The Astrophysical Journal, 937:14 (16pp), 2022 September 20 https://doi.org/10.3847/1538-4357/ac85b9

Possibilities and Limitations of Kinematically Identifying Stars from Accreted Ultra-faint Dwarf Galaxies

Kaley Brauer[1], Hillary Diane Andales[1], Alexander P. Ji[2], Anna Frebel[1], Mohammad K. Mardini[3,4], Facundo A. Gómez[5,6], and Brian W. O'Shea[7,8,9]
[1] Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
kbrauer@mit.edu
[2] Department of Astronomy and Astrophysics, University of Chicago, Chicago IL 60637, USA
[3] Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan
[4] Institute for AI and Beyond, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan
[5] Instituto de Investigación Multidisciplinar en Ciencia y Tecnología, Universidad de La Serena, Raúl Bitrán 1305, La Serena, Chile
[6] Departamento de Física y Astronomía, Universidad de La Serena, Av. Juan Cisternas 1200 N, La Serena, Chile
[7] Department of Computational Mathematics, Science and Engineering, Michigan State University, MI, 48824, USA
[8] Department of Physics and Astronomy, Michigan State University, MI, 48824, USA
[9] Facility for Rare Isotope Beams, Michigan State University, MI, 48824, USA
Received 2022 June 13; revised 2022 July 26; accepted 2022 July 29; published 2022 September 16

Abstract

The Milky Way has accreted many ultra-faint dwarf galaxies (UFDs), and stars from these galaxies can be found throughout our Galaxy today. Studying these stars provides insight into galaxy formation and early chemical enrichment, but identifying them is difficult. Clustering stellar dynamics in 4D phase space (E, L_z, J_r, J_z) is one method of identifying accreted structure that is currently being utilized in the search for accreted UFDs. We produce 32 simulated stellar halos using particle tagging with the Caterpillar simulation suite and thoroughly test the abilities of different clustering algorithms to recover tidally disrupted UFD remnants. We perform over 10,000 clustering runs, testing seven clustering algorithms, roughly twenty hyperparameter choices per algorithm, and six different types of data sets each with up to 32 simulated samples. Of the seven algorithms, HDBSCAN most consistently balances UFD recovery rates and cluster realness rates. We find that, even in highly idealized cases, the vast majority of clusters found by clustering algorithms do not correspond to real accreted UFD remnants and we can generally only recover 6% of UFDs remnants at best. These results focus exclusively on groups of stars from UFDs, which have weak dynamic signatures compared to the background of other stars. The recoverable UFD remnants are those that accreted recently, $z_{\rm accretion} \lesssim 0.5$. Based on these results, we make recommendations to help guide the search for dynamically linked clusters of UFD stars in observational data. We find that real clusters generally have higher median energy and J_r, providing a way to help identify real versus fake clusters. We also recommend incorporating chemical tagging as a way to improve clustering results.

Unified Astronomy Thesaurus concepts: Dwarf galaxies (416); Stellar kinematics (1608); Stellar dynamics (1596); Galaxy accretion (575); Clustering (1908)

1. Introduction

Throughout its formation history over billions of years, the Milky Way grew through mergers with many dwarf galaxies. The smallest and oldest of these accreted systems are the ultra-faint dwarf galaxies (UFDs), which were among the first galaxies in the universe (Frebel 2010; Simon 2019). These systems provide insight into the earliest stages of galaxy formation and are important components of the assembly history of the Milky Way.

Due to low star formation efficiency and quenching from reionization, UFDs preserve information about early chemical enrichment and can display clean signatures of important nucleosynthetic processes such as the rapid neutron-capture process (the *r*-process, which produces around half of the isotopes of the heaviest chemical elements; see Burbidge et al. 1957; Cameron 1957; Frebel 2018; Cowan et al. 2021). For example, the surviving UFD Reticulum II contains highly *r*-process enhanced stars, implying it was enriched by a prolific early *r*-process event such as a neutron star merger (Ji et al. 2016a, 2016b; Roederer et al. 2016). Tucana III and Grus II also exhibit *r*-process enhancement (Hansen et al. 2017, 2020). Satellite galaxies like these are located over 25 kpc away from the Sun (Drlica-Wagner et al. 2015), however, so studying their stars to learn about early chemical enrichment can be difficult.

Because the Milky Way was assembled hierarchically from many neighboring systems including UFDs, bona fide dwarf galaxy stars can also be found located throughout our Galaxy today, including near the Sun. Chemical tagging, i.e., using stellar chemical abundances to identify stars that formed together, is a promising way to identify dispersed UFD stars. Utilizing the Caterpillar simulation suite (Griffen et al. 2016) and a simple model for star formation and parameterized element enrichment, Brauer et al. (2019) suggested that the population of galactic metal-poor *r*-process enhanced halo stars could have largely originated in UFDs. This idea stems from both observations of surviving UFDs such as Reticulum II, and the kinematic studies of *r*-process stars (Roederer et al. 2018; Gudin et al. 2021) that appear to be chemically and dynamically linked. Further evidence in support of chemically tagging *r*-process enhanced halo stars remains limited due to a small sample size of known stars, but the *R*-Process Alliance (Hansen et al. 2018; Sakari et al. 2018; Ezzeddine et al. 2020;

Holmbeck et al. 2020) is continuing to discover more of these stars, which should soon provide a rich sample for study. Low-mass galaxies, especially UFDs, also host a higher percentage of metal-poor stars compared to higher-mass galaxies (e.g., Kirby et al. 2013). Chemical tagging with *r*-process elements and/or low-metallicity stars may thus help astronomers identify stars from UFDs.

Alongside chemical tagging, stellar dynamics also retain important information about the disrupted galaxies accreted by the Milky Way. In particular, the orbital actions and energy of a star are quasi-conserved quantities that can, in principle, be used to identify stars that were accreted together (see Section 2.3). While these quantities are not truly conserved in the Galaxy on long timescales, clustering in $E - L_z - J_r - J_z$ phase space (or a subset of this space) is a common, useful method to search for an accreted structure. Thanks to the Gaia mission, detailed 6D phase-space information is now available for millions of stars (Gaia Collaboration et al. 2018). This influx of data has already lead to a better understanding of the major mergers that the Milky Way experienced (e.g., the Gaia Sausage, Belokurov et al. 2018; Helmi et al. 2018; Sequoia, Myeong et al. 2019; Kraken, Kruijssen et al. 2019, 2020; Forbes 2020; and more, Naidu et al. 2020; Mardini et al. 2022). However, the low-mass galaxy mergers are far less understood because far fewer stars are contributed to the Galaxy from each accreted UFD, rendering the associated dynamic signatures far less pronounced and more difficult to isolate.

Currently, several groups are using kinematics to identify groups of stars that may have originated in UFDs. Roederer et al. (2018) explored the possibility of identifying groups of stars that possibly originated together in UFDs by clustering stars with *r*-process enhancement ("*r*-process stars") in dynamic phase space. Gudin et al. (2021) expanded on this idea with a much larger data set of 446 stars. Both papers found multiple dynamically linked groups of stars, suggesting that these groups may represent dissolved UFD remnants and that dynamic clustering is indeed a promising method to identify groups of stars from tidally disrupted UFDs. Similarly, Limberg et al. (2021) and Yuan et al. (2020) used clustering algorithms to identify dynamically linked groups among very metal-poor ([Fe/H] < -2) stars, several of which have similar dynamics to *r*-process enhanced stars.

This area of research is continuously expanding as more groups explore clustering with stellar dynamics—both with and without chemical tagging—as a means to identify possible groups of accreted stars from dwarf galaxies. And as astronomers continue to gather kinematics for millions of stars in our Galaxy, the search for these dwarf galaxy remnants is a difficult but worthwhile endeavor. It is unclear, however, to what degree we can trust the clusters identified by different clustering algorithms, and which clusters are most likely to correspond to real UFD remnants.

In this paper, we explore the possibilities and challenges of kinematically identifying stars from tidally disrupted UFDs in the Milky Way by analyzing a set of 32 cosmological zoom simulations of Milky Way-mass galaxies. Using the Caterpillar simulation suite (Griffen et al. 2016), we trace the tagged particles from accreted UFDs to $z = 0$ and test different clustering algorithms in dynamic phase space. Specifically, we explore what fraction of remnant UFDs can be expected to be recovered using basic clustering algorithms, which clustering algorithms work best and most reliably, and which identified dynamically linked groups are most likely to correspond to real UFD remnants. In this work, we focus exclusively on UFDs because prior work has investigated more massive accretion events (e.g., Wu et al. 2022), but UFDs remain poorly understood. While most cosmological simulations do not properly resolve UFDs, the Caterpillar simulation suite provides us with the unique ability to investigate many different Milky Way-mass galaxies forming in a cosmological context while resolving UFDs.

Section 2 describes how we created simulated stellar halos from dark matter (DM) cosmological simulations, focusing on the methodology of tagging DM particles as tracers of stellar material and measuring the corresponding dynamics at $z = 0$. Section 3 describes seven different clustering algorithms and how we test them on different data sets. Section 4 discusses our clustering results and their implications for kinematically identifying UFD remnants in real data sets. Section 5 discusses the properties of real clusters and how to identify which clusters are most likely to correspond to real accreted UFD remnants. Section 6 summarizes the takeaways for clustering observational data sets to best identify stars from accreted UFDs.

2. Simulated Stellar Halos

2.1. Cosmological Simulations

We simulate stellar halos using 32 DM-only cosmological simulations from the *Caterpillar Project* (Griffen et al. 2016). Each zoom-in simulation models the formation of a Milky Way-mass DM halo down to $z = 0$. The effective resolution is $16,\ 384^3$ particles of mass $3 \times 10^4\ M_\odot$ in and around the galaxies of interest, resolving subhalos down to total mass $\sim 10^6\ M_\odot$. We limit our analysis to simulated Milky Way-mass halos that experienced no recent major merger; all other aspects of the accretion history are unbiased.

The simulations are fully described in Griffen et al. (2016). The halos in the zoom-in simulations were selected from a larger, lower-resolution parent simulation with cosmological parameters from Planck 2013 ΛCDM cosmology: $\Omega_m = 0.32$, $\Omega_\Lambda = 0.68$, $\Omega_b = 0.05$, $\sigma_8 = 0.83$, $n_s = 0.96$, and $\mathrm{H} = 100\ h\ \mathrm{km\ s^{-1}\ Mpc^{-1}} = 67.11\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$ (Planck Collaboration et al. 2014). Initial conditions were constructed using `MUSIC` (Hahn & Abel 2011). DM subhalos were identified using a modified version of `ROCKSTAR` (Behroozi et al. 2013a; Griffen et al. 2016), and merger trees were constructed by `CONSISTENT-TREES` (Behroozi et al. 2013b). The halos were assigned a virial mass $M_{\rm vir}$ and radius $R_{\rm vir}$ using the evolution of the virial relation from Bryan & Norman (1998). For our cosmology, this corresponds to an overdensity of $\Delta_{\rm crit} = 104$ at $z = 0$.

2.2. Dark Matter Particles as Tracers of Stellar Material

Since the Caterpillar simulations do not directly simulate stars, we tag DM particles as tracers of the stellar material of each accreted galaxy. Stars form tightly bound to their halos and move within the same potential as the DM, so a fraction of the most bound DM particles is expected to trace the phase-space distribution of the stars (e.g., Bullock & Johnston 2005; Cooper et al. 2010). We refer to the tagged particles as *star particles* and trace their phase-space distribution down to $z = 0$.

There is debate over what fraction of DM particles should be tagged as tracers. The fraction generally ranges from the most bound 1%–3% (Cooper et al. 2010; Rashkov et al. 2012;

Bailin et al. 2014), to 5% (Le Bret et al. 2017; Cooper et al. 2017; Dooley et al. 2016), to 10% (De Lucia & Helmi 2008; Morinaga et al. 2019; Tumlinson 2010; Gómez et al. 2012). Cooper et al. (2017) finds that a fractions of 1%–10% all provide a good approximation to accreted halos of Milky Way analogs, implying that the results for accreted galaxies are holistically insensitive to the exact fraction. Our analysis in this paper focuses on UFDs ($M_{\rm halo} \leqslant 10^9 M_\odot$), so to ensure a sufficient number of particles to assess clustering, we tag the 5% most bound particles. At this resolution, each tagged particle in an accreted UFD corresponds to $\sim 10 M_\odot$ of stellar material. We note that having a single, fixed fraction is a simplifying assumption that breaks down in regions dominated by the baryonic potential and having significant angular momentum, such as the Milky Way disk (Cooper et al. 2017). However, given that we focus on dwarf galaxies in our analysis that are DM dominated and elliptical, assuming a fixed fraction is not a principal concern.

We tag 5% of the most bound DM particles at the snapshot where the accreted halo reaches its peak mass. Alternative methods include tagging the particles at the snapshot before the halo is accreted or *live* tagging where stellar mass is added at each snapshot while the galaxy is star forming. Our analysis focuses on small galaxies that are generally no longer forming stars at the time of their accretion, so we choose the peak mass as the snapshot at which to tag DM particles. We use a $M_{\rm star} \sim M_{\rm peak}$ relation to estimate the amount of stellar material represented by each tagged particle (Garrison-Kimmel et al. 2017). We note for completeness that live tagging would likely produce a more accurate phase-space distribution, but the significantly increased computational expense is beyond the scope of this work.

While particle tagging is an imperfect method, it has repeatedly been shown to qualitatively capture trends and produce accreted stellar populations with properties (e.g., metallicities, spatial distribution, velocity dispersions) in agreement with observations around the Milky Way (e.g., Cooper et al. 2017; Rashkov et al. 2012). Given that this study is concerned with the qualitative situations in which the kinematic clustering of accreted stars does or does not excel, particle tagging of DM cosmological simulations is an ideal technique as a means to explore such clustering effects in our set of many different Milky Way-mass simulations. Moreover, a simulation with a disk would result in enhanced tidal disruption and phase-space diffusion (Errani et al. 2017; Maffione et al. 2018), but because our results highlight the difficulty of identifying UFD remnants via clustering, our point is merely strengthened by our use of *N*-body simulations without an added disk potential.

2.3. Stellar Dynamics

We determine the dynamics of each accreted star particle (tagged DM particle) at $z = 0$. In axisymmetric galactic potentials, stellar orbits are described by three integrals of motion called the orbital actions: J_r, J_z, and J_ϕ (see Binney & Tremaine 2008, Section 3.5). Energy is another constant of motion for time-invariant potentials, which, while not independent of the orbital actions, are useful during clustering searches. These four quantities are not conserved in realistic, time-varying galactic potentials, and the galactic potentials in the Caterpillar simulations, for example, are approximately constant for only the last 5 Gyr or so ($z \lesssim 0.5$) (Griffen et al. 2016). Despite this, these quantities provide a useful phase space, in which to search for dynamically similar stars, which is currently being used by several groups in the search for stars from UFDs. We thus explore the possibilities of using these dynamics. These integrals of motion are defined as follows (Binney 2012):

1. E is the specific orbital energy, the total orbital energy of the star divided by its mass.
2. J_r is the orbital action that quantifies the oscillations of an orbit along the radial direction. J_r is nonnegative and increases for more eccentric orbits.
3. J_z is the orbital action that quantifies the oscillations about the equatorial plane. J_z is nonnegative and increases for orbits that rise more out of the equatorial plane.
4. J_ϕ is the azimuthal orbital action, equal to the angular momentum out of the equatorial plane ($J_\phi = L_z$).

To estimate orbital actions, one first needs an initial estimate of the gravitational potential. For each of our 32 simulations, we use the AGAMA software library (Vasiliev 2019) to construct an estimated axisymmetric gravitational potential. The potential is built via multipole expansion in spherical harmonics with $l_{\rm max} = 8$, using the locations and masses of all *N*-body particles at $z = 0$. We validate the estimated potential by comparing it to the value of the potential stored for each particle from the original Caterpillar simulation, confirming the same relative potential energy between particles. After constructing the axisymmetric potential, we use the galactocentric positions and velocities of each accreted star particle to compute the associated actions within AGAMA.

As an illustrative example, the $z = 0$ phase-space distribution for the accreted star particles in one of our simulations can be seen in Figure 1. The particles in these plots are colored based on the peak mass of the galaxy in which each of them formed: UFD ($M_* \leqslant 10^5 M_\odot$), Ursa Minor-mass ($M_* = 10^5$ to $10^6 M_\odot$), Sculptor-mass ($M_* = 10^6$ to $10^7 M_\odot$), and Fornax-mass ($M_* = 10^7$ to $10^8 M_\odot$). Note that this example galaxy did not accrete more massive dwarfs such as those with masses similar to that of the Large Magellanic Cloud.

In Figure 1, the particles from UFDs are only 9% of all the accreted particles within this radial cut, but they are still identifiable in the outskirts of the phase-space diagram because virtually all of the particles from more massive dwarfs are overlapped significantly in phase space. This implies we may be able to more easily identify some UFD remnants at, for example, high energy.

Specifically considering the particles from UFDs, in Figure 2, we show that any identifiable remnants are from relatively recent accretion events, while the most phase-mixed particles are from accretion events that occurred over 8 Gyr ago. This is to be expected, since more recent accretion events will have maintained a stronger dynamic signature at $z = 0$ compared to stars that have been relaxing in the stellar halo for many gigayears (e.g., Gómez et al. 2010).

2.4. The Different Data Sets We Consider

We consider how well clustering works for data sets with three different radial cuts at varying distances from the Sun:

1. *All accreted star particles, no radial cut*. This is a complete data set, which cannot be produced with real observations.

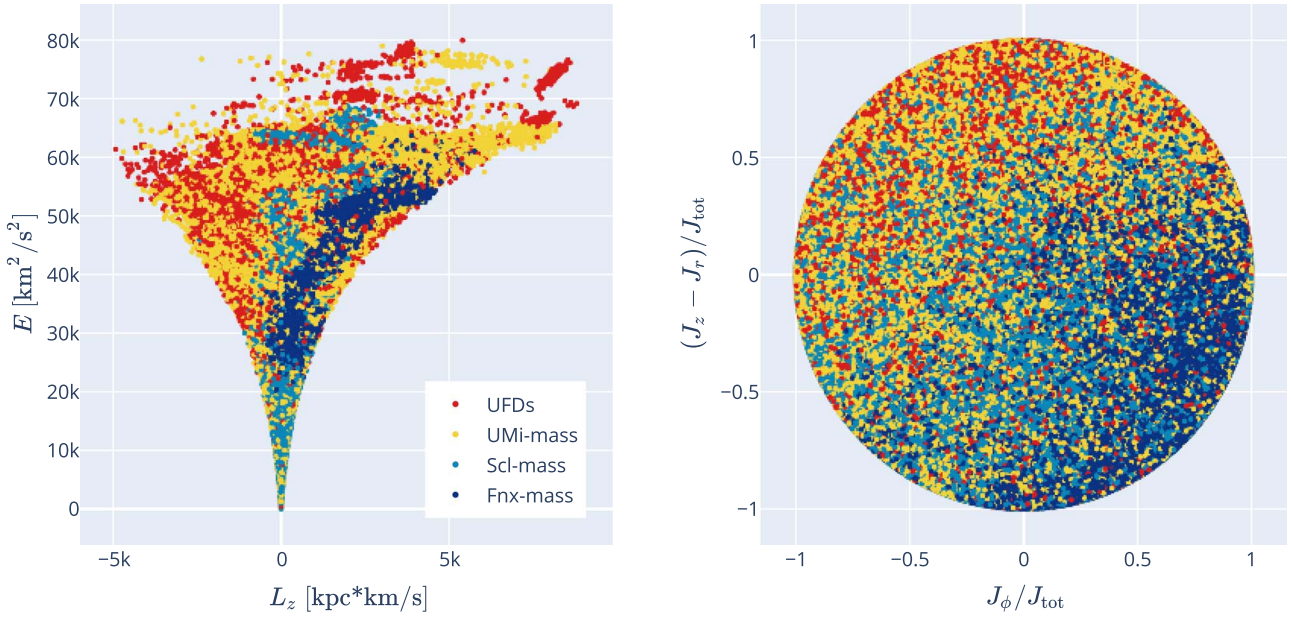


Figure 1. Left: $z = 0$ dynamics (energy and z-angular momentum) of all accreted star particles within 50 kpc of the Sun in simulation Cat-14, one of the 32 simulated Milky Way-mass galaxies. The color of each particle corresponds to the mass of the galaxy in which it formed. Star particles from the smallest galaxies, the UFDs, are seen in red. Right: J_r and J_z orbital actions for the same star particles.

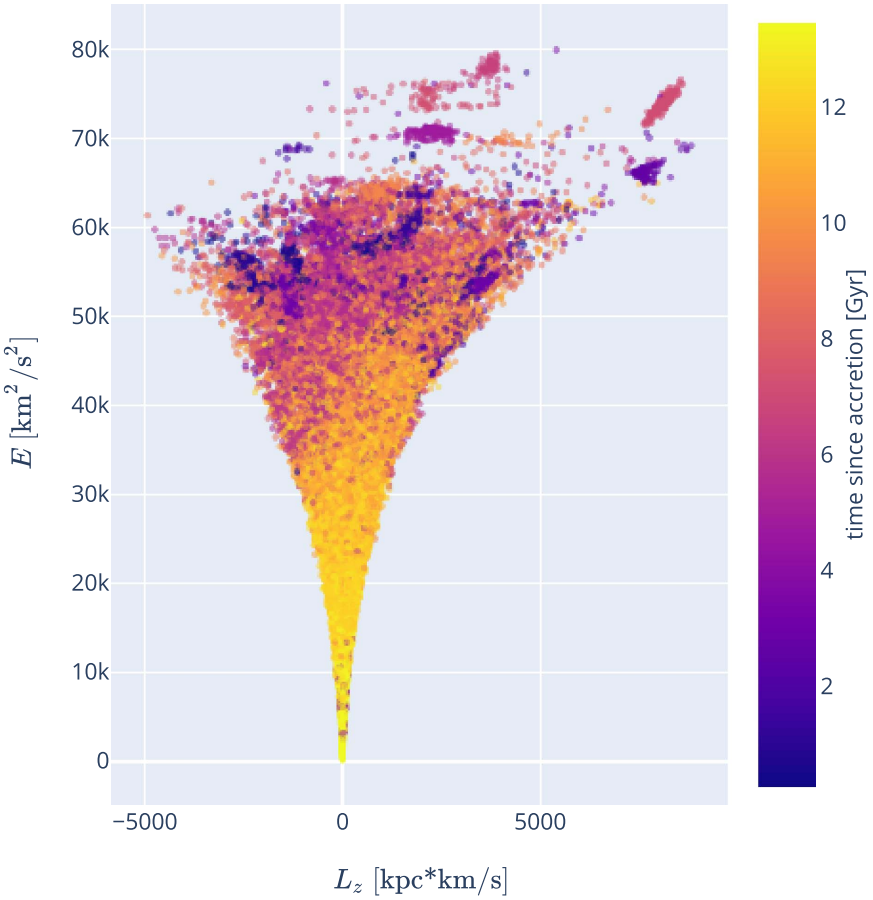


Figure 2. $z = 0$ dynamics of star particles that originated from UFDs in one of our simulations (with a 50 kpc radial cut; see Section 2.4). The color of each particle corresponds to how long ago it was accreted by the Milky Way-mass host galaxy. Stars that were accreted more recently are, generally, of higher energy and less phase-mixed. Over time, the stars mix more in the phase space and are less identifiable by clustering algorithms.

2. *All accreted star particles within 50 kpc of the Sun*. This is an idealistic data set that extends to roughly where the stellar halo drops off.
3. *All accreted star particles within 5 kpc of the Sun*. This is a more realistic data set that includes stars for which we can obtain decent parallax measurements from Gaia.

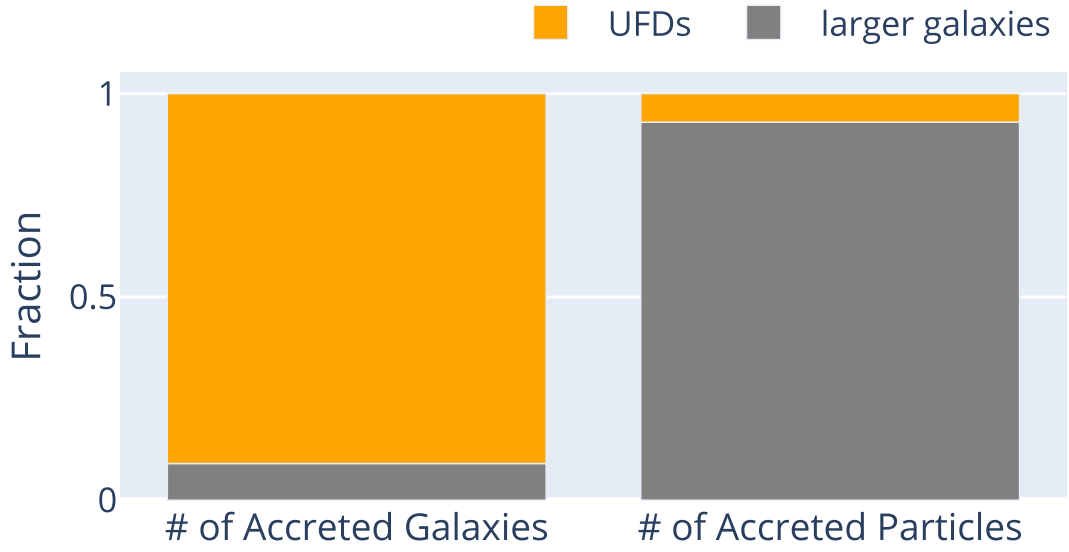


Figure 3. Across all 32 simulations, ∼91% of the galaxies accreted by Milky Way-mass galaxies are ultra-faint dwarf galaxies. These small galaxies only contribute ∼7% of the accreted star particles, however. These fractions are roughly constant with radial cut.

When applying these radial cuts, the location of the *Sun* in each simulation is a consistent, randomly chosen location in the equatorial plane 8 kpc from the galactic center.

We also consider data sets with the following:

1. *Only accreted star particles from UFDs*. This data set is idealistic. To pursue it observationally, one could focus on limiting to only stars with certain chemical signatures (e.g., low metallicity, *r*-process enhancement, deficiency in neutron-capture element abundances) and/or removing stars that are known to be associated with larger mergers.
2. *All accreted star particles*.

After matching each radial cut with both UFD-only and all-stars data sets, we have a total of six different data sets. Each data set includes the stellar dynamics from 32 simulations (although not all simulations are used when performing clustering analysis of the larger radial cuts due to computational limitations). We then quantify how well each clustering algorithm performs in these six situations.

For the data set without a radial cut, Milky Way-mass galaxies accrete on average 187^{+69}_{-65} UFDs. This is 91^{+1}_{-1}% of the total number of accreted systems that Milky Way-mass galaxies will ever accrete. Despite UFDs being the vast majority of accreted galaxies, though, they only contribute ∼7% of the accreted star particles. These fractions are shown in Figure 3. These results align with Monachesi et al. (2019), which estimated that the accreted stellar halo had only a handful of significant progenitors. For the data set with a 5 kpc radial cut, the total average number of accreted UFDs seen in the data set drops to 99^{+45}_{-30}, but the percentage representation remains the same. We note here that all uncertainty values provided represent 16th–84th percentile scatter across all the simulations.

3. Clustering Methodology

3.1. Clustering Algorithms

We apply seven different clustering algorithms on the 4D energy-action space of each simulated Milky Way–like halo. The algorithms studied in this work are HDBSCAN (Campello et al. 2015; McInnes et al. 2017), Gaussian mixture models (GMM; Dempster et al. 1977), agglomerative clustering (Ward 1963), K-means (Lloyd 1982; Arthur & Vassilvitskii 2006), affinity propagation (Frey & Dueck 2007), mean-shift (Comaniciu & Meer 2002; Derpanis 2005), and friends-of-friends (FoF; Huchra & Geller 1982; Press & Davis 1982; Davis et al. 1985; Gibbons 2020). Before running any clustering algorithms on our simulations, we normalize each of the 4D energy-action variables into the range [0, 1]. Here, we briefly comment on each of these algorithms.

HDBSCAN (Hierarchical DBSCAN) is a hierarchical extension of the density-based approach of DBSCAN. It measures the density around each point, constructs a hierarchical cluster tree based on this density information, and returns the clusters that are persistent across different density thresholds. As a result, it is sensitive to data sets having true groups at varying densities. It also scales well for massive data sets. Hunt & Reffert (2021) found that, compared to DBSCAN and GMM, it performs best at recovering open clusters in a massive sample of Gaia data. This was also the preferred clustering algorithm of Gudin et al. (2021) and Limberg et al. (2021), two papers that identified dynamically linked groups that may correspond to UFDs.

Agglomerative clustering forms clusters from the bottom up. It starts with each particle as its own cluster. The clusters that are separated by the least linkage distance (in our case, Euclidean distance) are then hierarchically merged until the preset number of clusters is reached. Because it has a time complexity of $O(n^3)$ and requires $\Omega(n^2)$ of memory, it is too slow and memory-intensive for large data sets.

K-means is a distance-based algorithm that returns a preset number of k clusters, each of equal variance. Starting with k randomly generated initial means, it first assigns each particle to the mean with the least sum-of-squares distance. Particles associated with the same mean form a cluster. The mean (or centroid) of each cluster and, consequently, cluster membership is then continually updated until convergence.

A GMM can be thought of as a generalization of K-means in that it returns the distance-based clusters that may be at different variances. It decomposes the sample into a mixture of a preset number of n Gaussian distributions and, upon convergence, returns the Gaussian components as separate clusters.

Unlike K-means, agglomerative clustering, and GMM, affinity propagation does not require a preset number of clusters before running. Its goal is to find *exemplars* or prototype particles that are representative of a cluster. First, each particle begins as a potential exemplar. Pairs of particles then pass *messages* to each other about the suitability of one particle to be the exemplar of the other. These messages are passed until a stable set of exemplars and, thus, clusters emerge.

Mean-shift is a centroid-based algorithm that treats each particle as a kernel with a preset bandwidth. It then performs a gradient ascent on the kernel peaks until convergence. Gómez et al. (2010) used mean-shift on the $E-L-L_z$ space of a mock Gaia catalog of the solar neighborhood and recovered roughly 50% of all satellite galaxies. We note that this differs from our results because this work focuses on a smaller quantity of larger-mass satellites as compared to our UFD-focused analysis.

FoF is commonly used to identify the gravitationally bound halos in cosmological simulations. Particles that are separated by a distance less than a preset linking length are linked as friends, forming a networked cluster of particles. Networks that have no mutual friends are designated as separate clusters. Helmi & Tim de Zeeuw (2000) applied this algorithm on the $E-L-L_z$ space of a mock Gaia catalog to identify simulated Milky Way accretion events.

Other groups have used custom clustering algorithms, e.g., StarGo (Yuan et al. 2018, 2020), Enlink (Sharma & Johnston 2009; Wu et al. 2022), and other hierarchical clustering techniques (Ruiz-Lara et al. 2022; Sofie Lövdal et al. 2022). We do not test all of these algorithms, but expect our UFD-focused results to holistically hold for them as well (see Section 4.5).

3.2. Hyperparameter Choices

All the algorithms included in this paper except affinity propagation require a preselected hyperparameter in order to begin clustering. To explore different hyperparameter choices, for each algorithm, we do the following:

1. We create a hyperparameter search space consisting of about 20 trial values. For instance, to select the `min_cluster_size` hyperparameter for HDBSCAN, we create a search space composed of integers from 3 to 20 inclusive, and for FoF we explore from 0.001 to 0.2.
2. We run the clustering algorithm with each trial hyperparameter on each simulation in each data set.
3. For every clustering run, we count the number of pure and complete clusters. A cluster is *pure* if $\geqslant\frac{2}{3}$ of the stars in that cluster accreted together from a UFD. A cluster is also *complete* if $\geqslant\frac{1}{2}$ of the stars from that accreted UFD are found together in that cluster.
4. For every simulation on which a particular hyperparameter is tested, we calculate a recovery rate and a realness rate. The recovery rate is defined as follows:

$$\frac{\text{number of pure and complete clusters}}{\text{number of accreted UFDs in the data set}}\times 100\%.$$

Meanwhile, the realness rate is defined as follows:

$$\frac{\text{number of pure clusters}}{\text{number of clusters found by the algorithm}}\times 100\%.$$

When calculating these rates, we only consider the clusters and remnants with at least 5 particles.

5. For each data set, we determine the optimal hyperparameter by assigning a score to each hyperparameter choice. To assign the score, normalize all of the recovery rates and realness rates using a min max scaler, and then add the normalized median recovery and realness rates together. The optimal hyperparameter thus balances the highest UFD recovery rate and the highest realness of its clusters.

We choose an optimal hyperparameter value for each algorithm on each data set. Since we are testing six algorithms that each require hyperparameters on six different data sets, we make a total of 36 optimized hyperparameter selections. A full list of the trial hyperparameter values are in Table 3 and the optimal choices are in Table 4.

3.3. Association of Different Observables with the True Cluster Labels

To help identify which observable variables are most likely to be important during clustering, we perform one-way analysis of variance (ANOVA) tests on the stellar kinematics of each simulation. The ANOVA test assesses the association between a categorical (e.g., the label of each true cluster) and a continuous variable (e.g., each of the kinematic variables; e.g., McDonald 2014; Gómez et al. 2014). If a given kinematic variable is strongly associated with the true cluster labels, it is likely to be important during clustering in situations where we do not know the true labels.

We use the `stats.f_oneway` ANOVA test from the `scipy` python package (Virtanen et al. 2020). This F-test analyzes whether the means of the continuous variable differs between groups. $F=$(variation between cluster means)/(variation within the clusters), so high F values for our data signify that a given observable varies more between clusters than within. For these tests, the clusters we are using are the true UFD remnant groups because we take the labels directly from the simulations. To quantify the level of the effect, we also calculate the ω^2 value of each test (e.g., Olejnik & Algina 2003). This metric is similar to R^2 in the context of regression analysis while also accounting for the degrees of freedom in the model. ω^2 can vary from -1 to $+1$; values far from zero imply a stronger effect.

The ANOVA test results are shown visually in Figures 4 and 5. Figure 4 shows the four axisymmetric actions we use in clustering. All four actions show correlation with the true cluster labels, with energy consistently being the most important observable. Figure 5 shows the correlations of other potentially useful observables, demonstrating the high correlation of total angular momentum, $L_{\rm tot}$. These results support our choice to cluster in E–J_r–J_z–J_ϕ phase space. They also imply that E–$L_{\rm tot}$ phase space can be useful to find UFD remnants in the cases where the full axisymmetric actions are unknown. This has been known previously (e.g., Helmi & Tim de Zeeuw 2000; Gómez et al. 2010).

Figure 5 shows that the total velocity is likely important at parallax-level cuts (e.g., 5 kpc), and the total distance from the galaxy's center is important for data sets with no radial cut. This is simply due to the relationship between velocity, radius, and total energy. All of the test results are summarized in the Notes for Table 2. As an additional check, we also include

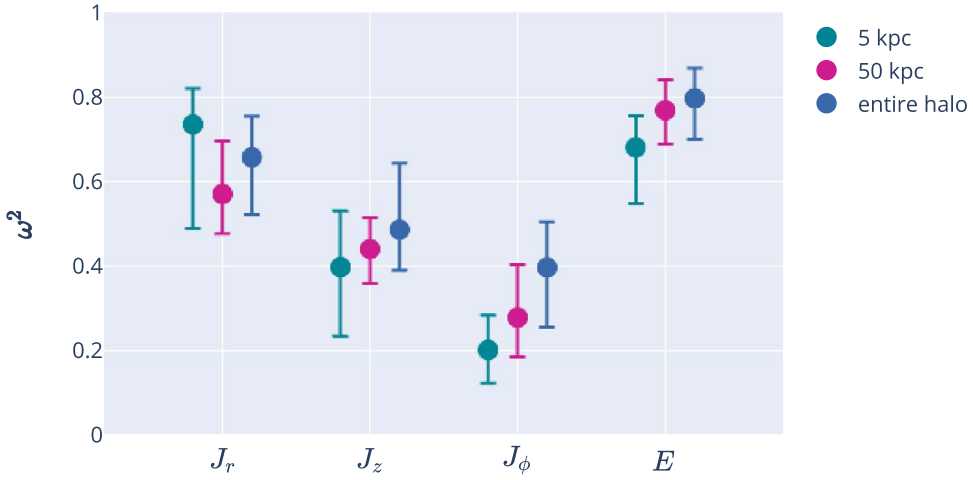


Figure 4. Strength of association between actions (J_r, J_z, J_ϕ, E) and the true cluster labels. Higher ω^2 values indicate a stronger association. E has the highest ω^2 values, implying it is the most important variable when seeking to find clustered stars that accreted together. Note that J_ϕ is defined in axisymmetric potentials to be equivalent to L_z.

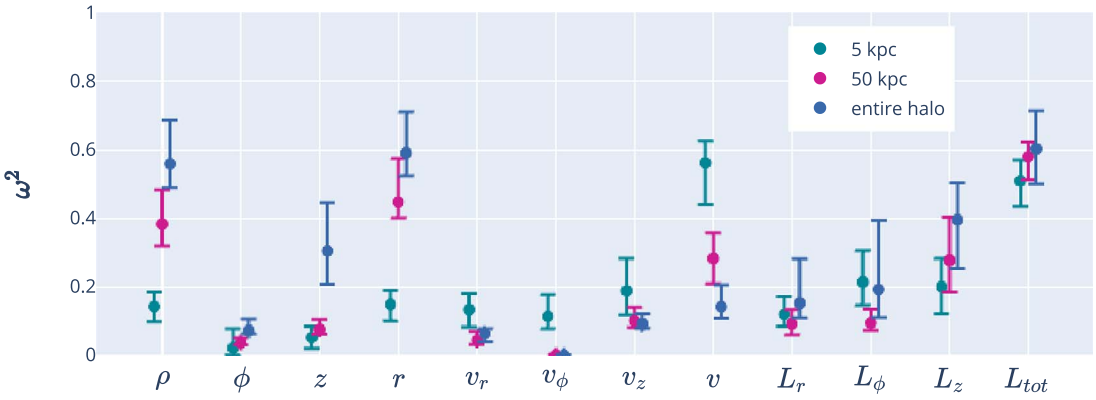


Figure 5. Strength of association between different kinematic observables and the true cluster labels. Higher ω^2 values indicate a stronger association. $L_{\rm tot}$ has consistently high ω^2 values, implying it is an important variable when seeking to find clustered stars that accreted together. ρ, ϕ, z and v_r, v_ϕ, v_z are the radius and velocity in cylindrical coordinates, respectively. The importance of r and v is due to their correlation with total energy.

ANOVA tests for $z_{\rm infall}$, the redshift at which the particles were accreted by the Milky Way. This variable perfectly aligns with the true cluster labels and thus should have $\omega^2 = 1$, which we find.

4. Quantifying the Abilities and Limitations of Clustering Algorithms

We run each of the clustering algorithms (HDBSCAN, GMM, agglomerative clustering, mean-shift clustering, K-means, FoF, and affinity propagation; see Section 3.1) on each simulation in each of the six data sets (see Section 2.4 for a definition). The hyperparameters of each algorithm are chosen as described in Section 3.2. All clustering is done in 4D energy-action space using E, L_z, J_r, and J_z as supported by the association results presented in Section 3.3. Given the seven algorithms, six data sets, up to 32 simulations per data set, and roughly 20 hyperparameter choices per algorithm, we ran over 10,000 clustering tests.

The results from these tests are largely a cautionary tale. All of these algorithms have significant limitations when it comes to identifying UFD remnant groups. Hence, in this section, we analyze the possibilities and limitations of the algorithms with a focus on how the results can inform the search for UFD remnants in real data sets since there currently exist no better methods to identify tidally disrupted UFDs from survey data. In future work, fully modeling the phase-space distribution of all accreted systems simultaneously could offer an alternative method to learn about accreted UFDs as compared to the current method of individually picking out a handful of dynamic clusters that may or may not correspond to UFDs. For now, though, kinematic clustering remains one of the few available methods.

The basic problem is that, due to phase-mixing and background, most star particles that accreted into the Milky Way-mass galaxies from the small UFD remnants overlap too much with other particles in phase space at $z = 0$ to be reliably identified as coherent remnant groups. This is true for all algorithms across all data sets. The clustering algorithms also frequently return clusters that do not correspond to any true UFD remnant (*false positives*). However, some algorithms work better than others, and some identified clusters are more likely to be real than others. We now give more details on algorithm usability.

4.1. Example Clustering Results

Figure 6 shows example clustering results from each of the seven algorithms. These results use a single Milky Way-mass simulation (simulation Cat-14) from one data set (accreted star

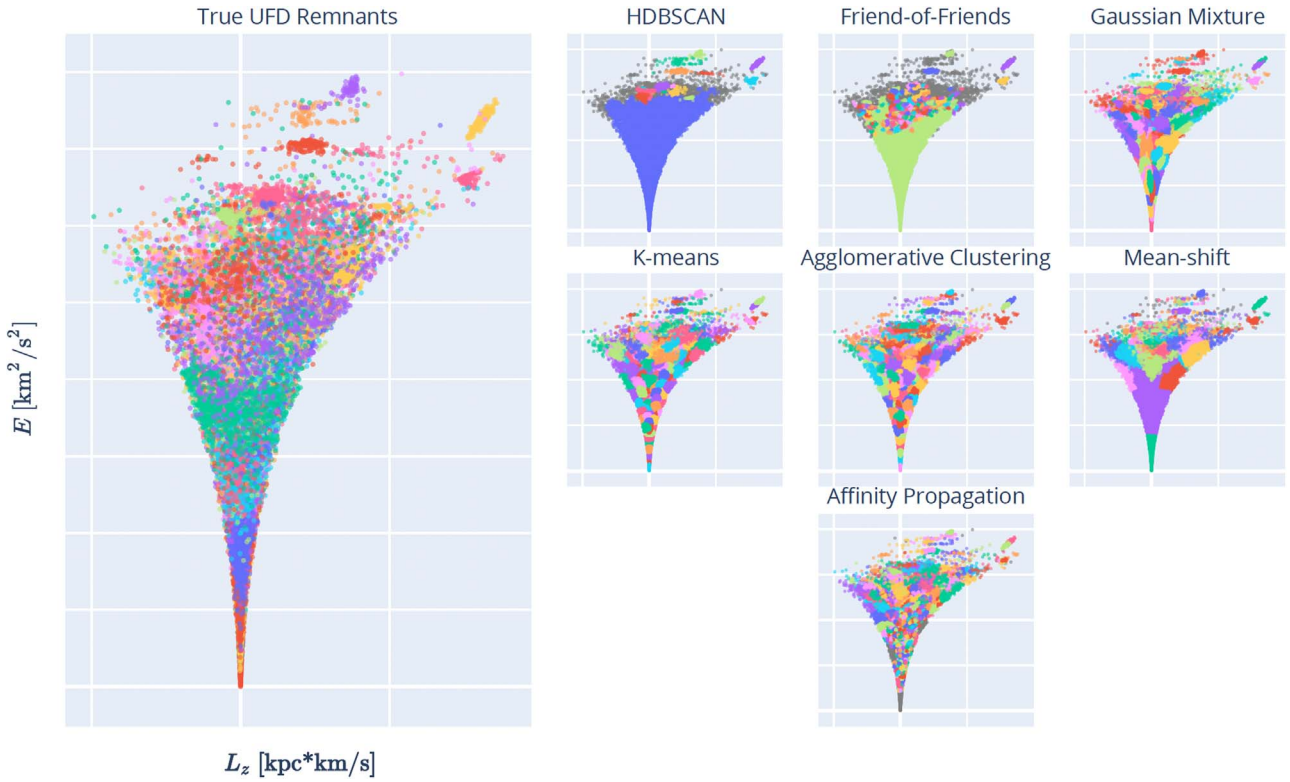


Figure 6. Example of clustering results for one simulation (Cat-14) from one data set (accreted star particles from UFDs within 50 kpc of the Sun). Far left: star particles from true UFD remnants in dynamic phase space. Many of the particles are phase-mixed. Right: results from each of the seven clustering algorithms tested in this paper. Most of the clusters found by these algorithms, especially those at lower energy, do not correspond to true UFD remnants.

Table 1
For the Example Simulation Shown in Figure 6, the Realness and Recovery Rates of Different Clustering Algorithms

Algorithm	Realness Rate	Recovery Rate
HDBSCAN	67% (12 pure clusters / 18 total clusters)	4% (5 pure & complete clusters)
Friends-of-Friends	34% (61 pure clusters / 176 total clusters)	5% (6 pure & complete clusters)
Gaussian Mixture Models	18% (29 pure clusters / 160 total clusters)	5% (6 pure & complete clusters)
K-Means	12% (27 pure clusters / 230 real clusters)	5% (6 pure & complete clusters)
Agglomerative Clustering	13% (32 pure clusters / 248 total clusters)	6% (8 pure & complete clusters)
Mean-Shift	22% (24 pure clusters / 100 total clusters)	3% (4 pure & complete clusters)
Affinity Propagation	5% (52 pure clusters / 989 total clusters)	2% (3 pure & complete clusters)

Note. The recovery rate is determined by comparing the number of pure & complete clusters to the total number of accreted UFDs in this simulation, 124 UFDs.

particles from UFDs within 50 kpc of the Sun). The left shows the true UFD remnants in phase space; each star particle is colored according to the UFD it was born in (note that each color repeats several times). The star particles in this example originated in 124 different UFDs. The panels on the right show how well each clustering algorithm performs. All clustering algorithms perform poorly in the high-density region of phase space and only consistently identify several isolated, high-energy clusters. These high-energy clusters do, in fact, correspond to real UFD remnants. The majority of the rest of the clusters found by these algorithms do not actually correspond to real UFD remnants. This is unsurprising given the high density of the overlapping structure in the high-density region.

For all of our clustering results, we use the metrics of *realness rate* and *recovery rate* to evaluate the findings. Realness rate is defined as the fraction of clusters that are *pure*, defined as clusters for which at least $2/3$ of the stars accreted together. Recovery rate is defined as the fraction of UFD remnants that are recovered. A remnant is recovered if (1) its stars are clustered into a pure cluster and (2) that cluster is *complete*, defined as clusters for which at least $1/2$ of the stars from a remnant are identified together in a single cluster. When determining these rates, we only consider the clusters or remnants with at least 5 particles. The purity and completeness thresholds ($2/3$ and $1/2$, respectively) are chosen with a stricter requirement on the *realness* of a cluster as our priority is identifying stars that accreted together. These thresholds can both be varied, though, and are simply chosen as example metrics. The holistic takeaways of this paper remain consistent even if you vary these thresholds.

As an illustrative example, the realness and recovery rates for each algorithm on the Cat-14 simulation are reported in

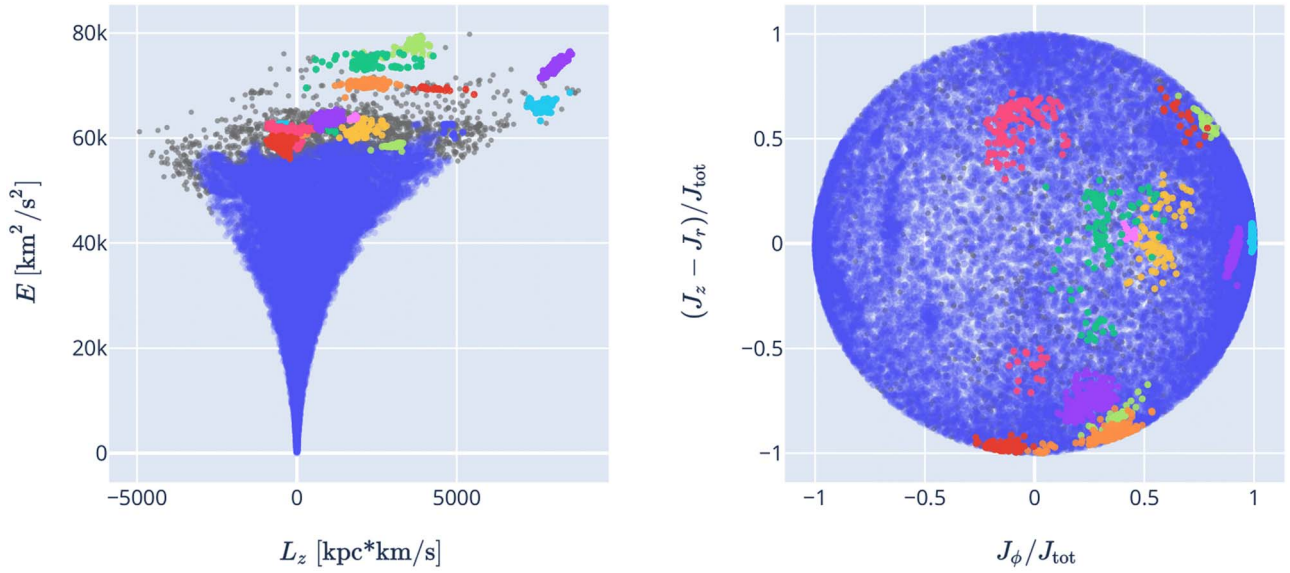
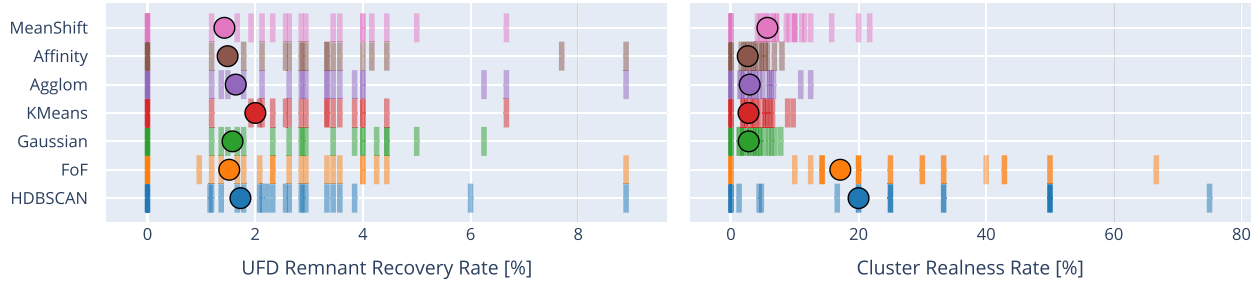
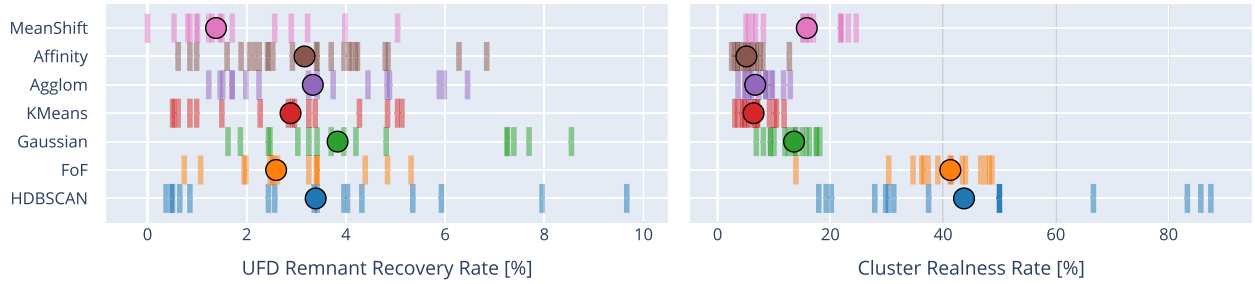


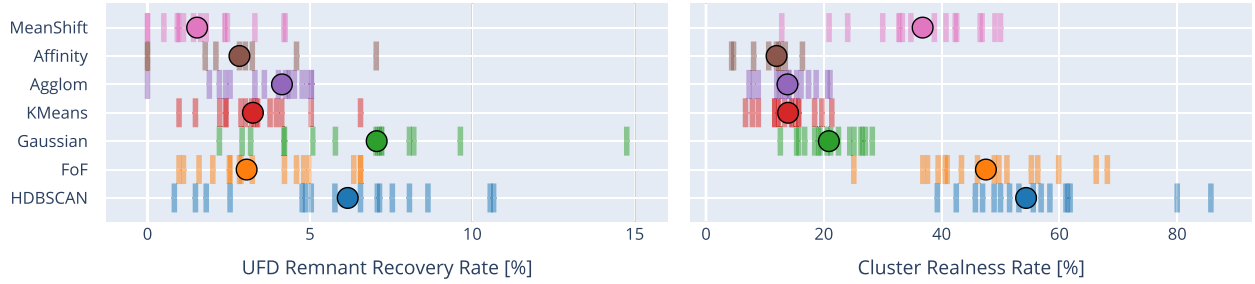
Figure 7. Example of the HDBSCAN clustering results for one simulation (Cat-14) from one data set (accreted star particles from UFDs within 50 kpc of the Sun). The clustering is done in 4D energy-action space (E, L_z, J_r, J_z). Gray points are particles not associated with any cluster. For this simulation, HDBSCAN finds eighteen clusters. Twelve of them are real groups of accreted UFD stars, and five of those twelve are fully *recovered* UFD remnants.



(a) UFD + 5 kpc data sets: All algorithms recover similarly low numbers of UFD remnants. HDBSCAN and FoF have the highest cluster realness rates.

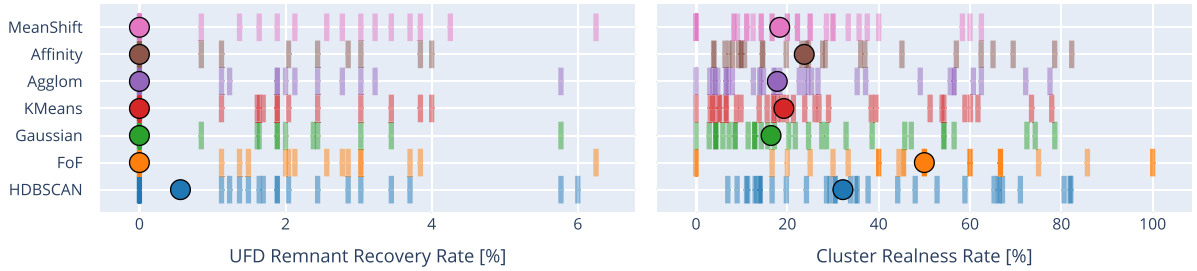


(b) UFD + 50 kpc data sets: Once again, HDBSCAN and FoF have the best balance of UFD remnant recovery and cluster realness rates.

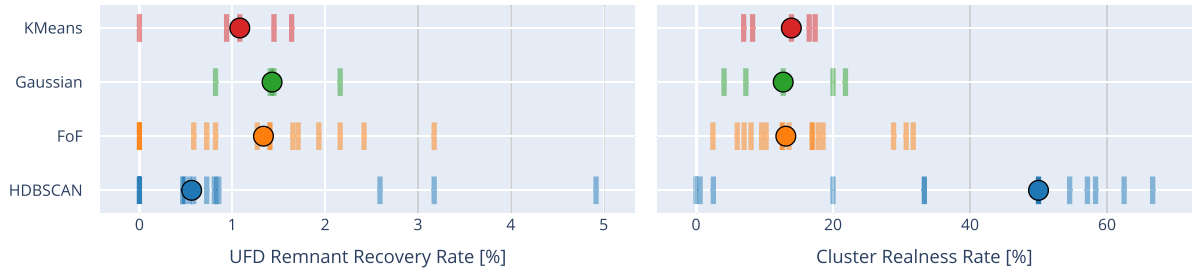


(c) UFD + entire halo data sets: HDBSCAN has the best balance of UFD recovery and cluster realness. FoF has a similar realness rate but recovers far fewer remnants.

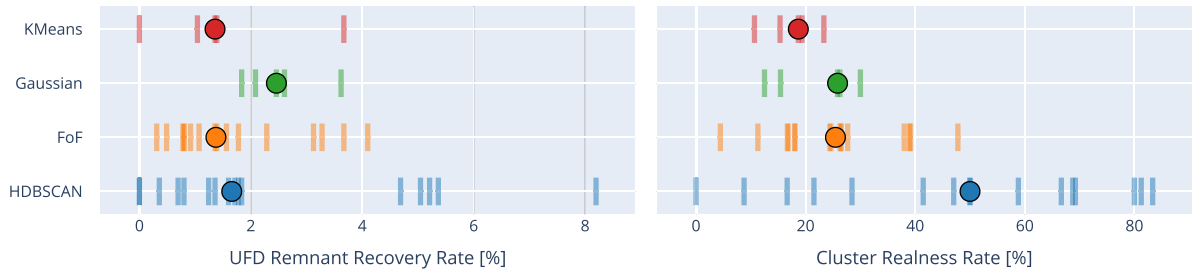
Figure 8. Results for the UFD-only data sets. Each line represents a single Milky Way-mass galaxy simulation, and each circle is the median rate across all simulations. See Section 2.4 for descriptions of the different data sets and Section 3.2 for definitions of recovery and realness rates. Generally, HDBSCAN and FoF perform better than the other algorithms and are also significantly faster. For results with the all-stars data sets, see Figure 9.



(a) All stars + 5 kpc data sets: Algorithms find real clusters accreted from dwarf galaxies, but almost none of them are UFD remnants.



(b) All stars + 50 kpc data sets: Recovery rates are once again low, but realness rates can be high as clusters from larger mass dwarfs are identified. HDBSCAN has highest realness rate, but all recovery rates are low.



(c) All stars + entire halo data sets: HDBSCAN once again has the best balance of recovery and realness.

Figure 9. Results for the all-stars data sets. Each line represents a single Milky Way-mass galaxy simulation, and each circle is the median rate across all simulations. On the large data sets, HDBSCAN and FoF are much faster than K-means and Gaussian mixture models.

Table 1. The example HDBSCAN results are shown in Figure 7.

4.2. Comparing Clustering Algorithms

Throughout this work, we test seven common clustering algorithms (described in Section 3.1). For the UFD-only data sets, we test all seven algorithms on every data set. For the all-stars data sets, the larger radial cuts (50 kpc and entire halo) are extremely large, so we only test the more scalable algorithms: HDBSCAN, FoF, GMM, and K-means.

The results for all UFD-only data sets are shown in Figure 8. Each line represents the results for a single Milky Way-mass galaxy simulation with the given radial cut. The median result for each algorithm is shown as a circle. The scatter in results across different simulations is significant because Milky Way-mass galaxies with a higher number of recent UFD accretions have higher rates. The results for the all-stars data sets are shown in Figure 9.

Even with UFD-only data sets, all algorithms have low UFD remnant recovery rates and cluster realness rates. The local radial cut, 5 kpc, has the worst results; the number of UFD remnants recovered from these simulations is frequently just one. Overall, all algorithms only recover about 2% of UFD remnants within 5 kpc of the Sun. HDBSCAN and FoF have the highest realness rates for their clusters, with around 20% of their clusters corresponding to tagged star particles that accreted together.

This clearly implies, in no uncertain terms, that the vast majority of clusters found by these algorithms do not actually represent any truly accreted groups!

In the larger data sets, the clustering algorithms perform better, recovering ∼3%–6% of UFD remnants and, for HDBSCAN and FoF, having a ∼40%–60% realness rate. Even with these idealized data sets and specially chosen hyperparameters, though, the rates are still low. We thus discuss how to identify real clusters versus false positives in Section 5.

Generally for the UFD-only data sets, HDBSCAN is the most reliable algorithm choice. FoF also often has a relatively high realness rate. These two algorithms are also the fastest choices for large data sets.

For the all-stars data sets (Figure 9), the realness rates are higher than the UFD-only data sets because the clusters of stars from the larger dwarf galaxy remnants are easier to identify than the small clusters of stars from UFD remnants. UFD remnant recovery rates are universally worse in the all-stars data sets, though, because the non-UFD stars act as significant

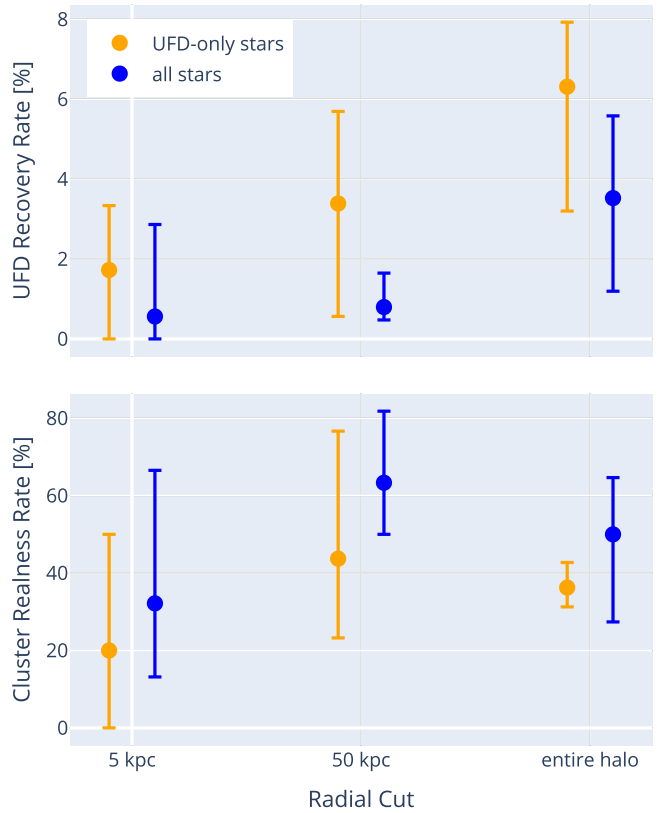


Figure 10. Comparing UFD remnant recovery rates and cluster realness rates with HDBSCAN for the data sets with only UFD stars and the data sets with all the stars. As expected, the UFD-only data sets result in higher UFD recovery rates. Realness rates for the all-stars data sets include real clusters from larger dwarfs, which are principally easier to identify, so overall realness rates are not improved by using a UFD-only data set. Error bars show 16%–84% scatter across all simulations.

noise during the search for UFD clusters. This is discussed in more detail in Section 4.3. Similar to the UFD-only data sets, HDBSCAN is once again generally a reliable choice to balance the recovery rates and realness rates in the all-stars data sets. For the largest data sets, computational constraints also become important, and HDBSCAN and FoF scale well computationally.

Overall, HDBSCAN tends to be the most reliable clustering algorithm across different data sets. Currently, it is also a popular clustering algorithm used in astronomy research (see Section 3.1). We thus focus on HDBSCAN for most of the rest of our main text.

4.3. Comparing UFD-only Data Sets to All-stars Data Sets

As discussed in Section 2.4, we have data sets with (1) only accreted star particles from UFDs and (2) all accreted star particles. The former data set is unrealistic because in real data we cannot know a priori which stars accreted from UFDs. The UFD-only data set can be imperfectly pursued observationally through the use of chemical tagging, however. Stars that formed in UFDs tend to have a lower metallicity distribution function, lower abundances in neutron-capture elements, and may preferentially have strong *r*-process enrichment (e.g., Kirby et al. 2013; Brauer et al. 2019; Gudin et al. 2021; Ji et al. 2016a). Additionally, as we identify the kinematic structures associated with larger-mass accretion events such as Gaia-Enceladus, removing those stars from observational data sets could also help toward creating a UFD-only data set. All these methods are imperfect, but as no more sophisticated and reliable methods exist to date to identify UFD stars, e.g., in observed survey data, we must do the best we can with the methods available to us.

In Figure 10, we demonstrate the need to find ways to exclude the stars from higher-mass accreted dwarfs if we hope to identify UFD remnants. At every radial cut, UFD remnant recovery rates are higher for UFD-only data sets. Realness rates are higher for all-stars data sets, but this is only because the structures from higher-mass dwarfs are principally easier to identify than those from UFDs and because pure clusters are generally more common for higher-mass dwarfs since they contribute more stars. This underscores how difficult it is to identify UFD structures even among UFD-only samples. If we hope to identify UFD remnants, though, pursuing the data sets with stars from UFDs will be, unsurprisingly, very beneficial.

4.4. Comparing Hyperparameter Choices

One downfall of most of these clustering algorithms is their dependence on hyperparameters. Each algorithm other than affinity propagation requires users to preselect a value for a hyperparameter, and it is generally not obvious which values are best. In this work, we already know the true labels, and thus have the unique privilege of selecting our hyperparameters to optimize our clustering results (see Section 3.2). For observational data sets, however, this is not possible.

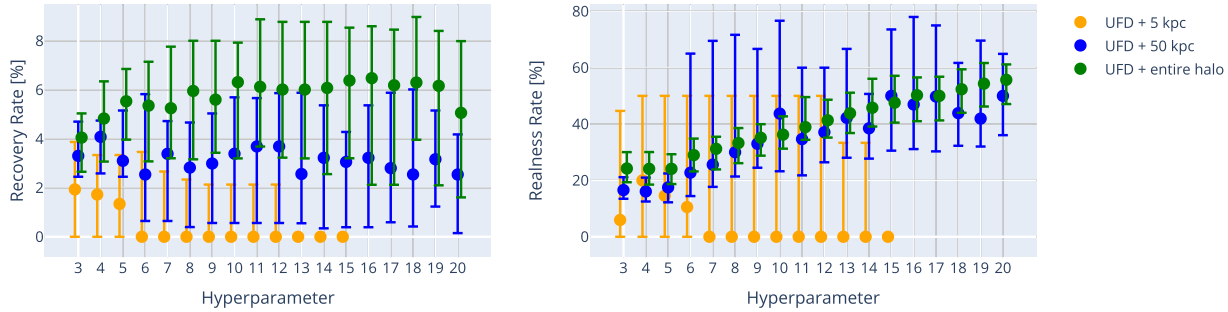
The results in all other subsections use optimal hyperparameter values. In this subsection, we vary the hyperparameter choices to illustrate how the results differ. Figure 11 shows the results for different hyperparameter choices of HDBSCAN and FoF. HDBSCAN requires an integer choice for min_cluster_size and thus has a smaller reasonable range of choices. The results can vary significantly with min_cluster_size choice, but generally the results are roughly stable across several integer choices. As expected, the best choice of min_cluster_size tends to increase for data sets with larger radial cuts. For FoF, we tested many possible choices for linking_length, and the results were more unstable than those for HDBSCAN.

Thus, for these data sets, the results from HDBSCAN are more stable with variations in hyperparameter choice. The hyperparameter choice is important for all algorithms, however. This remains a difficulty of automating the search for UFD remnants with these clustering algorithms. Some groups are developing algorithms without a hyperparameter dependence (e.g., Ruiz-Lara et al. 2022) to alleviate these concerns.

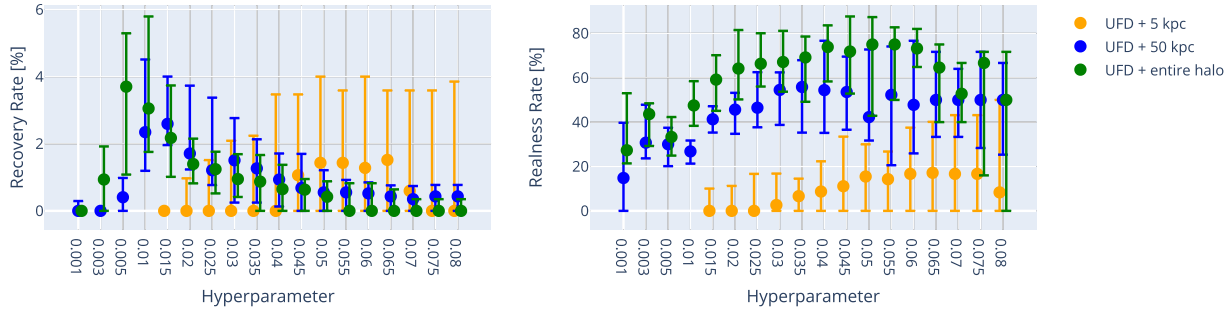
Still, for HDBSCAN, the hyperparameter value greatly affects the number of clusters. For too large of min_cluster_size, the algorithm finds no remnants. For example, for the 5 kpc data sets, min_cluster_size > 5 causes, on average, fewer than five total clusters returned by the algorithm, none of which are real UFD remnants. For the larger radial cuts, too small of min_cluster_size leads to too many clusters. For these data sets, min_cluster_size < 9 causes 200–2000 clusters while the number of recovered remnants remains constant or decreases. When selecting this hyperparameter, a balance must be struck to avoid the identification of an unreasonably small or large number of clusters in a given sample.

4.5. Why Clustering Algorithms Struggle

Due to their small size, the dynamic signatures of tidally disrupted UFDs are, over 90% of the time, weak and significantly



(a) Recovery rates and realness rates for different choices of min_cluster_size for HDBSCAN.



(b) Recovery rates for different choices of linking_length for FoF.

Figure 11. For these data sets, the FoF results differ more with varying hyperparameter choices than the HDBSCAN result. The hyperparameter choice is important for all algorithms, however. This causes additional difficulty when using these algorithms to identify UFD remnants. Error bars show 16%–84% scatter across all simulations.

outnumbered by other overlapping accreted structures. The limitations found in this paper are not unique to these clustering algorithms; we expect any clustering algorithm to struggle.

To illustrate this, we estimate signal-to-total ratios (similar to signal-to-noise ratios) for all the tidally disrupted UFD remnants in our data sets. Normalized histograms of the signal-to-total ratios from different data sets are shown in Figure 12. To determine these ratios, for each remnant, we draw a 4D sphere in phase space that is exactly large enough to enclose 50% of the particles from that remnant. We then compare the number of remnant particles in that volume to the total number of particles in that volume. The maximum value is thus 1 for the case where the tidally disrupted UFD is isolated from other particles. These ratios are similar to our purity metric, so we plot our purity threshold (67%) as a dotted line on Figure 12 for reference. We also note that the remnants are generally not spherical in 4D phase space, so this is merely an estimate.

For the vast majority of UFD remnants, the dynamic signature is completely washed out by the other particles in that volume. For UFD-only data sets, the typical remnant has a ratio of 1 UFD remnant particle to 30 other particles, 1:30. For the all-stars data sets, the typical remnant has a ratio of 1:1000. In the best case scenario, the UFD-only data set with the entire halo, only 8% of remnants have a signal-to-total ratio higher than our purity threshold of 67%.

The remnants with the highest signal-to-total ratios are the remnants that are successfully identified by the clustering algorithms. Most of the other remnants are simply too difficult to find in this dense 4D space, due to a combination of phase-mixing as the stellar dynamics relax over time and/or accreting with dynamics that are already similar to other star particles. We can thus optimize the clustering searches to try to find the greatest number of UFD remnants, but most will never be found by these methods. The ones that are kinematically identifiable are those that (1) accreted with outlying dynamics, e.g., higher energy than usual, and (2) recently accreted so that the star particles have not had time to phase-mix.

We also note that an additional difficulty of analyzing only star particles in the inner volume, e.g., our 5 kpc data sets, is that you cannot sample full satellites within this small volume. This issue is described in more detail in Gómez et al. (2010).

5. Properties of Real Clusters in Simulations

Even in the best cases, the clustering algorithms find many clusters that do not correspond to real accreted remnant groups. Hence, we compare the properties of real clusters versus *false positive* clusters to help inform which clusters are more likely to be real in observational data sets.

Figure 13 shows the E, L_z, J_r, and J_z of real recovered clusters (i.e., pure and complete clusters–clusters that correspond to an accreted UFD remnant) compared to the dynamics of clusters that do not correspond to UFD remnants. These results use HDBSCAN, but the plots are holistically similar for other algorithms. All dynamics are normalized relative to the median of all clusters in the sample. For each cluster, its energy (or L_z, J_r, J_z) is determined from the median of all the star particles in that cluster.

Compared to all clusters, the clusters that correspond to real UFD remnants have higher energy and axisymmetric actions. High energy and J_r are most important for distinguishing between real UFD clusters and all other clusters, especially in local (5 kpc) data sets. Of the action variables, L_z is the least important dynamic when determining which clusters are more likely to be real. This aligns with the results from the ANOVA tests in Figure 4.

Table 2
Results for the One-way ANOVA Tests

Variable	Radial Cut	F	ω^2	$p-$value	Variable	Radial Cut	F	ω^2	$p-$value
	5 kpc	2^{+1}_{-1}	$0.14^{+0.04}_{-0.05}$	$2e-10^{+6e-8}_{-2e-10}$		5 kpc	2^{+1}_{-1}	$0.12^{+0.05}_{-0.04}$	$6e-11^{+3e-4}_{-6e-11}$
ρ	50 kpc	186^{+93}_{-45}	$0.38^{+0.10}_{-0.06}$	$<1e-300$	L_r	50 kpc	28^{+23}_{-7}	$0.09^{+0.04}_{-0.03}$	$<1e-300$
	entire halo	460^{+407}_{-97}	$0.56^{+0.13}_{-0.07}$	$<1e-300$		entire halo	71^{+75}_{-21}	$0.15^{+0.13}_{-0.04}$	$<1e-300$
	5 kpc	1^{+1}_{-1}	$0.02^{+0.06}_{-0.02}$	$3e-2^{+4e-1}_{-3e-2}$		5 kpc	4^{+1}_{-1}	$0.21^{+0.09}_{-0.07}$	$2e-25^{+3e-13}_{-2e-25}$
ϕ	50 kpc	13^{+4}_{-3}	$0.04^{+0.01}_{-0.01}$	$<1e-300$	L_ϕ	50 kpc	31^{+20}_{-7}	$0.10^{+0.04}_{-0.02}$	$<1e-300$
	entire halo	33^{+13}_{-5}	$0.07^{+0.03}_{-0.01}$	$<1e-300$		entire halo	98^{+158}_{-45}	$0.19^{+0.20}_{-0.08}$	$<1e-300$
	5 kpc	2^{+1}_{-1}	$0.05^{+0.03}_{-0.03}$	$7e-4^{+2e-1}_{-7e-4}$		5 kpc	3^{+2}_{-1}	$0.20^{+0.08}_{-0.08}$	$2e-25^{+2e-8}_{-2e-25}$
z	50 kpc	27^{+9}_{-6}	$0.08^{+0.03}_{-0.01}$	$<1e-300$	L_z	50 kpc	110^{+101}_{-41}	$0.28^{+0.13}_{-0.09}$	$<1e-300$
	entire halo	161^{+147}_{-61}	$0.31^{+0.14}_{-0.10}$	$<1e-300$		entire halo	243^{+170}_{-101}	$0.40^{+0.11}_{-0.14}$	$<1e-300$
	5 kpc	2^{+2}_{-1}	$0.15^{+0.04}_{-0.05}$	$1e-10^{+3e-8}_{-1e-10}$		5 kpc	10^{+9}_{-3}	$0.51^{+0.06}_{-0.07}$	$2e-93^{+1e-42}_{-2e-93}$
r	50 kpc	244^{+137}_{-50}	$0.45^{+0.13}_{-0.05}$	$<1e-300$	$L_{\rm total}$	50 kpc	392^{+111}_{-69}	$0.58^{+0.04}_{-0.07}$	$<1e-300$
	entire halo	525^{+466}_{-63}	$0.59^{+0.12}_{-0.07}$	$<1e-300$		entire halo	600^{+382}_{-198}	$0.60^{+0.11}_{-0.10}$	$<1e-300$
	5 kpc	2^{+1}_{-1}	$0.14^{+0.05}_{-0.05}$	$2e-11^{+1e-6}_{-2e-11}$		5 kpc	22^{+30}_{-12}	$0.74^{+0.08}_{-0.25}$	$3e-166^{+6e-72}_{-3e-166}$
v_r	50 kpc	15^{+9}_{-4}	$0.05^{+0.03}_{-0.01}$	$<1e-300$	J_r	50 kpc	373^{+360}_{-122}	$0.57^{+0.13}_{-0.09}$	$<1e-300$
	entire halo	28^{+10}_{-9}	$0.07^{+0.02}_{-0.02}$	$<1e-300$		entire halo	760^{+565}_{-298}	$0.66^{+0.10}_{-0.14}$	$<1e-300$
	5 kpc	2^{+1}_{-1}	$0.11^{+0.06}_{-0.04}$	$2e-13^{+3e-4}_{-2e-13}$		5 kpc	7^{+8}_{-3}	$0.40^{+0.13}_{-0.16}$	$4e-55^{+1e-26}_{-4e-55}$
v_ϕ	50 kpc	1^{+1}_{-1}	$0.00^{+0.00}_{-0.00}$	$3e-4^{+10e-1}_{-3e-4}$	J_z	50 kpc	229^{+61}_{-68}	$0.44^{+0.07}_{-0.08}$	$<1e-300$
	entire halo	2^{+1}_{-1}	$0.00^{+0.00}_{-0.00}$	$7e-7^{+9e-1}_{-7e-7}$		entire halo	362^{+245}_{-98}	$0.49^{+0.16}_{-0.10}$	$<1e-300$
	5 kpc	3^{+2}_{-1}	$0.19^{+0.09}_{-0.07}$	$4e-21^{+1e-8}_{-4e-21}$		5 kpc	22^{+11}_{-9}	$0.68^{+0.07}_{-0.13}$	$6e-156^{+2e-77}_{-6e-156}$
v_z	50 kpc	35^{+15}_{-7}	$0.10^{+0.04}_{-0.02}$	$<1e-300$	E	50 kpc	934^{+755}_{-319}	$0.77^{+0.07}_{-0.08}$	$<1e-300$
	entire halo	43^{+14}_{-8}	$0.09^{+0.03}_{-0.01}$	$<1e-300$		entire halo	1489^{+1031}_{-567}	$0.80^{+0.07}_{-0.10}$	$<1e-300$
	5 kpc	12^{+7}_{-4}	$0.56^{+0.06}_{-0.12}$	$2e-98^{+5e-51}_{-2e-98}$		5 kpc	$(6^{+11}_{-5})\times10^{27}$	$1.00^{+0.00}_{-0.00}$	$<1e-300$
$v_{\rm total}$	50 kpc	117^{+39}_{-37}	$0.28^{+0.07}_{-0.07}$	$<1e-300$	$z_{\rm infall}$	50 kpc	$(5^{+5}_{-4})\times10^{27}$	$1.00^{+0.00}_{-0.00}$	$<1e-300$
	entire halo	67^{+38}_{-22}	$0.14^{+0.06}_{-0.03}$	$<1e-300$		entire halo	$(4^{+3}_{-3})\times10^{27}$	$1.00^{+0.00}_{-0.00}$	$<1e-300$

Note. Each continuous variable in the table is tested for its level of association to the true cluster labels. The ω^2 values estimate the strength of the association; a high ω^2 value (e.g., near 1) implies that this variable is likely to be important in clustering. Uncertainty values represent 16th–84th percentile scatter across the 32 simulations.

Based on these results, the clusters with high energy and high J_r are significantly more trustworthy. For example, the clusters with median energy higher than twice the median of all clusters in a local sample are pure *and* complete over 90% of the time. This is true for both UFD-only data sets and all-stars data sets.

The UFD remnants recovered in these real clusters are UFDs that, generally, accreted relatively recently. Figure 14 shows the median accretion redshift $z_{\rm accretion}$ for UFDs recovered by HDBSCAN compared to all unrecovered remnants. UFDs that were accreted at redshift $z=1$ and higher are virtually never recovered by any of these clustering algorithms. The dynamic signature of these small dwarfs is completely lost as the stars phase-mix in the dense region of action space, and the remnants are no longer identifiable. This is not surprising because energy and orbital actions are only truly conserved in static potentials, and the realistic, time-varying galactic potentials cause the stellar dynamics to relax over time.

As discussed in Section 4.5, for a UFD remnant to be reliably identified through kinematic clustering, it needs to both have had outlying dynamics at the time of accretion and also have a recent accretion time, $z_{accretion}\lesssim 0.5$, so that its stars have not had time to significantly phase-mix. Not all recently accreted UFD remnants are identifiable through kinematics (recently accreted UFDs can still end up in the dense regions of phase space; see Figure 2), but of the identifiable UFD remnants, virtually all are recently accreted.

6. Recommendations for Using Cluster Algorithms

Our study has clearly shown that using clustering algorithms with stellar dynamics to search for accreted UFD remnants is a challenging task, which, unfortunately, does not deliver reliable results a majority of the time.

Dynamically linked clusters identified by any clustering algorithms should thus not be blindly trusted but amply questioned and investigated, and results presented in a careful manner to avoid the presentation of numerically artificially created results. Case in point is our idealized situations in which we limit our data sets to only accreted UFD star particles and optimize our hyperparameter choices. The resulting UFD recovery rates are around ∼6% at best, and the majority of clusters found by all algorithms are not real. Only stars from fairly recently accreted UFDs ($z_{accretion}\lesssim 0.5$) can retain sufficiently strong dynamic signatures to be identified by these algorithms.

While these findings are unfortunate and must be taken into account in future searches, not all is lost. Clustering with stellar dynamics remains one of the few methods presently available to identify an accreted structure in observed Milky Way survey

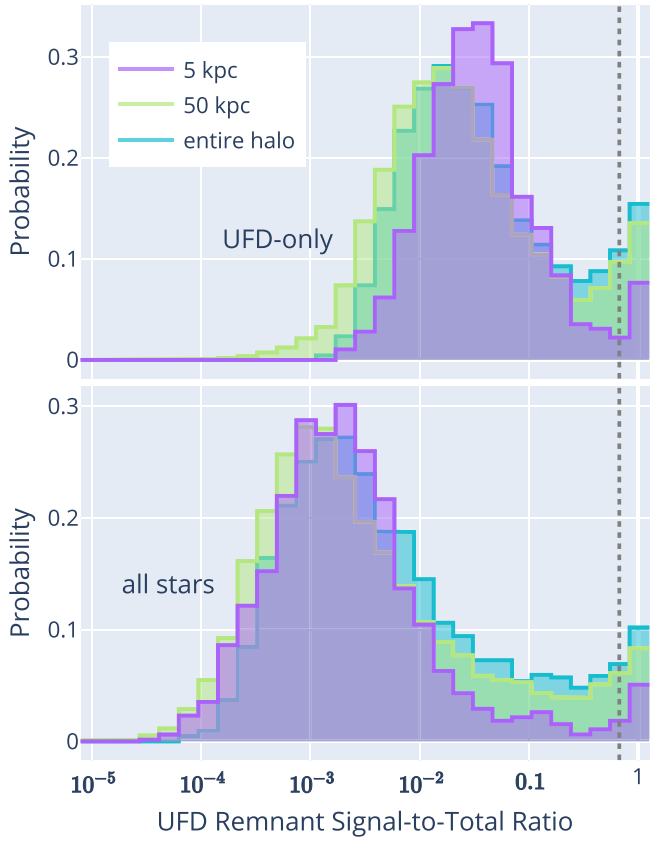


Figure 12. Normalized histograms of estimated signal-to-total ratios (similar to signal-to-noise ratios) for all UFD remnants in all data sets. For the vast majority of remnants, the ratio is tiny because the UFD remnant particles significantly overlap with all the other particles. The signal is very weak. For context, the dotted line shows 67%, our purity threshold. Depending on the data set, 92%–97% of UFD remnants have a signal-to-total ratio below this threshold. The all-stars data sets (bottom plot) have particularly low UFD signals—the median ratio is 1 UFD remnant particle to 1000 nonremnant particles.

data, and while not all UFDs can be found this way, identifying real remnants is possible.

To ensure that the results are as reliable and trustworthy as possible, we recommend the following for researchers:

1. Among these out-of-the-box clustering algorithms, choose HDBSCAN. Across our different data sets, HDBSCAN consistently balances the highest UFD remnant recovery rates and cluster realness rates. It is also more computationally scalable than all algorithms other than FoF.
2. UFD dynamic signatures are frequently weak, so incorporate chemical tagging when identifying groups of accreted stars. This can be done, for example, by focusing on low-metallicity stars and/or *r*-process enhanced stars. Successfully limiting a data set to UFD stars increases your remnant recovery rate by around 3 × on average. Chemical abundances can also be used to help validate dynamic clusters.
3. Assume most clusters identified by clustering algorithms do not correspond to real UFD remnants. Focus on the clusters with higher-than-average energy and J_r.
4. Recognize that only recently accreted UFDs in lower-density areas of phase space are consistently found by these clustering algorithms, so you generally only recover

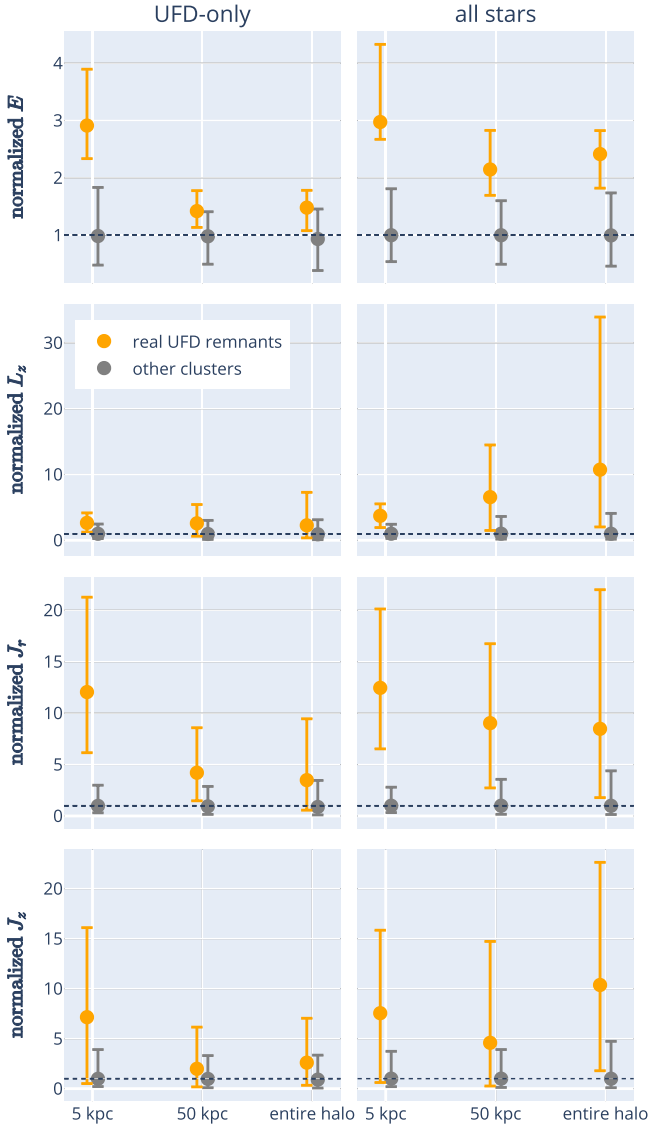


Figure 13. Median dynamics for clusters that correspond to real UFD remnants (i.e., pure and complete clusters) compared to other clusters. Clusters with higher actions are more likely to be real. Error bars show 16%–84% scatter across all clusters.

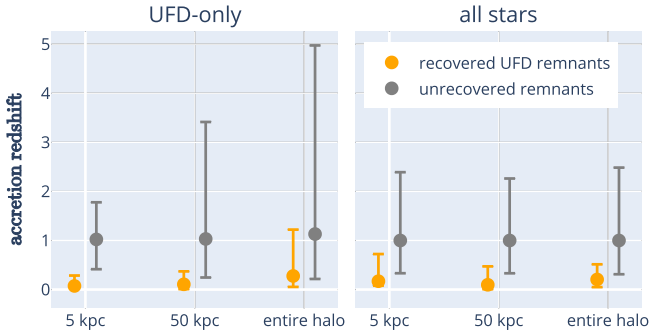


Figure 14. Median $z_{\rm accretion}$ (redshift at which a given dwarf galaxy was accreted) for recovered UFD remnants compared to all unrecovered remnants. As expected, the UFD remnants that are recovered by HDBSCAN (and other algorithms) were more recently accreted. Error bars show 16%–84% scatter across all remnants.

1%–6% of the UFD remnants in a given sample. The samples limited to the region around the Sun have lower recovery rates than the samples with larger radial cuts.

Table 3
Trial Hyperparameter Values for All Algorithms

Algorithm	Hyperparameter	Hyperparameter Search Space
HDBSCAN	min_cluster_size	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
K-means	n_clusters	60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250
Gaussian mixture models	n_clusters	
Agglomerative clustering	n_clusters	
Friends-of-friends	linking_length	0.001, 0.003, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05, 0.055, 0.06, 0.065, 0.07, 0.075, 0.08, 0.1, 0.125, 0.15, 0.175, 0.2
Mean-shift	bandwidth	0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2
Affinity propagation	…	…

Table 4
Optimal Hyperparameter Values for All Data Sets

		Chosen Hyperparameters					
		Only UFD particles			UFD and Non-UFD Particles		
Algorithm	Hyperparameter	5 kpc	50 kpc	no radial cut	5 kpc	50 kpc	no radial cut
HDBSCAN	min_cluster_size	4	10	15	3	10	19
K-means	n_clusters	60	230	180	200	230	150
Gaussian mixture models	n_clusters	90	160	210	70	220	240
Agglomerative clustering	n_clusters	80	250	250	170	too slow	too slow
Friends-of-friends	linking_length	0.065	0.015	0.01	0.065	0.015	0.01
Mean-shift	bandwidth	0.11	0.06	0.06	0.16	too slow	too slow
Affinity propagation	…	…	…	…	…	…	cdots

5. Vary your hyperparameter choices and consider the stability of the clustering results across several hyperparameter values. For HDBSCAN, the best hyperparameter values are the ones that produce fewer than several hundreds of clusters (in our samples, they require min_cluster_size $\gtrsim 9$ for our large radial cuts) and produce more than just a few clusters (in our samples, they require min_cluster_size $\lesssim 6$ for our 5 kpc radial cut). This will depend on your sample, so test different hyperparameter choices to avoid hyperparameters that result in an unreasonably large or small number of clusters.

K.B. acknowledges support from the United States Department of Energy grant DE-SC0019323. A.F. acknowledge support from NSF grant AST-1716251. F.A.G. acknowledges support from ANID FONDECYT Regular 1211370 and by the ANID BASAL project FB210003. F.A.G. also acknowledges funding from the Max Planck Society through a "Partner Group" grant.

This work made extensive use of the python libraries `numpy` (Harris et al. 2020), `scipy` (Virtanen et al. 2020), `sqlite3` (Häring 2006), and `plotly` (Inc., 2015).

ORCID iDs

Kaley Brauer https://orcid.org/0000-0002-8810-858X
Hillary Diane Andales https://orcid.org/0000-0002-2962-1391
Alexander P. Ji https://orcid.org/0000-0002-4863-8842
Anna Frebel https://orcid.org/0000-0002-2139-7145
Mohammad K. Mardini https://orcid.org/0000-0001-9178-3992
Facundo A. Gómez https://orcid.org/0000-0002-1947-333X
Brian W. O'Shea https://orcid.org/0000-0002-2786-0348

References

Arthur, D., & Vassilvitskii, S. 2006, k-means++: The Advantages of Careful Seeding. Technical Report, Stanford, 1027, https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf
Bailin, J., Bell, E. F., Valluri, M., et al. 2014, ApJ, 783, 95
Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013a, ApJ, 762, 109
Behroozi, P. S., Wechsler, R. H., Wu, H.-Y., et al. 2013b, ApJ, 763, 18
Belokurov, V., Erkal, D., Evans, N. W., Koposov, S. E., & Deason, A. J. 2018, MNRAS, 478, 611
Binney, J. 2012, MNRAS, 426, 1324
Binney, J., & Tremaine, S. 2008, Galactic Dynamics (2nd edn.; Princeton, NJ: Princeton Univ. Press)
Brauer, K., Ji, A. P., Frebel, A., et al. 2019, ApJ, 871, 247
Bryan, G. L., & Norman, M. L. 1998, ApJ, 495, 80

Bullock, J. S., & Johnston, K. V. 2005, ApJ, 635, 931
Burbidge, E. M., Burbidge, G. R., Fowler, W. A., & Hoyle, F. 1957, RvMP, 29, 547
Cameron, A. G. W. 1957, PASP, 69, 201
Campello, R. J., Moulavi, D., Zimek, A., & Sander, J. 2015, ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 10 (New York: Association for Computing Machinery), 1
Comaniciu, D., & Meer, P. 2002, ITPAM, 24, 603
Cooper, A. P., Cole, S., Frenk, C. S., Le Bret, T., & Pontzen, A. 2017, MNRAS, 469, 1691
Cooper, A. P., Cole, S., Frenk, C. S., et al. 2010, MNRAS, 406, 744
Cowan, J. J., Sneden, C., Lawler, J. E., et al. 2021, RvMP, 93, 015002
Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371
De Lucia, G., & Helmi, A. 2008, MNRAS, 391, 14
Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, J. R. Stat. Soc. Ser. B Methodol., 39, 1
Derpanis, K. G. 2005, Lecture Notes, 32
Dooley, G. A., Peter, A. H. G., Vogelsberger, M., Zavala, J., & Frebel, A. 2016, MNRAS, 461, 710
Drlica-Wagner, A., Bechtol, K., Rykoff, E. S., et al. 2015, ApJ, 813, 109
Errani, R., Peñarrubia, J., Laporte, C. F. P., & Gómez, F. A. 2017, MNRAS, 465, L59
Ezzeddine, R., Rasmussen, K., Frebel, A., et al. 2020, ApJ, 898, 150
Forbes, D. A. 2020, MNRAS, 493, 847
Frebel, A. 2010, AN, 331, 474
Frebel, A. 2018, ARNPS, 68, 237
Frey, B. J., & Dueck, D. 2007, Sci, 315, 972
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
Garrison-Kimmel, S., Bullock, J. S., Boylan-Kolchin, M., & Bardwell, E. 2017, MNRAS, 464, 3108
Gibbons, S. 2020, pyfof, GitHub, https://github.com/simongibbons/pyfof
Gómez, F. A., Coleman-Smith, C. E., O'Shea, B. W., Tumlinson, J., & Wolpert, R. L. 2012, ApJ, 760, 112
Gómez, F. A., Coleman-Smith, C. E., O'Shea, B. W., Tumlinson, J., & Wolpert, R. L. 2014, ApJ, 787, 20
Gómez, F. A., Helmi, A., Brown, A. G. A., & Li, Y.-S. 2010, MNRAS, 408, 935
Griffen, B. F., Ji, A. P., Dooley, G. A., et al. 2016, ApJ, 818, 10
Gudin, D., Shank, D., Beers, T. C., et al. 2021, ApJ, 908, 79
Hahn, O., & Abel, T. 2011, MNRAS, 415, 2101
Hansen, T. T., Holmbeck, E. M., Beers, T. C., et al. 2018, ApJ, 858, 92
Hansen, T. T., Marshall, J. L., Simon, J. D., et al. 2020, ApJ, 897, 183
Hansen, T. T., Simon, J. D., Marshall, J. L., et al. 2017, ApJ, 838, 44
Häring, G. 2006, SQLite3, v3.7.15. https://docs.python.org/3/library/sqlite3.html
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Helmi, A., Babusiaux, C., Koppelman, H. H., et al. 2018, Natur, 563, 85
Helmi, A., & Tim de Zeeuw, P. 2000, MNRAS, 319, 657
Holmbeck, E. M., Hansen, T. T., Beers, T. C., et al. 2020, ApJS, 249, 30
Huchra, J., & Geller, M. 1982, ApJ, 257, 423
Hunt, E. L., & Reffert, S. 2021, A&A, 646, A104
Inc., Plotly Technologies. 2015, Collaborative Data Science, Montreal, QC: Plotly Technologies Inc., https://plot.ly
Ji, A. P., Frebel, A., Chiti, A., & Simon, J. D. 2016a, Natur, 531, 610
Ji, A. P., Frebel, A., Simon, J. D., & Chiti, A. 2016b, ApJ, 830, 93
Kirby, E. N., Cohen, J. G., Guhathakurta, P., et al. 2013, ApJ, 779, 102
Kruijssen, J. M. D., Pfeffer, J. L., Chevance, M., et al. 2020, MNRAS, 498, 2472
Kruijssen, J. M. D., Pfeffer, J. L., Reina-Campos, M., Crain, R. A., & Bastian, N. 2019, MNRAS, 486, 3180
Le Bret, T., Pontzen, A., Cooper, A. P., et al. 2017, MNRAS, 468, 3212
Limberg, G., Rossi, S., Beers, T. C., et al. 2021, ApJ, 907, 10
Lloyd, S. 1982, ITIT, 28, 129
Maffione, N. P., Gómez, F. A., Cincotta, P. M., et al. 2018, MNRAS, 478, 4052
Mardini, M. K., Frebel, A., Chiti, A., et al. 2022, ApJ, 936, 78
McDonald, J. H. 2014, Handbook of Biological Statistics (3rd edn.; Baltimore, MD: Sparky House Publishing), 145, https://www.biostathandbook.com/HandbookBioStatThird.pdf
McInnes, L., Healy, J., & Astels, S. 2017, JOSS, 2, 205
Monachesi, A., Gómez, F. A., Grand, R. J. J., et al. 2019, MNRAS, 485, 2589
Morinaga, Y., Ishiyama, T., Kirihara, T., & Kinjo, K. 2019, MNRAS, 487, 2718
Myeong, G. C., Vasiliev, E., Iorio, G., Evans, N. W., & Belokurov, V. 2019, MNRAS, 488, 1235
Naidu, R. P., Conroy, C., Bonaca, A., et al. 2020, ApJ, 901, 48
Olejnik, S., & Algina, J. 2003, Psychol. Methods, 8, 434
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014, A&A, 571, A16
Press, W., & Davis, M. 1982, ApJ, 259, 449
Rashkov, V., Madau, P., Kuhlen, M., & Diemand, J. 2012, ApJ, 745, 142
Roederer, I. U., Hattori, K., & Valluri, M. 2018, AJ, 156, 179
Roederer, I. U., Mateo, M., Bailey, J. I. I., et al. 2016, AJ, 151, 82
Ruiz-Lara, T., Matsuno, T., Sofie Lövdal, S., et al. 2022, arXiv:2201.02405
Sakari, C. M., Placco, V. M., Farrell, E. M., et al. 2018, ApJ, 868, 110
Sharma, S., & Johnston, K. V. 2009, ApJ, 703, 1061
Simon, J. D. 2019, ARA&A, 57, 375
Sofie Lövdal, S., Ruiz-Lara, T., Koppelman, H. H., et al. 2022, arXiv:2201.02404
Tumlinson, J. 2010, ApJ, 708, 1398
Vasiliev, E. 2019, MNRAS, 482, 1525
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261
Ward, J. H., Jr 1963, Journal of the American Statistical Association, 58, 236
Wu, Y., Valluri, M., Panithanpaisal, N., et al. 2022, MNRAS, 509, 5882
Yuan, Z., Chang, J., Banerjee, P., et al. 2018, ApJ, 863, 26
Yuan, Z., Myeong, G. C., Beers, T. C., et al. 2020, ApJ, 891, 39