms (see Section 1). Following the analytic formalism in [77], Fig. 17 shows the SHMF assuming three different DM scenarios: CDM, WDM, and fuzzy dark matter (FDM). Crucially, it demonstrates that key properties of the DM particle (e.g., its mass) have significant effects on its ability to cluster at small scales and therefore modify the shape of the SHMF. Constraining the SHMF in the sub-galactic regime is therefore an indirect way to study the nature of DM and an important test of the CDM model.

5.2 Dark subhalo detection efforts

Cosmological N-body simulations have shown that the abundance of CDM subhalos is expected to increase significantly toward lower masses, following a power-law distribution. In CDM, subhalos can span a broad range of masses (down to Earth-mass [29, 27]) and are distributed throughout the galactic DM halo [31]. Although more massive subhalos ($> 10^8\,\mathrm{M}_\odot$) may host satellite galaxies, below this mass, the vast majority are expected

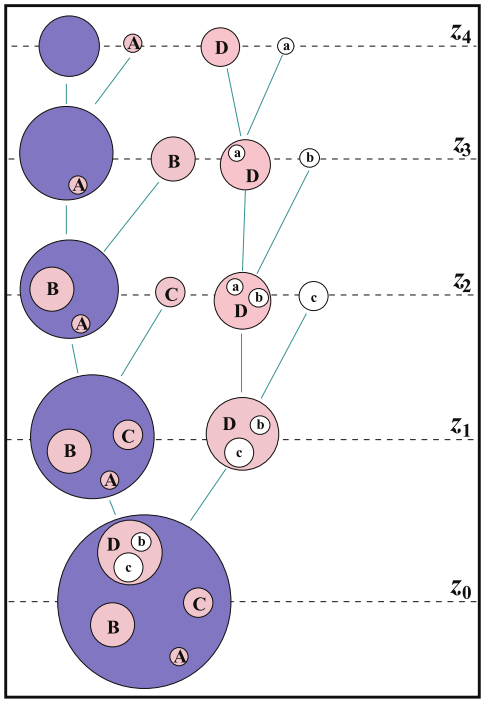


Figure 16: Hierarchy of subhalos, adapted from [76]. It depicts how small subhalos start to merge at earlier times (z_4) and finally end up as part of a large host halo at z_0.

to be devoid of baryonic content. This is thought to be a result of both their shallow gravitational potentials (i.e., their low masses) and reionization effects [78, 79]. The latter causes gas to heat up and evaporate from low-mass halos, significantly suppressing star formation due to insufficient gas retention.

Low-mass DM subhalos, which do not harbor any stars (i.e., dark subhalos), are therefore extremely challenging to detect with traditional observational methods, which rely on emitted or reflected EM radiation. Instead, the presence and dynamics of dark subhalos can only be inferred indirectly through the subtle gravitational interactions with the surrounding stellar medium. This is in contrast to studies of the smooth component (see Section 4), where the motions of stars (and gas) are strongly governed by the gravitational potential of the MW DM halo. Despite the challenges, several promising sub-galactic and extra-galactic probes have been proposed to constrain the low-mass end of the SHMF indirectly.

A promising direction among these probes is the search for gaps and kinematic disturbances in cold stellar streams [80, 81, 82]. This approach relies on the fact that stars in streams are organized into coherent elongated structures. As a subhalo passes a nearby stream, it perturbs its structure and is able to produce a gap in its density. The geometry and magnitude of this disturbance can then be used to infer the presence and mass of the suspected perturber. The advent of Gaia has provided a major boost to these detection methods, as it has facilitated the identification of both new streams and additional stars in previously known streams, thereby improving the resolution of likely perturbations. Notable examples of this detection approach include studies of the GD-1 stream [83, 84], where a subhalo with a mass of $5 \times 10^6\,\mathrm{M}_\odot$ has been suggested as a plausible culprit for

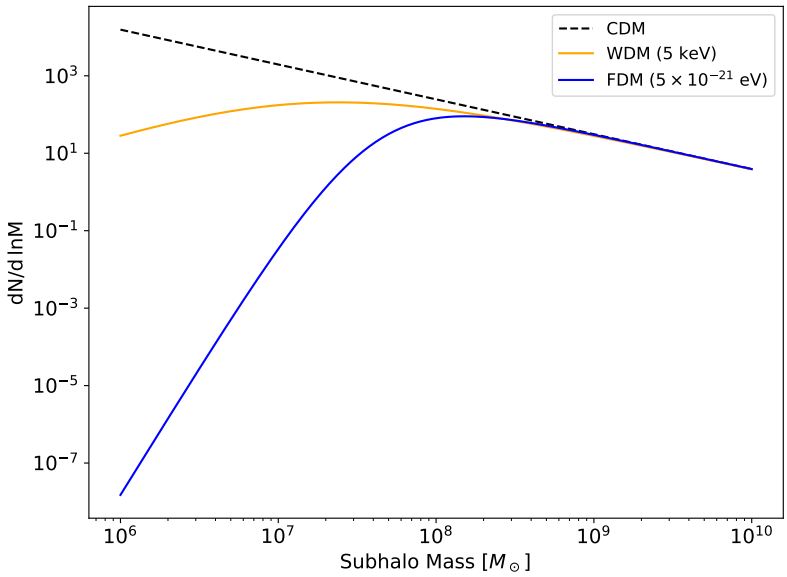


Figure 17: Analytic form of the SHMF for CDM (dashed), WDM (orange), and FDM (blue), following the work of [77]. The plot shows the expected number of subhalos per unit mass for each DM model. At the higher end of subhalo masses, the models predict similar abundances, making them observationally indistinguishable. At lower masses, the SHMF curves diverge due to model-dependent cutoffs in their power spectra. Due to having non-negligible thermal velocities, small-scale structure is suppressed for WDM through free-streaming. In the case of FDM, which is an ultralight bosonic DM model with a de Broglie wavelength $\lambda \sim$ kpc, formation of subhalos is suppressed due to quantum pressure effects.

perturbations in the stream.

Another proposed method involves pulsar timing arrays (PTAs). These are used to detect subhalos through subtle timing residuals in the arrival times of pulses from millisecond pulsars [85]. Specifically, depending on the transit of the subhalo with respect to Earth, the gravitational field of the perturber is expected to either advance or delay the pulses, thus modifying the observed frequency of the pulsar [86].

The previous two subhalo detection efforts are viable within our own Galaxy. Outside the MW, strong gravitational lensing (briefly introduced in Section 1) offers a third avenue. Instead of detecting subhalos individually, collective subhalo-induced perturbations can be inferred by analysing the shape of gravitational lenses [87]. That is, the population of subhalos inside a host halo imprints a deformation in the shape of the lens, which is treated as the residual of the main lens model.

In recent years, studies of DM subhalo inference in gravitational lenses have seen widespread adoption of ML approaches. For instance, [88] used a convolutional neural network (CNN) to determine whether an image contained DM substructure as part of a binary classification task. In [87], the authors employed a U-net architecture as part of an image segmentation task and found their method to be able to detect subhaloes of mass $> 10^{8.5}\,\mathrm{M}_\odot$ in the lensed images. In contrast to the latter examples, [89] studied and demonstrated the non-trivial performance of unsupervised ML methods by showing that the reconstruction loss of lenses with and without substructure has noticeably different distributions.

Similarly to the lensing-based approaches above, the dark subhalo detection studies presented in this thesis also leverage DL methods. Contextually, they are more aligned with other sub-galactic searches (e.g., PTAs, stellar streams) as they are targeted towards Gaia (or even LSST) data applications. By also keeping in mind DRs yet to come, the aim of the current thesis is to broaden the scope of these existing methods by laying the groundwork for future ML-based subhalo detection endeavors in the MW.

5.3 Deep learning

The abundance of data from current and future surveys encourages the adoption of novel data-driven methods as astronomical datasets are moving towards the exabyte-scale [90]. ML, and particularly DL, has become indispensable in modern astronomy, offering powerful tools to manage large datasets, extract subtle signals, and discover previously undetected structures in noisy, complex data. This section provides a brief overview of ML concepts, which underpin the analyses presented in both Publications I & III.

5.3.1 Rise of deep learning

The recent surge in the adoption of DL across multiple fields of research is attributable to several key developments. Based mainly on both [91, 92], we can summarize these within the following categories:

Algorithms Historically, deep neural networks were notoriously challenging to train due to the vanishing gradient problem[12], resulting in shallow neural networks with only a few layers. The issue therein lay in inefficient gradient backpropagation, which is needed to compute the updates to the model weights, allowing the model to learn. Advances in optimization algorithms such as RMSProp [93] and Adam [94] drastically improved training efficiency, enabling deeper architectures and more accurate models. These developments were further complemented by both better activation functions for neural layers and improved weight-initialization schemes [92].

Hardware The advent of GPUs has greatly accelerated DL research. Initially, GPUs found use primarily in the gaming industry, where they were developed to render computer game graphics [92]. Driven by market demand, investment from companies like Nvidia and AMD made GPUs efficient and abundantly available, facilitating the use of GPUs for general-purpose computation. In contrast to CPUs, GPUs can perform a massive number of computations in parallel, giving them much higher bandwidth. This enables the training of very large models in a reasonable amount of time and facilitates increased model complexity and scalability. Indeed, modern high-performance computing clusters are designed with this GPU-acceleration in mind. For instance, the LUMI-G hardware partition in the LUMI supercomputer located in Finland has 2978 compute nodes with 4 GPUs each [95].

Software The development of open source, high-level software libraries and application programming interfaces (APIs) has lowered the barrier for entry. The most popular DL libraries used today (Tensorflow [96], Keras [97], PyTorch [98]) require a relatively low level background in programming, making them more user-friendly for researchers outside of computer science.

[12] This problem occurs when the gradients of the loss function are washed out as they are propagated backwards through the layers of the model, inhibiting effective weight updates in earlier layers.

Data DL models thrive on data. Nowadays, data is becoming ever-abundant, with massive multi-dimensional datasets emerging in both academia and industry alike. Specifically in astronomy, we are seeing a significant increase in available data from current (e.g., Gaia [17], SDSS [99], APOGEE [100]) and future surveys (e.g., LSST [101], Euclid [102]). Owing to their data-intensive nature, training DL models has become increasingly easier, as modern infrastructure enables the distribution and use of large datasets in ways that were not feasible in the pre-Internet era.

5.3.2 Architecture of deep neural networks

At the heart of DL methods employed in this thesis lies the artificial neural network (ANN). The interest in ANNs in astronomy can be traced back 30 years (e.g., [103, 104]), although at the time, the methods were subject to intense skepticism [105]. One of the early successes of ANNs in astronomy is attributed to works regarding photometric redshift estimation [106]. Today, neural networks are being explored to solve problems on various astrophysical scales. These include, for instance, the inference of stellar ages from photometry [107], the detection and classification of strong gravitational lenses [88, 108, 109, 89, 110], galaxy image generation [111, 112], DM inference in the MW [113], and DM distribution reconstruction in cosmological simulations [114, 115][13].

The ANN is quite a robust algorithm due to the universal approximation theorem, which states that any sufficiently large network can approximate any continuous function to arbitrary precision [91]. In practice, the ANN is a group of interconnected nodes, through which data is propagated using individual chains of tensor operations. At each node, an affine transformation is applied to the input value along with an activation function as [91]

$$\vec{y} = \sigma(W\vec{x} + \vec{b}), \tag{11}$$

where W is a tensor containing the model weights and $\vec{b}$ is the bias tensor. Naturally, $\vec{x}$ refers to the input and $\vec{y}$ to the node output. The activation function σ is used to capture nonlinear patterns in the input data. Common choices include the rectified linear unit (ReLU), sigmoid, hyperbolic tangent (tanh), each with different characteristics in terms of gradient behavior and expressivity [91].

Figure 18 depicts a very simple case of a neural network, organized into four fully connected (or dense) layers: input, output, and two hidden layers between them. The number of hidden layers generally refers to the depth of the model, which is considered "deep" if it includes multiple hidden layers that perform non-linear feature transformations [117]. The input layer defines the expected dimensionality of the data samples. The set of variables that make up the input vector is generally referred to as the input features. For example, in Fig. 18, the input would be a vector composed of 6 features, which could be the 3D coordinates (x, y, z) and 3D velocities (v_x, v_y, v_z) of a single star. The final prediction of the network is given at the output layer. In the example of Fig. 18, all values propagated through the model are coalesced into a single value, which, depending on the problem at hand, could either be a probability of a star belonging to a particular class (classification) or a continuous value (regression).

Through iterative training, the ANN is able to learn complex patterns in the data by updating the model weight tensors, which contain the "knowledge" of the model [92]. This is done by coupling the training process with a feedback mechanism in the form of

[13] In reality, the landscape of ML applications in astronomy is becoming increasingly diverse. The interested reader can find a thorough overview of important developments in either [105] or [116].

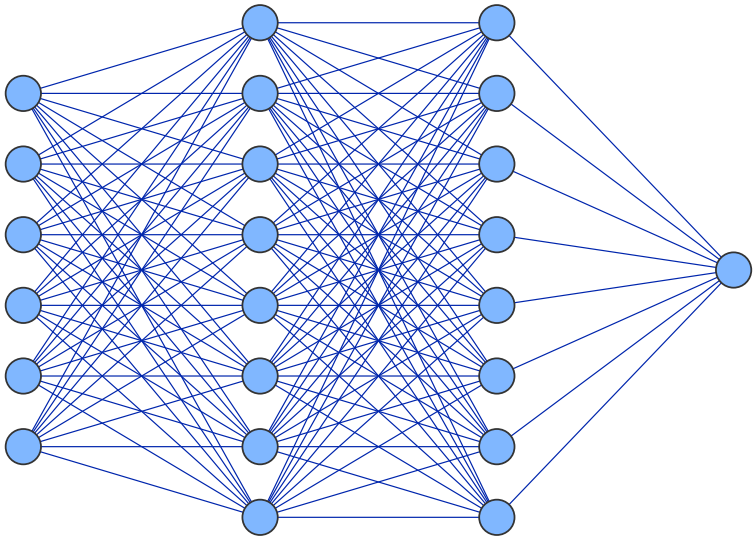


Figure 18: Schematic of a fully connected (dense) deep neural network. Each node in a layer is connected to every node in the subsequent layer. The network expects a 6-dimensional input vector, propagates it through two hidden layers with 8 neurons each, and produces a single scalar output at the final layer.

a loss function[14]. The loss function quantifies the success of a given model by computing the distance between predicted and true target values. While the true targets are used to compute the loss function after each forward pass, they are never exposed to the model during training [91].

An important property of the tensor operation chains in ML models is their differentiability. The derivatives are computed automatically by the model so that only the forward pass of the algorithm needs to be defined. This enables the model to use the chain rule to compute the gradient of the loss function with respect to the model parameters, which is in turn used by the optimizer to efficiently update the model weights. This process is known as backpropagation, and it is how the model is able to learn by iteratively tuning its weights to minimize the loss function and thus improve prediction performance.

5.3.3 Learning types

Depending on the availability and structure of the data, DL models can be trained using different learning paradigms. The most commonly used types are supervised and unsupervised learning, both of which are employed in this thesis.

In supervised learning, the model learns from input-output pairs, where each data point is associated with a known target value or label. The goal is to minimize a loss function that quantifies the discrepancy between the model's predictions and true labels. This approach is used in both Publications I & III, where labels of the training data are used explicitly to distinguish signal from background.

In unsupervised learning, the learning objective is defined in a way that does not require a labeled dataset. Instead, the model learns from the input data itself. In this thesis,

[14] Also known as an objective function.

unsupervised learning is employed in Publication I in the form of an autoencoder (AE), which is trained to reconstruct the phase-space vectors of unperturbed stars in the Latte galaxies (see also section 5.4).

Though not considered in the current thesis, other learning paradigms, such as semi-supervised and contrastive learning, have become increasingly popular in recent years [105]. These approaches lie between the fully supervised and unsupervised ends of the spectrum.

5.3.4 Hyperparameters

DL models depend not only on the structure of the network itself, but also on a variety of tunable parameters known as hyperparameters. These parameters control how the model learns and generalizes. Unlike the internal weights of the network, these must be set before training and are usually chosen through experimentation or prior knowledge. In DL models, some of the most common hyperparameters are:

Learning rate This controls the magnitude of parameter updates at each training step. If set too low, the model may converge very slowly or get stuck in a suboptimal state. If set too high, training can become unstable, and the model may fail to converge or overshoot optimal solutions.

Batch size This defines the number of samples processed before the network updates its internal parameters.

Elements of the network architecture For instance, number of hidden layers, number of units (neurons) per layer, choice of activation function.

Number of epochs The number of full passes (iterations) through the training data.

Regularization techniques Regularization plays an important role in ensuring model generalizability. In order to mitigate overfitting, techniques such as dropout, early stopping, and input normalization can be applied where appropriate.

Choosing appropriate hyperparameters is often an empirical process. In this thesis, hyperparameters were selected either through small experiments (Publication I) or the RandomSearch algorithm in the KerasTuner framework [118].

5.3.5 Evaluation and performance metrics

Evaluating the performance of an ML model requires appropriate metrics that reflect how well the model is able to differentiate between target and background samples. In this thesis, both supervised and unsupervised methods were used. Each approach has its own evaluation context, but a common set of performance metrics is used throughout.

The main metrics used in the works described in the current thesis are the true positive rate (TPR) and the false positive rate (FPR). The first is also known as sensitivity, and it records the fraction of samples that are correctly identified as signal. We can define it as:

$$TPR = \frac{TP}{TP + FN}, \tag{12}$$

where TP is the number of true positives and FN the number of false negatives.

The second metric describes the mislabeling of background samples as signal. It is defined as:

$$FPR = \frac{FP}{TN + FP}, \tag{13}$$

where FP refers to the number of false positives and TN the number of true negatives. When evaluating these quantities over all possible thresholds, their relationship is summarized by the receiver operating characteristic (ROC) curve. The ROC curve provides a global summary of the model's power in separating background and signal samples.

Using the integral of the ROC curve, we can assign a single score to a particular model denoted as the area under the curve (AUC). A model that is able to perfectly identify all signal samples with no false positives will exhibit an AUC of 0. In contrast, an AUC of 0.5 shows that the model is essentially random guessing and has no discriminative power. When discussing results of Publication III, the area over the curve (AOC) is also used, which is essentially equivalent to AUC through $AOC = 1 - AUC$.

In the case of unsupervised approaches (such as the AE in Publication I), the model is not optimized on a classification score directly. Rather, a reconstruction loss is minimized during training, and samples with different loss distributions are considered as anomalies. Although no labels are used in this training scenario, the reconstruction loss can still be used to construct ROC curves, reflecting the separability of designated background and signal distributions.

5.4 Effects of subhalos in MW-like simulations (Pub. I)

5.4.1 Motivation and scientific context

The perturbations induced by moving dark subhalos in the stellar phase-space are expected to be very subtle. It is clear that all such detection efforts depend on the availability of precise data across large volumes. After all, subhalos are expected to have arbitrary orbits and velocities and thus could be anywhere in the Galaxy. In addition to this, even if one knew the exact position of a particular dark subhalo, the magnitude of its gravitational effects could well be below the error limit of the telescope.

Fortunately, with the advent of Gaia, stellar measurements are becoming increasingly abundant and reliable, motivating data-driven searches of DM substructure in the Galaxy. However, due to the invisible nature of dark subhalos (no bound stars), the interpretation of likely DM signals from low-mass subhalos is hindered significantly as it is difficult to separate perturbations resulting from subhalos from other baryonic structures (e.g., spiral arms, giant molecular clouds, moving groups, etc.).

In this regard, realistic MW-like simulations (e.g., Latte) offer a controlled environment to test potential detection strategies. Since the real values of all star and DM particles are known precisely and at all times, this enables two capabilities. First, the information can be used to train ML models to learn DM subhalo-induced signals from the data. Second, since the positions of subhalos are known, the results can be validated in a relatively straightforward manner.

With the above in mind, the aim of this study was to inspect whether it is possible to detect these phase-space perturbations on a statistical basis in state-of-the-art galaxy simulations using DL techniques. The main body of this analysis was divided into two main parts. First, sensitivity to perturbations was studied in the Latte galaxies, representing an ideal setting, that is, without any observational effects. After this, we turned our focus to a more difficult task and studied how well we are able to infer the presence of subhalos in Gaia-like mock catalogs. Since the latter includes expected observational effects (e.g., measurement uncertainties, missing radial velocities, stellar extinction, observational reference frames), we could then compare the detection performance between the perfect, idealized situation with a more realistic expectation.

5.4.2 Subhalo identification and dataset preparation

In this study, we used the galaxies `m12f`, `m12i`, and `m12m` from the Latte suite of MW analogues and their Gaia DR2 mock catalogues, introduced in Section 3.2.1 and Section 3.2.2 of this thesis, respectively. Inside each galaxy, DM subhalos were identified with the Amiga Halo Finder (AHF) code [119], which was run only on DM particles. Subsequently, subhalos that were included in the analysis were required to satisfy the following mass condition:

$$3 \times 10^6 < M_{sub}/M_\odot < 4 \times 10^8. \tag{14}$$

This lower bound of this range is motivated by two key considerations. First, perturbations of subhalos with masses lower than $3 \times 10^6\,\mathrm{M}_\odot$ result in velocity changes that are on the order of $10^{-3}\,\mathrm{km/s}$ [120]. Secondly, according to [49], subhalos below this bound are not reliably resolved in the simulations [49]. In a study such as this, where we analyse the stellar content of galaxies on a star-by-star basis, the effects from subhalos lighter than our adopted limit would most likely not change the final result.

The higher bound of included subhalo masses is enforced by restraining our analysis to within 100 kpc from the GCs. This minimizes the risk of involving stars associated with dwarf galaxies, the detection of which was not within the scope of this work. Furthermore, we excluded any candidate dwarf galaxy halos whose associated star velocities are below the escape velocity.

After identifying the subhalos, the stars from both the idealised Latte and Ananke framework were divided into background and signal stars based on the distance to the respective subhalos. This was done by computing the Euclidean distance of each star particle to the nearest subhalo, where stars closer than 1 kpc were considered halo-associated and the rest background.

Specifically for the Gaia DR2-like mock catalogues, we removed the stellar disc of all galaxies by imposing a cut on the vertical coordinate and requiring that $|z| > 5$ kpc. While it is certainly true that this cut omits perturbations from baryonic components (e.g., spiral arms) from the analysis, the real reason was due to the excessive data volume within these datasets (see Section 3.2.2). In the end, the combined number of stars across all galaxies was on the order of 10^7 and 10^9 for the Latte and Ananke datasets, respectively.

5.4.3 ML methodology

In order to study the gravitational effects of dark subhalos identified in the Latte galaxies, we used the 6D phase-space parameters of the signal and background stars. The feature vector of a particular star is then defined as $\mathbf{X} \in \mathbb{R}^6$ consisting of their galactocentric coordinates (x, y, z) and the velocities in these respective directions (v_x, v_y, v_z).

Specifically, we adopted an anomaly detection approach by implementing a simple AE network, which learns the phase-space distribution of background samples via a lower-dimensional manifold. This can be done by designing a neural network consisting of two parts: an encoder and a decoder.

$$\begin{aligned} E(\mathbf{X}) &\rightarrow \mathbf{z} \in \mathbb{R}^D \\ D(\mathbf{z}) &\rightarrow \mathbf{X}' \in \mathbb{R}^6 \end{aligned} \tag{15}$$

In Eq. 15, $E(\mathbf{X})$ is the encoder, which embeds the input features of training samples to a latent space with lower dimensionality. The decoder $D(\mathbf{z})$ then reconstructs the data back to original dimensionality such that $D(E(\mathbf{X})) \rightarrow \mathbf{X}'$. Then, by defining the loss function between the original $\mathbf{X}$ and reconstructed feature vector $\mathbf{X}'$ as

$$L_b(\mathbf{X}_i) = \|\mathbf{X}_i - D(E(\mathbf{X}_i))\|\,, \tag{16}$$

and training the AE only on background stars, we can use the reconstruction loss $L_b(\mathbf{X}_i)$ as an empirical discriminator between background ($\mathbf{X}_{bkg}$) and signal ($\mathbf{X}_{sig}$) stars. Simply put, as the model learns to reconstruct the phase-space parameters of background stars, it is expected to have a harder time reconstructing signal stars as they are not involved during the training. In an ideal scenario, the distribution of $L_b(\mathbf{X}_{sig})$ is perfectly separable from $L_b(\mathbf{X}_{bkg})$.

Figure 19 depicts the AE architecture adopted in this study. Notably, both the encoder and decoder were implemented as feedforward networks containing two hidden layers with 128 neurons per layer. Through experimentation, we found that a latent space of $D = 3$ results in the best model performance. For each hidden layer, we adopted the scaled exponential linear unit (SELU) activation function [121].

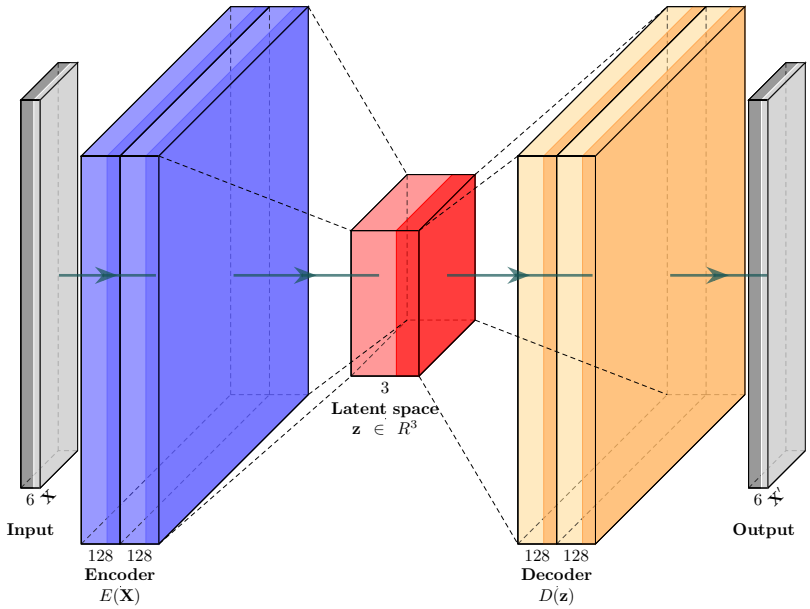


Figure 19: Architecture of the AE implemented in Publication III. The first layer expects a 6D input vector representing stellar phase-space features, which is then processed by an encoder ($E(\mathbf{X})$) with two hidden layers of 128 units each. The input is then compressed into a 3D latent space. The decoder ($D(\mathbf{z})$) mirrors the encoder, expanding the latent vector back to the previous dimensionality. Finally, the output layer reconstructs the original input, which is then used to compute the reconstruction loss.

With the purpose of cross-checking the performance of the AE, we also implemented a binary classifier which was applied to the same set of features. In contrast to the anomaly detection approach, this model used the signal and background labels of stars explicitly.

Similar ML models were used when analyzing the synthetic Gaia datasets, with minor changes made to the models resulting from the different training dataset formats between the two scenarios. For one, the synthetic Gaia stars were divided into patches with Healpy [122] in order to transform the dataset into a more manageable format. Further-

more, the shape of the synthetic Gaia dataset was different, consisting of the astrometric and spectroscopic observables that are usually provided in Gaia data: the parallax, right ascension, and declination with respect to the pixel center, proper motions in the right ascension and declination, and radial velocities. With each of these observables being also accompanied by its measurement uncertainty, it raised the total number of features per star particle to 12 (in contrast to 6 in Latte).

In order to avoid data leakage and maximize generalization to datasets unseen during training, we separated the galaxies by training on `m12m`, validating on `m12i`, and using `m12f` solely for testing in both Latte and synthetic Gaia scenarios. In the synthetic surveys, we therefore assigned all three LSR realizations of `m12m` to the training set, those of `m12i` to the validation set, and `m12f` LSRs to the testing set.

The ML models outlined above form the basis of the dark subhalo detection experiments conducted within the Latte MW analogues. The performance of the supervised and unsupervised approaches from this work is presented and discussed in Section 5.6.1.

5.5 Detection of stellar wakes (Pub. III)

5.5.1 Stellar wake phenomena

In addition to the dark subhalo detection methods introduced in Section 5.2 and the one outlined in the previous section, another promising avenue to look for low-mass subhalos is stellar wakes. These are localized phase-space perturbations imprinted in stellar populations from the passage of a heavy object. This phenomenon is closely related to dynamical friction [123], which is induced when a massive perturber moves through a homogeneous density field. As its gravity accelerates matter toward it, a region of enhanced density forms along the trajectory of the perturber, causing it to lose momentum over time. This effect has played a particularly significant role in studies of infalling satellites [124]. Recently, a popular testbed for stellar wake studies has been MW's largest satellite, the Large Magellanic Cloud (LMC) (e.g., [125, 126, 127, 128]). While studies such as [127] have modeled and visualized the LMC wake in N-body simulations, the first observation of a wake behind the LMC was described in [125].

Another important study in the context of stellar wakes has been [129], where the authors developed an analytic formalism to model perturbations in the stellar phase-space distribution. In contrast to studies looking at large MW satellites (e.g., LMC), this particular framework was aimed at detecting DM substructure in the form of low-mass subhalos.

Inspired by the works in both [127] and [129], the aim of the current study was to simulate stellar wakes induced by dark subhalos in a sub-galactic setting. In contrast to what was done in [127], we simulated wakes from low-mass subhalos with no gravitationally bound stellar counterpart. Furthermore, instead of an analytic approach (such as the one in [129]), we focused on the detection of dark subhalos from an ML perspective. At the time of writing, and to the best of our knowledge, a study regarding the detection of stellar wakes using DL methods has never been carried out before, making Publication III a first of its kind.

5.5.2 Wind tunnel simulations

Unlike those utilized in the study of Publication I, the simulations in Publication III were performed from the ground up. This allowed us to isolate and experiment with the subhalo-induced signal in an idealized and controlled setting. Along with the ML analysis, the simulations themselves constituted both a major part and result of this study, with their execution requiring careful planning and substantial computational power. Furthermore, considerable time was spent designing the setup and confirming the validity of the simu-

lation outputs.

Using the N-body gravity code PKDGRAV3 [44, 130], we simulated a periodic box with a side length of $L = 120\,\mathrm{kpc}$ which contained a moving perturber along with background DM and star particles. Although initial experimentation was done using a point mass, in the end, an extended potential was adopted to model the perturber. In the interest of comparing our results with those from [129], the subhalo was given a Plummer density profile as given by

$$\rho(r) = \frac{3\mathrm{M_{sh}}}{4\pi {R_s}^3}(1 + \frac{r^2}{{R_s}^2})^{-5/2}, \tag{17}$$

where r is the distance from the center of the halo, R_s the scale radius, and $\mathrm{M_{sh}}$ the total mass of the subhalo. The scale radius describes the characteristic length scale of the halo density and was computed as in [129, 131]

$$R_s = 1.62\,\mathrm{kpc} \times \left(\frac{\mathrm{M_{sh}}}{10^8\,\mathrm{M_\odot}}\right)^{1/2}. \tag{18}$$

In Figure 20, the scale radius for a subhalo of mass $M_{sh} = 5 \times 10^8\,\mathrm{M_\odot}$ is depicted as the small circle in the middle of the box.

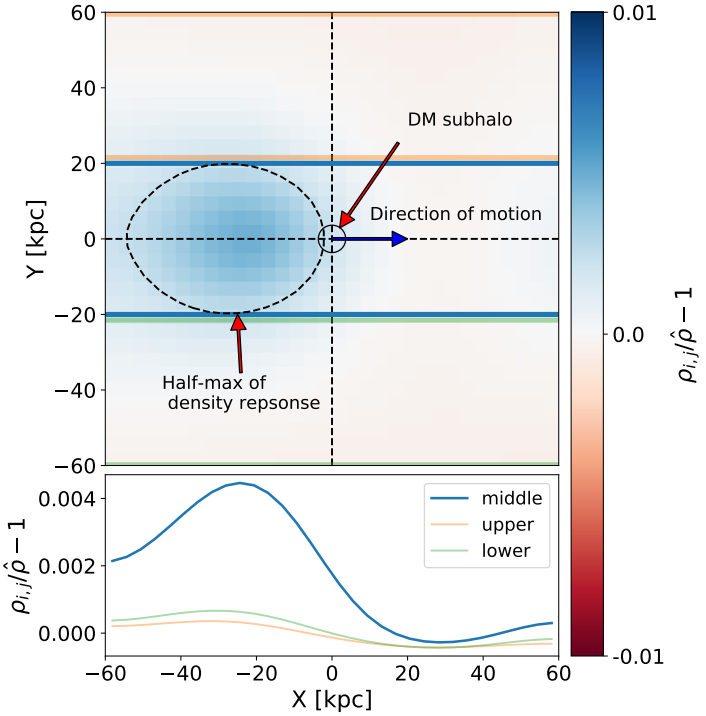


Figure 20: A depiction of the simulation setup in Publication III. Upper panel: The dark subhalo, situated in the middle of the box, is moving in the +X direction. The stellar wake is seen behind the direction of movement as the overdensity, whose half-max response is enclosed in the dashed ellipse. Lower panel: The radial density profile (along X) of the upper, middle, and lower regions defined in the Y-coordinate, corresponding to the orange, blue, and green lines, respectively.

So as to avoid unnecessary work which would result from simulating a moving ex-

tended potential, it was instead fixed at the box coordinates $X,Y,Z = (0,0,0)$, with background particles given a corresponding velocity boost in the -X direction to mimic the halo movement. While this means that the simulation actually takes place in the reference frame of the perturber, which experiences a wind[15] of stellar and DM particles, it is tantamount to a box-frame setting where the perturber is moving in the +X direction (see Fig. 20). The magnitude of the velocity boost was derived from both considerations of the circular velocity given the MW mass enclosed in a sphere with $R < 30\,\mathrm{kpc}$, and from what is reported in FIRE-2 simulations for subhalos with $> 10^7\,\mathrm{M}_\odot$.

The background particles were given velocities and mass values to mimic the conditions in the MW stellar halo at a specific galactocentric distance. We focused our detection efforts on the galactic stellar halo. This was in part due to the stellar halo being significantly easier to simulate than other components of the Galaxy. In any case, subhalo-induced signals in the stellar halo are expected to be more prominent than in e.g., the disc, which contains a myriad of baryonic structures, the effects of which would be challenging to differentiate from subhalos.

In the primary analysis of this study, we generated background initial conditions that mimic the stellar halo at a galactocentric distance of 30 kpc. In order to simulate the mass densities expected at this distance, DM and star particles were assigned an appropriate mass of $M_{\mathrm{DM}} \approx 1.3 \times 10^4\,\mathrm{M}_\odot$ and $M_{\mathrm{star}} \approx 1.3\,\mathrm{M}_\odot$, respectively. Later on, to test the robustness of the trained model, we also simulated the stellar halo at 50 kpc from the GC and modified the background parameters accordingly. Table 3 and Table 4 summarize important kinematic and mass parameters that were used to mimic the stellar halo in both of these scenarios.

Table 3: Velocity parameters of the subhalo and background particles adopted in the wind tunnel N-body simulations for two selected galactocentric distances in the stellar halo.

r [kpc]	σ_{DM} [km/s]	σ_{star} [km/s]	V_{sh} [km/s]
30	200	95	225
50	180	90	200

Table 4: Mass parameters of background DM and star particles adopted in the wind tunnel N-body simulations for two selected galactocentric distances in the stellar halo.

r [kpc]	ρ_{DM} $[\mathrm{M}_\odot/\mathrm{kpc}^3]$	N_{DM}	ρ_{star} $[\mathrm{M}_\odot/\mathrm{kpc}^3]$	N_{star}
30	10^6	512^3	10^2	512^3
50	$10^{5.5}$	512^3	10	512^3

5.5.3 Dataset generation

We simulated subhalos with three different masses: $5 \times 10^7\,\mathrm{M}_\odot$, $10^8\,\mathrm{M}_\odot$, $5 \times 10^8\,\mathrm{M}_\odot$. This mass range is consistent with that considered in the analysis described in Section 5.4, and similar reasoning was used. The detectability of subhalos with lower masses is not motivated by the current precision of observations, while more massive subhalos approach

[15] Hence why they are 'wind tunnel' simulations.

the regime of dwarf galaxies and were therefore outside the scope of this work. In this particular study, the lower mass threshold was increased to $5 \times 10^7\,\mathrm{M}_\odot$ after comparing the strength of the density response to the expected noise in our simulations. We concluded that the detection of subhalos lighter than this mass is unrealistic given our current setup.

From each simulation snapshot, 100 data samples were generated. This was done by sampling 1% of stars without replacement, corresponding to about 1.3M stars per sample. On the one hand, this procedure was intended as a heuristic to mimic the limited availability of observations in the stellar halo.[16] On the other hand, this allowed us to generate more training samples, which we were in a definite lack of due to the considerable computational challenge from running the simulations and generating datasets.

After sampling, star particles were binned along X- and Y-coordinates into image-like data samples in the form of 2D histograms with 32 pixels (bins) per side and 4 channels per sample. The channels referred to adopted training features describing the overdensity ($\bar{\rho}$), mean velocity on the X-Y plane ($\overline{V_{xy}}$), velocity dispersion (σ_{xy}), and velocity field divergence ($\nabla\vec{V}_{xy}$) in the box.

From considerations of dynamical friction, overdensity was the most obvious observable with which we expected to see a collective response in the background stellar distribution. The overdensity inside each bin was computed with the following equation:

$$\bar{\rho_{i,j}} = \frac{\rho_{i,j}}{\hat{\rho}} - 1, \tag{19}$$

where $\rho_{i,j}$ is the stellar mass density in a particular bin on the X-Y plane and $\hat{\rho}$ is the average mass density in the simulation box.

In the study of [127], it is also shown that the wake response to the perturber is seen not only in spatial but also in kinematic features. This is why, in addition to overdensity, we also implemented the selection of kinematic features in order to both maximize model performance and also to study the relative effectiveness between all adopted features.

The simulation box we implemented as part of this study was big (120 kpc along each axis), which is why we also split it into 3 different layers to inspect sensitivity in simulated regions not containing the perturber. These were the upper layer ($Y \in [20, 60]\,\mathrm{kpc}$) and the lower layer ($Y \in [-60, -20]\,\mathrm{kpc}$). This resulted in a dataset dimensionality of (N, 32, 32, 12), where N is the total number of samples, 32 the number of bins per axis[17], and 12 containing the training features (4 per layer).

In total, 48 unique random seeds were used to generate stellar wake simulations for each target mass and background scenario, resulting in 192 final simulations. The process was computationally expensive, with each simulation taking approximately 1.5 hours to complete. Afterwards, all simulation snapshots were post-processed and converted into an appropriate shape and format for training purposes, resulting in a total of 19200 samples across all target mass scenarios. The final ML dataset, containing stellar wake examples from three distinct subhalo masses as well as a background case (no subhalo), was compiled and uploaded to Zenodo.[18]

[16] Every snapshot contained 512^3 star particles, which, in the case of the stellar halo, is an amount one would not expect from current stellar surveys.

[17] We also experimented with different binning schemes (e.g., 16 or 64 bins), but noticed no significant impact on final performance.

[18] https://zenodo.org/records/12721089

5.5.4 ML methodology

In Publication III, the detection performance of stellar wakes was explored as a two-pronged classification task using physics-informed 2D histograms. The first part of this was formulated as a binary classification problem by asking whether there is a subhalo in the image. The second part was treated as a multiclass classification, focused on predicting the mass of the subhalo in the image.

In both ML scenarios, the entire simulated dataset was divided among training, validation, and testing using a 50%, 33%, and 17% split. In the case of the binary classifier, this corresponded to 4800, 3200, and 1600 samples, whereas for the multiclass classifier, this corresponded to 7200, 4800, and 2400 samples, respectively. The total number of samples included during training is greater for the latter because training was done on all signal targets simultaneously, with the background being excluded. In contrast, the binary classifier was trained by using all samples of a particular target mass case and background (no subhalo) samples.

Training and evaluation were repeated 30 times to capture the variance in model performance. Instead of a fixed number of epochs, we used early stopping with a patience of 5 epochs to automatically halt training when the validation loss stopped improving. At the start of each run, data samples derived from simulation snapshots were assigned to training, validation, and testing by considering a random permutation of the initial conditions' seed numbers used in simulations. During this procedure, care was taken to avoid data leakage and not to include samples originating from the same simulation seed in multiple ML datasets at the same time. That is, every time a model was trained, the following relation was expected to hold $k_{train} \cap l_{val} \cap m_{test} = \emptyset$, where k, l, and m refer to simulation seed numbers.

The exact network architecture was based on hypertuning by leveraging the RandomSearch algorithm of KerasTuner [118]. In hypertuning, different hyperparameters are defined as variables in a specific range. The RandomSearch algorithm then iteratively samples random values for the model architecture and evaluates it after training. The advantage of this is that it is automated, thus enabling the exploration of a wider range of all possible settings by using the final validation loss to find the best model. In comparison to doing this hyperparameter scan manually, it is also considerably faster.

Since the basis of our ML analysis consisted of image datasets, instead of a regular feedforward neural network, we opted for a model architecture inspired by the CNN. The reason for this is that standard fully connected networks exhibit a scaling issue when dealing with image-like datasets, as the number of model parameters required to propagate the data is already very large at the input layer, becoming increasingly so through the hidden layers [105]. Given the small size of our ML dataset (19 200 samples), we further opted for Harmonic layers [132] in lieu of conventional convolutional layers. Using the discrete discrete cosine transform (DCT), these layers transport the problem from the spatial to the frequency domain. Instead of learning filters to extract spatial correlation, the model then learns the individual weights for the DCT filters. For low volume ML datasets, the Harmonic layer is reported to perform better than the traditional CNN [133]. We experimented with both types of layers and found this to be the case.

5.6 Results and discussion

In this section, results from the ML-based subhalo detection methods presented in Publications I & III are summarized. Additionally, we discuss potential improvements to these methods and directions for future studies.

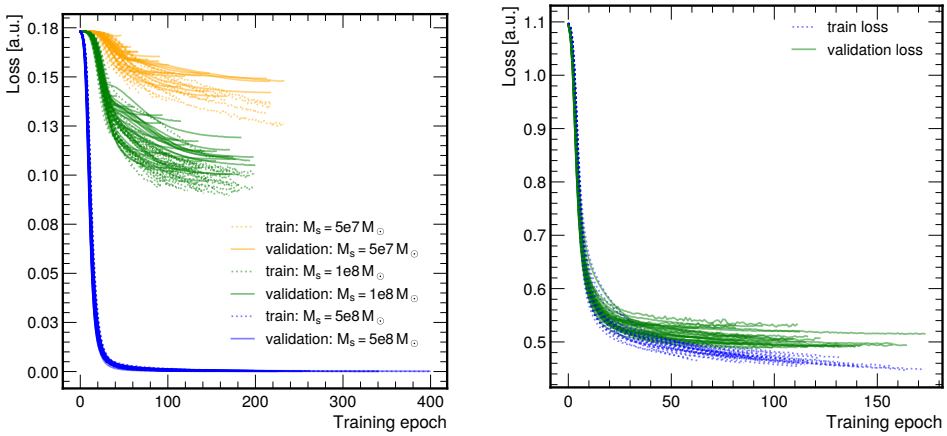


Figure 21: Training (dotted) and validation (solid) loss with respect to training epoch for both the binary classifier (left) and multiple mass hypothesis (right) scenarios, as shown in Publication III. Each model is trained and evaluated 30 times, resulting in an ensemble of loss curves for a particular mass test scenario. Left: Yellow, green, and blue lines show the loss for the three different mass scenarios $5 \times 10^7\,\mathrm{M}_\odot$, $10^8\,\mathrm{M}_\odot$, $5 \times 10^8\,\mathrm{M}_\odot$, *respectively. Right: In the multiple mass hypothesis case, training is done on all mass target cases simultaneously, with the blue dotted line showing the training and green solid line the validation loss.*

5.6.1 Subhalo perturbations in MW-like simulations

In Publication I, we implemented an AE neural network to detect phase-space anomalies in MW-like galaxies on a statistical basis. Specifically, in this setup, the reconstruction loss ($L_b(\mathbf{X}_i)$) between input $\mathbf{X}$ and reconstructed features $\mathbf{X}'$ was used as an empirical discriminator between signal and background star samples. Figure 22 (left panel) shows $L_b(\mathbf{X}_i)$ distributions for both subhalo-associated (signal) and background star particles in the `m12f` galaxy. It shows that star particles that are near subhalos exhibit, on average, a larger reconstruction loss than the background stars on which the model was trained, and thus the two populations are separable according to this metric. In order to check the robustness of this result, we compared performance when the model is trained in a setting where signal stars are chosen randomly from the dataset (right panel in Fig. 22). As expected, adopting a fake signal resulted in $L_b(\mathbf{X}_i)$ distributions that are indistinguishable from each other.

The overall performance of the anomaly detection approach in Latte galaxies is shown in Fig. 23. It summarizes the model sensitivity at all possible thresholds (See Section 5.3.5) in the form of ROC curves. With an AUC of 0.07 (AOC=0.93), we saw that in the idealized Latte scenario, the AE exhibits exceptional discriminative and generalization performance, being able to detect subhalo-associated star particles never used during training. In contrast, the supervised binary classifier model, which was trained explicitly on signal samples, showed performance that is trivial and not significantly better than random choice (AUC=0.48). Due to the strong class imbalance present in these datasets, this result was not entirely unexpected.

The unsupervised and supervised detection methods (described in Section 5.4) were also employed on the synthetic Gaia DR2 survey data. Unlike in the idealized Latte setting, these datasets include Gaia-like observational phase-space parameters and effects (see also Section 3.2.2). As before, the performance of both models in the mock survey scenario is summarized via ROC curves in Fig. 24. Since authors in [54] also report

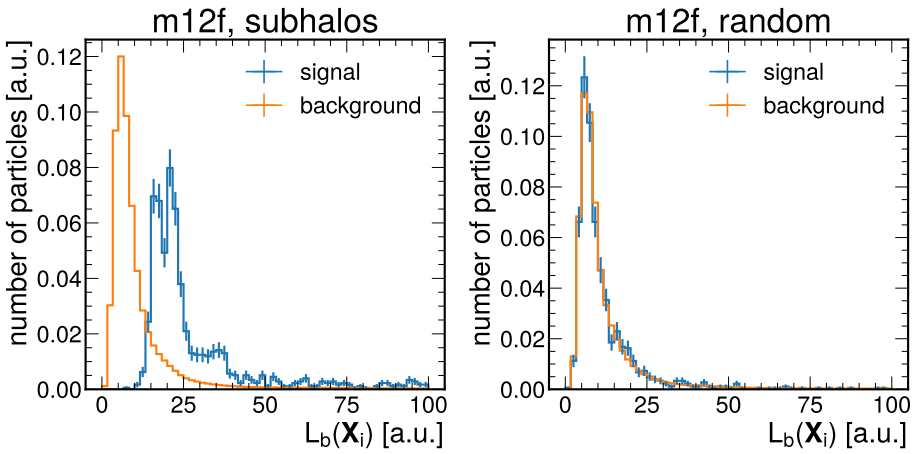


Figure 22: Reconstruction loss (L_b) distributions for signal (blue) and background (orange) stars in the `m12f` galaxy, as shown in Publication I. As a sanity check, the test dataset results, trained on the real signal (left), were compared with loss distributions from a model optimized on a fake signal (right).

underlying true values for mock observations, we were able to study how observational effects impact our final performance. For both ML scenarios, performance is depicted as three separate curves after having adopted either true values (thick solid line), error-convolved values (thin solid line), or error-convolved values with measurement uncertainties (dashed line).

In the synthetic Gaia scenario, we saw a significant reduction in detection performance for the anomaly detection approach (AUC=0.49), whereas the binary classification approach outperformed the AE with an AUC of 0.37. For both ML models, performance was seen to be increasing when using the true values of the mock observations, highlighted by performance differences from baseline of $\approx 32\%$ and $\approx 17\%$, respectively. Despite using observational values that are known exactly, results did not improve to the level seen in the idealized Latte scenario.

5.6.2 Detection performance of stellar wakes

The methodology in Publication I was aimed at detecting stellar phase-space perturbations on a star-by-star basis given a population of subhalos in the simulated galaxies. In contrast, Publication III represents a shift to a study of signals induced by individual dark subhalos based on characteristic stellar wake patterns found in groups of stars.

In this study, the ML datasets consisted of image-like samples encoding both spatial and kinematic features, computed to capture the stellar wake signal within the simulated stellar phase-space (see Section 5.5.3). As it was not known *a priori* which of the considered features would be most effective, we performed a series of experiments to determine their individual constraining power using the largest subhalo mass considered in this study – $5 \times 10^8\,\mathrm{M}_\odot$. In addition to this, we also inspected model performance when Gaussian smoothing is applied to the features before training. The results of this experimentation are summarized in Fig. 25, which displays the performance of the binary classifier trained on each of the four features separately in both smoothed and non-smoothed scenarios. Gaussian smoothing with a kernel size of $\sigma = 3$ substantially improved the performance for each feature. Among individual features, the most to least effective ranking

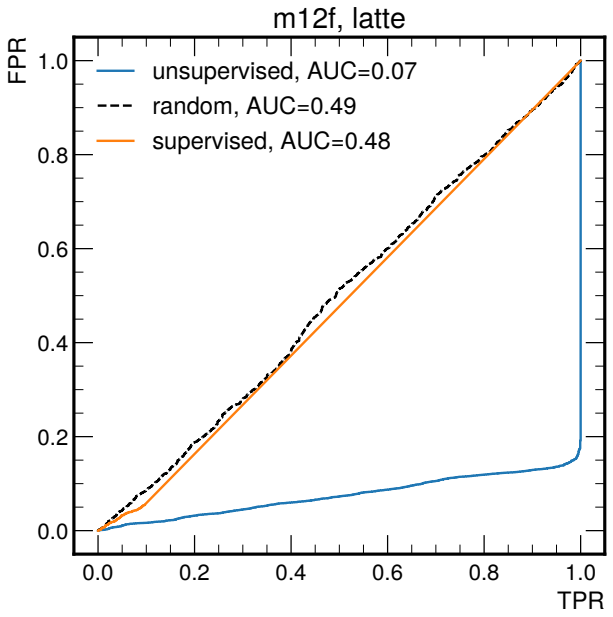


Figure 23: Supervised (orange) and unsupervised (blue) approach performance for the `m12f` galaxy in ideal conditions, as shown in Publication I. While the supervised model is comparable to random selection (dashed line), the unsupervised model showed excellent performance with an AUC of 0.07 (or AOC = 0.93).

was as follows: overdensity and velocity field divergence ($\langle \mathrm{AOC} \rangle = 0.99$), mean velocity ($\langle \mathrm{AOC} \rangle = 0.95$), and velocity dispersion ($\langle \mathrm{AOC} \rangle = 0.83$).

In addition to individual feature performance, we also experimented with different feature combinations. Interestingly, adding mean velocity and its dispersion did not improve model performance beyond what was already achieved using overdensity and velocity divergence.

For assessing the detection performance of dark subhalos with different masses, we adopted the feature combination of $\bar{\rho}$ and $\nabla \vec{V}_{xy}$ for both the binary classification and multiple mass hypothesis scenarios. Figure 26 summarizes the results of the former where the different ROC curves show the performance for subhalo masses of $5 \times 10^7\,\mathrm{M}_\odot$ (yellow), $10^8\,\mathrm{M}_\odot$ (green), and $5 \times 10^8\,\mathrm{M}_\odot$ (blue). The width of each ROC curve displays its standard deviation across 30 runs (see also Section 5.5.4). For subhalos with a mass of $5 \times 10^8\,\mathrm{M}_\odot$, we saw near-perfect classification with a TPR and FPR of 99% and 1%, respectively. For the intermediate ($\langle AOC \rangle = 0.77$) and lowest mass ($\langle AOC \rangle = 0.53$) targets, we saw a reduction in model performance, achieving a (TPR, FPR) of (74%, 35%) and (60%, 41%), respectively.

Similar results were seen in the multiple mass hypothesis scenario, the results of which are summarized by the confusion matrix in Fig. 27. In contrast to the binary classification task, the model here was trained only on the signal samples. Instead of a single probability, three scores were predicted for each sample, representing probabilities of belonging to a particular mass bin. Again, the model had very little difficulty recognizing samples containing the very massive subhalo $5 \times 10^8\,\mathrm{M}_\odot$, exhibiting trivial scatter in its prediction ($\sigma \approx 10$). For low and intermediate mass targets, the model showed less constraining power and tended to confuse between predicting $10^8\,\mathrm{M}_\odot$ and $5 \times 10^7\,\mathrm{M}_\odot$ subhalos. While

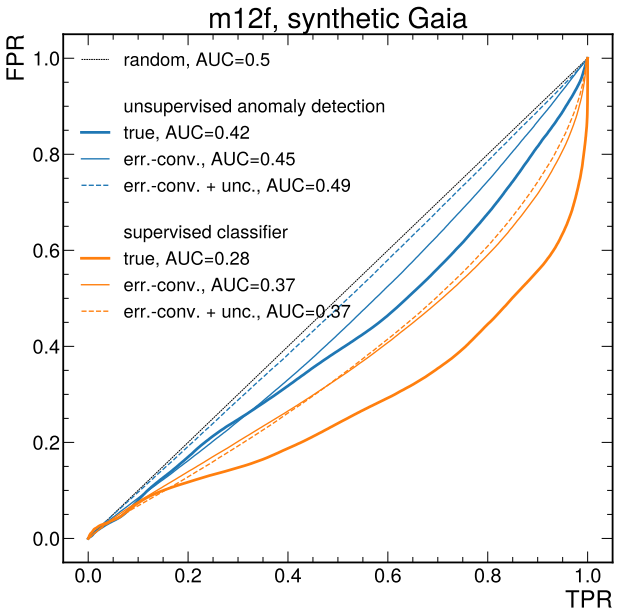


Figure 24: Synthetic Gaia dataset performance of supervised (orange) and unsupervised (blue) models as shown in Publication I. Solid, thick lines depict results when training and evaluation are done on true mock observations. Solid thin lines reflect the performance of either model when adopting stellar phase-space observables, which are convolved with a simple Gaia DR2-like error model. Dashed lines show the same as the previous, with observational uncertainties also considered during training and evaluation.

this is evident from larger numbers in the off-diagonal elements, it is also seen from the significantly larger scatter in both off- and on-diagonal components ($\sigma \approx 45$).

5.6.3 Limitations & future outlook

In Sections 5.6.1 and 5.6.2, main results from the ML-based dark subhalo detection studies were presented. Despite promising results, several limitations currently constrain the generalizability and sensitivity of these methods. In the following, key caveats of Publications I & III are summarized, and directions are proposed for future improvements.

In Publication I, we excluded the disk from the synthetic Gaia catalog due to computational constraints. Including disk stars in future analyses may, on the one hand, introduce additional noise due to baryonic substructure (e.g., spiral arms, bar), but, on the other hand, it also provides access to better stellar statistics and an opportunity to test the robustness of the detection method in a more complex environment. With improved computational resources and optimized data handling, the inclusion of the disk in this analysis could become tractable.

A dominant factor limiting the anomaly detection sensitivity in synthetic surveys was concluded to be the smearing introduced during the synthetic star sampling process in Ananke. As synthetic stars are generated from a 1D kernel, a smearing scale on the order of 0.7 kpc is introduced. This effect is also present in the velocity distributions of generated stars with a smearing which is approximately 10 km/s. As it stands, the magnitude of this smoothing in stellar kinematics is large enough to wash out phase-space substructure, especially at the scales relevant to low-mass subhalo ($\lesssim 10^8\,\mathrm{M}_\odot$) detection.

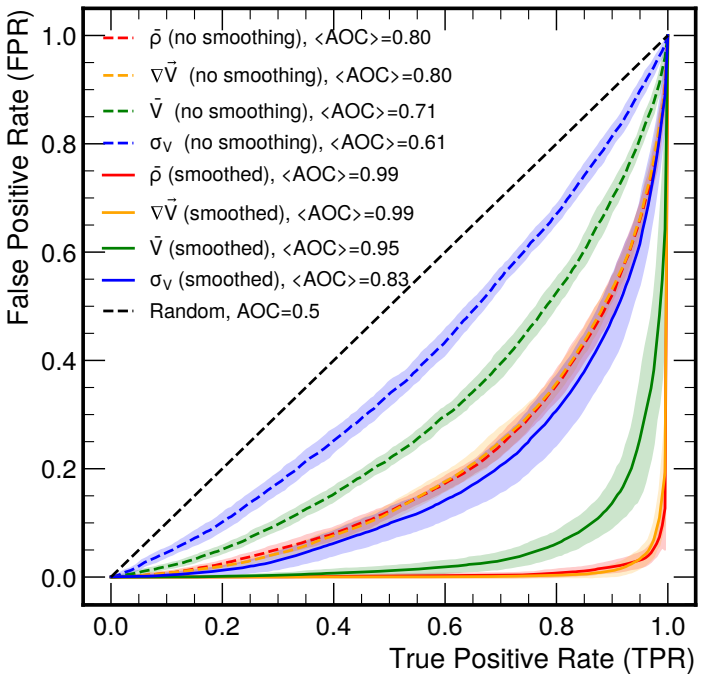


Figure 25: Binary classifier performance for all training features as adopted and shown in Publication III. Each curve represents 30 training and model prediction runs for dark subhalos with a mass of $5 \times 10^8\,\mathrm{M}_\odot$. The dashed curve shows performance for overdensity (red), velocity divergence in the X-Y plane (yellow), mean velocity in the X-Y plane (green), velocity dispersion (blue), when no Gaussian smoothing is applied prior to training. Solid lines show the same, but with smoothing applied to features before training. We saw significant improvement in model performance in all analyzed features compared to their non-smoothed counterparts.

Lastly, the labeling of signal stars was based on spatial proximity (< 1 kpc) to respective subhalos. As we saw from our study of stellar wakes, the perturbations from subhalos can potentially reach stars at much greater distances than originally considered in this study. Furthermore, it would also be worthwhile to consider additional signal criteria based on kinematics (e.g., velocity perturbations) instead of spatial proximity alone. More physically motivated labeling could boost training statistics in the synthetic catalogs and therefore improve model performance.

The work in Publication III was constrained by the limited availability of wake simulations due to the high computational cost of generating training data. As such, only a handful of subhalo masses and configurations were explored. From dedicated ablation studies, where we systematically modulated the amount of training data, we confirmed that the results improve with increased data availability, and no performance plateau with our current dataset can be observed. Given a larger training dataset, it would also allow us to explore additional ML models with objectives beyond classification (e.g., regression).

In terms of the stellar wakes themselves, several theoretical questions remain. For instance, we observed the spatial extent of the stellar wakes to be far larger than expected from analytic estimates (e.g., [129]). While it has been shown by [127] that the inclusion of self-gravity between simulated particles significantly (order of 10%) enhances the wake response, our experiments have shown that this effect alone cannot account for the enor-

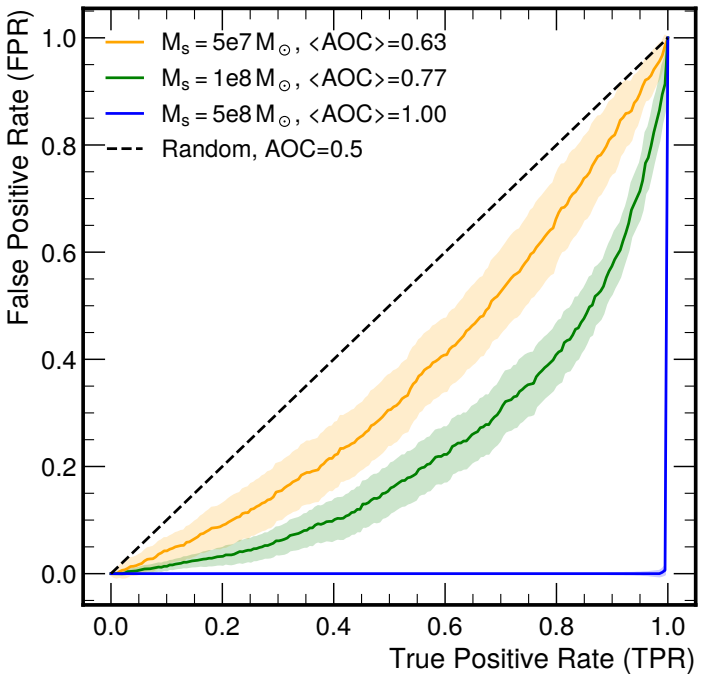


Figure 26: Binary classifier performance for subhalos of masses $5 \times 10^7\,\mathrm{M}_\odot$, $10^8\,\mathrm{M}_\odot$, $5 \times 10^8\,\mathrm{M}_\odot$ shown as the yellow, green, and blue bands, respectively. The width of each band depicts its standard deviation, which was computed after training and evaluating each mass target case scenario 30 times.

mity of the response. In addition to self-gravity, another major factor that could contribute to this is the integration time of the simulation. To a lesser extent, the choice of subhalo potential (e.g., Einasto, NFW, Plummer), ambient density, and velocity dispersion formed by background particles (stellar and DM), also modifies the characteristics of the wake, though none individual appear sufficient to explain the discrepancy. In reality, the unexpectedly large spatial extent of the wakes could stem from a combination of these effects or from other factors yet to be considered. Since the root cause remains unclear, a dedicated study could be useful to disentangle their individual contributions and clarify the origin of this discrepancy.

5.6.4 Towards observations

Both the anomaly and wake detection methods stand to benefit from next-generation surveys like LSST and Euclid. These will not only increase the number of observable halo stars but also provide more precise phase-space measurements and greater radial velocity coverage. Preparing methods for batch analyses of such data streams is an important next step. However, applying ML methods to real observations to constrain the SHMF is a challenging task, which calls for careful planning both in terms of the available training data and the broader methodological approach.

The synthetic Gaia DR2-like surveys used in Publication I rely on the Ananke framework [54], which adopts a simple error model specific to Gaia DR2. These mock catalogs omit stars that fall outside Gaia's magnitude limit of $\approx 20.7\,\mathrm{mag}$. Since the completion of Publication I, newer synthetic Gaia DR3 surveys [134] have become available within

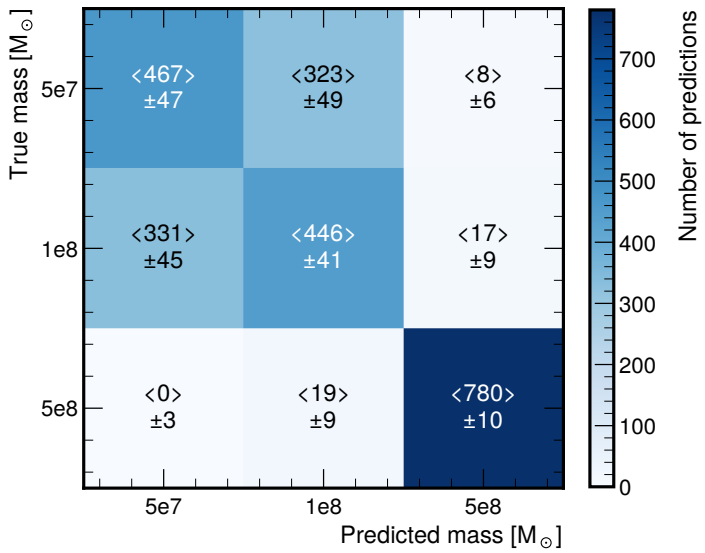


Figure 27: Confusion matrix summarizing the performance from the multiple mass hypothesis approach, as shown in Publication III. Best performance is seen for the heaviest mass test scenario ($5 \times 10^8\,\mathrm{M}_\odot$), while the model had difficulty differentiating between the intermediate ($10^8\,\mathrm{M}_\odot$) and low-mass $5 \times 10^7\,\mathrm{M}_\odot$ test samples.

Ananke. Notably, Gaia DR3 includes approximately 5 times more radial velocity measurements than the previous DRs. Given that the completeness of radial velocities was not the limiting factor of the study, the improvements from additional line-of-sight velocities are expected to be minimal. However, exploring the reduced observational errors remains an open question.

In anticipation of future stellar surveys with much deeper observational depths, an interesting follow-up study to Publication I would be to assess the impact of magnitude cuts and observational depth on subhalo detection performance. For instance, LSST is expected to reach a magnitude depth of $\approx 27.5\,\mathrm{mag}$ [101], which means it will be able to observe faint stars that are inaccessible to Gaia. It will also be able to observe stars with greater Heliocentric distances and will result in a much more detailed view of the MW stellar halo. Since the latter component was the focus of subhalo detection in the synthetic observations, it would be interesting to see how our method would fare in a combined Gaia-LSST observational framework.

In addition, the optimal methodology for detecting subhalo signals remains uncertain. The gravitational imprint left by subhalos is inherently weak and diffuse, particularly in the low-mass regime of interest. Currently, the two ML-based studies introduced in this section are methodologically distinct. One focuses on the statistical detection of individual subhalo-induced perturbations on a star-by-star basis (Publication I) and the other on coherent wake signatures from individual subhalos (Publication III). If the final goal is both detection and mass estimation of dark subhalos, then the approach of Publication I can be regarded as a preliminary tool, rather than a complete solution.

Given their complementary strengths, the two approaches could be integrated into a hybrid method operating across both spatially localized and extended subhalo signals. For instance, one could use anomaly detection to flag potential regions of interest, which would then be followed by wake-based modeling and pattern recognition to verify co-

herent features. Alternatively, the methodology of Publication I could be extended from point-based to group-based anomaly detection, potentially allowing for the identification of spatially correlated stellar wake responses to unseen perturbers.

Summary

The overarching aim of this thesis was divided into four major objectives (see Section 2), all of which were successfully achieved. Collectively, these objejctives tackled the two primary aspects of the MW DM halo - the smooth, virialised component and its substructure. Taken together, the analyses of the smooth halo and its substructure provide complementary insights into the distribution and clustering of DM in the MW. The smooth component constrains the gravitational potential of the Galaxy as a whole, while studies of substructure test predictions of the CDM paradigm on smaller scales.

Addressing the first aim of this thesis, we used Gaia DR3 data and an axisymmetric Jeans model within a Bayesian framework to derive the circular velocity curve of the MW. A key aspect of this study was the self-consistent treatment of various systematic uncertainties in the Sun's Galactocentric distance and parameters reflecting the spatial-kinematic morphology of the tracer sample. This approach enabled the derivation of a robust estimate for the local DM density at the Solar circle, $\rho_{\rm DM}(R_0) = \left(0.41^{+0.10}_{-0.09}\right)\ \mathrm{GeV/cm^3} = \left(0.011^{+0.003}_{-0.002}\right)\ \mathrm{M_\odot/pc^3}$, which is in good agreement with recent literature. Alongside the main scientific results, a significant part of this work involved developing `gaia-tools`, a Python repository that handles reference frame transformations and Gaia measurement error propagation effectively. Notably, the final circular velocity curve fit in this work was made computationally tractable by leveraging GPU acceleration.

The second major goal of this thesis was reached by exploring the detectability of DM substructure within the MW by developing ML-based approaches in both zoom-in hydrodynamical simulations of MW-mass galaxies and idealised N-body wind tunnel experiments. ML methods, together with the era of big astronomical data, are rapidly transforming how we approach data-driven problems in Galactic dynamics and DM inference. This thesis presented two complementary ML pipelines for subhalo detection, which can be considered proof-of-concept frameworks. While their current sensitivity is limited due to different factors (resolution limits, limited signal stars, observational uncertainties, etc.), they lay the groundwork for more sophisticated approaches that can be explored in future studies. Developing and stress-testing these methods in advance of next-generation data streams is essential. Modern galaxy simulations and synthetic stellar surveys provide valuable testbeds for this purpose, allowing us to explore detection limits, study systematic effects, and in this way guide future detection strategies.

As part of the third aim, we employed an AE neural network to study whether orbiting dark subhalos leave a detectable imprint in the phase-space distributions of stellar halo stars in MW-like simulated galaxies. While we found excellent performance when using the raw simulated galaxies, the separability of background and signal stars was significantly hindered in the mock Gaia DR2 surveys derived from them. The reason for this reduction in performance was concluded to be mainly from unphysical numerical effects stemming from the mock data generation procedure, and to a lesser extent, from various Gaia-like observational effects.

Building on the previous work, the fourth aim was realised by implementing an isolated and controlled simulation setup with which we were able to study the effects of individual subhalos on the stellar phase-space in greater detail. We simulated moving dark subhalos in the MW stellar halo with the following masses: $5 \times 10^7\,\mathrm{M_\odot}$, $10^8\,\mathrm{M_\odot}$, $5 \times 10^8\,\mathrm{M_\odot}$. Importantly, this work treated the subhalo-induced signal as a collective effect in populations of stars instead of a star-by-star approach adopted in the previous study. We used a CNN-inspired ML approach to study the detectability of stellar wakes in image-like datasets. Although severely limited by the size of training data ($\approx 10^4$ samples), we found non-trivial sensitivity down to a subhalo mass of $5 \times 10^7\,\mathrm{M_\odot}$ by sampling

just 1% of the stars in the stellar halo. This demonstrates the potential of ML- and stellar wakes-based probes of the SHMF.

The tools and results of this thesis contribute to the broader effort to constrain the nature of DM through its gravitational imprint in our Galaxy. Looking ahead, the approaches outlined in this thesis will benefit significantly from upcoming Gaia releases and other surveys such as LSST, which will, among other advantages, extend the available phase-space information to greater heliocentric distances. With the combination of additional observations, high-fidelity simulations, and next-generation ML methods, the approaches developed in this thesis can be further refined and extended to achieve more sensitive and robust constraints on the distribution of DM in the MW.

List of Figures

List of Tables

References

[1] A. Bazarov, M. Benito, G. Hütsi, R. Kipper, J. Pata, and S. Põder. Sensitivity estimation for dark matter subhalos in synthetic gaia dr2 using deep learning. *Astronomy and Computing*, 41:100667, 2022.

[2] Põder, Sven, Benito, María, Pata, Joosep, Kipper, Rain, Ramler, Heleri, Hütsi, Gert, Kolka, Indrek, and Thomas, Guillaume F. A bayesian estimation of the milky way's circular velocity curve using gaia dr3. *A&A*, 676:A134, 2023.

[3] Põder, Sven, Pata, Joosep, Benito, María, Alonso Asensio, Isaac, and Dalla Vecchia, Claudio. Detection of stellar wakes in the milky way: A deep learning approach. *A&A*, 693:A227, 2025.

[4] F. Zwicky. Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta*, 6:110–127, January 1933.

[5] F. Zwicky. On the Masses of Nebulae and of Clusters of Nebulae. , 86:217, October 1937.

[6] F. D. Kahn and L. Woltjer. Intergalactic Matter and the Galaxy. , 130:705, November 1959.

[7] Thornton Page. Masses of the double galaxies. , 64:53, March 1959.

[8] J. G. de Swart, G. Bertone, and J. van Dongen. How dark matter came to matter. *Nature Astronomy*, 1(3):0059, Mar 2017.

[9] H. I. Ewen and E. M. Purcell. Observation of a Line in the Galactic Radio Spectrum: Radiation from Galactic Hydrogen at 1,420 Mc./sec. , 168(4270):356, September 1951.

[10] Vera C. Rubin and W. Kent Ford, Jr. Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. , 159:379, February 1970.

[11] K. C. Freeman. On the Disks of Spiral and S0 Galaxies. , 160:811, June 1970.

[12] J. P. Ostriker, P. J. E. Peebles, and A. Yahil. The Size and Mass of Galaxies, and the Mass of the Universe. , 193:L1, October 1974.

[13] Jaan Einasto, Ants Kaasik, and Enn Saar. Dynamic evidence on massive coronas of galaxies. , 250(5464):309–310, July 1974.

[14] Barbara Ryden. *Introduction to Cosmology*. Cambridge University Press, 2 edition, November 2016.

[15] Gianfranco Bertone and Dan Hooper. History of dark matter. *Rev. Mod. Phys.*, 90:045002, Oct 2018.

[16] A. Arbey and F. Mahmoudi. Dark matter and the early Universe: a review. *Progress in Particle and Nuclear Physics*, 119:103865, July 2021. arXiv:2104.11488 [astro-ph, physics:hep-ph].

[17] Gaia Collaboration. The gaia mission*. *A&A*, 595:A1, 2016.

[18] Jason A. S. Hunt and Eugene Vasiliev. Milky Way dynamics in light of Gaia, January 2025. arXiv:2501.04075 [astro-ph].

[19] Anne M. Green. Dark Matter in Astrophysics/Cosmology. *SciPost Physics Lecture Notes*, page 37, January 2022. arXiv:2109.05854 [hep-ph].

[20] Anthony H. Gonzalez, Suresh Sivanandam, Ann I. Zabludoff, and Dennis Zaritsky. GALAXY CLUSTER BARYON FRACTIONS REVISITED. *The Astrophysical Journal*, 778(1):14, October 2013.

[21] European Space Agency. The Bullet Cluster (1E 0657-56). https://www.esa.int/ESA_Multimedia/Images/2007/07/The_Bullet_Cluster2, 2007. Accessed: 2025-06-16.

[22] Planck Collaboration. Planck 2018 results - i. overview and the cosmological legacy of planck. *A&A*, 641:A1, 2020.

[23] A. A. Penzias and R. W. Wilson. A Measurement of Excess Antenna Temperature at 4080 Mc/s. , 142:419–421, July 1965.

[24] C. L. Bennett et al. NINE-YEAR *WILKINSON MICROWAVE ANISOTROPY PROBE* (*WMAP*) OBSERVATIONS: FINAL MAPS AND RESULTS. *The Astrophysical Journal Supplement Series*, 208(2):20, September 2013.

[25] Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. *Astronomy & Astrophysics*, 641:A6, September 2020. arXiv:1807.06209 [astro-ph].

[26] Mark Vogelsberger, Federico Marinacci, Paul Torrey, and Ewald Puchwein. Cosmological simulations of galaxy formation, 2019.

[27] Jesús Zavala and Carlos S. Frenk. Dark matter haloes and subhaloes. 2019. Publisher: arXiv Version Number: 1.

[28] Matthew R. Buckley and Annika H.G. Peter. Gravitational probes of dark matter physics. *Physics Reports*, 761:1–60, October 2018.

[29] Alex Drlica-Wagner et al. Probing the Fundamental Nature of Dark Matter with the Large Synoptic Survey Telescope, April 2019. arXiv:1902.01055 [astro-ph, physics:hep-ex, physics:hep-ph].

[30] Michael Kuhlen, Mark Vogelsberger, and Raul Angulo. Numerical simulations of the dark universe: State of the art and the next decade, 2012.

[31] V. Springel et al. The Aquarius Project: the subhaloes of galactic haloes. *Monthly Notices of the Royal Astronomical Society*, 391(4):1685–1711, December 2008.

[32] ESA, editor. *The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission*, volume 1200 of *ESA Special Publication*, January 1997.

[33] Gaia Collaboration. Gaia early data release 3 documentation. https://www.cosmos.esa.int/web/gaia/dr3, 2020. Accessed: 2025-04-10.

[34] European Space Agency. Hipparcos catalogue summary. https://www.cosmos.esa.int/web/hipparcos/catalogue-summary. Accessed: 2025-06-19.

[35] D. Katz et al. *Gaia* Data Release 3: Properties and validation of the radial velocities. *Astronomy & Astrophysics*, 674:A5, June 2023.

[36] D. Katz et al. *Gaia* Data Release 2: Properties and validation of the radial velocities. *Astronomy & Astrophysics*, 622:A205, February 2019.

[37] European Space Agency. Gaia spacecraft exploded diagram. https://sci.esa.int/web/gaia/-/58253-gaia-spacecraft-exploded-diagram, 2016. Accessed: 2025-04-17.

[38] David W. Hogg. A likelihood function for the Gaia Data, April 2018. arXiv:1804.07766 [astro-ph].

[39] C. A. L. Bailer-Jones et al. Estimating Distances from Parallaxes. V. Geometric and Photogeometric Distances to 1.47 Billion Stars in Gaia Early Data Release 3. *The Astronomical Journal*, 161(3):147, March 2021.

[40] R. Andrae et al. Gaia Data Release 3: Analysis of the Gaia BP/RP spectra using the General Stellar Parameterizer from Photometry. 2022. Publisher: arXiv Version Number: 1.

[41] Gaia Collaboration. *Gaia* Early Data Release 3: The celestial reference frame (*Gaia* -CRF3). *Astronomy & Astrophysics*, 667:A148, November 2022.

[42] European Space Agency. Gaia Data Release 2 Documentation. https://gea.esac.esa.int/archive/documentation/GDR2/, 2018. Version 1.2, accessed on 2025-06-19.

[43] Robert J. J. Grand et al. Overview and public data release of the augmented Auriga Project: cosmological simulations of dwarf and Milky Way-mass galaxies, July 2024. arXiv:2401.08750 [astro-ph].

[44] Douglas Potter, Joachim Stadel, and Romain Teyssier. PKDGRAV3: Beyond Trillion Particle Cosmological Simulations for the Next Era of Galaxy Surveys, September 2016. arXiv:1609.08621 [astro-ph].

[45] *The encyclopedia of cosmology Volume 2, Numerical simulations in cosmology*. World Scientific, Singapore, 2018. OCLC: 1032684409.

[46] J. Onorbe et al. How to zoom: bias, contamination and Lagrange volumes in multi-mass cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 437(2):1894–1908, January 2014.

[47] Andrew R. Wetzel et al. RECONCILING DWARF GALAXIES WITH CDM COSMOLOGY: SIMULATING A REALISTIC POPULATION OF SATELLITES AROUND A MILKY WAY–MASS GALAXY. *The Astrophysical Journal Letters*, 827(2):L23, August 2016.

[48] Qirong Zhu et al. Baryonic impact on the dark matter distribution in Milky Way-sized galaxies and their satellites. *Monthly Notices of the Royal Astronomical Society*, 458(2):1559–1580, May 2016.

[49] Shea Garrison-Kimmel et al. Not so lumpy after all: modelling the depletion of dark matter subhaloes by Milky Way-like galaxies. *Monthly Notices of the Royal Astronomical Society*, 471:1709–1727, October 2017. ADS Bibcode: 2017MNRAS.471.1709G.

[50] Philip F. Hopkins et al. FIRE-2 simulations: physics versus numerics in galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 480:800–863, October 2018. ADS Bibcode: 2018MNRAS.480..800H.

[51] Shea Garrison-Kimmel et al. The origin of the diverse morphologies and kinematics of Milky Way-mass galaxies in the FIRE-2 simulations. *Monthly Notices of the Royal Astronomical Society*, 481(3):4133–4157, December 2018.

[52] Robyn E. Sanderson et al. Reconciling Observed and Simulated Stellar Halo Masses. *The Astrophysical Journal*, 869(1):12, December 2018.

[53] Megan Barry et al. The dark side of FIRE: predicting the population of dark matter subhaloes around Milky Way-mass galaxies. *Monthly Notices of the Royal Astronomical Society*, 523(1):428–440, May 2023.

[54] Robyn E. Sanderson et al. Synthetic Gaia Surveys from the FIRE Cosmological Simulations of Milky Way-mass Galaxies. *The Astrophysical Journal Supplement Series*, 246(1):6, January 2020.

[55] Pavel Kroupa. The Initial Mass Function of Stars: Evidence for Uniformity in Variable Systems. *Science*, 295(5552):82–91, January 2002. arXiv:astro-ph/0201098.

[56] Gilles Chabrier. Galactic Stellar and Substellar Initial Mass Function. *Publications of the Astronomical Society of the Pacific*, 115(809):763–795, July 2003. arXiv:astro-ph/0304382.

[57] Charles Keeton. *Principles of astrophysics: using gravity and stellar physics to explore the cosmos*. Undergraduate Lecture Notes in Physics. Springer, New York, 2014.

[58] Hannu Karttunen, Pekka Kröger, Heikki Oja, Markku Poutanen, and Karl Johan Donner, editors. *Fundamental Astronomy*. Springer Berlin Heidelberg, Berlin, Heidelberg, sixth edition edition, 2017.

[59] James Binney and Scott Tremaine. *Galactic Dynamics: Second Edition*. January 2008. Publication Title: Galactic Dynamics: Second Edition ADS Bibcode: 2008gady.book.....B.

[60] Fabio Iocco, Miguel Pato, and Gianfranco Bertone. Evidence for dark matter in the inner Milky Way. *Nature Physics*, 11(3):245–248, March 2015.

[61] Andrew R. Zentner and James S. Bullock. Halo Substructure and the Power Spectrum. *The Astrophysical Journal*, 598(1):49–72, November 2003.

[62] Catherine E Fielder et al. Illuminating dark matter halo density profiles without subhaloes. *Monthly Notices of the Royal Astronomical Society*, 499(2):2426–2444, October 2020.

[63] G. Kauffmann, S. D. M. White, and B. Guiderdoni. The formation and evolution of galaxies within merging dark matter haloes*. *Monthly Notices of the Royal Astronomical Society*, 264(1):201–218, September 1993.

[64] Ayuki Kamada, Manoj Kaplinghat, Andrew B. Pace, and Hai-Bo Yu. Self-Interacting Dark Matter Can Explain Diverse Galactic Rotation Curves. *Physical Review Letters*, 119(11):111102, September 2017.

[65] Gaia Collaboration. Gaia data release 3 - summary of the content and survey properties. *A&A*, 674:A1, 2023.

[66] Xiaowei Ou, Anna-Christina Eilers, Lina Necib, and Anna Frebel. The dark matter profile of the Milky Way inferred from its circular velocity curve, May 2023. arXiv:2303.12838 [astro-ph].

[67] Jo Bovy et al. THE MILKY WAY'S CIRCULAR-VELOCITY CURVE BETWEEN 4 AND 14 kpc FROM APOGEE DATA. *The Astrophysical Journal*, 759(2):131, November 2012.

[68] Anna-Christina Eilers, David W. Hogg, Hans-Walter Rix, and Melissa K. Ness. The Circular Velocity Curve of the Milky Way from 5 to 25 kpc. *The Astrophysical Journal*, 871(1):120, January 2019.

[69] J. I. Read. The Local Dark Matter Density. *Journal of Physics G: Nuclear and Particle Physics*, 41(6):063101, June 2014. arXiv:1404.1938 [astro-ph].

[70] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. , 125(925):306, March 2013.

[71] Ryosuke Okuta et al. Cupy: A numpy-compatible library for nvidia gpu calculations. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[72] Ekaterina V. Karukes, Maria Benito, Fabio Iocco, Roberto Trotta, and Alex Geringer-Sameth. A robust estimate of the Milky Way mass from rotation curve data. *Journal of Cosmology and Astroparticle Physics*, 2020(05):033–033, May 2020. arXiv:1912.04296 [astro-ph].

[73] María Benito, Fabio Iocco, and Alessandro Cuoco. Uncertainties in the Galactic Dark Matter distribution: An update. *Physics of the Dark Universe*, 32:100826, May 2021.

[74] Sven Põder. Galactic parameter estimation using spectroscopic data from the gaia space telescope. Master's thesis, Tallinn University of Technology, Tallinn, Estonia, 2021.

[75] European Space Agency (ESA). Gaia Data Release 4. https://www.cosmos.esa.int/web/gaia/dr4, 2025. Accessed: 2025-06-12.

[76] Carlo Giocoli, Giuseppe Tormen, Ravi K. Sheth, and Frank C. Van Den Bosch. The substructure hierarchy in dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, March 2010.

[77] María Benito, Juan Carlos Criado, Gert Hütsi, Martti Raidal, and Hardi Veermäe. Implications of Milky Way Substructures for the Nature of Dark Matter. *Physical Review D*, 101(10):103023, May 2020. arXiv:2001.11013 [astro-ph].

[78] Alejandro Benitez-Llambay and Carlos Frenk. The detailed structure and the onset of galaxy formation in low-mass gaseous dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 498(4):4887–4900, September 2020.

[79] Till Sawala, Carlos S. Frenk, Azadeh Fattahi, Julio F. Navarro, Tom Theuns, Richard G. Bower, Robert A. Crain, Michelle Furlong, Adrian Jenkins, Matthieu Schaller, and Joop Schaye. The chosen few: the low-mass haloes that host faint galaxies. *Monthly Notices of the Royal Astronomical Society*, 456(1):85–97, February 2016.

[80] R. A. Ibata, G. F. Lewis, M. J. Irwin, and T. Quinn. Uncovering cold dark matter halo substructure with tidal streams. *Monthly Notices of the Royal Astronomical Society*, 332(4):915–920, June 2002.

[81] Madison Walder, Denis Erkal, Michelle Collins, and David Martinez-Delgado. Probing the dark matter haloes of external galaxies with stellar streams, February 2024. arXiv:2402.13314 [astro-ph].

[82] Jo Bovy, Denis Erkal, and Jason L. Sanders. Linear perturbation theory for tidal streams and the small-scale CDM power spectrum. *Monthly Notices of the Royal Astronomical Society*, 466(1):628–668, April 2017. arXiv:1606.03470 [astro-ph].

[83] Adrian M. Price-Whelan and Ana Bonaca. Off the Beaten Path: Gaia Reveals GD-1 Stars outside of the Main Stream. *The Astrophysical Journal Letters*, 863(2):L20, August 2018.

[84] Ana Bonaca, David W. Hogg, Adrian M. Price-Whelan, and Charlie Conroy. The Spur and the Gap in GD-1: Dynamical Evidence for a Dark Substructure in the Milky Way Halo. *The Astrophysical Journal*, 880(1):38, July 2019. Publisher: The American Astronomical Society.

[85] E. R. Siegel, M. P. Hertzberg, and J. N. Fry. Probing dark matter substructure with pulsar timing. *Monthly Notices of the Royal Astronomical Society*, 382(2):879–885, December 2007.

[86] Vincent S. H. Lee, Andrea Mitridate, Tanner Trickle, and Kathryn M. Zurek. Probing Small-Scale Power Spectra with Pulsar Timing Arrays. *Journal of High Energy Physics*, 2021(6):28, June 2021. arXiv:2012.09857 [astro-ph, physics:hep-ph].

[87] Bryan Ostdiek, Ana Diaz Rivero, and Cora Dvorkin. Extracting the Subhalo Mass Function from Strong Lens Images with Image Segmentation. *The Astrophysical Journal*, 927(1):83, March 2022.

[88] Ana Diaz Rivero and Cora Dvorkin. Direct detection of dark matter substructure in strong lens images with convolutional neural networks. *Physical Review D*, 101(2):023515, January 2020.

[89] Stephon Alexander et al. Deep Learning the Morphology of Dark Matter Substructure. *The Astrophysical Journal*, 893(1):15, April 2020. arXiv:1909.07346 [astro-ph].

[90] Reinaldo R. Rosa. Data Science Strategies for Multimessenger Astronomy. *Anais da Academia Brasileira de Ciências*, 93(suppl 1):e20200861, 2021.

[91] Martin Erdmann, Jonas Glombitza, Gregor Kasieczka, and Uwe Klemradt. *Deep Learning for Physics Research*. WORLD SCIENTIFIC, 2021.

[92] François Chollet. *Deep learning with Python*. Manning Publications, Shelter Island, second edition edition, 2021. OCLC: on1289290141.

[93] Geoffrey Hinton. Csc321 lecture 6: Mini-batch gradient descent and optimization tricks. https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 2014. University of Toronto, CSC321: Introduction to Neural Networks and Machine Learning.

[94] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

[95] LUMI Consortium / CSC. GPU nodes – LUMI-G LUMI Supercomputer Documentation. https://docs.lumi-supercomputer.eu/hardware/lumig/, 2025. Accessed June 30, 2025.

[96] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[97] François Chollet et al. Keras. https://keras.io, 2015.

[98] Adam Paszke et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library, December 2019. arXiv:1912.01703 [cs].

[99] Juna Kollmeier et al. SDSS-V Pioneering Panoptic Spectroscopy. In *Bulletin of the American Astronomical Society*, volume 51, page 274, September 2019.

[100] Steven R. Majewski et al. The apache point observatory galactic evolution experiment (apogee). *The Astronomical Journal*, 154(3):94, August 2017.

[101] Željko Ivezić et al. Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, March 2019.

[102] Euclid Collaboration. Euclid - i. overview of the euclid mission. *A&A*, 697:A1, 2025.

[103] S. C. Odewahn, E. B. Stockwell, R. L. Pennington, R. M. Humphreys, and W. A. Zumach. Automated star/galaxy discrimination with neural networks. *The Astronomical Journal*, 103:318, January 1992.

[104] Miquel Serra-Ricart, Xavier Calbet, Lluis Garrido, and Vicens Gaitan. Multidimensional Statistical Analysis Using Artificial Neural Networks: Astronomical Applications. *The Astronomical Journal*, 106:1685, October 1993. Publisher: IOP ADS Bibcode: 1993AJ....106.1685S.

[105] Michael J. Smith and James E. Geach. Astronomia ex machina: a history, primer, and outlook on neural networks in astronomy. *Royal Society Open Science*, 10(5):221454, May 2023. arXiv:2211.03796 [astro-ph].

[106] Andrew E. Firth, Ofer Lahav, and Rachel S. Somerville. Estimating photometric redshifts with artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 339(4):1195–1202, March 2003.

[107] Ioana Ciucă et al. Unveiling the Distinct Formation Pathways of the Inner and Outer Discs of the Milky Way with Bayesian Machine Learning. *Monthly Notices of the Royal Astronomical Society*, 503(2):2814–2824, March 2021. arXiv:2003.03316 [astro-ph].

[108] Marco Chianese et al. Differentiable strong lensing: uniting gravity and neural nets through differentiable probabilistic programming. *Monthly Notices of the Royal Astronomical Society*, 496(1):381–393, July 2020.

[109] Ting-Yun Cheng et al. Identifying Strong Lenses with Unsupervised Machine Learning using Convolutional Autoencoder. *Monthly Notices of the Royal Astronomical Society*, 494(3):3750–3765, May 2020. arXiv:1911.04320 [astro-ph].

[110] C. E. Petrillo et al. Finding strong gravitational lenses in the Kilo Degree Survey with Convolutional Neural Networks. *Monthly Notices of the Royal Astronomical Society*, 472(1):1129–1150, November 2017.

[111] Michael J. Smith et al. Realistic galaxy image simulation via score-based generative models. *Monthly Notices of the Royal Astronomical Society*, 511(2):1808–1818, February 2022. arXiv:2111.01713 [astro-ph].

[112] Siamak Ravanbakhsh et al. Enabling Dark Energy Science with Deep Generative Models of Galaxy Images, November 2016. arXiv:1609.05796 [astro-ph].

[113] Sung Hak Lim, Eric Putney, Matthew R. Buckley, and David Shih. Mapping Dark Matter in the Milky Way using Normalizing Flows and Gaia DR3, May 2023. arXiv:2305.13358 [astro-ph].

[114] Core Francisco Park et al. 3D Reconstruction of Dark Matter Fields with Diffusion Models: Towards Application to Galaxy Surveys.

[115] Victoria Ono et al. Debiasing with Diffusion: Probabilistic reconstruction of Dark Matter fields from galaxies with CAMELS, March 2024. arXiv:2403.10648 [astro-ph].

[116] Nicholas M. Ball and Robert J. Brunner. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106, July 2010. arXiv:0906.2173 [astro-ph].

[117] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[118] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Keras Tuner. https://github.com/keras-team/keras-tuner, 2019.

[119] Steffen R. Knollmann and Alexander Knebe. AHF: AMIGA'S HALO FINDER. *The Astrophysical Journal Supplement Series*, 182(2):608–624, June 2009.

[120] Robert Feldmann and Douglas Spolyar. Detecting dark matter substructures around the Milky Way with Gaia. *Monthly Notices of the Royal Astronomical Society*, 446(1):1000–1012, January 2015.

[121] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-Normalizing Neural Networks. 2017. Publisher: arXiv Version Number: 5.

[122] Andrea Zonca et al. healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in Python. *Journal of Open Source Software*, 4(35):1298, March 2019.

[123] S. Chandrasekhar. Dynamical Friction. I. General Considerations: the Coefficient of Dynamical Friction. , 97:255, March 1943.

[124] W. A. Mulder. Dynamical friction on extended objects. *Astronomy and Astrophysics*, 117:9–16, January 1983. ADS Bibcode: 1983A&A...117....9M.

[125] Charlie Conroy et al. All-sky dynamical response of the Galactic halo to the Large Magellanic Cloud. *Nature*, 592(7855):534–536, April 2021.

[126] Nicolas Garavito-Camargo et al. Hunting for the Dark Matter Wake Induced by the Large Magellanic Cloud. *The Astrophysical Journal*, 884(1):51, October 2019.

[127] Hayden R. Foote et al. Structure, Kinematics, and Observability of the Large Magellanic Cloud's Dynamical Friction Wake in Cold versus Fuzzy Dark Matter. *The Astrophysical Journal*, 954(2):163, September 2023.

[128] K. J. Fushimi, M. E. Mosquera, and M. Dominguez. A determination of the LMC dark matter subhalo mass using the MW halo stars in its gravitational wake. 2023. Publisher: arXiv Version Number: 1.

[129] Malte Buschmann, Joachim Kopp, Benjamin R. Safdi, and Chih-Liang Wu. Stellar Wakes from Dark Matter Subhalos. *Physical Review Letters*, 120(21):211101, May 2018.

[130] Isaac Alonso Asensio, Claudio Dalla Vecchia, Douglas Potter, and Joachim Stadel. Mesh-free hydrodynamics in PKDGRAV3 for galaxy formation simulations. *Monthly Notices of the Royal Astronomical Society*, 519(1):300–317, December 2022. arXiv:2211.12243 [astro-ph, physics:physics].

[131] J. Diemand et al. Clumps and streams in the local dark matter distribution. *Nature*, 454(7205):735–738, 2008.

[132] Matej Ulicny, Vladimir A. Krylov, and Rozenn Dahyot. Harmonic networks for image classification. In *British Machine Vision Conference*, 2019.

[133] Matej Ulicny, Vladimir A. Krylov, and Rozenn Dahyot. Harmonic networks with limited training samples. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.

[134] Tri Nguyen et al. Synthetic Gaia DR3 surveys from the FIRE cosmological simulations of Milky-Way-mass galaxies, July 2023. arXiv:2306.16475 [astro-ph].

[135] María Benito, Konstantin Karchev, Rebecca K. Leane, Sven Põder, Juri Smirnov, and Roberto Trotta. Dark matter halo parameters from overheated exoplanets via bayesian hierarchical inference. *Journal of Cosmology and Astroparticle Physics*, 2024(07):038, jul 2024.

Acknowledgements

I am extremely grateful to my supervisors, Joosep Pata and María Benito, who took me on as a student and have been the guiding lights as I navigated the winding roads of my PhD. I have been exceptionally lucky to be mentored by both of them, as this thesis or the work it represents simply would not have been possible without their unending encouragement and guidance.

Also, I am grateful to Martti Raidal, Mario Kadastik, and the rest of the KBFI community for fostering a supportive research environment and providing me with many opportunities over the years. Andi Hektor deserves special thanks for introducing me to the KBFI family in the first place.

Throughout my PhD journey, I have been fortunate enough to collaborate with inspiring researchers and would like to thank all of my co-authors. I am especially grateful to Claudio Dalla Vecchia and Isaac Alonso Asensio, who generously hosted my visit to the IAC in Tenerife and very patiently taught me how to run the simulations, which were crucial to the completion of the work in this thesis.

I would also like to thank my fellow students, Kristjan, Laurits, Norman, Juan, and Martin, for being excellent brothers in arms, always offering advice and valuable feedback. The shared laughter and occasional karaoke nights have turned this long grind into a collection of fond memories.

I feel truly privileged to have a family that's had my back throughout, and I want to thank Liivika, Sven sen., Silvia, Sten L., Silver, Mariann, Anne, Ahto, Marika, Sten P. for their enduring support. I am eternally grateful to Liise, who has been a loving companion and stood by me through the highs and the lows, cheering me on every step of the way. My warmest thanks also go to Anne, Nicolai, August, Rose-Marie, Snupsu, and Watson for their kindness, encouragement, and for always making me feel at home at Liivaku.

To my wonderful friends Karl, Piret, Karl-Erik, Anna-Riin, Henn, Bum, Ain, Sten, Lenne, Nele, Indrek, thank you for always keeping me grounded, laughing, and (mostly) sane throughout this journey.

I owe special thanks to Erik Kesa, who mentored me in programming early on in my studies and unknowingly set me on a course that would eventually lead me here.

Finally, I acknowledge the financial support that made this work possible: the NVIDIA Hardware Grant Program, the EUCAPT travel grant, the Estonian Ministry of Education and Research (grant TK202), as well as the Estonian Research Council grants PSG938 and RVTT7.

Abstract

Probing the Milky Way's Dark Matter Halo in the Gaia and Machine Learning Era

According to the current theory of structure formation, galaxies reside within massive dark matter (DM) halos. Despite compelling evidence from cosmological and astrophysical observations, the fundamental nature of DM remains poorly understood. Since the properties of DM halos are a function of the microphysical nature of DM, characterising them can help test the validity of ΛCDM and narrow down proposed alternative DM models. This thesis focuses on two key aspects of the Milky Way's (MW) DM halo: (i) the smooth, virialised halo component that dominates the overall gravitational potential, and (ii) the substructure of the host halo comprised of a population of smaller dark subhalos, whose detection efforts remain challenging as these structures are not expected to host any stars.

To characterize the smooth halo, the circular velocity curve of the MW was reconstructed using a sample of red giant branch stars from Gaia DR3, spanning Galactocentric radii between 5 and 14 kpc. A key novelty of this study was the use of a Bayesian framework that marginalizes over major systematic uncertainties, including the Sun's galactocentric distance and spatial-kinematic morphology of the tracer sample. The resulting circular velocity curve is consistent with a flat profile, with an estimated circular velocity at the Solar circle ($R_0 = 8.277\,\mathrm{kpc}$) of $233 \pm 7\,\mathrm{km/s}$. The local DM density inferred with this approach is $\rho_{\mathrm{DM}}(R_0) = \left(0.41^{+0.10}_{-0.09}\right)\,\mathrm{GeV/cm^3} = \left(0.011^{+0.003}_{-0.002}\right)\,\mathrm{M_\odot/pc^3}$.

In addition to the above, this thesis covers two studies regarding the viability of machine learning-based approaches to probe the low-mass end of the subhalo mass function. The first employed an anomaly detection approach on high-resolution MW-like cosmological simulations and synthetic Gaia datasets derived from them to assess the statistical imprint of subhalos in realistic stellar halos. The second study made use of idealised N-body simulations and supervised deep learning techniques to detect stellar wakes induced by individual subhalos. Although various limitations currently hinder detection in MW-like simulations, we found that DM subhalos leave a detectable imprint in the phase-space of halo stars. When looking at stellar wakes in a controlled environment, we saw that we are able to achieve non-trivial detection performance for subhalos with masses as low as $5 \times 10^7\,\mathrm{M_\odot}$, while being severely limited by the amount of available training data.

In summary, the results of this doctoral thesis contribute to the current understanding of the MW's DM distribution. While keeping in mind the comprehensive datasets produced from current and future stellar surveys, the machine learning approaches developed in this work provide a foundation for future searches for DM substructure in the Galaxy. By extension, they constitute the first initial steps to use stellar wakes as probes of the particle nature of DM.

Kokkuvõte
Linnutee galaktika tumeaine halo uurimine Gaia ja masinõppe ajastul

Tänapäevase struktuuritekke teooria kohaselt paiknevad galaktikad massiivsete tumeaine halode sees, mille omadused sõltuvad suuresti tumeaine mikrofüüsikalisest iseloomust. Vaatamata tugevale kosmoloogilisele ja astrofüüsikalisele tõendusmaterjalile ei ole tumeaine fundamentaalne olemus galaktilisel ja subgalaktilisel skaalal üheselt mõistetud. Tumeaine halode täpne karakteriseerimine aitab seetõttu kontrollida ΛCDM kosmoloogilise mudeli kehtivust ning piirata pakutud tumeainet mudelite valikut.

Käesolev doktoritöö keskendub Linnutee tumeaine halo kahele põhikomponendile: (i) hajus halo komponent, mis laias laastus domineerib galaktika gravitatsioonipotentsiaali ja (ii) halo alamstruktuur, mis koosneb väiksematest alamhalodest, mille tuvastamine on keeruline, kuna need struktuurid ei sisalda tähti.

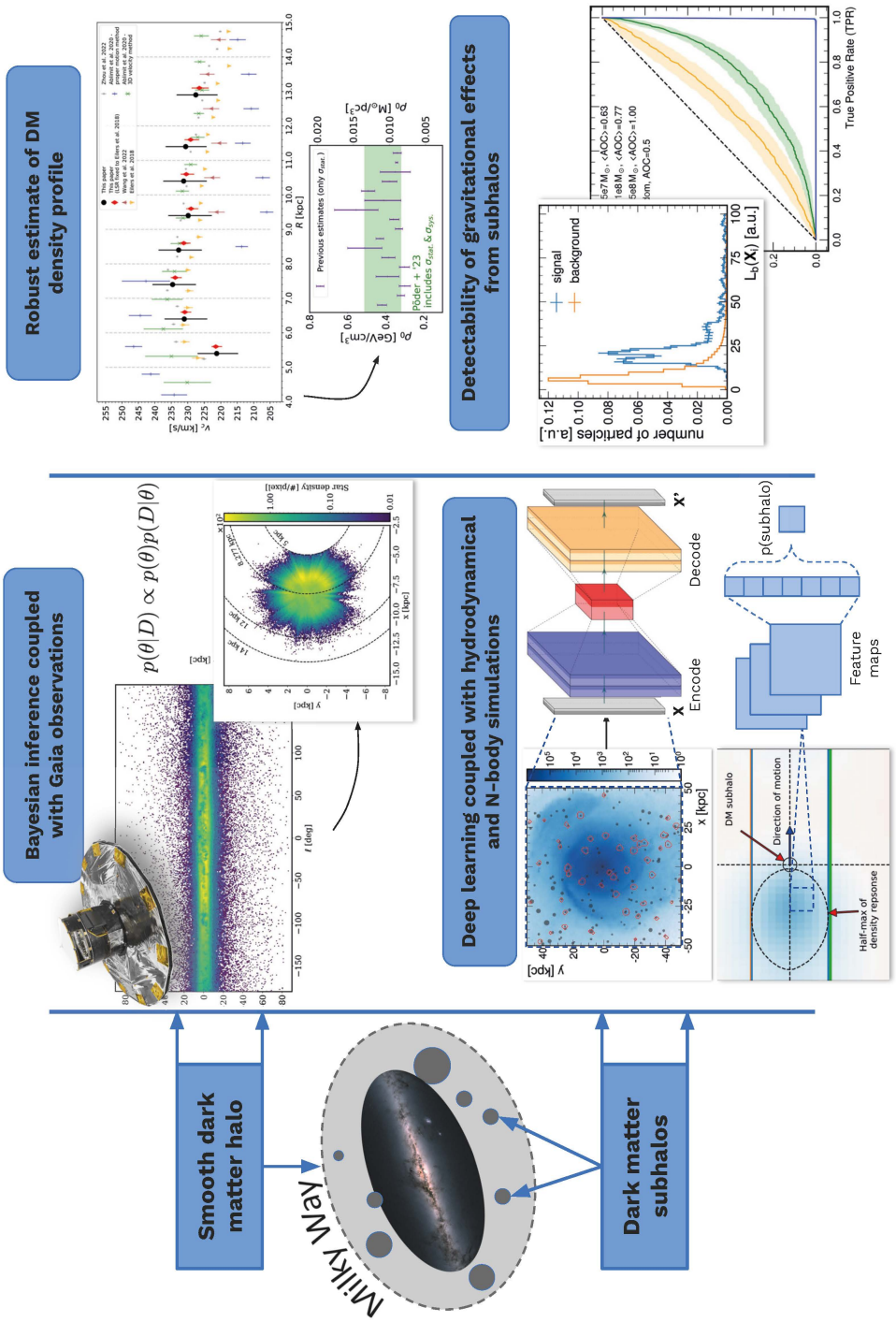
Halo hajuskomponendi iseloomustamiseks rekonstrueeriti Linnutee ringkiiruste kõver, kasutades selleks Euroopa Kosmoseagentuuri kosmoseteleskoobi Gaia 3. andmeväljalaset (DR3). Andmetest selekteeriti punastest hiidudest koosnev valim, mis paikneb galaktotsentrilises vahemikus 5–14 kpc. Valimi analüüsiks ning ringkiiruste tuletamiseks kasutati uudset Bayesiaanlikku lähenemist, mis arvestab lisaks statistilistele ka tähepopulatsiooniga seotud erinevaid süstemaatilisi määramatusi. Tulemuseks saadud ringkiiruste kõver on kooskõlaline lameda profiiliga, hinnanguline ringkiirus Päikese galaktotsentrilisel kaugusel ($R_0 = 8.277\,\mathrm{kpc}$) on $233 \pm 7\,\mathrm{km/s}$. Sellest tulenev lokaalne tumeaine tihedus on $\rho_{\mathrm{DM}}(R_0) = \left(0.41^{+0.10}_{-0.09}\right)\,\mathrm{GeV/cm^3} = \left(0.011^{+0.003}_{-0.002}\right)\,\mathrm{M_\odot/pc^3}$.

Lisaks eelnevale tutvustab doktoritöö kahte analüüsi, mis uurivad masinõppepõhiste lähenemiste rakendatavust alamhalode massifunktsiooni madala massipiirkonna uurimiseks. Esimene neist kasutas järelvalveta õppel baseeruvat anomaaliatuvastust, et hinnata alamhalode gravitatsiooniliste mõjutuste statistilist jälge Linnuteele sarnaste galaktikate tähtede halos. Selleks kasutati kõrglahutusega kosmoloogilisi simulatsioone ning nende põhjal loodud sünteetilisi Gaia andmestikke. Teine uuring tugines idealiseeritud N-keha simulatsioonidele ja järelvalvega masinõppele, et karakteriseerida ja tuvastada isoleeritud alamhalode põhjustatud häiritusi ümbritsevate tähtede faasiruumis – nn. tähejoomid. Antud doktoritöö tulemused näitavad, et tänapäevastes Linnutee-sarnastes simulatsioonides on alamhalode detekteerimine raskendatud erinevate numbriliste ja vaatluslike efektide tõttu. Sellegipoolest, nägime, et alamhalode läheduses olevate tähtede faasiruumi jaotus erineb tausta tähtede jaotusest. Idealiseeritud simulatsioone kasutades leidsime, et isegi mahult piiratud treeningandmestike kasutades saavutame individuaalsete alamhalode detekteerimisel massipiiriks ligikaudu $5 \times 10^7\,\mathrm{M_\odot}$.

Kokkuvõttes täiendavad antud doktoritöö tulemused olemasolevaid teadmisi Linnutee tumeaine jaotuse kohta. Pidades silmas tänavuste ja tulevaste vaatluslike andmestike mahukust, loovad töö käigus arendatud masinõppe meetodid aluse edasisteks alamstruktuuri otsinguteks.

Graphical Abstract

Probing the Milky Way's Dark Matter Halo in the Gaia and Machine Learning Era



Appendix 1

I

A. Bazarov, M. Benito, G. Hütsi, R. Kipper, J. Pata, and S. Põder. Sensitivity estimation for dark matter subhalos in synthetic gaia dr2 using deep learning. *Astronomy and Computing*, 41:100667, 2022

Astronomy and Computing 41 (2022) 100667

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Sensitivity estimation for dark matter subhalos in synthetic Gaia DR2 using deep learning

A. Bazarov [a], M. Benito [a,b], G. Hütsi [a], R. Kipper [b], J. Pata [a], S. Põder [a,*]

[a] NICPB, Rävala 10, Tallinn 10143, Estonia
[b] Tartu Observatory, University of Tartu, Observatooriumi 1, Tõravere 61602, Estonia

ARTICLE INFO

Article history:
Received 28 March 2022
Accepted 7 November 2022
Available online 14 November 2022

Keywords:
Machine learning
Dark matter
Dark subhalos
Gaia mission
Milky Way

ABSTRACT

The abundance of dark matter subhalos orbiting a host galaxy is a generic prediction of the cosmological framework, and is a promising way to constrain the nature of dark matter. In this paper, we investigate the use of machine learning-based tools to quantify the magnitude of phase-space perturbations caused by the passage of dark matter subhalos. A simple binary classifier and an anomaly detection model are proposed to estimate if stars or star particles close to dark matter subhalos are statistically detectable in simulations. The simulated datasets are three Milky Way-like galaxies and nine synthetic Gaia DR2 surveys derived from these. Firstly, we find that the anomaly detection algorithm, trained on a simulated galaxy with full 6D kinematic observables and applied on another galaxy, is nontrivially sensitive to the dark matter subhalo population. On the other hand, the classification-based approach is not sufficiently sensitive due to the extremely low statistics of signal stars for supervised training. Finally, the sensitivity of both algorithms in the Gaia-like surveys is negligible. The enormous size of the Gaia dataset motivates the further development of scalable and accurate data analysis methods that could be used to select potential regions of interest for dark matter searches to ultimately constrain the Milky Way's subhalo mass function, as well as simulations where to study the sensitivity of such methods under different signal hypotheses.

1. Introduction

Dark matter (DM) represents roughly 84% of the matter content in the Universe (Aghanim et al., 2020). However, unveiling its nature has proven a difficult endeavour, and none of the proposed candidates (from several extensions of the Standard Model to primordial black holes) have yet been detected. Cold DM is expected to form subhalos with masses many orders of magnitude below $10^8\,M_\odot$ (Blumenthal et al., 1984), which is roughly the mass above which galaxies can form (Kitayama and Yoshida, 2005; Read et al., 2006). The abundance of subhalos is dependent on the nature of DM. This dependency can be explained, on the one hand, by the effect of the properties of the DM on the linear matter power spectrum. If for cold DM the minimum halo mass might be as small as $10^{-12}\,M_\odot$ (Zybin et al., 1999; Bringmann, 2009), microscopic properties of the DM particle, e.g. non-negligible thermal velocities or quantum pressure, introduce a cut-off at the small scales in alternative DM scenarios. On the other hand, the nature of DM, e.g. thermal velocities or self-interactions, further impacts the non-linear growth of structures (Schneider et al., 2012; Vogelsberger et al., 2016). Detecting a dark subhalo would be the first direct evidence of DM clustering at small scales. Furthermore, constraints on the subhalo abundance would provide valuable information about the particle nature of DM.

Subhalos with masses lower than $10^8\,M_\odot$ are unable to form stars and remain dark (Kitayama and Yoshida, 2005; Read et al., 2006), thus hindering their detection. Strategies that aim to detect dark subhalos rely on measuring their gravitational signatures via stellar dynamics (Ibata et al., 2002; Yoon et al., 2011; Carlberg, 2012; Bovy et al., 2017; Banik et al., 2018; Bonaca et al., 2019; Benito et al., 2020; Feldmann and Spolyar, 2015; Buschmann et al., 2018), gravitational lensing (Hezaveh et al., 2016; Van Tilburg et al., 2018; Díaz Rivero et al., 2018; Gilman et al., 2019; Brehmer et al., 2019; Vattis et al., 2020) or pulsar timing (Siegel et al., 2007; Baghram et al., 2011; Clark et al., 2016; Kashiyama and Oguri, 2018; Delos and Linden, 2022) and, in the case of several DM candidates, e.g. Weakly Interacting Massive Particles (WIMPs), on detecting the flux of final stable particles produced by DM annihilation or decay (e.g. Buckley and Hooper (2010), Ackermann et al. (2012), Zechlin and Horns (2012), Moliné et al. (2017), Coronado-Blázquez et al. (2019b), Calore et al. (2019), Coronado-Blázquez et al. (2019a, 2021) and Mirabal and Bonaca (2021)). The goal of searches based on stellar

* Corresponding author.
E-mail address: sven.poder@kbfi.ee (S. Põder).

https://doi.org/10.1016/j.ascom.2022.100667
2213-1337/

dynamics is to detect perturbations in the phase-space distribution of Milky Way (MW) stars induced by gravitational effects of passing subhalos. We can look for these perturbations in stellar streams (Ibata et al., 2002; Yoon et al., 2011; Carlberg, 2012; Bovy et al., 2017; Banik et al., 2018; Bonaca et al., 2019) and in the disk or the halo stars (Feldmann and Spolyar, 2015; Buschmann et al., 2018). In the present work we investigate the usage of an anomaly detection and classification algorithms in the search for the imprint caused in halo stars by passing substructures. In this way, we exploit the increasing size of observational datasets and state-of-the-art techniques in deep learning.

In recent years, deep learning techniques have been applied in the search for substructures in our Galaxy (Ostdiek et al., 2020; Necib et al., 2020; Shih et al., 2021a). These detection methods assume that stars in the MW sharing a common origin should cluster in orbital properties and/or composition. Our search differs in that we aim to identify stars that, regardless of their origin, have their distribution in phase-space perturbed by the passage of a dark matter subhalo. For any identified star, it must be possible to test the halo hypothesis independently of the methodology used to select the candidates. One possibility could be to preselect the stars using a ML-based classifier, followed by detailed hypothesis tests using e.g. the orbital arc method (Kipper et al., 2020, 2021) or the stellar wakes technique (Buschmann et al., 2018).

The raw data that we used, which are described in Section 2, are three MW-like galaxies from the Latte suite of FIRE-2 simulations (Wetzel et al., 2016) and nine synthetic Gaia DR2 surveys generated from the simulated galaxies by means of the Ananke framework (Sanderson et al., 2020). First, we processed the synthetic Gaia datasets to correlate the position of stars and the dark subhalos, which were previously identified in the simulated galaxies. In Section 3, we estimate the detectability of the subhalo-associated stars using deep learning techniques. We conclude in Section 5.

2. Datasets

As our raw data, we used three MW-like galaxies from the Latte suite of FIRE-2 simulations (Wetzel et al., 2016; Garrison-Kimmel et al., 2017; Hopkins et al., 2018) (dubbed `m12f`, `m12i` and `m12m`) and nine synthetic Gaia DR2 surveys (Sanderson et al., 2020). This section describes these datasets and the processing we performed on them.

2.1. Milky way-like galaxies

We used the simulation snapshots at $z = 0$ of three MW-like galaxies,[1] namely `m12f`, `m12i` and `m12m` (Wetzel et al., 2016; Garrison-Kimmel et al., 2017; Hopkins et al., 2018). In the following we briefly describe how these MW analogues were obtained. For a complete description of this and the details of the N-body simulations we refer the interested reader to Wetzel et al. (2016) and references therein. The MW analogues were first identified in a DM-only cosmological simulation requiring that at $z = 0$: (i) their virial mass is in the range of $M_{200} = [1 - 2] \times 10^{12}\,\mathrm{M}_\odot$[2] (which agrees with recent measurements (Wang et al., 2020; Karukes et al., 2020; Shen et al., 2021)) and (ii) there is no neighboring halo of similar mass within $5R_{200}$. Three halos selected in this manner were then simulated using the zoom-in technique (Oñorbe et al., 2014). Simulations were run using the Gizmo gravity plus hydrodynamics code in meshless finite-mass (MFM) mode (Hopkins, 2015) and the FIRE-2 baryonic physics model (Hopkins et al., 2018). Dark matter particles in the zoom-in simulation have a mass of $m_{\mathrm{DM}} = 3.5 \times 10^4\,\mathrm{M}_\odot$, and the initial gas or star particle mass is $m_{\mathrm{gas}} = 7.1 \times 10^3\,\mathrm{M}_\odot$. Dark matter and stars have gravitational softening lengths $\epsilon_{\mathrm{DM}} = 20\,\mathrm{pc}$ and $\epsilon_{\mathrm{star}} = 4\,\mathrm{pc}$, respectively. The softening length of the gas is adaptive, and reaches a minimum value of $\epsilon_{\mathrm{gas,min}} = 1\,\mathrm{pc}$.

We identified DM subhalos in snapshots at $z = 0$ of the MW-like galaxies using the Amiga Halo Finder (AHF) code (Knollmann and Knebe, 2009). The AHF algorithm identifies bound DM structures by hierarchically clustering 3D positions of DM particles in the simulation. Following Garrison-Kimmel et al. (2017), AHF was run only on DM particles. We selected subhalos with more than 85 DM particles (corresponding to subhalos with masses $M_{\mathrm{sh}}^{\mathrm{DM}} > 3 \times 10^6\,\mathrm{M}_\odot$) since those substructures are reliably resolved in the simulation (Garrison-Kimmel et al., 2017). However, the MW is expected to have a population of subhalos with lower masses. The velocity changes in stars due to the gravitational encounter with a dark subhalo with a mass of $10^5\,\mathrm{M}_\odot$ are of the order of 10^{-3} km/s (Feldmann and Spolyar, 2015), which are well below the statistical uncertainties in observations of MW halo stars. Therefore, we argue that subhalos with masses smaller than $3 \times 10^6\,\mathrm{M}_\odot$ have a negligible impact on the current investigation of the feasibility of two simple algorithms that search for perturbations in each star independently of each other. Nonetheless, we leave a thorough investigation of the detection of subhalos unresolved in the simulation for a follow-up work.

Approximately $\simeq 10^3$ subhalos[3] for each MW-like galaxy remain as potentially observable. Fig. 1 shows the cumulative subhalo mass function normalized by the virial mass of the host halo (left panel) and the radial distribution of the subhalo population normalized by the virial radius (right panel). The virial masses are $M_{\mathrm{vir}}^{\mathrm{DM}} = 1.1 \times 10^{12}\,\mathrm{M}_\odot$, $0.8 \times 10^{12}\,\mathrm{M}_\odot$ and $1.0 \times 10^{12}\,\mathrm{M}_\odot$ for `m12f`, `m12i` and `m12m`, respectively.[4] In Fig. 2 we show the mass of the subhalos as a function of their galactocentric distance for `m12f`, `m12i` and `m12m`. It should be noted that no subhalos are identified below 14 kpc from the center of the galaxies, as previously noted in Garrison-Kimmel et al. (2017). Furthermore, 97%, 91% and 94% of the subhalos below 50 kpc for `m12f`, `m12i` and `m12m`, respectively, have masses lower than $1 \times 10^7\,\mathrm{M}_\odot$. The most massive subhalo below 50 kpc is identified at 20 kpc with $M_{\mathrm{sh}}^{\mathrm{DM}} = 3 \times 10^7\,\mathrm{M}_\odot$ for `m12f`, at 43 kpc with $M_{\mathrm{sh}}^{\mathrm{DM}} = 4 \times 10^7\,\mathrm{M}_\odot$ for `m12i` and at 43 kpc with $M_{\mathrm{sh}}^{\mathrm{DM}} = 2 \times 10^8\,\mathrm{M}_\odot$ for `m12m`. The depletion of the most massive dark subhalos in the inner 50 kpc of the MW-like galaxies conditions the ability to identify stars in the stellar halo which have been perturbed by the passage of a dark matter subhalo using the deep-learning techniques explored in this study.

2.2. Synthetic gaia surveys

The nine synthetic Gaia DR2 surveys were generated by applying the Ananke framework (Sanderson et al., 2020) to the three MW-like galaxies. Per simulated galaxy, three synthetic surveys were generated by adopting three local standards of rest (LSRs). Each synthetic survey contains approximately a billion mock stellar observations resembling Gaia DR2. We restrict our attention to stellar halo stars, applying a selection in true vertical distances $|z| > 5$ kpc. In this way, we remove disk stars that could suffer from disturbances induced, for example, by spiral arms, the Galactic bar or giant molecular clouds. We would like

[1] Taken from https://girder.hub.yt/#collection/5b0427b2e9914800018237da.

[2] Virial mass and virial radius follow the relation $M_{200} = \frac{4\pi}{3} 200 \rho_m R_{200}^3$, with ρ_m the average matter density of the Universe.

[3] AHF identifies 1298, 1001 and 1281 subhalos with $N_{\mathrm{DM}} > 85$ for `m12f`, `m12i` and `m12m`, respectively.

[4] In our work we have used the virial mass definition given by $M_{\mathrm{vir}} = \frac{4\pi}{3} 178 \rho_{\mathrm{crit}} R_{\mathrm{vir}}^3$, with ρ_{crit} the critical density of the Universe at $z = 0$.

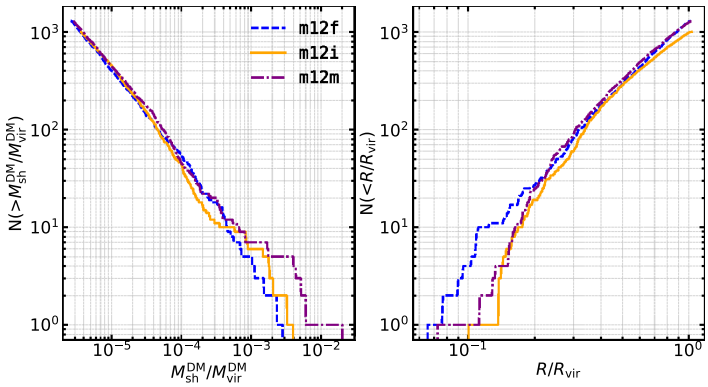


Fig. 1. The subhalo mass function scaled to the host halo virial mass (left) and the radial distribution of the subhalo population (right) for the three MW-like galaxies used in this work.

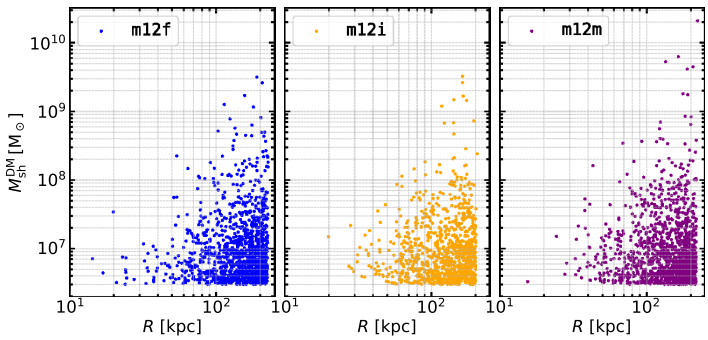


Fig. 2. Subhalo mass as a function of galactocentric distance. Each dot corresponds to a subhalo as identified by AHF for `m12f` (left), `m12i` (middle) and `m12m` (right) galaxies.

Table 1
Summary statistics of synthetic Gaia DR2 reduced catalogs used in this work (see text for more details).

		Stars with $\|z\| > 5$ kpc	With v_r [%]	Halo-associated stars [%]	Halo-associated stars with v_r [%]	Subhalos with associated stars
m12f	LSR0	216,446,024	0.42%	0.0291%	0.35%	73
	LSR1	182,538,592	0.44%	0.0291%	0.32%	76
	LSR2	204,017,261	0.44%	0.0306%	0.35%	71
m12i	LSR0	139,167,343	0.45%	0.0019%	0.41%	63
	LSR1	132,655,442	0.46%	0.0017%	0.41%	61
	LSR2	131,474,668	0.48%	0.0010%	0.23%	67
m12m	LSR0	170,255,144	0.47%	0.0013%	0.09%	67
	LSR1	156,093,757	0.47%	0.0016%	0.12%	71
	LSR2	161,369,511	0.47%	0.0013%	0.19%	68

to highlight that the disk was primarily excluded because the data volume was too large to cope with at this stage. Notice that this problem, however, does not affect the MW-like simulations where the number of star particles per galaxy is of the order of 10^7. It is not clear how our results would be affected if we were to include the disk in our analysis of the synthetic survey, and it is a question that we plan to address in a follow-up. After removing the disk, we are left with $O(10^8)$ mock stars for each LSR for the subsequent analysis. This reduced dataset consists of nearly 2 billion observed stars for the three different MW-like galaxies, three LSRs for each, correlated with potentially observable DM subhalo locations. Table 1 summarizes some statistics of this dataset.

Stars are tagged as halo-associated if their true distance to the central position of a subhalo is lower than 1 kpc. It is to be noticed that these halo-associated stars might not be bound to the subhalos (see next section). Fig. 3 shows the total number of stars associated to a subhalo as a function of the subhalo's mass for each LSR and each simulated galaxy. Within each galaxy, less than ∼10% of the subhalos contain associated stars, and 66%, 84% and 75% of this fraction contain less than 10 associated stars for `m12f`, `m12i` and `m12m` galaxies, respectively. Furthermore, approximately 40% of the halos that have associated stars contain only one star. The `m12f` galaxy has a larger percentage of subhalos which are associated to more than 100 stars compared to that of `m12i` or `m12f` galaxies. This is because the former galaxy has a larger fraction of subhalos below 30 kpc. We plot the projected stellar number densities for LSR 0 in Fig. 4, along with the halo locations and halo-associated observed stars.

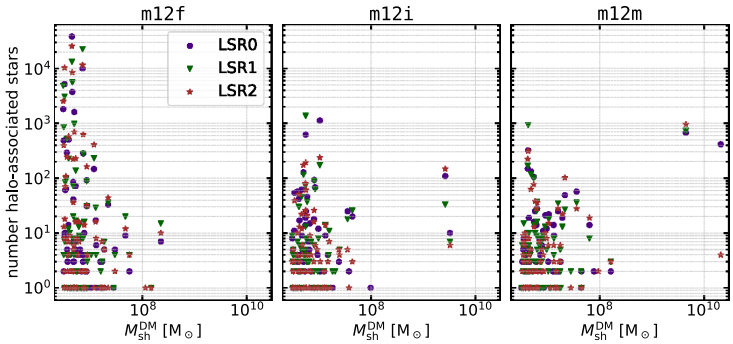


Fig. 3. Number of stars associated to a particular subhalo as a function of its mass for `m12f` (left), `m12i` (middle) and `m12m` (right) galaxies.

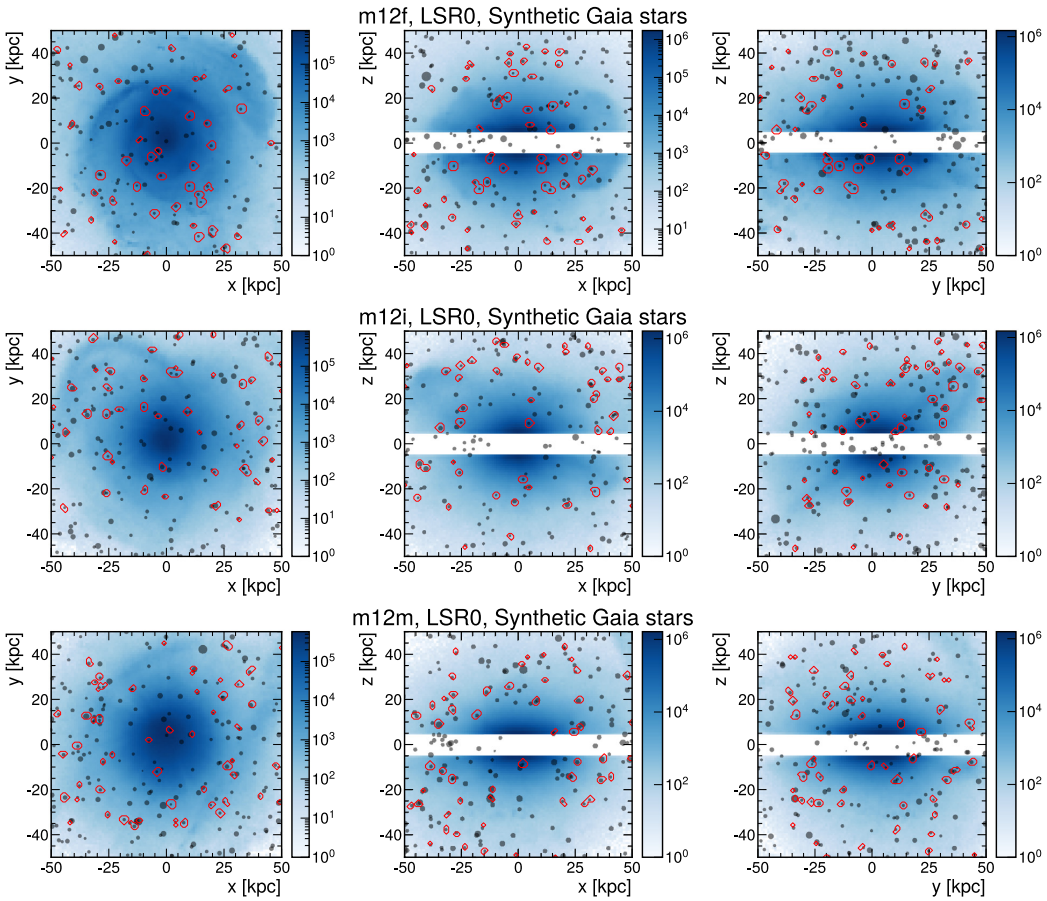


Fig. 4. Projected stellar number densities in the synthetic Gaia datasets in true galactocentric coordinates for the three MW-like galaxies, namely `m12f` (top row), `m12i` (middle row) and `m12m` (bottom row) for LSR0. The halo locations are shown in black, with the size of the markers being proportional to the halo's virial radius, while the regions with halo-associated observed stars are shown in red. The Sun's position for each case is $(x, y, z) = (0, 8.2\,\text{kpc}, 0)$.

3. Deep learning search of subhalo-associated stars

Dark subhalos perturb the positions and velocities of nearby stars. We wish to estimate if these kinematic imprints are detectable in MW-like galaxies and in synthetic Gaia data that accounts for observational uncertainties. Let us assume, without loss of generality, that the properties $\mathbf{X}$ of each star particle (or observed star) are drawn from the probability distribution $p(\mathbf{X}|\text{sig})$ or $p(\mathbf{X}|\text{bkg})$ if the star particle (observed star) has or has not been affected, respectively, by a dark subhalo at a given time in the Latte (Ananke) simulation. Then, if the probabilities are

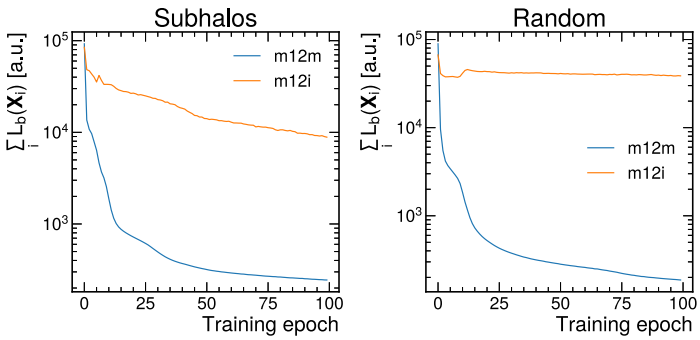


Fig. 5. Relative reconstruction loss with respect to the first epoch for the training (`m12m`) and validation (`m12i`) datasets based on the anomaly detection model over the training epochs. By construction, we expect the loss in `m12m` to decrease as the model fits the samples. The training is performed on star particles excluding the subhalo-associated particles (left) and by excluding the same number of random particles (right) as a check. A.u. stands for arbitrary units of the loss function. We note that the underlying signal distribution of the training and validation datasets are different, thus the numerical values of the loss functions are not necessarily directly comparable between the training and validation datasets. We observe no significant overtraining on the validation dataset.

known, the likelihood ratio $p(\mathbf{X}|\text{sig})/p(\mathbf{X}|\text{bkg})$ is the optimal discriminator between the two hypotheses for a given observation according to the Neyman–Pearson lemma (Neyman and Pearson, 1992). These probabilities are not known, however, and we only have simulated examples of either halo-associated or background star particles (observed stars). In the following, we investigate the possibility of using machine learning to define an approximate discriminator between the two hypotheses, and thus quantify the difference between the halo-associated stars and the background.

3.1. Detectability

As a starting point, we first focus on the Latte simulations, where for each star particle, the full six-dimensional phase-space coordinates, namely the three-dimensional Galactocentric Cartesian positions and velocities, are known. Unlike for the synthetic Gaia dataset, the disk is not excluded at this stage. In addition, here we only consider subhalos with galactocentric distances less than 100 kpc. We are then left with subhalos with masses smaller than 4×10^8 $M_\odot$, which reduces the probability of including stars associated with the halo of dwarf galaxies. This cut in radius, or equivalently in mass, does not strictly mean that luminous halos are excluded from our catalogue. For this reason, we have identified candidate dwarf galaxies as those subhalos that have more than one signal star with a relative velocity with respect to the subhalo smaller than the subhalo's escape velocity. In this manner we have identified 4, 6 and 9 subhalos for `m12f`, `m12i` and `m12m`, respectively. By removing these subhalos in the analyses presented in this section, our results are quantitatively the same.

For each star particle, we compute the Euclidean distance to the nearest dark subhalo d, and if it is below a threshold $d < d_{\max} = 1$ kpc, we identify the star particle as a halo-associated or signal particle. We then use an anomaly detection approach to estimate the strength of the subhalo signal (Baldi and Hornik, 1989; Sakurada and Yairi, 2014). For this purpose, the background-only likelihood $L_b(\mathbf{X}) \simeq p(\mathbf{X}|\text{bkg})$ is indirectly approximated using a so-called autoencoder neural network, and deviations from the background-only distribution are quantified.

Each star particle is characterized by the feature vector $\mathbf{X}$ containing its three-dimensional position and velocity, i.e. (x, y, z, v_x, v_y, v_z). Let us define an encoder $E(\mathbf{X})$ and a decoder $D(\mathbf{z})$ as

$$E(\mathbf{X}) \to \mathbf{z} \in \mathbb{R}^D \text{ and} \tag{1}$$

$$D(\mathbf{z}) \to \mathbf{X}' \in \mathbb{R}^6, \tag{2}$$

respectively, such that $D(E(\mathbf{X})) \to \mathbf{X}'$ approximates $\mathbf{X}$ for any given input via a lower-dimensional $D < 6$ representation. Both the encoder and decoder are implemented as feedforward neural networks, optimized by tuning the weights using only the background examples as follows:

$$\mathbf{D}, \mathbf{E} = \arg\min_{D,E} \sum_{i \in \text{bkg}} \|\mathbf{X}_i - D(E(\mathbf{X}_i))\| \,. \tag{3}$$

The neural network model parameters, such as the number of layers and neurons in each layer, the size of the lower-dimensional representation and the activation function, are chosen based on a small number of experiments rather than through a systematic hyperparameter optimization, which is left for a future study. We use two layers with 128 neurons for the encoder, the latent space $D = 3$, and two layers for the decoder, with again 128 neurons per layer. We use the scaled exponential linear unit (SELU) activation function for the hidden layers (Klambauer et al., 2017).

By construction, the encoder–decoder will tend to reconstruct well the background-like samples that it was optimized on. On the other hand, for any other $\mathbf{X}$ that is not distributed as $p(\mathbf{X}|\text{bkg})$, we would expect on average higher values for the reconstruction loss $L_b(\mathbf{X}_i) = \|\mathbf{X}_i - D(E(\mathbf{X}_i))\|$. Therefore, we can use the distribution $L_b(\mathbf{X})$, optimized only on the background particles, as an empirical discriminator between the background and signal samples. We have checked this approach by defining a fake signal consisting of a random sub-population of stars irrespective of the dark subhalo locations. In this case, no detectable difference between the main sample and the random subpopulation of stars is expected with this method.

We optimize the model on the `m12m` galaxy, while cross-checking the performance on the `m12i` galaxy. This ensures that the model is not simply memorizing the locations of the halos, as the result in this case would not be generalizable to other galaxy simulations. Training is carried out for 100 iterations (epochs) over the full dataset using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $l = 10^{-4}$ and a minibatch size of 10^5 star particles. We show the evolution of the total reconstruction loss over training epochs in Fig. 5 for both the real subhalo signal (left panel) and the fake signal cases (right panel). We observe that the model converges for the training dataset `m12m` and exhibits in general stable behavior for the validation dataset `m12i`.

Fig. 6 shows the $L_b(\mathbf{X}_i)$ distributions for the signal and background stars for the `m12f` dataset never used for training. It is

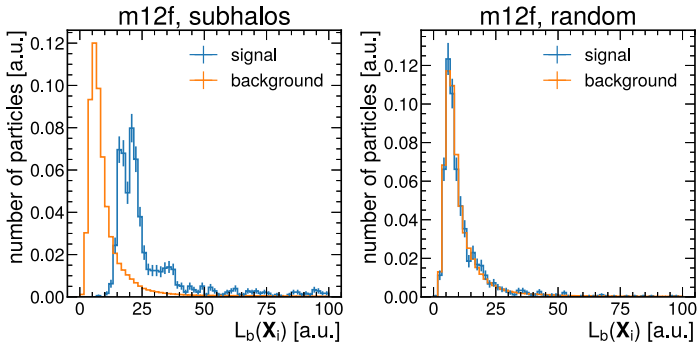


Fig. 6. The distribution of the reconstruction loss L_b for the `m12f` galaxy that was never used in the training procedure. We show the distributions of the halo-associated (signal) and the rest (background) of the star particles separately. On the left, the model was trained and evaluated with the subhalo-associated as the signal, while on the right, a random set of stars was denoted as the signal as a check.

observed that for the real subhalo signal case (left panel), there is a distinction between the distribution of halo-associated (signal) and non-halo associated (background) stars, with the signal stars having on average higher values of the reconstructed distribution. No such distinction is observed for the model trained and tested on the random subset (right panel), as would be expected.

We quantify the performance of the anomaly detection in terms of the true positive and false positive rates. The true positive rate (TPR) gives the fraction of signal stars that are correctly identified as signal particles at a particular threshold t (i.e. a given value of L_b),

$$\mathrm{TPR}(t) = \frac{N_{\mathrm{sig}}(L_b > t)}{N_{\mathrm{sig}}}. \tag{4}$$

Contrary, the false positive rate (FPR) is the fraction of background stars that are incorrectly identified as signal, namely

$$\mathrm{FPR}(t) = \frac{N_{\mathrm{bkg}}(L_b > t)}{N_{\mathrm{bkg}}}. \tag{5}$$

Fig. 7 shows the FPR versus TPR while scanning over t for the real (solid blue) and random (dashed black) signal cases. Using the unsupervised anomaly detection model, we see that at a TPR $\simeq$ 80%, the FPR is $\simeq$15% (i.e. 80% of the signal stars are correctly identified while we misclassify 15% of the background stars as signal), presenting a significant improvement over a random selection.

We cross-check the anomaly detection approach against a simple binary classifier, where the signal model is used explicitly, but which is thus limited by the available statistics for the halo-associated stars. Contrary, in the anomaly-detection based approach, the star labels based on the proximity to a dark subhalo were only used to exclude the signal samples from the optimization. The supervised classification model uses the signal sample labels directly, i.e. the optimization target is the star label $y_i = \{0, 1\}$ for background and signal stars, respectively. Thus, it can be used to determine the upper limit detectability for this particular signal model, assuming training statistics are not a limiting factor.

The binary star classification model is defined as a parametric function using a deep neural network

$$\Phi(\mathbf{X}_i|\mathbf{w}) \rightarrow y_i' \in [0, 1], \tag{6}$$

which can be optimized by tuning the weights $\mathbf{w}$ to minimize a classification loss function. As before, the hyperparameters of the neural network were chosen based on a manual optimization, rather than a dedicated hyperparameter scan which is left for a subsequent study. We use two hidden layers with 256 elements each, the SELU activation function and dropout with a coefficient of $p = 0.3$. The dropout regularization (Srivastava et al., 2014) limits the amount of overtraining. Finally, we use the focal loss, which is a modification of the binary cross entropy loss, originally proposed for rare object detection in Lin et al. (2017), and is defined as the following sum over the total number of star particles N_{star} in the dataset:

$$L = \sum_{i=1}^{N_{\mathrm{star}}} -y_i\alpha(1-y_i')^{\gamma}\log(y_i') - (1-y_i)(1-\alpha)y_i'^{\gamma}\log(1-y_i'), \tag{7}$$

where α and γ are empirical factors that adjust the weight of easy-to-classify background-like examples in the loss. We choose $\alpha = 0.25, \gamma = 2.0$ based on the defaults introduced in Lin et al. (2017). By this construction, the model output y_i' for star i is a continuous value between 0 and 1 that can be interpreted as a test statistic for the star being labeled as signal.

As before, we use `m12m` for the optimization, `m12i` for the validation, while `m12f` is used for testing. As shown by the orange line in Fig. 7, the supervised binary classifier has a performance comparable to random selection on this dataset. The negligible sensitivity of the classifier compared to the autoencoder is to be expected due to the very low number of independent signal stars (a few thousand stars per galaxy associated with less than a hundred subhalos).

3.2. Feasibility in synthetic Gaia survey

In this section, we investigate if the halo-associated stars are detectable in the synthetic Gaia surveys derived from the very same simulations, that is, under the effects of extinction, partial measurement of the radial velocity v_r and measurement errors. This is done by searching for dark subhalos on the reduced synthetic surveys described in Section 2.2. The goal is again to select candidate stars which are likely to be perturbed by a nearby dark subhalo, such that they could potentially be further analyzed with more detailed approaches. The search for dark subhalos in the reduced Gaia-like catalogs differs from the previous section on two fronts. On the one hand, the mock observed stars in the synthetic Gaia datasets are divided into patches using the hierarchical pixelization algorithm HEALPY (Zonca et al., 2019; Górski et al., 2005) with a pixel level 6. This allows to process the data in manageable subsets in a physically meaningful fashion. In addition, as the halo-associated stars are located in well-defined, localized regions in the sky, we avoid using the absolute right ascension and declination coordinates to unfairly bias the model. Instead, we compute the positional information with respect to

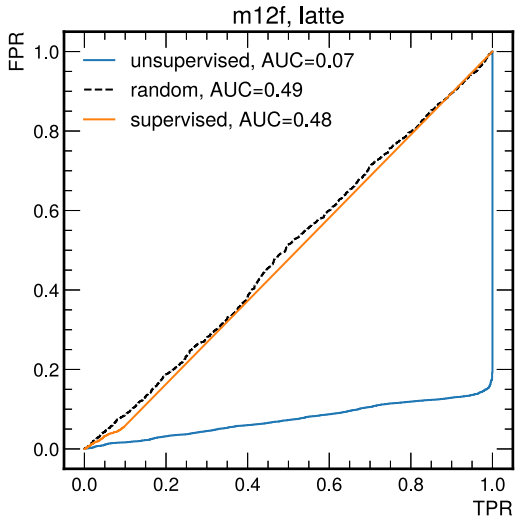


Fig. 7. The true positive rate (horizontal axis) vs. the false positive rate (vertical axis). The blue line depicts the classification performance when the reconstruction loss L_b is used as a discriminator between the signal and background labels on the star particles from the `m12f` galaxy. The orange line shows the binary classifier performance as evaluated on the same galaxy. By construction, we observe no distinction for random stars, while the unsupervised model distinguishes halo-associated stars from the background using a combination of the position and velocity information. The bad performance of the supervised classifier is expected as the number of signal stars is very low compared to the background.

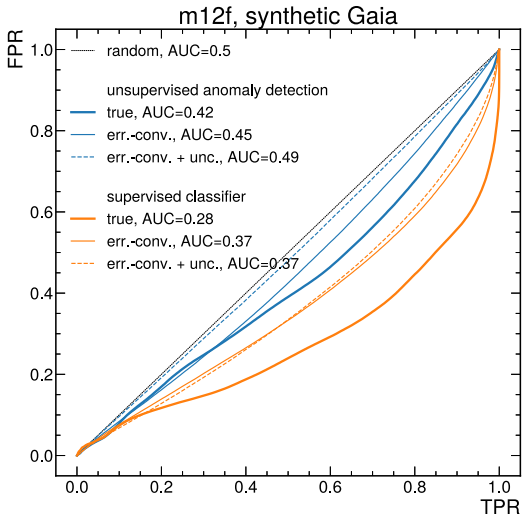


Fig. 8. The true positive rate versus the false positive rate after evaluating the anomaly-detection (blue) and binary classification (orange) models on the synthetic Gaia dataset `m12f`. We compare the model sensitivity with true inputs and error-convolved inputs with and without knowledge of the uncertainties.

the pixel center. For a Gaia DR2-like dataset, a pixel can contain up to $\simeq 2 \times 10^4$ stars.

On the other hand, the input feature of each observed star is different. For each synthetic dataset realization $g \in \{\texttt{m12f}, \texttt{m12f}, \texttt{m12m}\}$, $l \in \{\texttt{LSR0}, \texttt{LSR1}, \texttt{LSR2}\}$, we then have a list of (star observation, label) pairs

$$\mathrm{D}_{g,l} = [(\mathbf{X}_i, y_i), \ldots].$$

Each stellar feature vector $\mathbf{X}$ consists of the following astrometric observables:

- parallax p [mas],
- the right ascension with respect to the pixel center $\Delta\alpha$ [deg],
- the declination with respect to the pixel center $\Delta\delta$ [deg],
- the proper motion in the right ascension direction (multiplied by $\cos\delta$) μ_α^* [mas/year],
- the proper motion in the declination μ_δ [mas/year] and
- the radial velocity v_r [km/s].

These observables are available with estimated uncertainties, resulting in 12 input features. Features which are not always measured, such as the radial velocities, are filled with a placeholder value (numerically set to zero) for a consistent numerical treatment in the neural network model.

The anomaly detection model was trained for 200 epochs, while the classification model for 50 epochs. As before, we use `m12m` for the optimization, `m12i` for the validation, while `m12f` is used for testing. The training and testing is done on stars from all three LSRs simultaneously. Overall, as summarized in Table 1, the optimization, testing and final evaluation is carried out on nearly 1.5 billion mock stars, of which less than 0.01% are identified as signal, resulting in extreme class imbalance as well as an overall low number of independent signal samples.

The sensitivity of the anomaly detection and classifier methods for identifying halo-associated stars in the synthetic Gaia dataset can be seen in Fig. 8. As for the Latte runs, the `m12f` dataset was never used in the optimization. We observe that the binary classification distinguishes between the halo-associated and background stars at a non-negligible level, with a FPR of $\simeq$35% at a TPR of $\simeq$50%. On the other hand, the anomaly detection approach, where we only attempt to learn the background distribution, does not differ significantly from a purely random selection in the synthetic survey.

4. Discussion

Based on the results presented in the last section, we conclude that the halo-associated star particles in the galaxy simulations adopted in this study have a distinguishable distribution in 6-dimensional phase-space (positions and momenta). The anomaly detection method is able to correctly identify halo-associated stars, where as the supervised binary classification does not perform well in the Latte simulations due to the very low signal statistics. When going from the idealized simulation to the synthetic survey, we observe, on the one hand, a mildly better performance for the supervised binary classifier (AUC 0.48 $\rightarrow$ 0.37). On the other hand, the unsupervised anomaly detection performs significantly worse (AUC 0.07 $\rightarrow$ 0.45).

It is important to note that even though the synthetic survey is derived from the Latte simulation, the results of both cases are not comparable, as multiple experimental and observational effects affect the synthetic dataset. In particular, there are the following two uncontrollable (at least for us) and random phenomena involved. First, the synthetic survey performance is limited by the sampling process that generates a population of synthetic stars from each star particle in the simulation. Thus introducing a smearing scale that might dilute the signal. Second, you can get lucky (or unlucky) with the selection of LSRs. This introduces an observational bias since, by chance, more or less subhalos might be included in the footprint of the synthetic survey. On top of these, there is firstly the effect of simulated

data uncertainties and secondly the fact that in each synthetic survey less than 0.5% of the stars have measured radial velocities (see Table 1), thus reducing the kinematic data from 6 to 5 dimensions. Finally, in the synthetic survey, the stellar disc was excluded in order to reduce the volume of data.

We studied the effect of simulated measurement errors and the inclusion of radial velocities by redoing the synthetic Gaia analysis using the *true* astrometric inputs. As seen in Fig. 8, the performance improves when changing from error-convolved to true values, but does not arrive to the one in the Latte simulation. Therefore, neither data uncertainties nor the absence of measured radial velocities have a major impact in driving the performance difference between the Latte simulation and the synthetic surveys. Furthermore, as also seen from Fig. 8, providing information on the simulated uncertainties of the error-convolved values to the model does not significantly affect the supervised classification sensitivity, while it somewhat reduces the sensitivity of the autoencoder-based anomaly detection approach. The latter may be attributable to the increased difficulty of encoding 12 instead of 6 inputs.

Note that the name "*true* inputs" is misleading, since each synthetic star in phase-space has been sampled from a one-dimensional kernel centered on the generating star particle in position and velocity space (Sanderson et al., 2020). We argue that the main difference in performance might be caused by this sampling process that introduces a smearing scale of the order of 0.7 kpc in position and roughly 10 km/s in velocity.[5] Since the change of a star's velocity due to the encounter with a subhalo of mass $M_{\rm sh}$ scales as $\sim 0.5\text{–}1\ \mathrm{km\,s^{-1}}\left(M_{\rm sh}/10^8\,\mathrm{M_\odot}\right)^{2/3}$ (Feldmann and Spolyar, 2015) and kinematic perturbations are partially washed out below the smearing scale, the sampling process causes significant changes in the phase-space distribution of synthetic stars. It is therefore expected to dominate over the luck factor and the removal of the stellar disk in explaining the difference in performance between the idealized Latte simulation and the Gaia-like surveys.

Finally, we would like to highlight that to thoroughly investigate the above conclusion we need a set of dedicated simulations, where each possible effect can be turned on in sequence and can be easily disentangled. Given the scope of the additional studies required, we are studying this in a follow-up paper.

5. Summary and conclusions

Machine learning (ML) techniques, either alone or combined with classical methods, have been demonstrated to be helpful in uncovering new structures in Gaia-scale datasets (e.g. Necib et al. (2020)). Dark subhalos are among the most challenging substructures to search for. In this paper, we study the detectability of dark subhalos by means of ML in three MW-like galaxies and in nine synthetic Gaia DR2 surveys. Rather than attempting to pinpoint the exact subhalo locations and determine their properties, we attempt to identify candidate stars that are likely to be close to a subhalo on a statistical basis.

We have first correlated star particles in the simulated galaxies and mock stars in the synthetic catalogs with the position of dark subhalos found by the Amiga halo Finder (AHF). In Section 3.1 we then tested the feasibility of an anomaly detection and a binary classification algorithm against simulated galaxies to detect the phase-space imprint in stellar halo stars of nearby subhalos. The first algorithm builds a likelihood function of the background star particles and is able to correctly identify 80% of signal stars while misclassifying as signal 15% of background particles. On the other hand, the binary classifier does not perform well due to the very low signal statistics. We concluded that the distribution function of the 6-dimensional phase-space coordinates of signal and background star particles are distinguishable in the MW-like galaxies used in this work. Therefore, on a statistical basis, position and velocity information can be combined into a statistical discriminator for the halo-associated signal.

Finally, we have tested the feasibility of our algorithms in Gaia DR2-like surveys in Section 3.2. The anomaly detection approach has no sensitivity to distinguish between signal and background stars, while the binary classification algorithm is able to select 50% of signal stars while wrongly identifying 15% of background stars as signal. Although the binary classification shows a mild sensitivity, overall both approaches are of limited effectiveness in the synthetic Gaia survey. Although the results above are not directly comparable, our hypothesis is that the sampling process that generates a population of synthetic stars from each star particle mainly causes the difference in performance between the Latte simulation and the synthetic surveys. A thorough investigation of this conclusion is left to future work, leaving the use of ML-based tools as a new way to quantitatively study the effects of dynamical perturbations of DM subhalos as the main message of this paper.

A number of subsequent improvements to the methodology are possible. In the above analysis, all the observed stars were treated independently of each other. Local correlations, density or clustering were not taken into account, which could potentially limit the sensitivity of the method used so far. As an example, novel approaches based on density-based clustering have been employed for open clusters (Castro-Ginard et al., 2018) and may be interesting to study here for dark subhalos. Clustering can also be combined with unsupervised deep learning for anomaly detection (Mikuni and Canelli, 2021). Another possible approach is the direct search for overdensities by comparing signal and sideband regions, which has been so far demonstrated for stellar streams, but could potentially be studied also for dark subhalos (Shih et al., 2021b). Furthermore, in order to understand the potential sensitivity of the method, simulated datasets with a known DM distribution were used. However, the halo distribution in these is fixed, and the number of actual simulated halos in the potentially visible region is limited. Additional simulated datasets with a varying halo distribution could be helpful to establish sensitivity dependence of a potential method on halo mass and distance from the galactic center.

Finally, Rubin/LSST will provide a deeper map of the Galactic stellar halo of the Milky Way compared to that of Gaia. For main-sequence stars, Rubin/LSST is expected to achieve a tangential velocity precision of $\mathcal{O}(10$ km/s) up to Galactocentric distances of 20–30 kpc, increasing up to $\sim$300 km/s at distances of roughly 60 kpc (Abell et al., 2009; Ivezić et al., 2019). For the MW-like galaxies adopted in our analysis, we find dark subhalos with masses $\sim 2\times 10^7\,\mathrm{M_\odot}$ ($\sim 2\times 10^8\,\mathrm{M_\odot}$) within 20–30 kpc (60 kpc) from the Galactic center. These subhalos induced velocity changes in the neighboring stars of $\sim$0.5 km/s ($\sim$2 km/s). Since these velocities changes are below the kinematic precision expected to be achieved with the Rubin/LSST, it will be a challenge to detect the kinematic effect of subhalos with this upcoming telescope.

CRediT authorship contribution statement

A. Bazarov: Software, Investigation. **M. Benito:** Conceptualization, Methodology, Software, Supervision, Writing. **G. Hütsi:** Methodology, Supervision, Writing. **R. Kipper:** Methodology, Supervision, Writing. **J. Pata:** Conceptualization, Methodology, Software, Supervision, Data curation, Writing. **S. Põder:** Software, Investigation, Writing.

[5] The smearing scales are defined as the standard deviation of the differences between the position or velocity of each halo-associated synthetic star and that of the parent star particle from which it was generated.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Estonian Research Council, Estonia grants PSG700, PRG803, PRG1006, PUTJD907 and MOBTT5, MOBTP187, and by the European Regional Development Fund through the CoE program grant TK133. We would like to thank Martti Raidal for suggesting to us the idea of using large-scale ML approaches on the Gaia dataset, and for advice and support throughout this project. We also thank the referees for their time, effort and constructive input.

References

Abell, P.A., et al., 2009. (Lsst science, LSST project). In: LSST Science Book. Version 2.0 arXiv:0912.0201.

Ackermann, M., Albert, A., Baldini, L., Ballet, J., Barbiellini, G., Bastieri, D., Bechtol, K., Bellazzini, R., Blandford, R.D., Bloom, E.D., Bonamente, E., Borgland, .A.W., Bottacini, E., Brandt, T.J., Bregeon, J., Brigida, M., Bruel, P., Buehler, R., Burnett, T.H., Caliandro, G.A., Cameron, R.A., Caraveo, P.A., Casandjian, J.M., Cecchi, C., Charles, E., Chiang, J., Ciprini, S., Claus, R., Cohen-Tanugi, J., Conrad, J., Cutini, S., de Palma, F., Dermer, C.D., Digel, S.W., Silva, E.d.C.e., Drell, P.S., Drlica-Wagner, A., Essig, R., Falletti, L., Favuzzi, C., Fegan, S.J., Focke, W.B., Fukazawa, Y., Funk, S., Fusco, P., Gargano, F., Germani, S., Giglietto, N., Giordano, F., Giroletti, M., Glanzman, T., Godfrey, G., Grenier, I.A., Guiriec, S., Gustafsson, M., Hadasch, D., Hayashida, M., Hou, X., Hughes, R.E., Johnson, R.P., Johnson, A.S., Kamae, T., Katagiri, H., Kataoka, J., Knödlseder, J., Kuss, M., Lande, J., Latronico, L., Lee, S.H., Lionetto, A.M., Garde, M.Llena., Longo, F., Loparco, F., Lovellette, M.N., Lubrano, P., Mazziotta, M.N., McEnery, J.E., Michelson, P.F., Mitthumsiri, W., Mizuno, T., Moiseev, A.A., Monte, C., Monzani, M.E., Morselli, A., Moskalenko, I.V., Murgia, S., Naumann-Godo, M., Norris, J.P., Nuss, E., Ohsugi, T., Okumura, A., Orlando, E., Ormes, J.F., Ozaki, M., Paneque, D., Pelassa, V., Pierbattista, M., Piron, F., Pivato, G., Porter, T.A., Rainò, S., Rando, R., Razzano, M., Reimer, A., Reimer, O., Ritz, S., Sadrozinski, H.F.W., Sehgal, N., Sgrò, C., Siskind, E.J., Spinelli, P., Strigari, L., Suson, D.J., Tajima, H., Takahashi, H., Tanaka, T., Thayer, J.G., Thayer, J.B., Tibaldo, L., Tinivella, M., Torres, D.F., Troja, E., Uchiyama, Y., Usher, T.L., Vandenbroucke, J., Vasileiou, V., Vianello, G., Vitale, V., Waite, A.P., Wang, P., Winer, B.L., Wood, K.S., Yang, Z., Zalewski, S., Zimmer, S., 2012. Search for dark matter satellites using Fermi-LAT. Astrophys. J. 747, 121. doi:10.1088/0004-637X/747/2/121, arXiv:1201.2691.

Aghanim, N.Planck.Collaboration., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A.J., Barreiro, R.B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J.P., Bersanelli, M., Bielewicz, P., Bock, J.J., Bond, J.R., Borrill, J., Bouchet, F.R., Boulanger, F., Bucher, M., Burigana, C., Butler, R.C., Calabrese, E., Cardoso, J.F., Carron, J., Challinor, A., Chiang, H.C., Chluba, J., Colombo, L.P.L., Combet, C., Contreras, D., Crill, B.P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J.M., Di Valentino, E., Diego, J.M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, H.K., Fantaye, Y., Farhang, M., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Fraisse, A.A., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga, K., Génova-Santos, R.T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K.M., Gratton, S., Gruppuso, A., Gudmundsson, J.E., Hamann, J., Handley, W., Hansen, F.K., Herranz, D., Hildebrandt, S.R., Hivon, E., Huang, Z., Jaffe, A.H., Jones, W.C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T.S., Knox, L., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lagache, G., Lamarre, J.M., Lasenby, A., Lattanzi, M., Lawrence, C.R., Le Jeune, M., Lemos, P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P.B., Lilley, M., Lindholm, V., López-Caniego, M., Lubin, P.M., Ma, Y.Z., Macías-Pérez, J.F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P.G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J.D., Meinhold, P.R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M.A., Molinari, D., Montier, L., Morgante, G., Moss, A., Natoli, P., Nørgaard-Nielsen, H.U., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Peiris, H.V., Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J.L., Rachen, J.P., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño Martín, J.A., Ruiz-Granados, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E.P.S., Sirignano, C., Sirri, G., Spencer, L.D., Sunyaev, R., Suur-Uski, A.S., Tauber, J.A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valenziano, L., Valiviita, J., Tent, B.Van., Vibert, L., Vielva, P., Villa, F., Vittorio, N., Wandelt, B.D., Wehus, I.K., White, M., White, A., Zonca, A., 2020. Planck 2018 results. VI. Cosmological parameters. Astron. Astrophys. 641, A6. doi:10.1051/0004-6361/201833910, arXiv:1807.06209.

Baghram, S., Afshordi, N., Zurek, K.M., 2011. Prospects for detecting dark matter halo substructure with pulsar timing. Phys. Rev. D 84, 043511. doi:10.1103/PhysRevD.84.043511, arXiv:1101.5487.

Baldi, P., Hornik, K., 1989. Neural networks and principal component analysis: Learning from examples without local minima. Neural Netw. 2, 53–58. doi:10.1016/0893-6080(89)90014-2.

Banik, N., Bertone, G., Bovy, J., Bozorgnia, N., 2018. Probing the nature of dark matter particles with stellar streams. J. Cosmol. Astropart. Phys. 2018, 061. doi:10.1088/1475-7516/2018/07/061, arXiv:1804.04384.

Benito, M., Criado, J.C., Hütsi, G., Raidal, M., Veermäe, H., 2020. Implications of Milky Way substructures for the nature of dark matter. Phys. Rev. D 101, 103023. doi:10.1103/PhysRevD.101.103023, arXiv:2001.11013.

Blumenthal, G.R., Faber, S.M., Primack, J.R., Rees, M.J., 1984. Formation of galaxies and large-scale structure with cold dark matter. Nature 311, 517–525. doi:10.1038/311517a0.

Bonaca, A., Hogg, D.W., Price-Whelan, A.M., Conroy, C., 2019. The spur and the gap in GD-1: Dynamical evidence for a dark substructure in the milky way halo. Astrophys. J. 880, 38. doi:10.3847/1538-4357/ab2873, arXiv:1811.03631.

Bovy, J., Erkal, D., Sanders, J.L., 2017. Linear perturbation theory for tidal streams and the small-scale CDM power spectrum. Mon. Not. R. Astron. Soc. 466, 628–668. doi:10.1093/mnras/stw3067, arXiv:1606.03470.

Brehmer, J., Mishra-Sharma, S., Hermans, J., Louppe, G., Cranmer, K., 2019. Mining for dark matter substructure: Inferring subhalo population properties from strong lenses with machine learning. Astrophys. J. 886, 49. doi:10.3847/1538-4357/ab4c41, arXiv:1909.02005.

Bringmann, T., 2009. Particle models and the small-scale structure of dark matter. New J. Phys. 11, 105027. doi:10.1088/1367-2630/11/10/105027, arXiv:0903.0189.

Buckley, M.R., Hooper, D., 2010. Dark matter subhalos in the Fermi first source catalog. Phys. Rev. D 82, 063501. doi:10.1103/PhysRevD.82.063501, arXiv:1004.1644.

Buschmann, M., Kopp, J., Safdi, B.R., Wu, C.L., 2018. Stellar wakes from dark matter subhalos. Phys. Rev. Lett. 120, 211101. doi:10.1103/PhysRevLett.120.211101, arXiv:1711.03554.

Calore, F., Hütten, M., Stref, M., 2019. Gamma-ray sensitivity to dark matter subhalo modelling at high latitudes. Galaxies 7 (90), doi:10.3390/galaxies7040090, arXiv:1910.13722.

Carlberg, R.G., 2012. Dark matter sub-halo counts via star stream crossings. Astrophys. J. 748, 20. doi:10.1088/0004-637X/748/1/20, arXiv:1109.6022.

Castro-Ginard, A., Jordi, C., Luri, X., Julbe, F., Morvan, M., Balaguer-Núñez, T., 2018. A new method for unveiling open clusters in gaia-new nearby open clusters confirmed by dr2. Astron. Astrophys. 618, A59.

Clark, H.A., Lewis, G.F., Scott, P., 2016. Investigating dark matter substructure with pulsar timing– I. Constraints on ultracompact minihaloes. Mon. Not. R. Astron. Soc. 456, 1394–1401. doi:10.1093/mnras/stv2743, arXiv:1509.02938. Erratum:; 2017. Mon. Not. Roy. Astron. Soc. 464, 2468.

Coronado-Blázquez, J., Doro, M., Sánchez-Conde, M.A., Aguirre-Santaella, A., 2021. Sensitivity of the Cherenkov Telescope Array to dark subhalos. Phys. Dark Univ. 32, 100845. doi:10.1016/j.dark.2021.100845, arXiv:2101.10003.

Coronado-Blázquez, J., Sánchez-Conde, M.A., Di Mauro, M., Aguirre-Santaella, A., Ciucă, I., Domínguez, A., Kawata, D., Mirabal, N., 2019a. Spectral and spatial analysis of the dark matter subhalo candidates among Fermi Large Area Telescope unidentified sources. J. Cosmol. Astropart. Phys. 2019, 045. doi:10.1088/1475-7516/2019/11/045, arXiv:1910.14429.

Coronado-Blázquez, J., Sánchez-Conde, M.A., Domínguez, A., Aguirre-Santaella, A., Di Mauro, M., Mirabal, N., Nieto, D., Charles, E., 2019b. Unidentified gamma-ray sources as targets for indirect dark matter detection with the Fermi-Large Area Telescope. J. Cosmol. Astropart. Phys. 2019, 020. doi:10.1088/1475-7516/2019/07/020, arXiv:1906.11896.

Delos, M.S., Linden, T., 2022. Dark matter microhalos in the solar neighborhood: Pulsar timing signatures of early matter domination. Phys. Rev. D 105, 123514. doi:10.1103/PhysRevD.105.123514, arXiv:2109.03240.

Díaz Rivero, A., Dvorkin, C., Cyr-Racine, F.Y., Zavala, J., Vogelsberger, M., 2018. Gravitational lensing and the power spectrum of dark matter substructure: Insights from the ETHOS N -body simulations. Phys. Rev. D 98, 103517. doi:10.1103/PhysRevD.98.103517, arXiv:1809.00004.

Feldmann, R., Spolyar, D., 2015. Detecting dark matter substructures around the Milky Way with Gaia. Mon. Not. R. Astron. Soc. 446, 1000–1012. doi:10.1093/mnras/stu2147, arXiv:1310.2243.

Garrison-Kimmel, S., Wetzel, A., Bullock, J.S., Hopkins, P.F., Boylan-Kolchin, M., Faucher-Giguère, C.A., Kereš, .D., Quataert, E., Sanderson, R.E., Graus, A.S., et al., 2017. Not so lumpy after all: modelling the depletion of dark matter subhaloes by milky way-like galaxies. Mon. Not. R. Astron. Soc. 471, 1709–1727.

Gilman, D., Birrer, S., Treu, T., Nierenberg, A., Benson, A., 2019. Probing dark matter structure down to 10^7 solar masses: flux ratio statistics in gravitational lenses with line-of-sight haloes. Mon. Not. R. Astron. Soc. 487, 5721–5738. doi:10.1093/mnras/stz1593, arXiv:1901.11031.

Górski, K.M., Hivon, E., Banday, A.J., Wandelt, B.D., Hansen, F.K., Reinecke, M., Bartelmann, M., 2005. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. Astrophys. J. 622, 759–771. doi:10.1086/427976, arXiv:arXiv:astro-ph/0409513.

Hezaveh, Y., Dalal, N., Holder, G., Kisner, T., Kuhlen, M., Perreault Levasseur, L., 2016. Measuring the power spectrum of dark matter substructure using strong gravitational lensing. J. Cosmol. Astropart. Phys. 2016, 048. doi:10.1088/1475-7516/2016/11/048, arXiv:1403.2720.

Hopkins, P.F., 2015. A new class of accurate, mesh-free hydrodynamic simulation methods. Mon. Not. R. Astron. Soc. 450, 53–110. doi:10.1093/mnras/stv195, arXiv:1409.7395.

Hopkins, P.F., Wetzel, A., Kereš, C.A., Quataert, E., Boylan-Kolchin, M., Murray, N., Hayward, C.C., Garrison-Kimmel, S., Hummels, C., et al., 2018. Fire-2 simulations: physics versus numerics in galaxy formation. Mon. Not. R. Astron. Soc. 480, 800–863.

Ibata, R.A., Lewis, G.F., Irwin, M.J., Quinn, T., 2002. Uncovering cold dark matter halo substructure with tidal streams. Mon. Not. R. Astron. Soc. 332, 915–920. doi:10.1046/j.1365-8711.2002.05358.x, arXiv:astro-ph/0110690.

Ivezić, v., et al., (LSST), 2019. LSST: from science drivers to reference design and anticipated data products. Astrophys. J. 873, 111. doi:10.3847/1538-4357/ab042c, arXiv:0805.2366.

Karukes, E.V., Benito, M., Iocco, F., Trotta, R., Geringer-Sameth, A., 2020. A robust estimate of the Milky Way mass from rotation curve data. J. Cosmol. Astropart. Phys. 2020, 033. doi:10.1088/1475-7516/2020/05/033, arXiv:1912.04296.

Kashiyama, K., Oguri, M., 2018. Detectability of small-scale dark matter clumps with pulsar timing arrays. arXiv:1801.07847.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv e-prints, arXiv:1412.6980.

Kipper, R., Tenjes, P., Tempel, E., de Propris, R., 2021. Non-equilibrium in the solar neighbourhood using dynamical modelling with Gaia DR2. Mon. Not. R. Astron. Soc. 506, 5559–5572. doi:10.1093/mnras/stab2104, arXiv:2106.07076.

Kipper, R., Tenjes, P., Tuvikene, T., Ganeshaiah Veena, P., Tempel, E., 2020. Quantifying torque from the milky way bar using gaia dr2. Mon. Not. R. Astron. Soc. 494, 3358–3367.

Kitayama, T., Yoshida, N., 2005. Supernova explosions in the early universe: Evolution of radiative remnants and the halo destruction efficiency. Astrophys. J 630, 675–688. doi:10.1086/432114, arXiv:astro-ph/0505368.

Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S., 2017. Self-normalizing neural networks. Adv. Neural Inf. Process. Syst. (30).

Knollmann, S.R., Knebe, A., 2009. Ahf: Amiga's halo finder. Astrophys. J. Suppl. Ser. 182 (608).

Lin, T., Goyal, P., Girshick, R.B., He, K., Dollá, P., 2017. Focal loss for dense object detection. CoRR abs/1708.02002 arXiv:1708.02002.

Mikuni, V., Canelli, F., 2021. Unsupervised clustering for collider physics. Phys. Rev. D 103, 092007.

Mirabal, N., Bonaca, A., 2021. Machine-learned dark matter subhalo candidates in the 4FGL-DR2: search for the perturber of the GD-1 stream. J. Cosmol. Astropart. Phys. 2021, 033. doi:10.1088/1475-7516/2021/11/033, arXiv:2105.12131.

Moliné, Á., Sánchez-Conde, M.A., Palomares-Ruiz, S., Prada, F., 2017. Characterization of subhalo structural properties and implications for dark matter annihilation signals. Mon. Not. R. Astron. Soc. 466, 4974–4990. doi:10.1093/mnras/stx026, arXiv:1603.04057.

Necib, L., Ostdiek, B., Lisanti, M., Cohen, T., Freytsis, M., Garrison-Kimmel, P.F., Wetzel, A., Sanderson, R., 2020. Evidence for a vast prograde stellar stream in the solar vicinity. Nat. Astron. 4, 1078–1083. doi:10.1038/s41550-020-1131-2, arXiv:1907.07190.

Neyman, J., Pearson, E.S., 1992. On the problem of the most efficient tests of statistical hypotheses. In: Breakthroughs in Statistics. Springer, pp. 73–108.

Oñorbe, J., Garrison-Kimmel, S., Maller, A.H., Bullock, J.S., Rocha, M., Hahn, O., 2014. How to zoom: bias, contamination and Lagrange volumes in multimass cosmological simulations. Mon. Not. R. Astron. Soc. 437, 1894–1908. doi:10.1093/mnras/stt2020, arXiv:1305.6923.

Ostdiek, B., Necib, L., Cohen, T., Freytsis, M., Lisanti, M., Garrison-Kimmmel, S., Wetzel, A., Sanderson, R.E., Hopkins, P.F., 2020. Cataloging accreted stars within Gaia DR2 using deep learning. Astron. Astrophys. 636, A75. doi:10.1051/0004-6361/201936866, arXiv:1907.06652.

Read, J.I., Pontzen, A.P., Viel, M., 2006. On the formation of dwarf galaxies and stellar halos. Mon. Not. Roy. Astron. Soc 371, 885–897. doi:10.1111/j.1365-2966.2006.10720.x, arXiv:astro-ph/0606391.

Sakurada, M., Yairi, T., 2014. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: MLSDA'14.

Sanderson, R.E., Wetzel, A., Loebman, S., Sharma, S., Hopkins, P.F., Garrison-Kimmel, S., Faucher-Giguère, C.A., Kereš, D., Quataert, E., 2020. Synthetic gaia surveys from the FIRE cosmological simulations of milky way-mass galaxies. Astrophys. J. 246, 6. doi:10.3847/1538-4365/ab5b9d, arXiv:1806.10564.

Schneider, A., Smith, R.E., Macciò, .A.V., Moore, B., 2012. Non-linear evolution of cosmological structures in warm dark matter models. Mon. Not. R. Astron. Soc. 424, 684–698. doi:10.1111/j.1365-2966.2012.21252.x, arXiv:1112.0330.

Shen, J., Eadie, G.M., Murray, N., Zaritsky, D., Speagle, J.S., Ting, Y.S., Conroy, C., Cargile, P.A., Johnson, B.D., Naidu, R.P., Han, J.J., 2021. The mass of the milky way from the H3 survey. doi:10.3847/1538-4357/ac3a7a, arXiv e-prints, arXiv:2111.09327.

Shih, D., Buckley, M.R., Necib, L., Tamanas, J., 2021a. Via machinae: Searching for stellar streams using unsupervised machine learning. Mon. Not. R. Astron. Soc. doi:10.1093/mnras/stab3372, arXiv:2104.12789.

Shih, D., Buckley, M.R., Necib, L., Tamanas, J., 2021b. Via machinae: Searching for stellar streams using unsupervised machine learning. Mon. Not. R. Astron. Soc. doi:10.1093/mnras/stab3372, arXiv:https://academic.oup.com/mnras/advance-article-pdf/doi/10.1093/mnras/stab3372/41281394/stab3372.pdf. stab3372.

Siegel, E.R., Hertzberg, M.P., Fry, J.N., 2007. Probing dark matter substructure with pulsar timing. Mon. Not. R. Astron. Soc. 382, 879. doi:10.1111/j.1365-2966.2007.12435.x, arXiv:astro-ph/0702546.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Van Tilburg, K., Taki, A.M., Weiner, N., 2018. Halometry from astrometry. JCAP 2018, 041. doi:10.1088/1475-7516/2018/07/041, arXiv:1804.01991.

Vattis, K., Toomey, M.W., Koushiappas, S.M., 2020. Deep learning the astrometric signature of dark matter substructure. arXiv e-prints, arXiv:2008.11577.

Vogelsberger, M., Zavala, J., Cyr-Racine, F.Y., Pfrommer, C., Bringmann, T., Sigurdson, K., 2016. ETHOS - an effective theory of structure formation: dark matter physics as a possible explanation of the small-scale CDM problems. Mon. Not. R. Astron. Soc. 460, 1399–1416. doi:10.1093/mnras/stw1076, arXiv:1512.05349.

Wang, W., Han, J., Cautun, M., Li, Z., Ishigaki, M.N., 2020. The mass of our Milky Way. Sci. China Phys., Mech., Astron. 63, 109801. doi:10.1007/s11433-019-1541-6, arXiv:1912.02599.

Wetzel, A.R., Hopkins, P.F., Kim, J.h., Faucher-Giguère, C.A., Kereš, D., Quataert, E., 2016. Reconciling dwarf galaxies with Λcdm cosmology: Simulating a realistic population of satellites around a milky way-mass galaxy. Astrophys. J. 827, L23. doi:10.3847/2041-8205/827/2/L23, arXiv:1602.05957.

Yoon, J.H., Johnston, K.V., Hogg, D.W., 2011. Clumpy streams from clumpy halos: Detecting missing satellites with cold stellar structures. Astrophys. J. 731, 58. doi:10.1088/0004-637X/731/1/58, arXiv:1012.2884.

Zechlin, H.S., Horns, D., 2012. Unidentified sources in the Fermi-LAT second source catalog: the case for DM subhalos. J. Cosmol. Astropart. Phys. 11, 050. doi:10.1088/1475-7516/2012/11/050, arXiv:1210.3852. Erratum:; 2015. JCAP 02, E01.

Zonca, A., Singer, L., Lenz, D., Reinecke, M., Rosset, C., Hivon, E., Gorski, K., 2019. Healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in python. J. Open Source Softw. 4 (1298), doi:10.21105/joss.01298.

Zybin, K.P., Vysotsky, M.I., Gurevich, A.V., 1999. The fluctuation spectrum cut-off in a neutralino dark matter scenario. Phys. Lett. A 260, 262–268. doi:10.1016/S0375-9601(99)00434-X.

Appendix 2

II

Põder, Sven, Benito, María, Pata, Joosep, Kipper, Rain, Ramler, Heleri, Hütsi, Gert, Kolka, Indrek, and Thomas, Guillaume F. A bayesian estimation of the milky way's circular velocity curve using gaia dr3. *A&A*, 676:A134, 2023

A&A 676, A134 (2023)
https://doi.org/10.1051/0004-6361/202346474

Astronomy & Astrophysics

A Bayesian estimation of the Milky Way's circular velocity curve using *Gaia* DR3

Sven Põder[1,2], María Benito[3,1], Joosep Pata[1], Rain Kipper[3], Heleri Ramler[3], Gert Hütsi[1], Indrek Kolka[3], and Guillaume F. Thomas[4,5]

[1] NICPB, Rävala 10, Tallinn 10143, Estonia
e-mail: sven.poder@kbfi.ee
[2] Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia
[3] Tartu Observatory, University of Tartu, Observatooriumi 1, Tõravere 61602, Estonia
e-mail: mariabenitocst@gmail.com
[4] Instituto de Astrofísica de Canarias, C/Vía Láctea s/n, 38205 La Laguna, Tenerife, Spain
[5] Universidad de La Laguna, Dpto. Astrofísica, Avenida Astrofísico Francisco Sánchez, 38206 La Laguna, Tenerife, Spain

Received 21 March 2023 / Accepted 27 June 2023

ABSTRACT

Aims. Our goal is to calculate the circular velocity curve of the Milky Way, along with corresponding uncertainties that quantify various sources of systematic uncertainty in a self-consistent manner.
Methods. The observed rotational velocities are described as circular velocities minus the asymmetric drift. The latter is described by the radial axisymmetric Jeans equation. We thus reconstruct the circular velocity curve between Galactocentric distances from 5 kpc to 14 kpc using a Bayesian inference approach. The estimated error bars quantify uncertainties in the Sun's Galactocentric distance and the spatial-kinematic morphology of the tracer stars. As tracers, we used a sample of roughly 0.6 million stars on the red giant branch stars with six-dimensional phase-space coordinates from *Gaia* Data Release 3 (DR3). More than 99% of the sample is confined to a quarter of the stellar disc with mean radial, rotational, and vertical velocity dispersions of $(35 \pm 18)\,\mathrm{km\,s^{-1}}$, $(25 \pm 13)\,\mathrm{km\,s^{-1}}$, and $(19 \pm 9)\,\mathrm{km\,s^{-1}}$, respectively.
Results. We find a circular velocity curve with a slope of $0.4 \pm 0.6\,\mathrm{km\,s^{-1}\,kpc^{-1}}$, which is consistent with a flat curve within the uncertainties. We further estimate a circular velocity at the Sun's position of $v_c(R_0) = 233 \pm 7\,\mathrm{km\,s^{-1}}$ and that a region in the Sun's vicinity, characterised by a physical length scale of ∼1 kpc, moves with a bulk motion of $V_{\mathrm{LSR}} = 7 \pm 7\,\mathrm{km\,s^{-1}}$. Finally, we estimate that the dark matter (DM) mass within 14 kpc is $\log_{10} M_{\mathrm{DM}}(R < 14\,\mathrm{kpc})/M_\odot = \left(11.2^{+2.0}_{-2.3}\right)$ and the local spherically averaged DM density is $\rho_{\mathrm{DM}}(R_0) = \left(0.41^{+0.10}_{-0.09}\right)\,\mathrm{GeV\,cm^{-3}} = \left(0.011^{+0.003}_{-0.002}\right)\,M_\odot\,\mathrm{pc^{-3}}$. In addition, the effect of biased distance estimates on our results is assessed.

Key words. Galaxy: kinematics and dynamics – Galaxy: disk – methods: statistical

1. Introduction

The rotation of stars and gas in galactic discs has been extensively used as a kinematical tracer of matter distribution of external galaxies and our own Galaxy, the Milky Way (MW) (Kuzmin 2022). Several recent studies (Mróz et al. 2019; Eilers et al. 2019; Chrobáková et al. 2020; Ablimit et al. 2020; Khanna et al. 2023; Gaia Collaboration 2023; Wang et al. 2022; Zhou et al. 2023) have measured the stellar disc rotation in the MW using stellar data from the *Gaia* satellite (Gaia Collaboration 2016). These studies differ in the samples of stars used as a tracer and/or in the methodology and assumptions employed. Moreover, some of the cited studies provided rotational (azimuthal) velocities, whereas the others presented circular ones. The former are direct measurements and no underlying assumptions are made with respect to the shape or time dependence of the MW's gravitational potential. On the other hand, circular velocities assume a stationary gravitational potential that exhibits axial symmetry. Moreover, these are the velocities that should be used to derive the total or dynamical matter distribution. The modelling assumptions can therefore bias the determination of the total and dark matter content in our Galaxy.

The amount of phase space data currently available is large, and statistics is generally not the limiting factor for studies of the dynamics of the MW stellar disc. The limiting factor is instead systematic errors, such as the Sun's Galactocentric distance or the adoption of incorrect modelling assumptions. In this paper, we present a Bayesian inference approach to estimate the circular velocity curve of our Galaxy that allows the straightforward incorporation of systematic and statistical sources of uncertainty, which are both treated as nuisance parameters. In this way, we provide, for the first time, a circular velocity curve with errorbars that self-consistently include uncertainties in the Sun's Galactocentric distance and in the spatial-kinematic structure of the stellar disc. Therefore, our uncertain knowledge about astrophysical parameters is propagated through Bayes' theorem to the determination of the circular velocity curve and, subsequently, to the estimation of the dark matter density profile in our Galaxy. Taking into account the uncertainties about how dark matter is distributed in the MW is essential for interpreting the results of dark matter particle searches (see e.g. Benito et al. 2017).

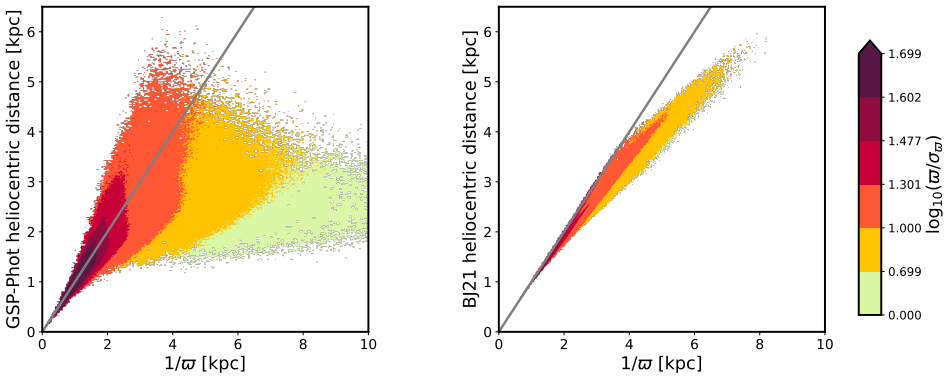


Fig. 1. Bias in different distance estimates. Left: GSP-Phot heliocentric distances as a function of parallax inverse. The colour bar shows the mean parallax quality inside each pixel. The quality cut used in this paper corresponds to $\log_{10}(\varpi/\sigma_\varpi) > 1.301$. Right: same but for photogeo heliocentric distances by Bailer-Jones et al. (2021).

The Bayesian inference approach presented here is a flexible method that models the observed rotational or azimuthal velocity at a given Galactocentric distance as the difference between the circular velocity and the asymmetric drift component. The latter velocity component is obtained from the stationary, axisymmetric radial Jeans equation under the assumption of symmetry above and below the Galactic plane. Observed and modelled azimuthal velocities are then compared by means of the Bayes theorem. The paper is divided as follows. Section 2 describes the data; Sect. 3 presents the Bayesian methodology. The results are presented in Sect. 4, and we conclude in Sect. 5.

2. Data

2.1. Red giant branch stars

We used astrometric data and radial velocities from *Gaia* data release 3 (DR3) for 665 660 stars in the red giant branch (RGB). These stars are old and have relatively large velocity dispersion. Thus, they are less susceptible to perturbations. They are also bright enough to have measured radial velocities out to large distances. Specifically, we used the same sample of almost six million RGB stars as in Gaia Collaboration (2023) to which we performed additional spatial and kinematic cuts.

In the following we briefly describe how the RGB sample was obtained and we refer readers to the original paper Gaia Collaboration (2023) for a thorough description. Red giants are identified as stars with effective temperatures between 3000 K and 5500 K and surface gravity satisfying the condition $\log g < 3$. Both stellar parameters are provided as data products in *Gaia* DR3 (Andrae et al. 2023). Using these first selection criteria, we obtained 11 576 957 sources. We then selected RGB stars with good astrometric data as measured by the fidelity parameter f_a given in Rybizki et al. (2022) and we removed stars with $f_a \leq 0.5$, thus remaining with 6 586 329 stars. After this, we performed a cut in height above and below the Galactic plane, $|z| < 0.2$ kpc (which removes ca. 4.5M stars), a cut in Galactocentric distances $5\,\mathrm{kpc} < R < 14\,\mathrm{kpc}$ (removing ca. 84k stars), and a cut in heliocentric velocity $|\boldsymbol{v} - \boldsymbol{v}_\odot| < 210\,\mathrm{km\,s^{-1}}$[1]. The latter cut removed roughly 28k stars. The z and velocity cuts are applied to remove halo stars (Helmi et al. 2018; Thomas et al. 2019). The cut in height also avoids large density and velocity gradients in the z-coordinate, thus making the derivatives with respect to z in the Jeans equation negligible (see Eq. (3)). We note that the scale height of the thin disc is roughly 250 pc (Bland-Hawthorn & Gerhard 2016).

We used GSP-Phot distances (Andrae et al. 2023) instead of the inverse of the parallax due to noisy parallax measurements. As shown in the left panel of Fig. 1, the GSP-Phot distances are significantly underestimated at large distances from the Sun (see also Andrae et al. 2023; Fouesneau et al. 2023). This would lead to artificially lower circular velocities and thus to a steeper slope of the estimated circular velocity curve. To reduce this bias, we imposed a tight constraint on the quality of the parallax measurements, namely $\varpi/\sigma_\varpi > 20$. This cut-off removed approximately 1.3M stars and suppressed the systematic underestimation of the distances, in fact leading to a slight overestimation. To assess the dependence of our results on inaccurate distances, we further performed the same analysis using a less stringent quality cut on parallaxes and 'photogeo' distances from Bailer-Jones et al. (2021; henceforth referred to as BJ21). Photogeo distances suffer from underestimation when including measurements with $\varpi/\sigma_\varpi \lesssim 20$ (see the right panel of Fig. 1). In the end, we are left with a final sample of 665 660 stars which is shown in Fig. 2.

2.2. Galactocentric frame

2.2.1. Galactocentric transformation

We transformed RGB stars from a heliocentric to a Galactocentric reference frame. We used a right-handed Galactocentric coordinate system (x, y, z), with the Sun located at negative x, y pointing in the direction of the Galactic rotation and z towards the North Galactic Pole. In order to define this new frame and perform the transformation correctly, we assumed solar orbital parameters (R_0, z_0, $U_\odot$, $V_\odot$, $W_\odot$) from contemporary literature. First, we treated the Galactocentric distance R_0 as a nuisance parameter of the analysis in order to account for uncertainties in its determination. In particular, we used a uniform prior range $R_0 \in [7.8{-}8.5]$ kpc that encompasses recent estimates with their corresponding uncertainties (Do et al. 2019; GRAVITY Collaboration 2021, 2019, 2020, 2022; Abuter et al. 2018; Leung et al. 2022). Our intention was not to constrain the value of R_0, but rather to show how the uncertainty in this parameter, which is encoded in the prior, propagates into the circular velocity curve. For this reason, we remained agnostic

[1] $\boldsymbol{v}_\odot$ is defined in Eq. (1).

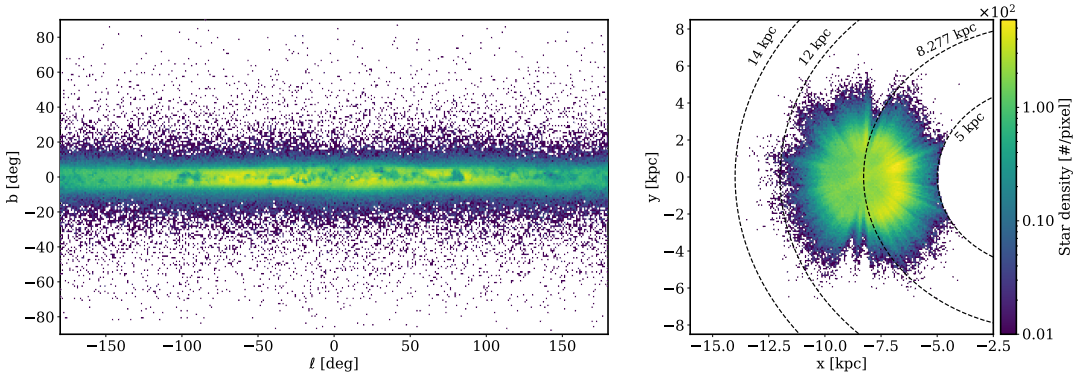


Fig. 2. Spatial distribution of the final red giants sample. Left: distribution in the Galactic longitude (ℓ) and latitude (b) plane. The colour bar shows the number density of stars per pixel. Each pixel has a size of 1 degree in both latitude and longitude. Right: same but projected into the Galactic plane. Each pixel has a size of 0.05 kpc and 0.06 kpc in the x and y coordinates, respectively. In this figure, the Galactic centre is located at (0, 0), the Sun is located at $(-8.277, 0)$ and the rotation of the Galaxy is clockwise. We added dashed, black circles at 5 kpc, $R_0 = 8.277$ kpc, 12 kpc and 14 kpc to ease visualisation. For the transformation to Galactocentric coordinates, we adopt $R_0 = 8.277$ kpc, $z_0 = 25$ pc and $\boldsymbol{v}_\odot = (11.1, 251.5, 8.59)\,\mathrm{km\,s^{-1}}$.

about the actual value of R_0 and adopted a uniform distribution that encompasses the most recent R_0 estimations within 2σ uncertainty. Second, for the height of the Sun over the Galactic plane z_0, we assumed a value of 25 pc (Jurić et al. 2008)[2]. The transformation from spherical coordinates in ICRS to a Galactocentric Cartesian setting was done as described in Johnson & Soderblom (1987) and Hobbs et al. (2018).

Finally, the radial velocity measurements from *Gaia* also allow us to construct the full 3D space velocities and then transform them to a new frame by using the Sun's Galactocentric velocity. Under the assumption that Sg A* is at rest at the Galactic centre, the y and z components of the total solar velocity vector are derived from the proper motion of Sg A*, as measured in Reid & Brunthaler (2020), in combination with the adopted value of R_0. On the other hand, for the x-component of the velocity, we adopted the value from Schönrich et al. (2010). We have not corrected this value for the offset in radial velocity between the radio-to-infrared reference frames determined by the GRAVITY Collaboration (GRAVITY Collaboration 2019, 2020, 2021, 2022) as suggested in Drimmel & Poggio (2018). There are many possible sources for this systematic offset and, in any case, it is compatible with zero at the 2σ level (GRAVITY Collaboration 2022). In this way, we obtain the following vector

$$\boldsymbol{v}_\odot = \begin{bmatrix} U_\odot \\ V_\odot \\ W_\odot \end{bmatrix} = \begin{bmatrix} 11.1 \\ 251.5 \times \left(\frac{R_0}{8.277\,\mathrm{kpc}}\right) \\ 8.59 \times \left(\frac{R_0}{8.277\,\mathrm{kpc}}\right) \end{bmatrix} \mathrm{km\,s^{-1}}, \quad (1)$$

where $U_\odot$, $V_\odot$, $W_\odot$ correspond to the velocity components in the Galactocentric x, y, and z-directions, respectively. As $U_\odot$ corresponds to the radial motion of the Sun towards the Galactic centre, we implicitly assumed that the LSR has no such motion.

Using *Gaia* measurements for right ascension, declination, and radial velocity, and the GSP-Phot distances with a quality parallax cut of $\varpi/\sigma_\varpi > 20$, we transformed the proper motions and radial velocities first to Cartesian velocities in a similar way as was done in Johnson & Soderblom (1987) and Hobbs et al. (2018). Finally, we switched to Galactocentric cylindrical coordinates $(R, \phi, z, v_R, v_\phi, v_z)$.

The left panel of Fig. 3 shows the mean observed rotational velocities in the Galactic plane. The right panel of the same figure depicts the mean axisymmetric radial and azimuthal velocity dispersions. Figure 4 shows the mean azimuthal and radial velocities. In both figures uncertainties are calculated by bootstrap resampling and are given by half the interval between the 16th and 84th percentiles of the corresponding distribution. As shown in the right panel of the last figure, the bulk motion of the stars in the radial direction exhibits an oscillatory pattern with an amplitude of roughly $5\,\mathrm{km\,s^{-1}}$. This was already reported in *Gaia* DR2 (Gaia Collaboration 2018) and might be a kinematic signature of the spiral arms or the result of interaction with a perturber. We leave for future work a careful study of the origin of this intriguing oscillation in the radial velocities.

2.2.2. Binning

We treated the Sun's Galactocentric distance as a free parameter in our analysis. Varying R_0 translates into a variation of the R coordinate of the RGB stars. For this reason, we distributed the RGB sample in bins defined in the adimensional coordinate $x = R/R_0$. This mitigates the shift of the red giants' R coordinate when varying R_0. Thus, for different R_0 values, a given x bin contains approximately the same stars (Benito et al. 2019). Furthermore, we note that by marginalising over the azimuthal coordinate ϕ, the rotational velocity in the Galactic disc is treated as a purely radially dependent observable.

In total, we defined eight bins from $x = 5/8.277$ to $x = 14/8.277$ with a step of $\Delta x = 1/8.277$. Observed rotational (azimuthal) velocities inside each bin approximately follow a Gaussian distribution as shown in Fig. 5. In this figure, we plot the distribution of observed velocities inside two bins: the bin where the distribution deviates most from a Gaussian and a randomly selected bin. The rotational velocity inside each bin is defined by the mean. The associated uncertainties are calculated by bootstrap resamplings and are given by half the interval between the 16th and 84th percentiles of the velocity distributions.

[2] By adopting an alternative Z_0 value of 0 pc, the change in the central circular velocities is smaller than 1%.

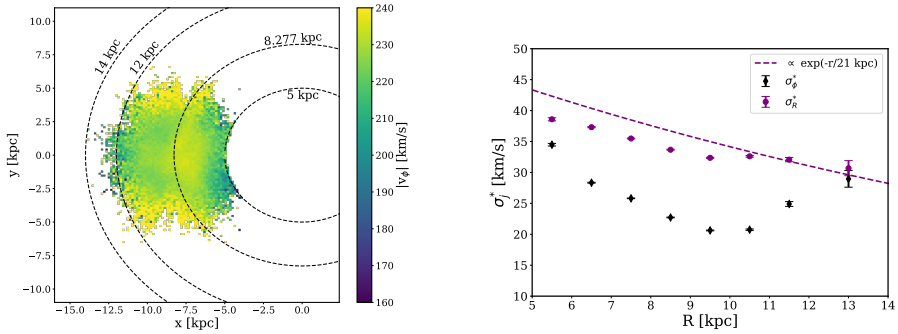


Fig. 3. Distribution of the mean observed rotational velocity v_ϕ inside x–y pixels for our final sample of RGB stars (left). The size of the bin is 150 pc in both x and y coordinates, respectively. We added dashed, black circles at 5 kpc, $R_0 = 8.277$ kpc, 12 and 14 kpc to ease visualisation. Square root of the radial and azimuthal diagonal components of the velocity dispersion tensor in radial bins of 1 kpc (right).

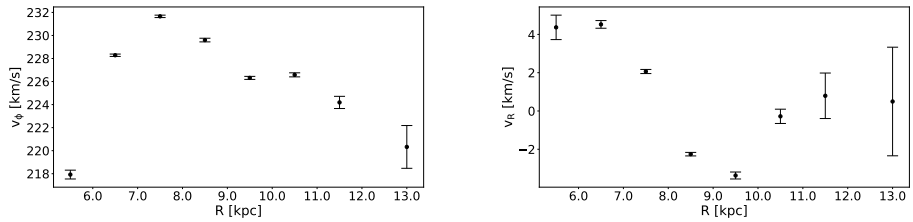


Fig. 4. Mean azimuthal (left) and radial (right) velocities as a function of Galactocentric distance R. As in Fig. 3, errorbars are calculated via bootstrap (see text for details).

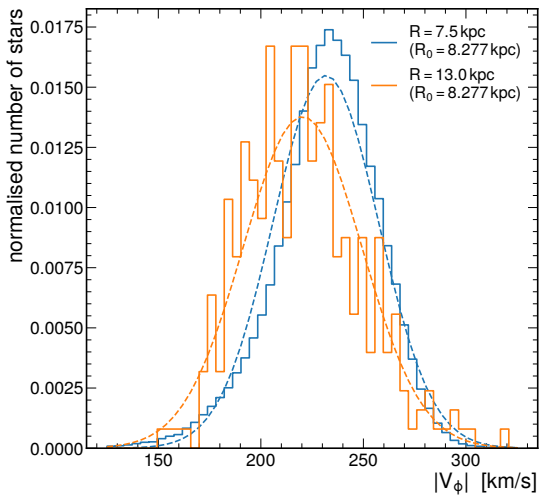


Fig. 5. Distribution of v_ϕ inside the last radial bin where the distribution of values deviates most from a Gaussian distribution (orange) and a randomly selected bin (blue). The dashed lines depict the best-fitted Gaussian of the rotational velocity inside each particular bin.

3. Methodology

3.1. Axisymmetric kinematic model

Inside each radial bin, we modeled the mean rotational or azimuthal velocity as

$$v_{\phi\mathrm{model}} = v_\mathrm{c} - v_\mathrm{a}, \tag{2}$$

where v_c is the circular velocity or the velocity of a star moving in a circular orbit and v_a is the asymmetric drift. The latter accounts for the diffusion of stars in phase-space as the stars orbit the Galaxy and streaming or bulk motions inside the disc. Neglecting large scale non-axisymmetric features, the asymmetric drift component can be obtained from the radial Jeans equation under the assumption that the MW is in a steady-state and has an axisymmetric gravitational potential (Binney & Tremaine 2008). If we further expect the density distribution to be symmetric with respect to the Galactic plane, the Jeans equation takes the following form

$$\frac{R}{\nu}\frac{\partial\left(\nu\overline{v_R^2}\right)}{\partial R} + R\frac{\partial\overline{v_R v_z}}{\partial z} + \overline{v_R^2} - \overline{v_\phi^2} + v_\mathrm{c}^2 = 0. \tag{3}$$

Substituting $\overline{v_\phi^2} = \sigma_\phi^{*2} + (\overline{v_\phi})^2$[3] and assuming $\overline{v_R^2} = \sigma_R^{*2}$, we further obtain

$$\sigma_\phi^{*2} - \sigma_R^{*2} - \frac{R}{\nu}\frac{\partial\left(\nu\sigma_R^{*2}\right)}{\partial R} - R\frac{\partial\overline{v_R v_z}}{\partial z} = v_c^2 - (\overline{v_\phi})^2. \tag{4}$$

We have checked the latter assumption explicitly and observed that the measured values of $\overline{v_R^2}$ are at least two orders of magnitude larger than the measured $\overline{v_R}^2$.

In traditional approaches, circular velocities in radial bins are calculated by plugging numbers into Eq. (4). Circular velocity values between bins are correlated as the radial dependence of the density profile and radial velocity dispersion are described

[3] To avoid confusion between uncertainties and the components of the velocity-dispersion tensor, we add the superscript $*$ to the latter as in Gaia Collaboration (2023).

by exponential functions. In the proposed new approach, we add Eq. (2) to transform the problem into an inference procedure. In this way, we can introduce free parameters such as scale lengths of the exponential functions and/or the Sun's Galactocentric distance R_0. The fitting procedure introduces an additional correlation between the bins. Nonetheless, we found that if we fix all free parameters of the analysis (i.e. scale lengths and R_0), the traditional and new approaches give the same circular velocity curve.

By expanding the difference of squares on the right hand side and introducing Eq. (2), we arrive to the following expression for the asymmetric drift

$$v_\mathrm{a} = \frac{\sigma_R^{*2}}{v_\mathrm{c} + \overline{v_\phi}} \left[\frac{\sigma_\phi^{*2}}{\sigma_R^{*2}} - 1 - \frac{\partial \ln \nu}{\partial \ln R} - \frac{\partial \ln(\sigma_R^{*2})}{\partial \ln R} - \frac{R}{\sigma_R^{*2}} \frac{\partial \overline{v_R v_z}}{\partial z} \right]. \quad (5)$$

The sum $v_\mathrm{c} + \overline{v_\phi}$ in the denominator is often approximated as $\approx 2v_\mathrm{c}$ (Binney & Tremaine 2008). Nonetheless, we leave it as it is and estimate $\overline{v_\phi}$ as the mean rotational velocity inside each bin. In addition to this, the diagonal components of the velocity-dispersion tensor σ_ϕ^{*2} and σ_R^{*2} are also calculated directly from the data, and they correspond to the variance of the azimuthal and radial velocity in the bin respectively. For the 3rd component inside the brackets of Eq. (5), the number density distribution ν is described by an exponential profile, namely $\nu \propto \exp(-R/h_R)$ with h_R the disc scale radius. Notice that in the 4th component we describe σ_R^* as $\sigma_R^* \propto \exp(-R/h_\sigma)$, where h_σ is the scale length of the radial velocity dispersion. Finally, after taking the derivatives with respect to $\ln R$ in (5), we are left with the following equation

$$v_\mathrm{a} = \frac{\sigma_R^{*2}}{v_\mathrm{c} + \overline{v_\phi}} \left[\frac{\sigma_\phi^{*2}}{\sigma_R^{*2}} - 1 + R\left(\frac{1}{h_r} + \frac{2}{h_\sigma} \right) \right]. \quad (6)$$

We neglect the last term of (5) in our axisymmetric treatment of the rotation curve derivation as $\overline{v_R v_z} \approx 0$. This is motivated because the radial and vertical motions are expected to decouple for circular orbits near the disc when the velocity ellipsoid is aligned with the Galactic plane (Bovy, in prep.). In reality, however, this is not necessarily true (e.g. some general models are provided by Tempel & Tenjes 2006; Kipper et al. 2016). In any case, our cut in z-coordinate $|z| < 0.2$ kpc minimises gradients in the vertical direction. Furthermore, the inclusion of this term changes the final circular velocity at the percent level, as shown in Eilers et al. (2019).

3.2. Circular velocity fitting

We used the axisymmetric kinematic model described in the previous section to derive the circular velocity $v_{\mathrm{c},j}$ in each radial bin j. We approached this as a Bayesian inference problem, wherein we used a Markov chain Monte Carlo (MCMC) algorithm to sample the posterior probability of our model parameters θ, namely the circular velocities $\{v_{c,j}\}$, h_R, h_Θ and R_0. According to Bayes' theorem, the posterior distribution of a set of model parameters θ given a particular set of data D can be defined as

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}, \quad (7)$$

where $p(\theta)$ is the prior distribution function that contains a priori knowledge about the parameters, $p(D)$ is the Bayesian evidence which is an irrelevant normalisation constant in this context. Moreover, the likelihood function takes the form

$$p(D|\theta) = -\prod_j \left[\frac{1}{\sqrt{2\pi\sigma_{\overline{v_{\phi,j}}}}} \exp\left(\frac{\left(\overline{v_{\phi,j}} - v_{\phi\,\mathrm{model},j}(\theta)\right)^2}{\sigma^2_{\overline{v_{\phi,j}}}} \right) \right], \quad (8)$$

where j is iterated over R bins. We note that this equation assumes that rotational and azimuthal velocities in each of the bins are independent of each other. The terms $\overline{v_{\phi,j}}$ and $\sigma^2_{\overline{v_{\phi,j}}}$ are obtained from the data and are the mean and variance of the azimuthal velocity in the j-th bin respectively.

For the prior distribution in (7), we defined flat priors, where the circular velocities are allowed within a range of $[-400, 400]\,\mathrm{km\,s^{-1}}$. In addition to the velocities, we defined naive priors for the scale length parameters $h_R = 3 \pm 1$ kpc and $h_\sigma = 21 \pm 1$ so to encompass values in the literature (Eilers et al. 2019; Bland-Hawthorn & Gerhard 2016). As mentioned previously, the Galactocentric distance R_0 was also treated as a free parameter of the analysis and was given a uniform prior within $[7.8, 8.5]$ kpc. Since the solar Galactocentric velocities $V_\odot$ and $W_\odot$ are scaled with R_0 (see Eq. (1)), the chosen prior is reflected both in the median value and error bar of the circular velocity in a given bin.

Having defined our likelihood and prior functions to use in the fitting, we set up our MCMC algorithm using the python package `emcee` (Foreman-Mackey et al. 2013). The parameter space of our model was explored using 48 independent walkers. All in all, we used 13 parameters in the fitting, where the first ten were circular velocities $v_{\mathrm{c},j}$ of the radial bins and the rest were the Sun's Galactocentric distance and the scale length terms.

By treating the Sun's Galactocentric distance as a free parameter of the analysis we were required to repeat the coordinate and velocity transformation at each step in the MCMC. In addition to this, we also had to propagate the covariance information of each star resulting in each step of the MCMC being computationally expensive. In order to bring down the iteration time, we used `numpy` (Harris et al. 2020) and `cupy` (Okuta et al. 2017), which make it possible to implement the calculations on both CPU and GPU in an efficient vectorised form.

The use of GPUs was particularly well motivated, since the parameter and uncertainty propagation routines in our code consist largely of matrix operations with relatively large arrays. GPU-accelerated computing libraries (such as `cupy`) take advantage of the fact that modern GPUs have significantly more threads than a CPU and are thus better at parallelising certain computation routines than their CPU-counterparts. In the end, both `numpy` and `cupy` were utilised simultaneously and the MCMC routine easily parallelised across the available CPUs and GPU devices where the most computationally demanding aspects of the pipeline were handled by the latter. The full data was analysed by using two CPU cores per GPU and with a total of six GPUs the computation time for each step was brought down to ≈11 s. This translates into a 6-fold speed increase when compared to running the code with just a single GPU and a 164-fold increase when running solely on CPUs. Using a single GPU for the full dataset described in this work, quickly leads to either out of memory issues or extremely long runtimes and thus it must be noted that the feasibility of the analysis was heavily dependent on the availability of multiple GPU devices and CPU cores. Our RGB sample of roughly 0.6 million RGB stars and the code used in our analysis can be found in zenodo[4] and online[5], respectively.

[4] https://zenodo.org/record/8014011
[5] https://github.com/HEP-KBFI/gaia-tools

Table 1. Measured circular velocities v_c, also plotted in Fig. 6.

x	R [kpc]	v_c [km s^{-1}]	σ_v^- [km s^{-1}]	σ_v^+ [km s^{-1}]	v_a [km s^{-1}]	σ_{va} [km s^{-1}]
0.66	5.5	221.3	6.5	5.7	3.3	6.1
0.79	6.5	231.0	6.9	6.0	2.8	6.5
0.91	7.5	234.6	7.1	6.1	2.9	6.6
1.03	8.5	232.7	7.1	6.1	3.1	6.6
1.15	9.5	229.8	7.1	6.1	3.5	6.6
1.27	10.5	231.2	7.0	6.2	4.6	6.6
1.39	11.5	230.6	6.3	6.1	6.4	6.2
1.57	13.0	227.5	6.5	5.8	7.2	6.4

Notes. We quote the median of each x bin, as the fitting is done using this adimensional variable, and the corresponding R value is calculated for $R_0 = 8.277$ kpc (GRAVITY Collaboration 2022).

4. Results

4.1. Circular velocity curve

The circular velocity curve of our sample of RGB stars is summarised in Table 1 and shown in Fig. 6. Inside each bin, we quote the median of the 1D marginalised posterior probability distribution obtained in the MCMC fitting. For the error bars, we quote the 16th and 84th percentile of the distribution. We would like to highlight that, in the classical approach for calculating the circular velocity curve, circular velocities in radial bins are calculated by plugging values into Eq. (4). In the proposed new approach, we add a simple kinematic model (given by (2)) on top of the Jeans equation, thus transforming the problem into an inference procedure. This allows to introduce nuisance parameters, such as the h_R, h_σ and R_0, and propagate their corresponding uncertainties (regardless of whether we have normal or non-normal errors) into the final circular velocity curve via Bayes theorem. The fitting procedure may, nonetheless, introduce additional correlations between the radial bins. For this reason, we checked that, if we fix the nuisance parameters R_0, h_R and h_σ, the central values of the circular velocities obtained with the new approach (MCMC analysis) coincide with the values obtained by the classical or traditional technique.

In Fig. 6, we compare our result to others from the literature. Our circular velocity curve is in agreement with the one estimated in Ablimit et al. (2020) using the 3D velocity vector method on $\sim 10^3$ classical Cepheids. However, for $R > 8$ kpc, we obtain larger circular velocities than those calculated by the same authors but using the proper motions of ~600 classical Cepheids. The former and latter samples have around 370 Cepheids in common and both results show that modelling assumptions and/or tracer samples can induce differences in the estimated circular velocities of at least 10%. We note that these changes are larger than the estimated uncertainties in this work, which are in the $\lesssim$3% level. Our error bars include statistical uncertainties, which are negligible owing to the large data sample. They further include uncertainties in the spatial-kinematic morphology of the tracer stars (scale radius of the density profile h_R and of the velocity distribution h_σ) and in the Sun's galactocentric distance. Circular velocities show a mild sensitivity to h_R, specially for values $h_R \leq 2.5$ kpc and, at least within the prior range explored in our analysis, h_σ and circular velocity central values are independent. On the contrary, the adopted value of R_0 strongly affects the final circular velocities and it is the main source of systematic uncertainties (from the ones studied in this analysis).

In addition, our estimated circular velocities are also compatible with those obtained in Eilers et al. (2019), Wang et al. (2022) and Zhou et al. (2023), due to our large uncertainties compared to those estimated in these articles. If we fixed the Sun's galactocentric distance and total velocity in the azimuthal direction to the values adopted in the former article, namely $R_0 = 8.122$ kpc and $V_\odot = 245.8$ km s^{-1}, the estimated error bars on the circular velocities are reduced and our results are incompatible with Eilers et al. (2019) analysis at 1σ for our fiducial distance estimates (see Sect. 4.4 for a comparison using the circular velocity curve obtained with other distance estimates). This shows that R_0 is the main source of uncertainty in the reconstruction of the circular velocity curve. Moreover, if for the fixed R_0 case the prior range in the scale length h_σ is increased by a factor of three, the results remain unchanged. In contrast, increasing the prior range in the scale length h_R by the same factor, the median values decrease by less than 2% and the size of error bars remains roughly the same.

The circular velocity curve in Wang et al. (2022) was obtained by describing, by means of the radial axisymmetric Jeans equation, the dynamics of all *Gaia* DR3 stars within the region $160^\circ < \ell < 200^\circ$ and $|Z| < 3$ kpc that have measured radial velocities. All stars are thus described by the same asymmetric drift. However, younger stars are expected to have a smaller asymmetric drift than an older population of stars. In fact, Kawata et al. (2019) estimated $v_a(R_0) = 0.28 \pm 0.20$ km s^{-1} using young classical Cepheids, whether we obtain, as expected, the central larger value $v_a(R_0) = 3 \pm 7$ km s^{-1} for older RGB stars. Figure 7 shows the asymmetric drift as a function of Galactocentric distance for $R_0 = 8.277$ kpc. The asymmetric drift mildly increases with distance to the Galactic centre with a slope of 0.59 ± 0.12 km s^{-1} kpc^{-1}.

Our estimated value of the circular velocity at the Sun's position, namely 233 ± 7 km s^{-1}, is compared in Table 2 with other estimates from the literature. We found that the estimated gradient of the curve is extremely sensitive to the radial interval included in its inference. If we remove the first two radial bins where the circular velocities increase, the obtained value is -1.1 ± 0.3 km s^{-1} kpc^{-1}, which points to a smooth decrease of the circular velocities with Galactocentric distances. If we rather include all radial bins, the estimated value of the slope is 0.4 ± 0.6 km s^{-1} kpc^{-1}, which describes a flat circular velocity curve within the uncertainties. Mróz et al. (2019) and Eilers et al. (2019) included all radial intervals for the determination of the slope, and found decreasing slopes that do not agree with the latter value. This may point out to the presence

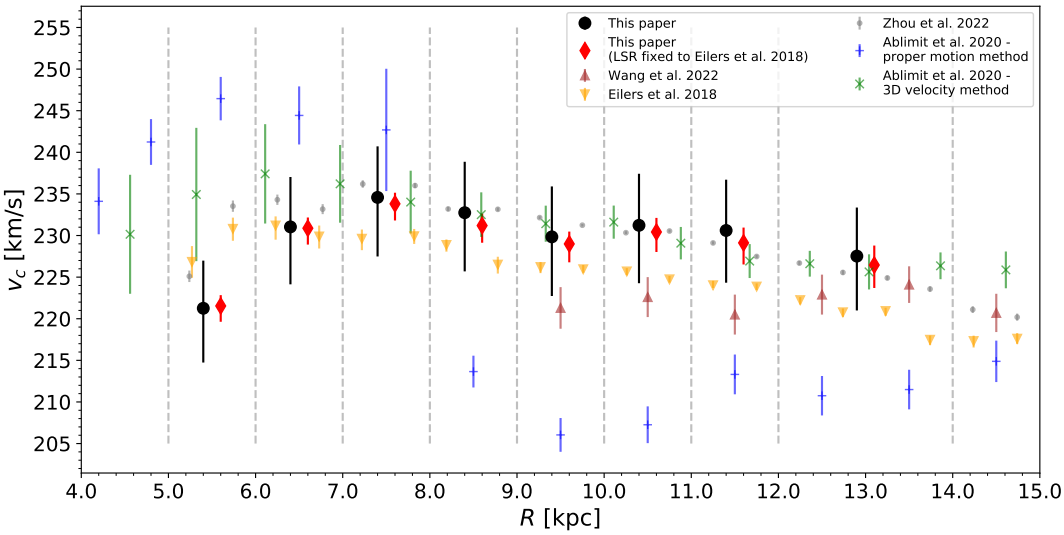


Fig. 6. Circular velocity curve obtained from the MCMC. Grey dashed lines have been plotted to indicate the position of each radial bin. In black (with circles) we show the circular velocities as obtained in this paper where the error bars correspond to the 16th and 84th percentile of the circular velocity posterior distribution in a particular bin. We adopted $R_0 = 8.277$ kpc to convert the adimensional coordinate x into Galactocentric distance R.

Table 2. Circular velocity at the solar location $v_c(R_0)$ as measured by different methods.

Source	$v_c(R_0)$ [km s^{-1}]	R_0 [kpc]
This work	233 ± 7	8.277
Zhou et al. (2023)	234.04 ± 0.08(stat.) ± 1.36(sys.)	8.122 ± 0.031
Kipper et al. (2021)	228.4 ± 3.5	8.3
Eilers et al. (2019)	229.0 ± 0.2	8.122 ± 0.031
Kawata et al. (2019)	236 ± 3	8.2 ± 0.1
Bobylev (2017)	231 ± 6	8
Huang et al. (2016)	240 ± 6	8.34
Bovy et al. (2012)	218 ± 6	$8.1^{+1.2}_{-0.1}$

Notes. In the last column, we quote the value of the Sun's galactocentric distance that was adopted in each of the referenced articles.

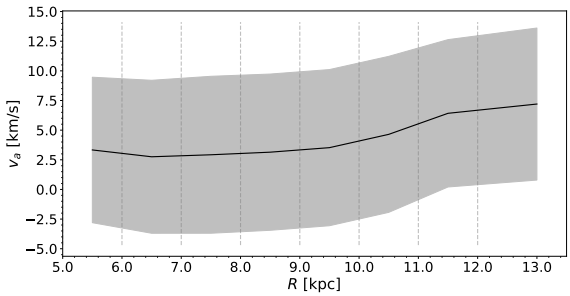


Fig. 7. Asymmetric drift profile. The black line shows the asymmetric drift correction in each radial bin and the shaded region depicts its propagated uncertainty. We note that the uncertainties are fully correlated between the bins.

of systematic biases in the distance estimations, as described in Sect. 4.4.

4.2. Smooth dark matter halo

In the region analysed in this article, mean radial velocities are <5% of the azimuthal or rotational velocities. According to Chrobáková et al. (2020), when the ratio of radial to azimuthal velocities is smaller than 10%, circular velocities obtained through the radial axisymmetric and stationary Jeans equation provide an unbiased estimator for the centrifugal velocities that balance the averaged radial gravitational force. According to this result, our circular velocity curve then provides an unbiased estimation of the spherically averaged dynamical mass distribution within 14 kpc. Although this conclusion will be tested in an upcoming paper where we assess the effects of modelling assumptions, such as the axial symmetry condition, and the effect of Galactic substructure, using our current results we estimate that the DM mass is

$$\log_{10}\left[M_{\rm DM}(R < 14\,{\rm kpc})/M_\odot\right] = 11.2^{+2.0}_{-2.3}. \quad (9)$$

Furthermore, we find that the local spherically averaged DM density is

$$\rho_{\rm DM}(R_0) = \left(0.41^{+0.10}_{-0.09}\right)\,{\rm GeV\,cm^{-3}} = \left(0.011^{+0.003}_{-0.002}\right)\,M_\odot\,{\rm pc}^3. \quad (10)$$

These estimates were obtained by fitting the observed circular velocities to the velocities predicted by Newtonian gravity for the baryonic components (stellar bulge, disc and gas) and the DM halo. For each baryonic component we adopted a set of three-dimensional density profiles, originally compiled in Iocco et al. (2015). The stellar bulge mass is constrained by microlensing towards the Galactic centre and the stellar disc is normalised by the stellar surface density at the Sun's position. We describe the DM distribution using a generalised Navarro-Frenk-White density profile Zhao (1996). We compare observed and predicted velocities using the Bayesian prescription presented in Karukes et al. (2019, 2020). The estimates provided are Bayesian model averages that include uncertainties in the Sun's Galactocentric distance, the three-dimensional density profile of bulge and disc stars, and the stellar mass of the Galaxy.

We would like to highlight that our estimate of the local DM density is compatible, within 1σ uncertainties, with recent estimates of this quantity using the circular velocity curve method (Eilers et al. 2019; Mróz et al. 2019; de Salas et al. 2019; Lin & Li 2019; Karukes et al. 2019; Sofue 2020). In addition, it is compatible with local estimates using the vertical Jeans equation (Salomon et al. 2020; Guo et al. 2020; Nitschai et al. 2020) and with the most-recent local estimate using a novel machine learning approach by Lim et al. (2023). Thus reinforcing the conclusion about the spherical shape of the inner ~ 15 kpc of the DM halo, obtained by modelling stellar streams (Koposov et al. 2010; Bowden et al. 2015) and the kinematics of halo stars Wegg et al. (2019).

4.3. The local standard of rest and the solar peculiar velocity

The total Galactocentric azimuthal velocity of the Sun can be used to derive the solar peculiar velocity when we incorporate knowledge about the circular velocity at its position. In particular, the total azimuthal velocity is often decomposed as

$$V_\odot = v_c(R_0) + V_{\odot,\mathrm{LSR}}, \tag{11}$$

where the last term is the Sun's peculiar motion in the local standard of rest (LSR). The treatment of the solar velocity as shown in Eq. (11) assumes that the LSR moves in a circular orbit about the Galactic centre and therefore, it coincides with the rotational standard of rest (RSR) in which stars move on circular orbits in the azimuthally averaged gravitational potential. However, in recent years it has been shown that the stellar disc exhibits bulk motions at the kiloparsec scale (Bovy et al. 2015; Williams et al. 2013; Khanna et al. 2023). In the presence of these large scale streaming motions, the LSR, which is defined as the reference frame of a local population of stars with zero velocity dispersion, might not coincide with the RSR. Considering this, the total azimuthal velocity of the Sun can be decomposed as (Drimmel & Poggio 2018)

$$V_\odot = v_c(R_0) + V_{\mathrm{LSR}} + V_{\odot,\mathrm{LSR}}, \tag{12}$$

where V_{LSR} is the velocity of the LSR with respect to the RSR. This difference of velocity between the LSR and RSR might account for the discrepancy between locally-derived estimations of the Sun's peculiar motion (i.e. using the Strömberg relation) and globally-measured values using a sample of tracers in a larger volume around the Sun. In fact, Bovy et al. (2012) concluded that the LSR itself might not be on a circular orbit and it is rotating $V_{\mathrm{LSR}} \approx 12\,\mathrm{km\,s^{-1}}$ faster than the actual RSR. This is in agreement with the recently reported value in Khanna et al. (2023) of $\approx 10\,\mathrm{km\,s^{-1}}$. On the other hand, Bland-Hawthorn & Gerhard (2016) estimated $V_{\mathrm{LSR}} = 0 \pm 15\,\mathrm{km\,s^{-1}}$.

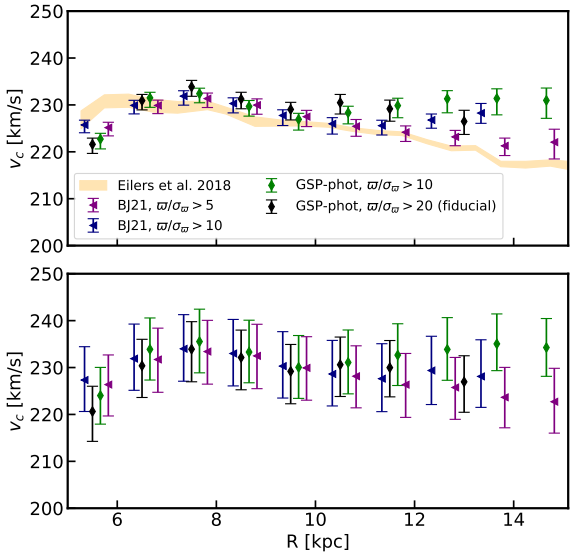


Fig. 8. Circular velocity curve for different distances estimates as explained in the main text. In the top panel, we show the results where fixing the Sun's galactocentric distance to $R_0 = 8.122$ kpc and leaving free the scale lengths of the radial spatial density and velocity dispersion of the tracer RGB sample. The orange band shows encompass the estimated circular velocity curve in Eilers et al. (2019) within 1σ. The results when additionally leaving R_0 as a free parameter are shown in the bottom panel.

Assuming $V_{\odot,\mathrm{LSR}} = 12.24\,\mathrm{km\,s^{-1}}$ as measured in Schönrich et al. (2010) from the Hipparcos Catalogue, we obtain $V_{\mathrm{LSR}} = 7 \pm 7\,\mathrm{km\,s^{-1}}$. Our estimate is still statistically compatible with zero streaming motion, but nevertheless strengthens the hypothesis that a region around the Sun, with a characteristic length scale of 1 kpc, exhibits a bulk motion in the azimuthal direction of the order of $10\,\mathrm{km\,s^{-1}}$.

4.4. Cautionary tale about distances

Our study is based on GSP-Phot distances (Andrae et al. 2023), which have been shown to systematically underestimate the distance beyond 2 kpc from the Sun (Fouesneau et al. 2023). This bias is also shown in the left panel of Fig. 1, as the two-dimensional distribution of the estimated distance versus inverse parallax is not symmetric with respect to the 1:1 line, as expected from a Gaussian noise model for parallax measurements, but is more populated to the right of this line. Fouesneau et al. (2023) find that imposing a cut-off in the quality of the parallax measurement of $\varpi/\sigma_\varpi > 10$ yields reliable heliocentric distances up to 10 kpc. And Andrae et al. (2023) point out that a strict parallax quality cut-off of $\varpi/\sigma_\varpi > 20$ provides reliable distances. For our sample of RGB stars, the first cut-off eliminates the systematic underestimation of the GSP-Phot distances, although a slight overestimation of distances appears (see left panel of Fig. 1). For this reason, as our fiducial run, we adopted the last strict cut, which alleviates the mild overestimation. However, in this section we describe how our results would change if we had adopted the less stringent cut-off in parallax quality or rather

Table 3. Summary of the results when having R_0, h_R and h_σ as free parameters and adopting different distance estimates.

Distance estimate	Slope (first 2 bins removed) [km s^{-1} kpc^{-1}]	ρ_0 [GeV cm^{-3}]	$M_{\rm DM}(R < 14\,{\rm kpc})$ [$M_\odot$]	$V_{\rm LSR}$ [km s^{-1}]
BJ21 + $\varpi/\sigma_\varpi > 5$	-0.9 ± 0.3 (-1.58 ± 0.08)	$0.37^{+0.08}_{-0.07}$	$9.9^{+1.6}_{-1.9}$	7 ± 7
BJ21 + $\varpi/\sigma_\varpi > 10$	-0.3 ± 0.3 (-1.0 ± 0.3)	$0.39^{+0.09}_{-0.08}$	$10.8^{+2.0}_{-1.7}$	6 ± 7
GSP-Phot + $\varpi/\sigma_\varpi > 10$	0.6 ± 0.3 (0.2 ± 0.3)	$0.43^{+0.07}_{-0.06}$	$11.6^{+1.8}_{-2.0}$	6 ± 7
GSP-Phot + $\varpi/\sigma_\varpi > 20$	0.4 ± 0.6 (-1.1 ± 0.3)	$0.41^{+0.10}_{-0.09}$	$11.2^{+2.0}_{-2.3}$	7 ± 7

Notes. Namely, photogeo distances from Bailer-Jones et al. (2021) with a cut in parallax of $\varpi/\sigma_\varpi > 5$ and $\varpi/\sigma_\varpi > 10$, and GSP-Phot distances with quality parallax cuts of $\varpi/\sigma_\varpi > 10$ and $\varpi/\sigma_\varpi > 20$ (fiducial case). The second column shows the estimated slope of the circular velocity curve when a straight line is fitted to all bins (and in brackets the value obtained by eliminating the first two radial bins as the circular velocities increase within ~5−7 kpc), the local DM density and DM mass within 14 kpc are shown in the third and forth columns, respectively. We quote the velocity of the LSR in the last column.

used photogeo distances from Bailer-Jones et al. (2021) (hereinafter referred to as BJ distances).

A bias in distance estimates has a noticeable impact in the results of our analysis. In particular, underestimated distances lead to an overestimation of the circular velocity curve gradient, and thus to an underestimation of the DM content, and vice versa in the case of overestimated distances. We note that the bias in the estimated slope is more pronounced the greater the actual slope. In order to assess the effect of biased distances, we performed our MCMC analysis using four different distance estimates: BJ distances with a cut-off of $\varpi/\sigma_\varpi > 5$ and $\varpi/\sigma_\varpi > 10$, and GSP-Phot distances with $\varpi/\sigma_\varpi > 10$ and $\varpi/\sigma_\varpi > 20$. For each of these distance estimates, the top panel of Fig. 8 shows the resultant circular velocity curve while fixing the Sun's galactocentric distance and total velocity in the azimuthal direction to the values adopted in Eilers et al. (2019), namely $R_0 = 8.122$ kpc and $V_\odot = 245.8$ km s^{-1}, and leaving h_R and h_σ free. The bottom panel of the same figure depicts circular velocities while letting R_0 as an additional free parameter. From this figure, it is clear that the inclusion of uncertainties in R_0 makes the four circular velocities compatible within uncertainties. Furthermore, as we increase the cut-off in the parallax quality from 5 to 10 for BJ distances, the declining of the curve becomes less steep, thus increasing the DM mass of the Galaxy. On the other hand, by increasing the cut-off from 10 to 20 for the GSP-Phot distances, we are alleviating the mild overestimation of distances and the positive gradient of the curves becomes shallower, reducing the DM mass. Table 3 summarises these results.

5. Summary and conclusions

We estimated the circular velocity curve from 5 kpc to 14 kpc from the Galactic centre using 665 660 RGB stars that are approximately located in one quarter of the stellar disc with 6D phase-space information as measured by *Gaia* DR3, and GSP-Phot distance estimates. We determined the circular velocity curve by describing observed rotational velocities, in adimensional radial bins, as the difference between the circular velocity and the asymmetric drift. The latter given by the stationary and axisymmetric radial Jeans equations, under the further assumption of reflection symmetry above and below the Galactic plane. In the traditional approach, one simply plugs values into the Jeans equation. In our approach, by describing the observed rotational velocity as the circular velocity minus the asymmetric drift, we transformed the problem into an inference procedure. In particular, observed and model rotational velocities were fitted using a Bayesian inference approach that incorporates systematic and statistical uncertainties as nuisance parameters. This allowed us to propagate into the final results uncertainties of different nature. In particular, our relative uncertainties are ~3% and, apart from the statistics, account for uncertainties in the Sun's galactocentric distance (which is the main source of uncertainty) and uncertainties in the spatial-kinematic morphology of the stellar disc.

We studied the effect of biased distances on our results and showed, as expected, that underestimated distances lead to steeper (negative) slopes and thus to an underestimation of the dark matter content in the Galaxy. This may explain some recent findings of significantly declining circular velocity curves and, consequently, lower spherically averaged local DM densities than those purely local values obtained using stars in the Solar neighbourhood. Owing to the spherical shape of the DM halo in the inner ~15 kpc of the Galaxy, these two sets of estimates should converge.

Acknowledgements. We thank the referee for her/his constructive comments and for pointing out the biases in the distance estimates. This has undoubtedly opened up many avenues for further studies. The authors would like to thank A. Cuoco, G. Battaglia and E. Fernández Alvar for fuitful discussions and comments. This work was supported by the Estonian Research Council grants PRG1006, PSG700, PRG803, PSG864, MOBTP187, PRG780, MOBTT86 and by the European Regional Development Fund through the CoE program grant TK133. This research has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. The authors gratefully acknowledge the support of Nvidia, whose technology played a critical role in the success of this research. G.F.T. acknowledges support from the Agencia Estatal de Investigación del Ministerio de Ciencia en Innovacón (AEI-MICIN) under grant number CEX2019-000920-S and the AEI-MICIN under grant number PID2020-118778GB-I00/10.13039/501100011033

References

Ablimit, I., Zhao, G., Flynn, C., & Bird, S. A. 2020, ApJ, 895, L12
Abuter, R., Amorim, A., Anugu, N., et al. 2018, A&A, 615, L15
Andrae, R., Fouesneau, M., Sordo, R., et al. 2023, A&A, 674, A27
Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, AJ, 161, 147
Benito, M., Bernal, N., Bozorgnia, N., Calore, F., & Iocco, F. 2017, J. Cosmol. Astropart. Phys., 2017, 007
Benito, M., Cuoco, A., & Iocco, F. 2019, J. Cosmol. Astropart. Phys., 2019, 033
Binney, J., & Tremaine, S. 2008, Galactic Dynamics: Second Edition
Bland-Hawthorn, J., & Gerhard, O. 2016, ARA&A, 54, 529
Bobylev, V. V. 2017, Astron. Lett., 43, 152
Bovy, J., Allende Prieto, C., Beers, T. C., et al. 2012, ApJ, 759, 131
Bovy, J., Bird, J. C., Pérez, A. E. G., et al. 2015, ApJ, 800, 83
Bowden, A., Belokurov, V., & Evans, N. W. 2015, MNRAS, 449, 1391
Chrobáková, Ž., López-Corredoira, M., Sylos Labini, F., Wang, H. F., & Nagy, R. 2020, A&A, 642, A95

de Salas, P. F., Malhan, K., Freese, K., Hattori, K., & Valluri, M. 2019, J. Cosmol. Astropart. Phys., 2019, 037
Do, T., Hees, A., Ghez, A., et al. 2019, Science, 365, 664
Drimmel, R., & Poggio, E. 2018, Res. Notes Am. Astron. Soc., 2, 210
Eilers, A.-C., Hogg, D. W., Rix, H.-W., & Ness, M. K. 2019, ApJ, 871, 120
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASA, 125, 306
Fouesneau, M., Frémat, Y., Andrae, R., et al. 2023, A&A, 674, A28
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Katz, D., et al.) 2018, A&A, 616, A11
Gaia Collaboration (Drimmel, R., et al.) 2023, A&A, 674, A37
GRAVITY Collaboration (Abuter, R., et al.) 2019, A&A, 625, L10
GRAVITY Collaboration (Abuter, R., et al.) 2020, A&A, 636, L5
GRAVITY Collaboration (Abuter, R., et al.) 2021, A&A, 647, A59
GRAVITY Collaboration (Abuter, R., et al.) 2022, A&A, 657, L12
Guo, R., Liu, C., Mao, S., et al. 2020, MNRAS, 495, 4828
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357
Helmi, A., Babusiaux, C., Koppelman, H. H., et al. 2018, Nature, 563, 85
Hobbs, D., Lindegren, L., Bastian, U., et al. 2018, Gaia DR2 documentation Chapter 3: Astrometry
Huang, Y., Liu, X.-W., Yuan, H.-B., et al. 2016, MNRAS, 463, 2623
Iocco, F., Pato, M., & Bertone, G. 2015, Nat. Phys., 11, 245
Johnson, D. R. H., & Soderblom, D. R. 1987, AJ, 93, 864
Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864
Karukes, E. V., Benito, M., Iocco, F., Trotta, R., & Geringer-Sameth, A. 2019, J. Cosmol. Astropart. Phys., 2019, 046
Karukes, E. V., Benito, M., Iocco, F., Trotta, R., & Geringer-Sameth, A. 2020, J. Cosmol. Astropart. Phys., 2020, 033
Kawata, D., Bovy, J., Matsunaga, N., & Baba, J. 2019, MNRAS, 482, 40
Khanna, S., Sharma, S., Bland-Hawthorn, J., & Hayden, M. 2023, MNRAS, 520, 5002
Kipper, R., Tenjes, P., Tihhonova, O., Tamm, A., & Tempel, E. 2016, MNRAS, 460, 2720
Kipper, R., Tenjes, P., Tempel, E., & de Propris, R. 2021, MNRAS, 506, 5559
Koposov, S. E., Rix, H.-W., & Hogg, D. W. 2010, ApJ, 712, 260
Kuzmin, G. G. 2022, ArXiv e-prints [arXiv:2201.04136]
Leung, H. W., Bovy, J., Mackereth, J. T., et al. 2022, MNRAS, 519, 948
Lim, S. H., Putney, E., Buckley, M. R., & Shih, D. 2023, ArXiv e-prints [arXiv:2305.13358]
Lin, H.-N., & Li, X. 2019, MNRAS, 487, 5679
Malkin, Z. M. 2013, Astron. Rep., 57, 128
Mróz, P., Udalski, A., Skowron, D. M., et al. 2019, ApJ, 870, L10
Nitschai, M. S., Cappellari, M., & Neumayer, N. 2020, MNRAS, 494, 6001
Okuta, R., Unno, Y., Nishino, D., Hido, S., & Loomis, C. 2017, in Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)
Reid, M. J., & Brunthaler, A. 2020, ApJ, 892, 39
Rybizki, J., Green, G. M., Rix, H.-W., et al. 2022, MNRAS, 510, 2597
Salomon, J.-B., Bienaymé, O., Reylé, C., Robin, A. C., & Famaey, B. 2020, A&A, 643, A75
Schönrich, R., Binney, J., & Dehnen, W. 2010, MNRAS, 403, 1829
Sofue, Y. 2020, Galaxies, 8, 37
Tempel, E., & Tenjes, P. 2006, MNRAS, 371, 1269
Thomas, G. F., Laporte, C. F. P., McConnachie, A. W., et al. 2019, MNRAS, 483, 3119
Wang, H. F., Chrobáková, Ž., López-Corredoira, M., & Labini, F. S. 2022, ApJ, 942, 12
Wegg, C., Gerhard, O., & Bieth, M. 2019, MNRAS, 485, 3296
Williams, M. E. K., Steinmetz, M., Binney, J., et al. 2013, MNRAS, 436, 101
Zhao, H. 1996, MNRAS, 278, 488
Zhou, Y., Li, X., Huang, Y., & Zhang, H. 2023, ApJ, 946, 73

Appendix A: MCMC results

Figure A.1 shows the marginalised two-dimensional and one-dimensional posterior distributions for three different MCMC runs: first, (R_0, h_R, h_σ) are free nuisance parameters (black), second, (h_R, h_σ) are not fixed and R_0 is fixed to the value $R_0 = 8.122$ kpc (red) and finally, the output of the MCMC when all nuisance parameters are fixed to the values $R_0 = 8.277$ kpc, $h_R = 3$ kpc, $h_\sigma = 21$ kpc.

For the first run (i.e. R_0, h_R and h_σ are free parameters), the movement of circular velocities is driven by changes in R_0. Since we are not considering strong priors on R_0, a strong positive correlation between this parameter and the circular velocities is observed. The circular velocity curve is sensitive to the Sun's Galactocentric distance R_0 (Benito et al. 2019), nonetheless, R_0 is constrained in the literature much better by different types of analysis than the circular velocities (see Malkin 2013 for a review of techniques). Therefore, we do not aim to restrict R_0, but to assess the impact of its uncertain value on the circular velocity curve. For this reason, we adopt as a prior a uniform distribution that encompasses the most recent determinations of R_0 within 2σ uncertainties. This is a conservative approach that does not favour any particular estimate. We would like to also emphasise that the actual value of R_0 may be troublesome. For example, the LMC is causing differences in reflex motion in distinct parts of the Galaxy causing the definition of centre to be vague or we do not know to what extent the different definitions of centre (e.g. local isopotential curve, SMBH position) affect estimations via the Jeans equations. One of the results of our analysis is that, given the scatter in the most recent determination of R_0, this parameter represents the main source of uncertainty in the calculation of circular velocities in radial bins.

On the contrary, circular velocities show a mild sensitivity to h_R and h_σ. Neither of these scale lengths can be constrained by our analysis and are simply treated as nuisance parameters whose prior range is defined by observational determinations (see e.g. Eilers et al. 2019 and references therein).

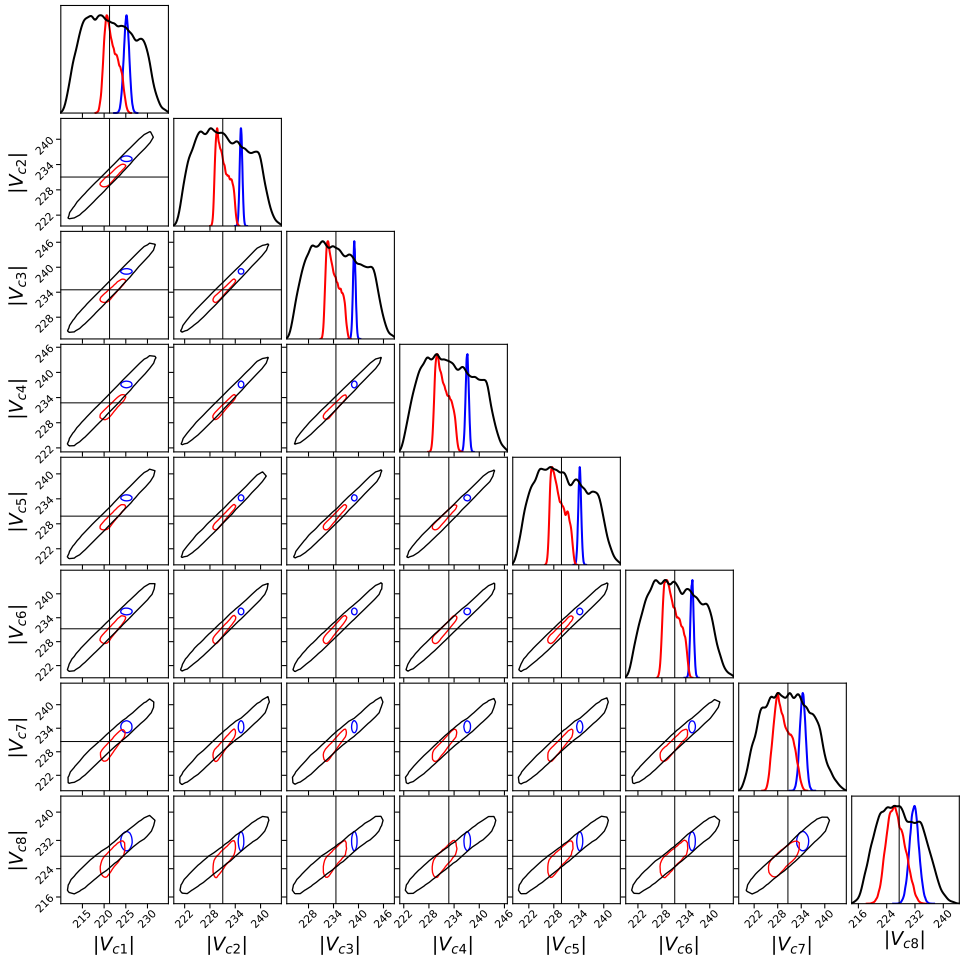


Fig. A.1. One and two-dimensional marginalised posterior distributions of the circular velocity. The figure shows distributions from three different parameter setups: all nuissance parameters free (black), only scale parameters free (red), all nuissance parameters fixed (blue), see text for more details. The contours delimit regions of 2-σ probability and the best fit circular velocity of the black posteriors is demarked with horisontal and vertical lines. The diagonal contains the normalised 1D posteriors of the circular velocities from the three different runs.

Appendix 3

III
Põder, Sven, Pata, Joosep, Benito, María, Alonso Asensio, Isaac, and Dalla Vecchia, Claudio. Detection of stellar wakes in the milky way: A deep learning approach. *A&A*, 693:A227, 2025

A&A, 693, A227 (2025)
https://doi.org/10.1051/0004-6361/202451480

Astronomy & Astrophysics

Detection of stellar wakes in the Milky Way: A deep learning approach

Sven Põder[1,2,⋆], Joosep Pata[1], María Benito[3,⋆], Isaac Alonso Asensio[4,5], and Claudio Dalla Vecchia[4,5]

[1] National Institute of Chemical Physics and Biophysics (NICPB), Rävala 10, Tallinn 10143, Estonia
[2] Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia
[3] Tartu Observatory, University of Tartu, Observatooriumi 1, Tõravere 61602, Estonia
[4] Instituto de Astrofísica de Canarias, c/ Vía Láctea s/n, 38205 La Laguna, Tenerife, Spain
[5] Departamento de Astrofísica, Universidad de La Laguna, Av. Astrofísico Francisco Sánchez s/n, 38206 La Laguna, Tenerife, Spain

Received 12 July 2024 / Accepted 5 December 2024

ABSTRACT

Context. Due to poor observational constraints on the low-mass end of the subhalo mass function, the detection of dark matter (DM) subhalos on sub-galactic scales would provide valuable information about the nature of DM. Stellar wakes, induced by passing DM subhalos, encode information about the mass (properties) of the inducing perturber and thus serve as an indirect probe for the DM substructure within the Milky Way.

Aims. Our aim is to assess the viability and performance of deep learning searches for stellar wakes in the Galactic stellar halo caused by DM subhalos of varying mass.

Methods. We simulated massive objects (subhalos) moving through a homogeneous medium of DM and star particles with phase-space parameters tailored to replicate the conditions of the Galaxy at a specific distance from the Galactic centre. The simulation data was used to train deep neural networks with the purpose of inferring both the presence and mass of the moving perturber. We then investigated the performance of our deep learning models and identified the limitations of our current approach.

Results. We present an approach that allows for quantitative assessment of subhalo detectability in varying conditions of the Galactic stellar and DM halos. We find that our binary classifier is able to infer the presence of subhalos in our generated mock datasets, showing non-trivial performance down to a mass of 5×10^7 $M_\odot$. In a multiple-hypothesis case, we are also able to discern between samples containing subhalos of different mass. By simulating datasets describing subhalo orbits at different Galactocentric distances, we tested the robustness of our binary classification model and found that it performs well with data generated from different initial physical conditions. Based on the phase-space observables available to us, we conclude that overdensity and velocity divergence are the most important features for subhalo detection performance.

Key words. methods: data analysis – Galaxy: kinematics and dynamics

1. Introduction

The standard Lambda cold dark matter (ΛCDM) scenario successfully describes the behaviour of dark matter (DM) on extra-galactic scales (Einasto 2010; Zavala & Frenk 2019). Studies of structure formation (Gramann 1990), galaxy clustering (Darragh-Ford et al. 2023), supernova luminosities (Perivolaropoulos & Skara 2022), and cosmic microwave background (CMB) correlation functions (Planck Collaboration VII 2020) have left little room for deviations from the CDM model at these scales.

A key prediction of ΛCDM that is yet to be confirmed is the abundance of DM subhalos on sub-galactic scales. In fact, studies of Milky Way-like galaxy simulations show that the subhalo mass function (SHMF), which is the abundance of subhalos per unit mass, follows a power law well below the galactic scale (e.g. Springel et al. 2008). In the absence of convincing observational evidence for small-scale DM clustering below subhalo masses of $\approx 10^9$ $M_\odot$, other alternative DM models (warm dark matter, self-interacting dark matter, fuzzy dark matter, etc.) are also allowed. These models impose a cut-off in the SHMF below a specific mass threshold and thus change the expected abundance of dark subhalos orbiting galaxies (Ostdiek et al. 2022; Zavala & Frenk 2019). Constraining the low-mass end of the SHMF is therefore an important test of the CDM scenario, as deviations from the expected SHMF behaviour could be explained by alternative DM models (e.g. Benito et al. 2020). This, however, is not an easy endeavour, as subhalos less massive than 10^{8-9} $M_\odot$ are not expected to host any stars due to their small mass and reionisation effects (Sawala et al. 2015; Benitez-Llambay & Frenk 2020) – they are dark subhalos.

In recent years, the expected count of low-mass subhalos inside a Milky Way (MW)-sized galaxy in the CDM scenario has been revised, although uncertainties remain. Garrison-Kimmel et al. (2017) have reported that the inclusion of baryonic physics actually suppresses the size of the expected subhalo population when compared to DM-only simulations. The previous work has been improved on by Barry et al. (2023), who used the FIRE-2 simulations (described in Hopkins et al. 2018) and found that at least 20 dark subhalos of mass $>10^6$ $M_\odot$ should exist within <=30 kpc of the Galactic centre. Given that these theoretical predictions are yet to be robustly validated by empirical observation, the MW presents itself as an ideal laboratory for probing the low-mass end of the SHMF.

⋆ Corresponding authors; sven.poder@kbfi.ee; mariabenitocst@gmail.com

Beyond the Local Group, investigating dark subhalos can be effectively pursued by observing the perturbations they impart on strongly lensed images of distant galaxies and quasars (e.g. Wagner-Carena et al. 2024). In recent years, deep learning methods have proven to be valuable tools in this endeavour, owing to their effectiveness in image classification tasks (for a thorough overview, see e.g. Varma et al. 2020 and the references therein). Inside the Galaxy, a promising method to probe the low-mass end of the SHMF involves searching for gaps or density fluctuations in the distribution of cold stellar streams (Bonaca & Price-Whelan 2024). For example, Bonaca et al. (2019) studied the interaction of the GD-1 stream with a massive perturber whose mass range they found to be of 10^6–10^8 $M_\odot$. With improved measurement data, the mass-detection limit via stream perturbations could be as low as 10^5 $M_\odot$ (Bovy et al. 2016). Another method to detect DM substructure in the Galaxy via pulsar timing array measurements, proposed in Siegel et al. (2007), promises the detection of even lower masses.

This work focuses on the detection of stellar wakes in the MW – arguably the least studied phenomenon for DM detection in the literature. The underlying concept is built on the notion that when a massive object moves through a field of stars, it experiences dynamical friction (Chandrasekhar 1943), as it perturbs the phase-space of the surrounding stellar medium (Mulder 1983). Through gravitational interactions with the perturber, stars are pulled towards it and in time cause a relative overdensity opposite to the direction of the perturber's movement (see e.g. Weinberg 1986, who described this effect analytically in the context of infalling satellites). In recent years, there has been growing interest in exploring the effects of dynamical friction-induced wakes as a promising avenue for investigating DM substructure. In the work of Buschmann et al. (2018), the authors developed an analytical likelihood formalism to use these perturbations in the stellar phase-space and infer the mass of the DM halo passing through the stars.

A popular test bed for the detection of stellar wakes has been the MW's largest satellite – the Large Magellanic Cloud (LMC; see e.g. Garavito-Camargo et al. 2019; Tamfal et al. 2021; Rozier et al. 2022; Foote et al. 2023). Rozier et al. (2022) studied the response of a static MW to the LMC's infall using linear response theory. More recently, Foote et al. (2023) studied the wake produced by the infall of the LMC using idealised wind tunnel simulations in the context of both CDM and fuzzy dark matter. Notably, they observed that the self-gravity of the DM wake amplifies the extent of the stellar wake, particularly for subhalos with masses of the order of 10^{11} $M_\odot$. Going beyond simulations, Conroy et al. (2021) observed for the first time the density wake trailing behind the orbit of the LMC using data from *Gaia*'s Early Data Release 3. Their work was expanded on by Fushimi et al. (2024), who used the wake to estimate the mass of the LMC's DM halo with the method proposed in Buschmann et al. (2018).

In our work, we focus our attention on perturbers less massive than the LMC and thus broaden the scope of Foote et al. (2023). Furthermore, we expand on the work in Buschmann et al. (2018), as we include the effects of self-gravity in our study of stellar wakes. This study also builds on our previous research Bazarov et al. (2022), which demonstrated the discernible impact of dark subhalos on the phase-space distribution of stars in simulated MW-like galaxies. To address the limitations of our previous work, we investigated dark subhalos in the MW using wind tunnel simulations, which afford greater control over the signal induced by dark subhalos. As in Bazarov et al. (2022), we tackled the problem in a data-driven way using machine learning (ML) in lieu of classical likelihood methods. The reason for this choice is that the latter become intractable as simulation complexity increases - and even more so in the case of real data with uncertainties.

The structure of the paper is as follows: in Sect. 2, we describe the numerical and physical details of our simulation setup. In Sect. 3, we discuss how we generated the mock data and set up our deep learning approach. Section 4 contains the performance results of the models described in the previous section. Section 5 outlines the key limitations of the current work and discusses future directions, while Sect. 6 summarises the main conclusions.

2. Wind tunnel simulations

In the following, we describe our simulations of an extended object orbiting at 30 kpc from the Galactic centre. Adopting a spherically symmetric gravitational potential and total mass M, this perturber experiences a stationary wind of simulation particles with a bulk velocity $-v$. We note that in the reference frame of the box, the simulation is equivalent to a setting where a perturber with constant velocity v moves through a homogeneous medium of field particles with constant mass density ρ and isotropic Maxwellian velocity distribution.

2.1. Perturber setup

This physical setup was simulated using `Pkdgrav3` (Potter et al. 2017; Alonso Asensio et al. 2023), which is a highly versatile cosmological N-body gravity code. Although generally used to simulate phenomena on cosmological scales, such as large-scale structure formation, it can also be used to accurately study the dynamics of systems down to planetesimal scales (see e.g. Alonso Asensio et al. 2023 and the references therein). In our work, we used `Pkdgrav3` to simulate a massive perturber moving through a homogeneous medium of background particles in a box with equal side lengths of $L = 120$ kpc and periodic boundary conditions in all directions (X, Y, Z). The coordinates of the box were defined in the range $x, y, z \in [-60, 60]$ kpc, and therefore any particle that is at -60 kpc and moving in the -X direction reappears at +60 kpc once it crosses the boundary. The simulation takes place in the rest frame of the perturber, which is stationary in the centre of the box at coordinates (0, 0, 0). To simulate the perturber's motion, we introduced a wind of stellar and DM simulation particles moving from right to left with some bulk velocity $-v$ along the X-axis. The magnitude of this velocity was approximated by assuming a circular orbit for the perturber and taking into account the total dynamical mass of the MW enclosed in the region where the Galactocentric distance is $R < 30$ kpc. In the work of Karukes et al. (2020), the mass of MW at this range is found to be approximately 3×10^{11} $M_\odot$, resulting in a circular orbital speed of $\sim$200 km s^{-1} at 30 kpc. In the FIRE-2 simulations Barry et al. (2023), the tangential velocity of DM subhalos with masses larger than 10^7 $M_\odot$ at the radius of 30 kpc from the Galactic centre is somewhere closer to 250 km s^{-1}. With all of this in mind, we chose a fiducial perturber velocity of 225 km s^{-1}, which is in the middle of these estimates.

Following Buschmann et al. (2018), the perturber is described by a Plummer density profile. This choice allows us to make a clearer comparison between their results and ours. The density of a Plummer sphere as a function of r is given by

$$\rho(r) = \frac{3M_{\rm sh}}{4\pi {R_s}^3}\left(1 + \frac{r^2}{{R_s}^2}\right)^{-5/2}, \quad (1)$$

Table 1. Physical parameters adopted in the wind tunnel-like N-body simulations for two different locations in the stellar halo.

Case	r [kpc]	v [km/s]	$\rho_{\rm DM}$ [$M_\odot$/kpc^3]	$\sigma_{\rm DM}$ [km/s]	$N_{\rm DM}$	$\rho_{\rm star}$ [$M_\odot$/kpc^3]	$\sigma_{\rm star}$ [km/s]	$N_{\rm star}$
Case 1	30	225	10^6	200	512^3	10^2	95	512^3
Case 2	50	200	$10^{5.5}$	180	512^3	10	90	512^3

where R_s is the Plummer scale radius, $M_{\rm sh}$ is the subhalo total mass and r represents the radial distance from the subhalo's centre. In the same way as in Buschmann et al. (2018); Diemand et al. (2008), we adopted the following equation for the computation of R_s,

$$R_s = 1.62\,{\rm kpc} \times \left(\frac{M_{\rm sh}}{10^8\,M_\odot}\right)^{1/2}. \quad (2)$$

We chose to run our simulation with a range of mass options (in addition to simulations with no subhalo present) in order to gauge how our ML model's performance changes with respect to the mass of the perturber. For the purposes of this study, we adopted the following subhalo masses: $5\times10^7\,M_\odot$, $10^8\,M_\odot$ and $5\times10^8\,M_\odot$. We did not implement subhalos below $5\times10^7\,M_\odot$ as the stellar wakes produced by perturbers smaller than this are not resolved in the simulations. Likewise, subhalos more massive than $5\times10^8\,M_\odot$ could host dwarf galaxies and are therefore outside the scope of this work.

2.2. Background particles and initial conditions

The background star and DM particles were defined in two grids superimposed on each other but shifted in x and y by $L/(2\times512)$. Although this initialisation is not realistic, we did not expect any spurious structures to form due to the sufficiently large velocity dispersion of the background particles.

For the background, we assumed a total mass density of $10^6\,M_\odot\,{\rm kpc}^{-3}$ for DM and $10^2\,M_\odot\,{\rm kpc}^{-3}$ for stars. These values roughly match the mass densities of the DM and smooth stellar halo components at 30 kpc from the Galactic centre. The stellar halo of the MW, with a mass of approximately 4 to $7\times10^8\,M_\odot$, comprises of distinct smooth and clumpy components, each contributing roughly equally to the total mass (Bland-Hawthorn & Gerhard 2016; Deason et al. 2019). This study focuses on discerning the influence of dark subhalos within the smooth, virialised portion of the stellar halo, deferring the exploration of detecting stellar wakes within the portion that remains incompletely phase-mixed to future studies. Figure 1 shows the mass density profiles of the virialised stellar and DM halos. The former is obtained by fitting the Einasto mass density profile (Einasto 1972) as reconstructed in Hernitschek et al. (2018) to its total mass. For the total mass we adopted three different values, namely $2\times10^8\,M_\odot$, $4\times10^8\,M_\odot$ and $7\times10^8\,M_\odot$. The DM halo is described by a generalised Navarro-Frenk-White (gNFW) density profile (Zhao 1996). From the figure it is clear that the mass density of the DM is always higher than that of the stars, and that this difference increases rapidly with Galactocentric distance.

In order to simulate the above-mentioned ambient densities, we generated $N_{\rm bkg} = 2\times512^3$ particles, and divided them equally into DM and stellar particle types. The mass values assigned to the two particle types were scaled to satisfy the ratio of total stellar mass to the total DM mass in the Galactic halo. This means that given the number resolution of 2×512^3, the star particles were assigned a mass $M_{\rm stars}\approx1.3\,M_\odot$ whereas the DM particles

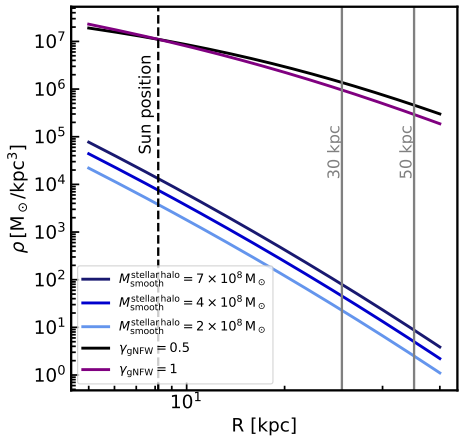


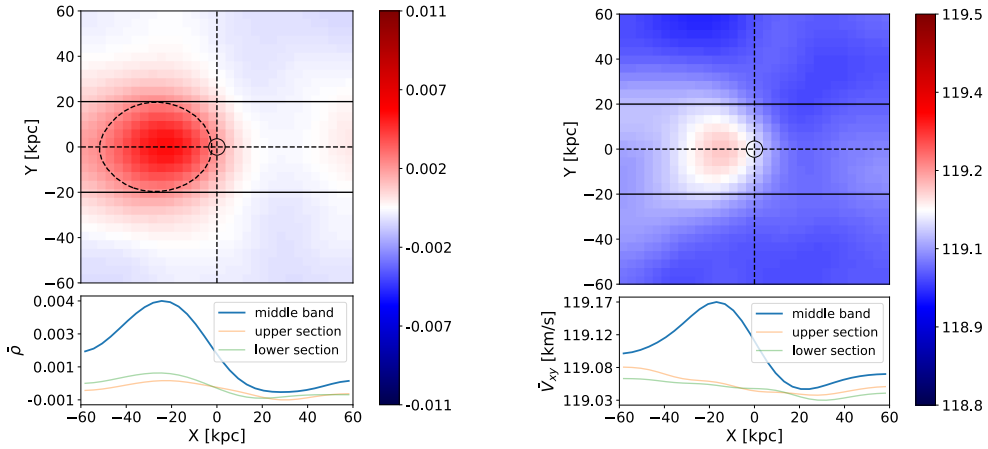
Fig. 1. Mass density profile of the virialised stellar and DM halos in the MW. Both gNFW profiles assume a scale-radius and local DM density values of $R_s = 20$ kpc and $\rho_0 = 0.011\,M_\odot/{\rm pc}^3$ (Benito et al. 2021), respectively.

were initialised with $M_{\rm DM}\approx1.29\times10^4\,M_\odot$. For the softening of both particle types, we adopted a widely used approach in the literature (Potter et al. 2017) by setting the softening length to 1/50 of the mean inter-particle separation such that $\epsilon_{\rm bkg}$=3.72 pc.

We used a 3D isotropic Maxwellian velocity distribution for the velocities of both particles $\boldsymbol{v}_{\rm DM}$ and $\boldsymbol{v}_{\rm star}$. In practice, the velocity components of each Cartesian direction of the DM and star particles were generated by sampling from a 1D Gaussian distribution centred at zero and with standard deviation $\sigma_{\rm DM}$ and $\sigma_{\rm star}$, respectively (see values in Table 1). In order to find a reasonable DM particle velocity dispersion ($\sigma_{\rm DM}$), we turned to cosmological simulations of MW-sized galaxies and the reported DM dispersion profiles reported therein. In particular, we looked at studies using data from both the Aquarius Project (Navarro et al. 2010) and the FIRE-2 simulations (McKeown et al. 2022), and deemed a reasonable DM dispersion at 30 kpc to be $\sigma_{\rm DM} = 200\,{\rm km\,s^{-1}}$. The choice of the velocity dispersion for the stellar particles ($\sigma_{\rm star}$) was motivated by the Galactocentric velocity dispersion profile of halo stars obtained in Deason et al. (2012). As we intend to reproduce the physical conditions of the Galactic halo at 30 kpc from the Galactic centre, we assumed a value of $\sigma_{\rm star} = 95\,{\rm km\,s^{-1}}$. The physical parameters adopted for this case are summarised in Table 1.

2.3. Stellar wakes

Figure 2 shows the star particles of a subhalo simulation with mass $5\times10^8\,M_\odot$ after an integration time of approximately 195 Myr. At this particular time stamp, the stellar wind has moved a distance of about half the length of the box, giving the wake sufficient time to form. At the same time we took care



(a) Overdensity

(b) Mean $X-Y$ speed [km/s]

(c) Dispersion $X-Y$ speed [km/s]

(d) Divergence $X-Y$ velocity [km/s/kpc]

Fig. 2. Stellar phase-space feature maps shown in the reference frame of the simulation box (i.e. frame at which the perturber is moving from left to right with an initial speed of 225 km/s) extracted from a simulation with a perturber of mass $5 \times 10^8\,M_\odot$ after an integration time $t = 194.94$ Myr. We note that in the simulation box's reference frame, the average (unperturbed) 3D velocity of the wind is 0 km/s, and its average (unperturbed) speed in the $X-Y$ plane is 119 km/s. The maps are generated from data contained in a z-slice of $z \in [-20, 20]$ kpc. Panel a: Overdensity. Panel b: mean speed. Panel c: speed dispersion. Panel d: divergence. The panels show the Gaussian-smoothed features projected onto the $X-Y$ plane. Inside the dashed contour of panel a, we show the half-max region of the overdensity. Each subfigure includes a lower plot that shows each Y-band's radial profile along the X-axis. The perturber is situated in the middle of the histogram, with the black circle depicting its scale radius. We observed that the wake effects are seen in all four of the phase-space features.

to avoid snapshots at later times where the simulation particles, having already interacted with the perturber, cross the boundaries on the left and reappear on the right. The reason for this was to prevent any unphysical effects arising from the wake interacting with itself from manifesting in our data. Therefore, we used simulation snapshots at this particular point in integration time to plot the wake and later generate ML datasets (see Sect. 3). The figure is plotted from data that lies in a slice of $z \in [-20, 20]$ kpc and it is binned spatially along x and y into 2D histograms with a bin size of 3.75 kpc. For better visibility, we summed the results from ten different simulations with the same subhalo mass and took the mean across these simulations.

In the figure, we show 2D histograms of four phase-space features of the stellar particles: relative overdensity, mean speed, speed dispersion and velocity divergence. To compute the last three features, we used the velocity of the stellar particles in the X-Y plane. Under each figure, we also show how the particular feature varies over the X-coordinate as three profiles that average the quantity in different Y-bands of the simulation box. Instead of raw histograms, we show Gaussian smoothed variants that aim

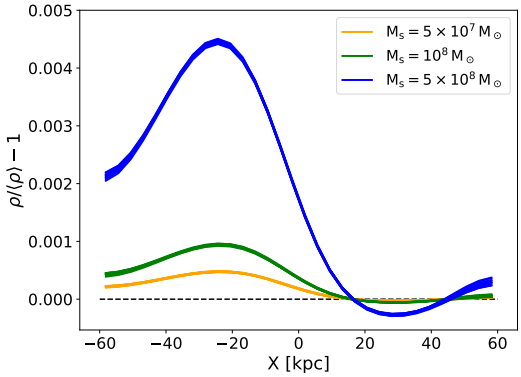


Fig. 3. Background-subtracted overdensity response profiles averaged across $Y \in [-20, 20]$ kpc. Each coloured band corresponds to a different subhalo mass and consists of profiles from 12 simulations; each have a different initial random seed. The figure demonstrates that the amplitude of the density response scales with the subhalo mass.

to reduce the overall noise in the figure while preserving the most important features of the wakes. For example, in Fig. 2, features in the radial profile of the velocity divergence become more discernible compared to its unsmoothed counterpart. As also shown in Foote et al. (2023), the divergence exhibits a dip behind the formed wake.

The wake is most clearly visible in the upper left panel of Fig. 2 as an overdense region. The half-max region (i.e. region in which overdensity exceeds half of the maximum, denoted as a dashed line) extends from the middle of the box in the -X direction and is contained between $Y \in [-20, 20]$ kpc. The overdensity inside a particular bin (i,j) is computed with the following equation:

$$\bar{\rho_{i,j}} = \frac{\rho_{i,j}}{\hat{\rho}} - 1, \tag{3}$$

where $\rho_{i,j}$ is the mass density inside bin (i,j) and $\hat{\rho}$ is the average stellar mass density in the simulation box.

We inspected how the mass of the subhalo affects the maximal overdensity response in the stellar medium by running the simulation with identical initial conditions both with the subhalo and without. In Fig. 3 we show the Gaussian-smoothed density response of different mass halos after having subtracted the background-only simulation fluctuations exactly from the halo case. We observed that the density peak scales with the subhalo mass and as such we expect the signal to be considerably lower as we explore the detectability of masses lower than $5 \times 10^8\,M_\odot$. Interestingly, while we observed that the amplitude of the maximum overdensity is a function of the halo mass, we did not see a similar correlation for the relative position of the maximum. In fact, we saw the same wake maximum location in the X-coordinate for both the lowest and highest mass halos with the difference being only in the response amplitude. We checked and confirmed that this density peak location is dependent on the subhalo velocity. In particular, we looked at a case where we simulate conditions that mimic the stellar halo at 50 kpc from the Galactic centre (Case 2 in Table 1). In this case, where the perturber is moving $25\,\mathrm{km\,s^{-1}}$ slower than in our baseline simulations, we observed that the peak of the density profile is shifted closer to the location of the subhalo. Our simulations also suggest that the physical size of the stellar wake is considerably larger than what is expected from Buschmann et al. (2018). Similar wake characteristics have also been shown in Foote et al. (2023). Be as it may, we leave the study of the discrepancy between expected wake size from theory and simulation to future investigations.

3. Deep learning approach

In this section we introduce our mock data generation procedure and the deep learning model used to detect the stellar wakes caused by subhalos of varying masses. In this first approach, we studied the extent to which we are able to detect a subhalo of any particular mass, formulating the detection as a binary classification problem. A given set of N star particles, each described by position and velocity vectors $(\mathbf{p}, \mathbf{v})$ can be described by a $N \times 6$ array $X \in \mathbb{R}^{N\times 6}$. In general, the ideal discriminator between the subhalo and no subhalo hypothesis is the ratio

$$D(X) = \frac{L(X|\text{subhalo})}{L(X|\text{no subhalo})},$$

where the likelihoods $L(X|\text{subhalo})$ and $L(X|\text{no subhalo})$ are unknown in practice. We therefore approximated $D(X)$ with the output of a binary discriminator model $\tilde{D}(X)$ that is optimised on simulation samples that implicitly follow the unknown likelihoods.

3.1. Dataset generation

We used the wind tunnel simulations described in Sect. 2 to generate mock datasets for the purpose of training and evaluating our ML model. In addition to running simulations with a subhalo mass of $5 \times 10^8\,M_\odot$ (as shown in Sect. 2.3), we also produced simulations with two additional subhalo mass configurations ($5 \times 10^7\,M_\odot$ and $10^8\,M_\odot$), as well as a configuration where no subhalo is present. We ran the simulation for each mass configuration listed above 48 times using unique random seed settings to draw varying particle velocities from their respective distributions and thus generated additional statistically independent data.

The full dataset of all simulations was divided into samples, each sample consisting of approximately 1.3×10^6 star particles, corresponding to 1% of the number of simulated star particles, 512^3. The samples served as the basis of our analysis, as we aimed to distinguish samples from simulations where a subhalo was present with respect to simulations where there was no subhalo. In a real survey, a single sample could represent a candidate collection of stars of the survey (a region of interest) for which one wishes to infer the likelihood of a subhalo being present.

Each sample array X now consists of $1.3 \times 10^6 \times 6 \simeq 8 \times 10^6$ values – the phase-space properties of all star particles. One approach would be to feed the raw kinematic data of each sample directly to a model for classification. However, this would result in very large datasets required for model training, and may be nonoptimal due to having to learn from raw data and not inserting any physics priors to speed up the process. The alternative approach is to define effective observables, computed from the raw star particle kinematics based on a physics *ansatz*. To summarise the kinematics of a large set of stars, we first started with observables based on 2D histograms by projecting each sample to Cartesian axes in position and velocity. These projections were produced by equally slicing the simulation box into three slices along the Z-coordinate and binning them into 2D histograms with 32 bins along the x- and y-coordinates. Based on the star

particles in each bin along X and Y, we computed the following four features:

– bin overdensity with respect to the background density,
– mean speed in the X-Y plane,
– speed dispersion in the X-Y plane, and
– velocity divergence in the X-Y plane.

The feature histograms for a particular simulation are shown in Fig. 2. After slicing and binning, each sample is thus defined by 3×4 channels such that each sample is reduced from 8×10^6 raw observables to $32 \times 32 \times 12 \simeq 12 \times 10^3$ effective observables.

Before training, we used the Gaussian filter from SciPy's `ndimage` module (Virtanen et al. 2020) on our samples. This filter is designed to smooth the value of each pixel by an amount that is based on the values of its neighbouring pixels. While this smoothing effect blurs the image removing sharp edges, it also works to reduce the overall noise. In our case, we found that a Gaussian kernel of three helps reduce the Poisson noise in the histograms while also preserving the most important features of the wake. The effect that this filter has on the underlying histograms is visible in Fig. 2.

During each training session, we adopted a split of 50, 33, and 17% to divide the simulation data into statistically independent training, validation and testing sets. The training set was used for optimising the model, the validation for the hyperparameter tuning, and the testing set for the final results. For a particular target mass case, we then have 2400 training samples, 1600 validation and 800 testing samples. The derived ML dataset used in our work can be found in zenodo[1].

3.2. Binary classifier

In order to learn the difference between background and subhalo perturbed images, the model has to be provided adequately labelled data to train on. We adopted the simplest labelling possible, where samples derived from simulations containing a subhalo were given an integer label of '1', whereas background simulations (no subhalo) were assigned a label of '0'. As our physics-based observables are in the form of 2D histograms or equivalently images, convolutional neural networks (CNNs, or convnets) were a natural first choice for the model. The CNN has found wide use in most computer vision domains and has been a major contributor to the rise in popularity of deep learning methods in the past decade (Chollet 2021). In our case, we were dealing with images of 32×32 bins (pixels), with 12 features (channels) per pixel, as described above. As the dataset generation is based on a complex N-body simulation and we were limited by computational budget, our training dataset consists of only a few thousand samples, putting us in a small dataset regime. For this reason, we adopted methods that are specifically developed for image classification based on small datasets. In particular, we used harmonic networks (Ulicny et al. 2019a), which use a windowed discrete cosine transform (DCT), to perform a harmonic decomposition of the input features and thus reducing the sensitivity to input noise.

The harmonic layer is different from a standard convolutional layer as it does not learn filters for extracting spatial correlation, but instead operates in the frequency domain and learns the weights of the DCT filters. According to the work presented in Ulicny et al. (2019b), these layers perform better in the case of small datasets when compared to traditional CNNs, which we have confirmed in our dataset directly.

[1] https://zenodo.org/records/12721089

Table 2. Hyperparameter selection for our binary classifier.

Hyperparameter	Range	Final value
Number of z-slices	[1, 2, 3]	3
Filters	[4, 128]	32
Learning rate	[1e-8, 1e-2]	1.9602e-06
Dropout	[0, 0.6]	0.49259
Activation	[relu, selu]	relu
Kernel of 1st layer	[3, 10]	9
Kernel of 2nd layer	[1, 3]	2
Extra layers	[0, 3]	1
Filter expansion	[1, 16]	2

The model was trained with the Adam (Kingma & Ba 2017) optimiser and binary focal cross entropy loss function (Lin et al. 2017) to give larger weight to hard-to-classify samples. The model was implemented using Keras (Chollet et al. 2015) and TensorFlow (Abadi et al. 2015).

The choice of the exact architecture and the number of layers and filters per layer was based on hyperparameter tuning. We used the RandomSearch in the KerasTuner (O'Malley et al. 2019) framework for parameter tuning. The scanned hyperparameters, their initial ranges, and the final values are shown in Table 2, along with the evolution of the loss in Fig. 4. As we divided our simulation box into three slices in the Z-coordinate, the 'number of z-slices' in Table 2 refers to how many of these slices we included in the training. Similar to traditional 2D convolutional layer, the 'filters' hyperparameter configures the output dimension of the layer. In the table, we show the output dimension of the first layer, which we increased two-fold after each successive harmonic layer. 'Learning rate', 'dropout', and 'activation' correspond to the step size of the loss function, the amount of regularisation used after each layer, and which activation function we used. We also show the explored range of kernel sizes for the first two Harmonic layers. In addition to this, we experimented with adding additional layers to these baseline layers showing this as 'extra layers' in the table. Finally, the last parameter in the table is a scalar factor, with which we expanded the output dimensionality of the second-to-last fully connected layer in our model. The full hyperparameter tuning took about 20 hours on one Nvidia RTX2070S.

With the aim of producing statistically independent simulations and training samples, we adopted a unique random seed every time we drew the initial conditions before running any simulation. Due to fluctuations in the simulations, we expect variability in the training performance. In order to assess the effect that this has on our model's performance, we trained the model 30 times for each adopted subhalo mass target. Every run we picked a random permutation of train, validation and testing samples such that the sets of their origin simulations using seeds k, l and m obey $k_{train} \cap l_{val} \cap m_{test} = \emptyset$. This allowed us to separate training and testing samples during training, average any metrics relevant to the model performance across the training runs, and also report the error bars.

During each training run, we used early stopping to halt training after validation loss has not decreased during the last 5 epochs. With a constant learning rate of $\simeq 2 \times 10^{-5}$, the total training time for a particular mass case (30 runs) adds up to about 1 hr on one Nvidia RTX2070S. We show the training and validation loss progression in Fig. 4, where each set of coloured lines corresponds to a particular subhalo target case. As expected, the approximate final loss value plateaus differ for each

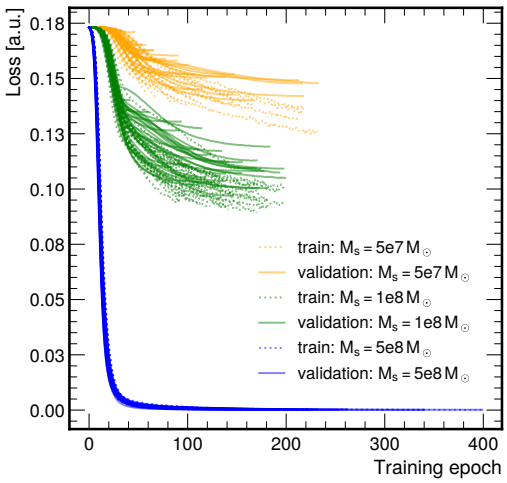


Fig. 4. Training and validation loss of the binary classifier model after running the model 30 times. The training loss of the model measures the discrepancy between the predicted outputs of the model and the actual targets in the training dataset. The validation loss depicts the model performance on the validation data and is therefore a measure of how well the model generalises to unseen data.

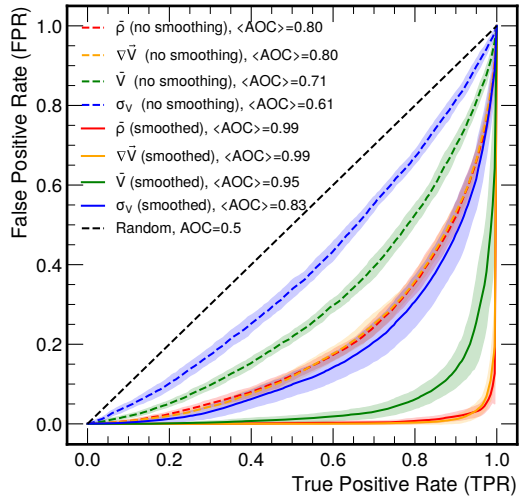


Fig. 5. Binary classifier performance for $5 \times 10^8\,M_\odot$ when training on images generated from the middle slice ($Z \in [-20, 20]$ kpc) of the simulation box. The different coloured bands depict the performance when training on different features: red – overdensity, yellow – divergence, green – mean speed, and blue – speed dispersion. Model performance where training was done on Gaussian smoothed features is depicted by the solid lines, whereas dashed lines show when training was done on unsmoothed features. We observed that best performance is achieved by using smoothed features out of which overdensity and velocity divergence are most effective.

subhalo mass target and we saw that the model's training difficulty decreases as the subhalo mass increases. We also observed that the final training and validation loss values exhibit more scatter in the case of the lighter subhalo masses. When using datasets with a subhalo of $5 \times 10^8\,M_\odot$, the training is more stable, as both losses plateau at smaller values and show a much smaller variance between training runs. We present our binary classifier's final results and discuss its detection performance in Sect. 4.

4. Results

4.1. Effect of spatial and kinematic training features

We used the binary classifier model described in Sect. 3.2 to study which physical observables or their combination would be most useful for detecting the stellar wakes. We quantified the performance using the receiver operating characteristic (ROC) curves and the area over the curve (AOC). The ROC curves represent the model's sensitivity (true positive rate, TPR) and specificity (false positive rate, FPR) across all possible threshold settings. To assess the physics and model performance, the model was trained and evaluated 30 times on independent sets of training and testing datasets. Below we summarise results from testing different feature engineering and selection options.

– Gaussian smoothing: we inspected how the model performance is affected when the Gaussian filter (introduced in Sect. 3.1) is applied to the training features. Figure 5 depicts performance for the target mass $5 \times 10^8\,M_\odot$ when training our baseline binary classifier model (detailed in Sect. 3.2) separately on each of the four features introduced in Sect. 3.1. Solid lines show results when training is done on Gaussian smoothed features whereas dashed lines show results when smoothing is turned off. We found that in the case of smoothed features, we see a significant improvement (≈25−35%) in model performance when compared to their non-smoothed counterparts.
– Individual feature performance: from the same figure, we observed that overdensity and divergence (AOC = 0.99 for both) seem to be the most effective training features, followed by mean speed (AOC = 0.95) and lastly by speed dispersion (AOC = 0.83). We repeated the same exercise for a lower mass target case ($10^8\,M_\odot$) to see whether these conclusions are affected by the mass of the simulated subhalo but our results remained qualitatively the same. Namely, in terms of AOC values, mean dispersion yields 0.58, mean speed in X-Y 0.68 (17.24% increase) and divergence yields a value score of 0.73 (a further 7.35% increase). We also checked that by using the Cartesian velocity component v_x instead of the mean speed (v_{xy}), there is no statistically significant difference in performance between the two.
– Combining kinematic features: we studied how our model performance is affected when combining different kinematic features. For this purpose, we performed three training runs: first we trained only on divergence, then added dispersion, and finally including mean speed. This enabled us to quantify the difference between performance when using one, two or three kinematic features. We observed AOC values of 0.73, 0.71 and 0.71 respectively. While all features exhibit positive constraining power when used individually, we did not observe a stacking effect in the overall performance of the model when other kinematic features are combined with divergence. We concluded from these results that divergence, as expected, already contains much of the information present in the other two features.
– Combining kinematic and spatial features: we found that training on overdensity and velocity divergence yields the best classification performance, with no significant improvement when adding the other two kinematic features. For

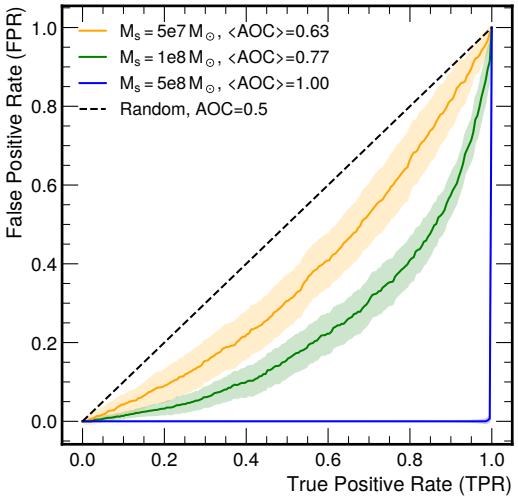


Fig. 6. Receiver operating characteristic curves for the binary classifier trained on datasets with varying subhalo masses. The curves represent subhalos with masses 5×10^7 $M_\odot$ (<AOC>=0.63), 10^8 $M_\odot$ (<AOC>=0.77), and 5×10^8 $M_\odot$ (<AOC>=1.00) solar masses. The width of the bands represents the standard deviation of the curves when training and evaluating the model 30 times. The median AOC values indicate the classifier's performance in distinguishing between background (no subhalo) and the presence of a subhalo. The performance of the binary classifier scales with the mass of the subhalo.

this two-feature combination and the 10^8 $M_\odot$ mass case, the model correctly identifies 74% of signal samples at optimal threshold while misclassifying background samples at a rate of 35%. In contrast, using only overdensity information resulted in 70 and 40%, respectively This highlights the added value of kinematic information in distinguishing signal from background.

– Miscellaneous feature experiments: in addition to the above, we also tried a few other options with hopes to increase model performance. For example, we looked at using relative velocities (normalising the kinematic features around 0 analogously to Eq. (3)). Furthermore, since we slice our data into three equal slices in the Z-coordinate, we also investigated whether we should keep or exclude the outer layers during training. Using small experiments we confirmed that adopting relative velocities instead of absolute ones and the inclusion of the upper and lower Z-slice do not improve our results. Consequently, we decided to keep absolute values and proceeded with training only on data from the middle slice of the box, which contains the majority of the wake.

Our findings described above may suggest that the detection of subhalo masses considered in the current work is achievable with either positional or kinematic data but a combination of both yields strongest results. As we found that including both overdensity and divergence during training results in the best performance, we decided to adopt this feature combination for all ML models used in this study.

4.2. Binary classification performance

We present the performance of our binary classifier for our chosen target cases in Fig. 6. For a particular target case, we show the median ROC (solid lines in Fig. 6) as well as the standard deviation across multiple training runs (shown as the shaded area in Fig. 6). We saw that at all tested masses, the model is able to distinguish between samples from the background and subhalo simulations better than random choice. Furthermore, as expected, subhalos with higher masses and thus a more prominent wake are detected with higher accuracy. This demonstrates that there is sufficient residual information to distinguish the presence of a subhalo in the wake of the stellar particles down to $M = 5 \times 10^7$ $M_\odot$ under the ideal conditions. As was already suggested in the training loss curves of Fig. 4, we observed that for $M = 5 \times 10^8$ $M_\odot$, the scatter appears to be negligible across the runs, but the same cannot be said about the other target cases.

The variance in the ROC scatter and median AOC was investigated in dedicated ablation studies by changing the size of the training dataset. For the lower-mass subhalo target of 5×10^7 $M_\odot$, the binary classifier's performance, trained on 25, 50, 75, and 100% of the available data, resulted in AOC values of 0.586 ± 0.118, 0.609 ± 0.083, 0.621 ± 0.064, and 0.628 ± 0.059, respectively. In addition, we also investigated how architectural changes from our baseline model impact our results and found that the current configuration is optimal within the tested set of configurations. These studies confirmed our hypothesis that our results are most significantly affected by the amount of available training data. That is, by increasing the number of statistically independent samples, the training becomes both more stable and accurate.

The remedy for this issue might seem trivial (i.e. generate more data), but in practice, running more simulations after a certain point becomes cumbersome as it would require implementing data reduction techniques or access to substantial computing resources. In recent years, the development of deep generative models and emulators have begun to push the boundaries in terms of fast data generation for different simulation-based inference problems (see e.g. in Ramesh et al. 2022; Hemmati et al. 2022), which may be interesting to explore in future studies. We expect that with increased simulation datasets, our results can be significantly improved and, at the same time, the effect of uncertainties on the detectability can be studied.

4.3. Multiple mass hypothesis testing

In addition to studying the ability to infer the presence of a subhalo in our samples, we also investigated how well we can discern between the different subhalo mass cases in a multiple-hypothesis case. This time, instead of using both background and signal samples, we trained exclusively on samples containing all three signal cases, labelling them from lowest mass to highest as 0, 1 and 2. As before, we trained the model 30 times with early stopping, where at the start of each run, we picked a random permutation of simulation seeds for training, validation, and testing. Figure 7 shows the training and validation loss curves for all training runs. We observed that for each iteration of the dataset shuffle, both training and validation losses decrease smoothly over time and start to plateau at around 100 training epochs.

Instead of a single prediction score, each test sample was given three scores, each of which represents the probability of belonging to a particular target class. In each of the cases, the model is able to discern between samples in the testing dataset when there is a clear difference between the prediction distribution of the samples actually belonging to the particular target case with respect to the rest. We could then summarise the accuracy of our model, that is, how well it is able to discriminate between these distributions with a confusion matrix in Fig. 8.

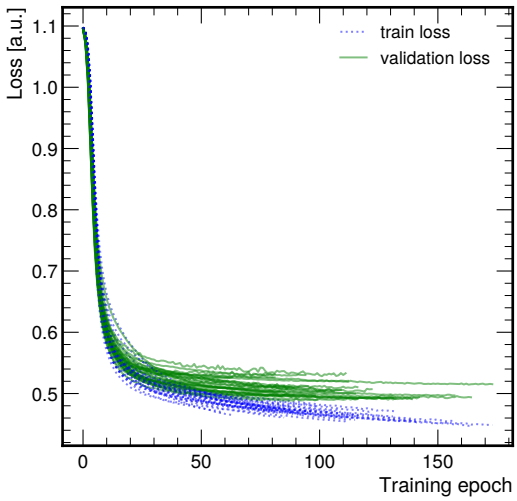


Fig. 7. Training (blue) and validation (green) loss of the multiple mass hypothesis classifier model after running the model 30 times with early stopping.

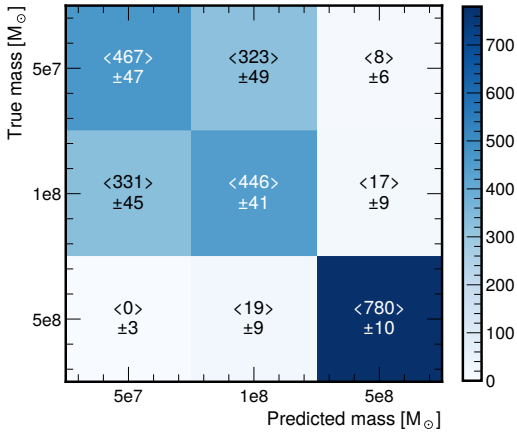


Fig. 8. Multiple mass hypothesis performance summarised in terms of the predicted and true mass of the subhalo in test samples. Each element in the confusion matrix is characterised by the mean number predictions and the standard deviation across 30 runs of training and evaluating.

Ideally we would like to maximise the values of elements on the main diagonal that depict the number of instances the model is able correctly predict the mass of the subhalo. The off-diagonal elements show mismatches between predicted and true labels and thus indicate which targets are harder to classify for the model. Since we run the model 30 times, we show the prediction count values of each element in the matrix by computing the mean and standard deviation across all runs. We note that since we average many training runs, we do not expect the counts across columns to sum to the total number of samples (800) in each target mass test dataset. We do however expect this sum to be within the standard error across the runs.

Similar to what we saw in the binary classifier analysis of Sect. 4.2, we again observed that the model performs best in the case of the heaviest subhalo mass (5×10^8 M$_\odot$). In this case the model was able to identify the correct mass of approximately 780 samples with a small scatter in the mean number of predictions (±10) and mislabel 20 samples as other targets. For the lower masses of $M = 5 \times 10^7$ M$_\odot$ and $M = 10^8$ M$_\odot$, the task was more challenging as we observed a larger scatter in correct predictions counts (±47 and ±41) as well as a tendency to mislabel the samples between these two. In both cases, about 300 samples were mislabelled. Since wake effects created by a subhalo of mass $M = 5 \times 10^7$ M$_\odot$ are considerably smaller than those created by 5×10^8 M$_\odot$, similar performance between the lower mass cases points to a difficulty in identifying intermediate mass samples.

4.4. Detection performance 50 kpc from the Galactic centre

On top of inspecting the subhalo detection performance at 30 kpc from the Galactic centre, we also looked at what would happen if the perturber was in orbit at 50 kpc, which is roughly the distance to the LMC. While Foote et al. (2023) studied wakes created by LMC-sized subhalos, in this study we are interested in effects created by much smaller subhalos. We then ran our simulation again using our intermediate subhalo mass, $M = 10^8$ M$_\odot$, and as before, we configured the background and perturber phase-space parameters to literature-informed values (summarised in Table 1). By modulating these different parameters we expect the density response of the perturber to change. For example, the Chandrasekhar dynamical friction equation (Chandrasekhar 1943) suggests that at a constant perturber velocity, the reduction in the ambient velocity dispersion will result in a larger deceleration (i.e. density wake) of the perturber. This classical dynamical friction equation, however, does not take into account the effects of self-gravity, and applies only to specific idealised conditions. Furthermore, the combined effects of all background and subhalo parameter changes (e.g. subhalo velocity, stellar and DM mass density, etc.) on the actual amplitude and extent of the stellar wake are not easily estimated beforehand and are thus interesting to explore. We leave a full investigation of the relationship between simulation phase-space parameters and wake observables for a future work and continue with results from our ML analysis.

Using data from the 50 kpc simulations, we derived new ML samples in exactly the same way as was described in Sect. 3.1. This way we ensure that the performance comparison between these two cases is done on a fair basis. Without making any changes to the binary classification model, we trained the model again with the same setup, and we represent the results from these runs in Fig. 9 with a green band. We observed a performance similar to the first case for these new samples. This shows that our binary classification model is able to learn from a completely new and independent dataset and that our previous results are not case specific. One physical interpretation of the similarity between the two cases could be that 20 kpc is too small a distance for the phase-space parameters to change enough to have an impact on our detection model. In other words, the slopes of, for example, mass density and velocity dispersion profiles are too small and perhaps the Galactocentric distance should be even larger. The usefulness of (small mass) subhalo simulations in environments out to >100 kpc is another question as the lack of stellar observations with adequate precision discourages the detection of the subhalo induced wake effects.

In addition to the above, we also looked at the detection performance when evaluating the new data (subhalo orbit at 50 kpc) on a previous model that was trained on data when the subhalo was 30 kpc from the Galactic centre. We show the results as the grey band on Fig. 9. We again observed similar performance as

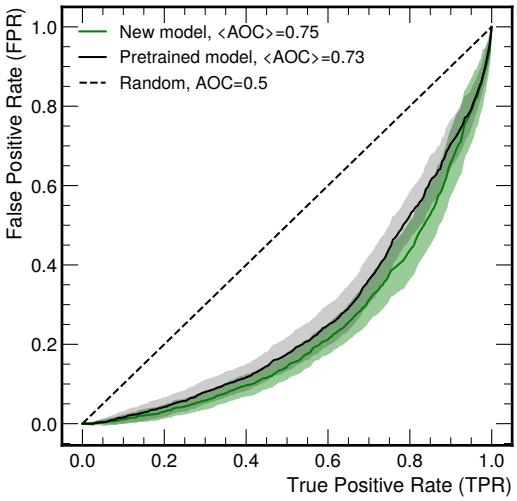


Fig. 9. Binary classification results of $M_s = 10^8\,M_\odot$ when the orbit is placed at 50 kpc from the Galactic centre. Green shows the performance of the new model, which is trained and evaluated on simulation data from 50 kpc. The black line shows the model performance when training is done on data describing a subhalo orbit at 30 kpc from the Galactic centre.

before, which is a good indication that our model is able to generalise to new conditions. As expected, the AOC of the black ROC curve is smaller as in the case of the green band the model was trained on the new dataset and is therefore better tailored to make predictions on it.

5. Discussion

The physical setup of the idealised simulations described in this work can certainly be improved upon in many aspects. For example, the current setup does not include the gravitational potential of the Galaxy or the effect of tidal stripping. Also, it would be interesting to see how the stellar wakes and their detection performance is affected when using different density profiles (e.g. Navarro-Frenk-White, Einasto, etc.) for the subhalo.

In a future study, we also plan to investigate how our results are affected by the inclusion of observational effects. Specifically, we would like to relate our data from ideal simulations to real surveys (e.g. *Gaia*; Gaia Collaboration 2016) by studying the detectability at varying error levels in an observational frame of reference. Due to their large spatial extension, we do not expect to detect stellar wakes in their entirety. However, we know from Bazarov et al. (2022) that we are able to observe a signal when looking at regions near subhalos on a star-by-star basis. In any case, acquiring intuition on the actual physical scales of the stellar wake phenomena could be an important step as we start look for and identify suitable regions of interest from real survey data.

In addition to creating mock datasets in a new frame of reference, the ML models will also need to be adapted. In the current work, we used three overdensity images (slices) per sample for training. This means that in order to evaluate already trained models with new data, the input needs to conform to the same dimensionality that is (N, 32, 32, 2). Creating similar Z-slices in an observational setting is not as straightforward as was the case in our idealised box simulation setup. In our case, the thickness of the slices is chosen arbitrarily to divide the simulation region into three equal slices wherein the middle layer contains the subhalo and majority of the stellar wake. In a mock dataset of an observational region of interest, the decomposition of data into slices along the line of sight might not be justified altogether as the position of subhalo is not localised and the direction of motion is arbitrary with respect to the coordinate axes.

Even though we have achieved very good classification performance on samples containing very heavy subhalos (i.e. $M_s > 10^8\,M_\odot$), we still have room for improvement in identifying lower mass target cases. One direction to tackle this would be to consider alternative ML approaches and architectures as the models described in the current work are certainly not exhaustive. For example, it would also be interesting to see how well one would be able to predict the subhalo mass as part of a regression model setup.

While other methods are possible, we find that in our case, the key limiting factor is the amount of available training data. By retraining our binary classifier while modulating the amount of available training data we have seen that with more data we achieve increased AOC values and smaller scatter in the ROC curves. The data problem is something that could be overcome with access to considerably larger computing resources or finding alternative ways to generate simulation data faster (e.g. emulators, generative models, etc.). Since we have not reached a performance plateau, it is difficult to give estimates on sufficient training dataset sizes. We would like to note that the ultimate goal, which is the focus of a future study, is how well our model generalises to observational data rather than trying to learn the simulation data perfectly.

6. Conclusions

Constraining the SHMF in the sub-galactic mass regime is an important endeavour in order to understand more about the particle nature of DM. Theoretically predicted dark subhalos are extremely difficult to detect, as their presence can only be inferred from gravitational effects on the surrounding stellar medium. In this paper, we studied the strength of the DM subhalo-induced gravitational signal by investigating how well we can detect individual stellar wakes induced by orbiting subhalos in the stellar halo.

We implemented wind tunnel simulations with self-gravity enabled using `Pkdgrav3`, replicating the ambient phase-space conditions of DM and stars at 30 and 50 kpc from the Galactic centre. Interestingly, we observed stellar wakes in line with those for larger perturber masses as described in Foote et al. (2023) but significantly more spatially extended than those in Buschmann et al. (2018). The former study finds that for perturbers with masses of $O(10^{11}\,M_\odot)$, the inclusion of self-gravity increases the magnitude of the density response by roughly 10% while also significantly extending the length of the overdensity and kinematic wake. In the latter work, self-gravity was not considered, but we found in our simulations that although the removal of self-gravity reduces the spatial extension of the wake, this omission alone does not fully account for the difference.

We then derived mock datasets by binning the simulated data into 2D histograms and computing different physical observables in each bin to be used as training features. The phase-space features that we implemented were the overdensity, mean speed on the X-Y plane (V_{xy}), and its dispersion and divergence. We find that by applying a Gaussian smoothing filter on the features prior to training, a significant increase in classification performance can be observed. Even though all the considered

features showed a non-trivial constraining power when used exclusively, we find that the combination of overdensity and velocity divergence is equivalent to using all four features. This became evident, as including additional kinematic features did not significantly improve classification performance when divergence was already included in the training dataset. In any case, these findings suggest that stellar wakes may best be found in ongoing or future stellar surveys by using a combination of positional and kinematic information, which in our study exhibited comparable constraining power.

Finally, we divided our ML approach into two parts. First, we investigated how well we are able to infer the presence of different mass subhalos in the generated images. We implemented a binary classification model that we then trained and evaluated on our three target mass cases: 5×10^7 $M_\odot$, 10^8 $M_\odot$, and 5×10^8 $M_\odot$. We saw that for all the chosen target cases, we are able to infer the presence of a subhalo at a rate that is better than random. As expected, we observed that the performance follows a hierarchical trend such that more massive subhalos exhibit more signal and are easier to detect. We also investigated our binary classification model's performance, having simulated a subhalo of mass 10^8 $M_\odot$ at 50 kpc from the Galactic centre. Using this new simulation data, we compared the classification performance of a model that was trained on a newly derived ML dataset against the pretrained model at 30 kpc. We found similar results in both cases and saw that our model's performance is generalisable to data from simulations with different physical conditions.

We also studied the classification between different subhalo masses in a multiple-hypothesis case. We find that the model is able to recognise and correctly label subhalos of mass 5×10^8 $M_\odot$ about 97% of the time, demonstrating a potential capability to constrain subhalo masses.

This work is summarised as follows:

- We used ML to evaluate how effectively we can detect individual stellar wakes induced by DM subhalos in the MW's stellar halo.
- Our simulated stellar wakes are in line with Foote et al. (2023) but significantly more spatially extended than previously reported in the literature. We found that the inclusion or omission of self-gravity does not fully account for the difference.
- In the context of detection performance, we found that
  - Gaussian smoothing plays a crucial role, improving AOC values by approximately 25–35%.
  - The combination of overdensity and velocity divergence results in maximal performance, achieving a TPR of 60, 74, and 99% and an FPR of 41, 35, and 1% for the 5×10^7 $M_\odot$, 10^8 $M_\odot$, and 5×10^8 $M_\odot$ mass cases, respectively.
  - Training only on overdensity reduces the performance to a TPR of 70 and 97% and an FPR of 40 and 5% respectively for the 10^8 $M_\odot$ and 5×10^8 $M_\odot$ subhalo cases.
  - With the amount of training data available (4800 samples), the 5×10^8 $M_\odot$ subhalo is perfectly identifiable when using only a 1% fraction of all star particles present in the snapshot (i.e. 1.3 million star particles).
  - Detection performance for smaller subhalos is significantly reduced, with the amount of available training data being the key limiting factor.
  - We found that our performance remains effectively unchanged when varying the subhalo's position relative to the Galactic centre within 50 kpc, demonstrating generalizability to data under different physical conditions.
  - In a multi-class classification scenario, the model performed best for the heaviest subhalo mass (5×10^8 $M_\odot$), correctly classifying around 97% of these samples.

The ML approach presented in this work serves as a proof of concept for detecting stellar wakes in the MW stellar halo. While demonstrating significant constraining power for the subhalo masses considered in this study, the current model can be further refined to increase its robustness and applicability. Therefore, we hope that the work presented in this paper encourages further studies in this domain, ultimately aiding in the exploration of stellar wakes through current and future stellar surveys.

Acknowledgements. We would like to express our gratitude to the referee for their constructive feedback, which has significantly contributed to the improvement of the results presented in this work. We thank Hayden Foote for valuable discussions. We would like to thank Martti Raidal for pointing us towards using ML approaches to study dark subhalo detection in the MW. This work was supported by the Estonian Research Council grants PSG938, PSG864 and PRG1006, the Estonian Ministry of Education and Research (grant TK202) and the European Union's Horizon Europe research and innovation programme (EXCOSM, grant No. 101159513). We further acknowledge the support of the European Consortium for Astroparticle Theory in the form of an Exchange Travel Grant. I.A.A. and C.D.V. thank the Ministerio de Ciencia e Innovación (MICINN) for financial support under research grant PID2021-122603NB-C22. The results and figures presented in this work were made possible thanks to the following software libraries: Matplotlib (Hunter 2007), NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), HDF5 for Python (Collette 2013) and Jupyter (Kluyver et al. 2016). This research was supported by an academic grant of an A100 GPU from Nvidia.

References

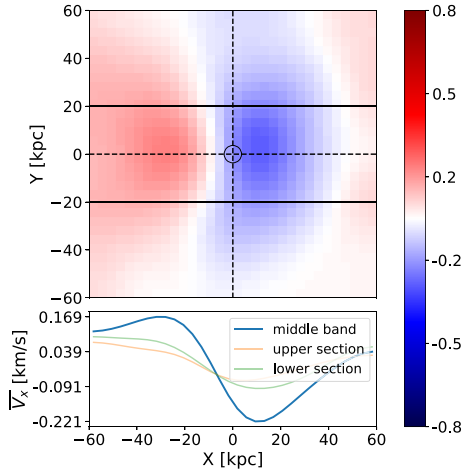
Abadi, M., Agarwal, A., Barham, P., et al. 2015, arXiv e-prints [arXiv:1603.04467]
Alonso Asensio, I., Dalla Vecchia, C., Potter, D., & Stadel, J. 2023, MNRAS, 519, 300
Barry, M., Wetzel, A., Chapman, S., et al. 2023, MNRAS, 523, 428
Bazarov, A., Benito, M., Hütsi, G., et al. 2022, Astron. Comput., 41, 100667
Benitez-Llambay, A., & Frenk, C. 2020, MNRAS, 498, 4887
Benito, M., Criado, J. C., Hütsi, G., Raidal, M., & Veermäe, H. 2020, Phys. Rev. D, 101, 103023
Benito, M., Iocco, F., & Cuoco, A. 2021, Phys. Dark Univ., 32, 100826
Bland-Hawthorn, J., & Gerhard, O. 2016, ARA&A, 54, 529
Bonaca, A., Hogg, D. W., Price-Whelan, A. M., & Conroy, C. 2019, ApJ, 880, 38
Bonaca, A., & Price-Whelan, A. M. 2024, arXiv e-prints [arXiv:2405.19410]
Bovy, J., Erkal, D., & Sanders, J. L. 2016, MNRAS, 466, 628
Buschmann, M., Kopp, J., Safdi, B. R., & Wu, C.-L. 2018, Phys. Rev. Lett., 120, 211101
Chandrasekhar, S. 1943, ApJ, 97, 255
Collette, A. 2013, Python and HDF5 (USA: O'Reilly Media)
Chollet, F. 2021, Deep Learning with Python, 2nd edn. (USA: Manning Pubns Co)
Chollet, F. et al. 2015, Keras, https://keras.io
Conroy, C., Naidu, R. P., Garavito-Camargo, N., et al. 2021, Nature, 592, 534
Darragh-Ford, E., Mantz, A. B., Rasia, E., et al. 2023, MNRAS, 521, 790
Deason, A. J., Belokurov, V., Evans, N. W., et al. 2012, MNRAS, 425, 2840
Deason, A. J., Belokurov, V., & Sanders, J. L. 2019, MNRAS, 490, 3426
Diemand, J., Kuhlen, M., Madau, P., et al. 2008, Nature, 454, 735
Einasto, J. 1972, PhD thesis, Tartu Observatory, Estonia
Einasto, J. 2010, Dark Matter arXiv e-prints [arXiv:0901.0632]
Foote, H. R., Besla, G., Mocz, P., et al. 2023, ApJ, 954, 163
Fushimi, K. J., Mosquera, M. E., & Dominguez, M. 2024, A&A, 688, A147
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Garavito-Camargo, N., Besla, G., Laporte, C. F. P., et al. 2019, ApJ, 884, 51
Garrison-Kimmel, S., Wetzel, A., Bullock, J. S., et al. 2017, MNRAS, 471, 1709
Gramann, M. 1990, MNRAS, 244, 214
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357
Hemmati, S., Huff, E., Nayyeri, H., et al. 2022, ApJ, 941, 141
Hernitschek, N., Cohen, J. G., Rix, H.-W., et al. 2018, ApJ, 859, 31
Hopkins, P. F., Wetzel, A., Kereš, D., et al. 2018, MNRAS, 480, 800
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90

Karukes, E., Benito, M., Iocco, F., Trotta, R., & Geringer-Sameth, A. 2020, J. Cosmol. Astropart. Phys., 2020, 033
Kingma, D. P., & Ba, J. 2017, arXiv e-prints [arXiv:1412.6980]
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, eds. F. Loizides, & B. Schmidt (Amsterdam: IOS Press), 87
Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. 2017, in Proceedings of the IEEE international conference on computer vision, 2980
McKeown, D., Bullock, J. S., Mercado, F. J., et al. 2022, MNRAS, 513, 55
Mulder, W. A. 1983, A&A, 117, 9
Navarro, J. F., Ludlow, A., Springel, V., et al. 2010, MNRAS, 402, 21
O'Malley, T., Bursztein, E., Long, J., et al. 2019, Keras Tuner, https://github.com/keras-team/keras-tuner
Ostdiek, B., Rivero, A. D., & Dvorkin, C. 2022, ApJ, 927, 83
Perivolaropoulos, L., & Skara, F. 2022, New Astron. Rev., 95, 101659
Planck Collaboration VII. 2020, A&A, 641, A7
Potter, D., Stadel, J., & Teyssier, R. 2017, Comput. Astrophys. Cosmol., 4, 2
Ramesh, P., Lueckmann, J.-M., Boelts, J., et al. 2022, arXiv e-prints [arXiv:2203.06481]
Rozier, S., Famaey, B., Siebert, A., et al. 2022, ApJ, 933, 113
Sawala, T., Frenk, C. S., Fattahi, A., et al. 2015, MNRAS, 456, 85
Siegel, E. R., Hertzberg, M. P., & Fry, J. N. 2007, MNRAS, 382, 879
Springel, V., Wang, J., Vogelsberger, M., et al. 2008, MNRAS, 391, 1685
Tamfal, T., Mayer, L., Quinn, T. R., et al. 2021, ApJ, 916, 55
Ulicny, M., Krylov, V. A., & Dahyot, R. 2019a, in British Machine Vision Conference
Ulicny, M., Krylov, V. A., & Dahyot, R. 2019b, in 2019 27th European Signal Processing Conference (EUSIPCO), 1
Varma, S., Fairbairn, M., & Figueroa, J. 2020, arXiv e-prints [arXiv:2005.05353]
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nat. Methods, 17, 261
Wagner-Carena, S., Lee, J., Pennington, J., et al. 2024, arXiv e-prints [arXiv:2404.14487]
Weinberg, M. D. 1986, ApJ, 300, 93
Zavala, J., & Frenk, C. S. 2019, Galaxies, 7, 81
Zhao, H. 1996, MNRAS, 278, 488

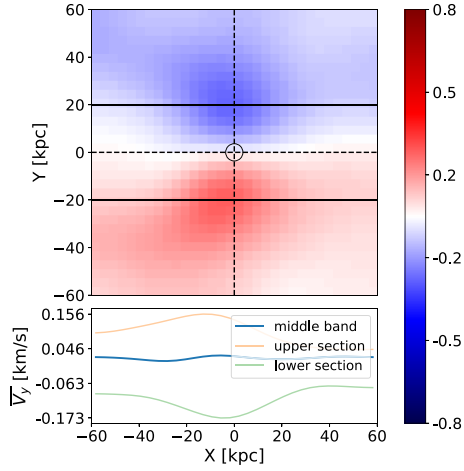
Appendix A: V_x and V_y velocity maps

Figure A.1 shows the V_x and V_y velocity maps of star particles in a simulation containing a subhalo of mass 5×10^8 $M_\odot$. In the same way as in Fig. 2 of the main text, stars of z-slice $z \in [-20, 20]$ kpc are binned in a 2D histogram with 32 bins on both axes. Inside each bin, the velocity components of the star particles are summed and averaged across ten simulations. In this way, the kinematic signatures of the wake become much clearer in the figures.

We note that the velocity scales of Figs. 2b and A.1 differ due to the fact that in the former we show the mean speed in the X-Y plane, whereas in the latter we show maps of the velocity components (V_x, V_y) separately. Since the stellar velocities (V_x, V_y, V_z) are drawn from distributions centred on 0 km/s in the reference frame of the simulation box, the mean velocities in A.1 also naturally average to around 0 km/s.



(a) Mean V_x velocity [km/s]



(b) Mean V_y velocity [km/s]

Fig. A.1: Stellar velocity maps of V_x (a) and V_y (b) in a simulation containing a subhalo of mass 5×10^8 $M_\odot$.

Curriculum Vitae

1. Personal data

Name	Sven Põder
Date and place of birth	12 October 1995 Tallinn, Estonia
Nationality	Estonian

2. Contact information

Address	National Institute of Chemical Physics and Biophysics, High Energy and Computational Physics, Rävala pst 10, 10143 Tallinn, Estonia
E-mail	sven.poder@kbfi.ee

3. Education

2021–...	Tallinn University of Technology, Applied Physics, PhD studies
2019–2021	Tallinn University of Technology, Applied Physics, MSc *cum laude*
2016–2019	Tallinn University of Technology, Engineering Physics, BSc

4. Language competence

Estonian	native
English	fluent
Russian	beginner

5. Professional employment

2021– ...	National Institute of Chemical Physics and Biophysics, Junior Researcher
2018–2021	Metatellus, Software Developer

6. Workshops and Schools

2025	Dynamical Tracers of the Nature of Dark Matter (online)
2025	Moving your AI training jobs to LUMI workshop, Espoo, Finland
2023	Comprehensive General LUMI Course, Tallinn, Estonia
2022	International Doctorate Network in Particle Physics, Astrophysics and Cosmology (IDPASC), Olomouc, Czech Republic
2022	MITP Workshop “Feebly Interacting Sectors Impact on Cosmology & Astrophysics” (online)
2021	COST-MW PhD School: Stellar spectroscopy and Astrophysical parameterisation from Gaia to Large Spectroscopic surveys (online)
2021	Star Clusters: the Gaia Revolution (online)
2020	Internship, National Institute of Chemical Physics and Biophysics
2019	CERN Summer School, CERN, Geneva, Switzerland
2018	Internship, Transformative AI, Tallinn, Estonia

7. Computer skills

- Operating systems: Windows, Linux, iOS
- Document preparation: LaTeX
- Programming languages: Python, C#, Bash
- Scientific packages: NumPy, SciPy, Pandas, Matplotlib, Astropy

8. Honours and awards

- 2023, EuCAPT Exchange Travel Grant
- 2022, NVIDIA Academic Hardware Grant Program

9. Supervision

- 2025, Ele-Liis Evestus, BSc thesis (defended), TalTech - "A chemo-kinematic data analysis of the Milky Way disk"

10. Defended theses

- 2021, "Galactic parameter estimation using spectroscopic data from the Gaia space telescope", MSc, supervisors Dr. María Benito & Dr. Andi Hektor, National Institute of Chemical Physics and Biophysics
- 2019, "A study of the proper motions of red giants in the Large Magellanic Cloud using data from the Gaia space telescope", supervisor Dr. Andi Hektor, National Institute of Chemical Physics and Biophysics

11. Field of research

- Astrophysics
- Machine learning

12. Scientific work
Papers

1. A. Bazarov, M. Benito, G. Hütsi, R. Kipper, J. Pata, and S. Põder. Sensitivity estimation for dark matter subhalos in synthetic gaia dr2 using deep learning. *Astronomy and Computing*, 41:100667, 2022
2. Põder, Sven, Benito, María, Pata, Joosep, Kipper, Rain, Ramler, Heleri, Hütsi, Gert, Kolka, Indrek, and Thomas, Guillaume F. A bayesian estimation of the milky way's circular velocity curve using gaia dr3. *A&A*, 676:A134, 2023
3. María Benito, Konstantin Karchev, Rebecca K. Leane, Sven Põder, Juri Smirnov, and Roberto Trotta. Dark matter halo parameters from overheated exoplanets via bayesian hierarchical inference. *Journal of Cosmology and Astroparticle Physics*, 2024(07):038, jul 2024
4. Põder, Sven, Pata, Joosep, Benito, María, Alonso Asensio, Isaac, and Dalla Vecchia, Claudio. Detection of stellar wakes in the milky way: A deep learning approach. *A&A*, 693:A227, 2025

Conference presentations

1. **S. Põder**. "Detection of stellar wakes in the Milky Way: A deep learning approach", IAUS 397: Exploring the Universe with Artificial Intelligence (UniversAI): 2–6 June 2025, Athens, Greece.
2. **S. Põder**. "Data Driven Dark Matter Searches in the Milky Way", LLM retreat for PhD students and supervisors, 27–28 November 2024, Nelijärve, Estonia. *Awarded a prize*.
3. **S. Põder**. "Searching for Dark Matter Subhalos in the Milky Way using Deep Learning", 4th CERN Baltic Conference (CBC 2024), 15–17 October 2024, Tallinn, Estonia.
4. **S. Põder**. "Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning", CLUES Workshop 2024, 10–14 June 2024, Warsaw, Poland.
5. **S. Põder**. "Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning", Tuorla-Tartu meeting 2024, 6–8 May 2024, Turku, Finland.
6. **S. Põder**. "Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning", 1st European AI for Fundamental Physics Conference (EuCAIFCon), 30 April–3 May 2024, Amsterdam, Netherlands.
7. **S. Põder**. "A Bayesian Estimation of the Milky Way's Circular Velocity Curve using Gaia DR3", Kashiwa Dark Matter Symposium 2023, 5–8 December 2023, Tokyo, Japan (online).
8. **S. Põder**. "The Milky Way's dark matter halo: A Bayesian Estimation of the Milky Way's Circular Velocity Curve using Gaia DR3", 3rd CERN Baltic Conference (CBC 2023), 9–11 October 2023, Riga, Latvia.
9. **S. Põder**. "A Bayesian Estimation of the Milky Way's Circular Velocity Curve using Gaia DR3", Third EuCAPT Annual Symposium, 31 May–2 June 2023, CERN, Switzerland (online).
10. **S. Põder**. "Searching for Dark Matter Subhalos in the Milky Way using Deep Learning", Kashiwa Dark Matter Symposium 2022, 29 November–2 December 2022, Tokyo, Japan (online).
11. **S. Põder**. "Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning", 2nd CERN Baltic Conference (CBC 2022), 10–12 October 2022, Vilnius, Lithuania.

Elulookirjeldus

1. Isikuandmed

Nimi	Sven Põder
Sünniaeg ja -koht	12. oktoober 1995, Tallinn, Eesti
Kodakondsus	Eesti

2. Kontaktandmed

Aadress	Keemilise ja Bioloogilise Füüsika Instituut, Kõrge Energia ja Arvutusfüüsika Laboratoorium, Rävala pst 10, 10143 Tallinn, Eesti
E-post	sven.poder@kbfi.ee

3. Haridus

2021–…	Tallinna Tehnikaülikool, Rakendusfüüsika, doktoriõpe
2019–2021	Tallinna Tehnikaülikool, Rakendusfüüsika, MSc *cum laude*
2016–2019	Tallinna Tehnikaülikool, Tehniline füüsika, BSc

4. Keelteoskus

eesti keel	emakeel
inglise keel	kõrgtase
vene keel	algtase

5. Teenistuskäik

2021–…	Keemilise ja Bioloogilise Füüsika Instituut, nooremteadur
2018–2021	Metatellus, tarkvaraarendaja

6. Töötoad ja suvekoolid

2025	Dynamical Tracers of the Nature of Dark Matter (veebis)
2025	Moving your AI training jobs to LUMI workshop, Espoo, Soome
2023	Comprehensive General LUMI Course, Tallinn, Eesti
2022	International Doctorate Network in Particle Physics, Astrophysics ja Cosmology (IDPASC), Olomouc, Tšehhi
2022	MITP Workshop “Feebly Interacting Sectors Impact on Cosmology & Astrophysics” (veebis)
2021	COST-MW PhD School: Stellar spectroscopy and Astrophysical parameterisation from Gaia to Large Spectroscopic surveys (veebis)
2021	Star Clusters: the Gaia Revolution (veebis)
2020	Praktika, Keemilise ja Bioloogilise Füüsika Instituut
2019	CERN Suvekool, CERN, Genf, Šveits
2018	Praktika, Transformative AI, Tallinn, Eesti

7. Arvutioskus

- Operatsioonisüsteemid: Windows, Linux, iOS
- Kontoritarkvara: LaTeX
- Programmeerimiskeeled: Python, C#, Bash
- Teadustarkvara paketid: NumPy, SciPy, Pandas, Matplotlib, Astropy

8. Autasud

- 2023, EuCAPT Exchange Travel Grant
- 2022, NVIDIA Academic Hardware Grant Program

9. Juhendamised

- 2025, Ele-Liis Evestus, bakalaureusetöö (kaitstud), TalTech - "Linnutee galaktika ketta keemiline ja kinemaatiline analüüs"

10. Kaitstud lõputööd

- 2021, „Galaktiliste parameetrite hindamine kasutades Gaia kosmoseteleskoobi spektroskoopilisi andmeid“, MSc, juhendajad Dr. María Benito ja Dr. Andi Hektor, Keemilise ja Bioloogilise Füüsika Instituut
- 2019, „Punaste hiidude omaliikumise uurimine Suures Magalhãesi Pilves kasutades Gaia kosmoseteleskoobi andmeid“, BSc, juhendaja Dr. Andi Hektor, Keemilise ja Bioloogilise Füüsika Instituut

11. Teadustöö põhisuunad

- Astrofüüsika
- Masinõpe

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-428-5 (PDF)