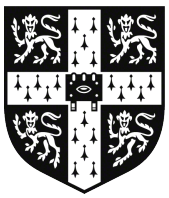


UNIVERSITY OF CAMBRIDGE

A4 Galactic Archaeology - Coursework Report

Jacob Tutt (JLT67)

Department of Physics, University of Cambridge

April 4, 2025

Word Count: 2999

1 Introduction

The vast spatial extent and long orbital periods of the Milky Way's halo allow it to serve as a fossil record of its hierarchical assembly. By preserving the coherence of a progenitor system's debris over gigayear timescales, it enables the reconstruction of past accretion events [13] and acts as a tracer of the galaxy's dark matter distribution at large radii [15].

Deason and Belokurov [7] highlighted that the study of the galactic halo has long been hampered by incomplete and insufficient sample sizes, sky coverage and chemo-dynamical information. However, as we enter the era of big data astronomy, Gaia [16] and its complementary spectroscopic surveys promise unprecedented insight into the remnants of past mergers. As datasets grow in size and dimensionality, so does the motivation of developing automated, data-driven techniques.

This paper leverages the astrometric measurements of 1.8 billion sources from Gaia EDR3 [19] to explore potential methods for automating stellar halo substructure detection. We aim to increase our sensitivity to distant features by optimising the selection of Red Giant Branch (RGB) stars, whose intrinsic brightness enables their astrometry to be reliably measured at large Galactocentric distances. We investigate both well-established clustering algorithms, as well as introduce a novel method designed for efficient substructure identification in noisy, large-scale datasets.

2 Investigation of Documented Substructures

Initially, the distribution of known substructures was investigated to contextualise the effects of data cuts described in section 3. We examined their Galactic latitude and longitude (Figure 1) using the globular cluster catalog from Vasiliev [20] and Milky Way satellites from Nadler et al. [14], extending the analysis to the proper motion distribution (Figure 2). Early findings highlighted the risk of excluding a significant number of substructures with overly strict proper motion cuts. Although latitude restrictions ($abs(b) < 10$) may remove a subset of globular clusters, these are largely concentrated in the bulge [4] and thus lie outside this investigation's scope.

3 Red Giant Branch Data

3.1 Initial Acquisition

Initial samples were retrieved from the Gaia archive using ADQL via both `Astroquery` and the web portal. Standard quality cuts ensured initial high-quality data, including Gaia's recommended RUWE < 1.4 to remove poorly resolved/ blended astrometry. A parallax cut of $\varpi < 0.1$ mas removed nearby stars (within $\sim$10 kpc) to reduce bulge contamination. This was later concluded

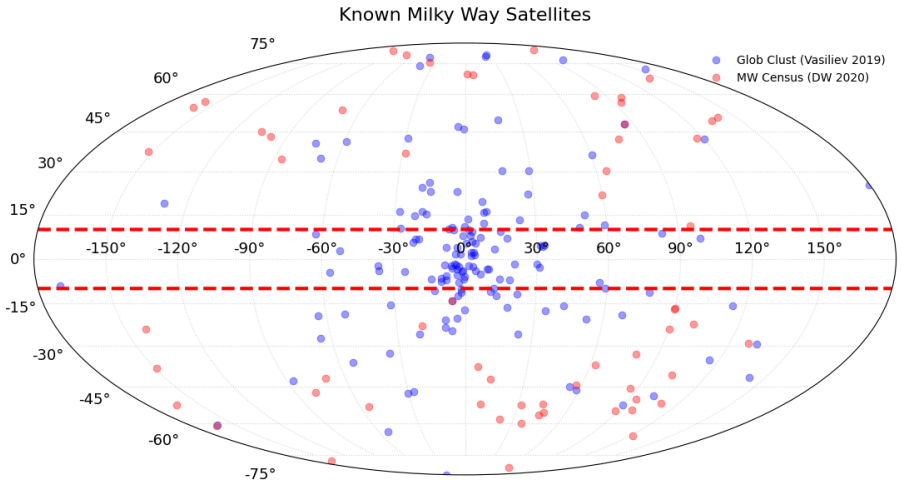


Figure 1: Distribution in Galactic latitude–longitude of known substructures.

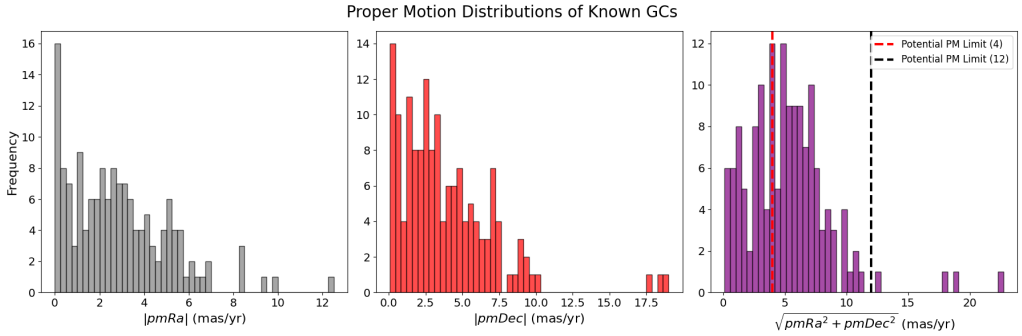


Figure 2: Distribution in proper motion of known Globuluar Clusters.

insufficient and a stricter Galactic latitude cut of $|b| > 10°$ was adopted. A broad G-band magnitude range (10–20.5) was initially used to prevent prematurely excluding bright stars due to extinction, and further refined after correction (subsection 3.3).

To further isolate halo stars, a total proper motion ($\sqrt{\text{PMRA}^2 + \text{PMDEC}^2}$) cut was applied. Two thresholds, 4 mas/yr and 12 mas/yr, were adopted to prevent over-restrictive filtering (demonstrated in Figure 2). Additionally, each star required that the photogeometric distance estimates (Bailer-Jones et al. [2]) be available, and Gaia's random index was employed for downsampling where necessary, to stay within memory constraints without biasing the sample.

3.2 Corrections

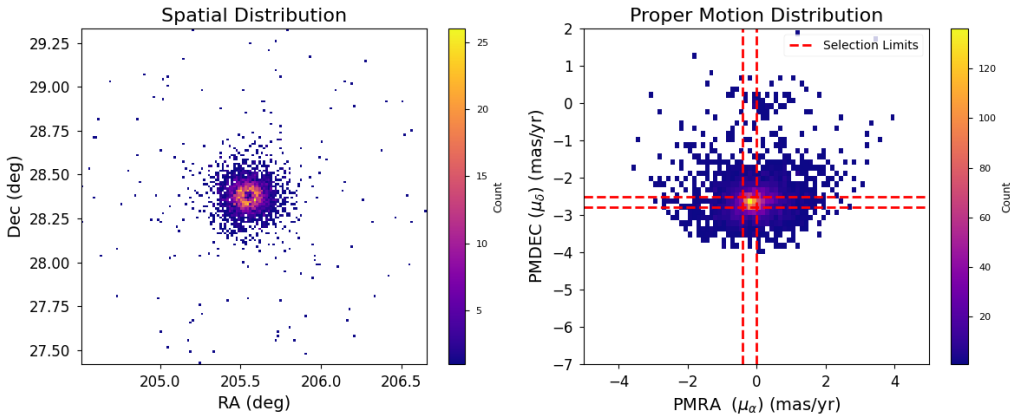
On import, all data was corrected for line-of-sight extinction in Gaia's G, B_P, and R_P bands. This used the dust maps provided by Schlegel, Finkbeiner, and Davis [17] and coefficients from Casagrande et al. [6] for the band-specific calibration. The Sun's peculiar motion was deemed negligible for distant halo dynamics, consistent with Helmi, Amina et al. [10]. However, subsection 3.5 transforms units into a Galactocentric frame and computes integrals of motion to enable a more accurate reconstruction of progenitors.

3.3 Red Giant Branch Refinement

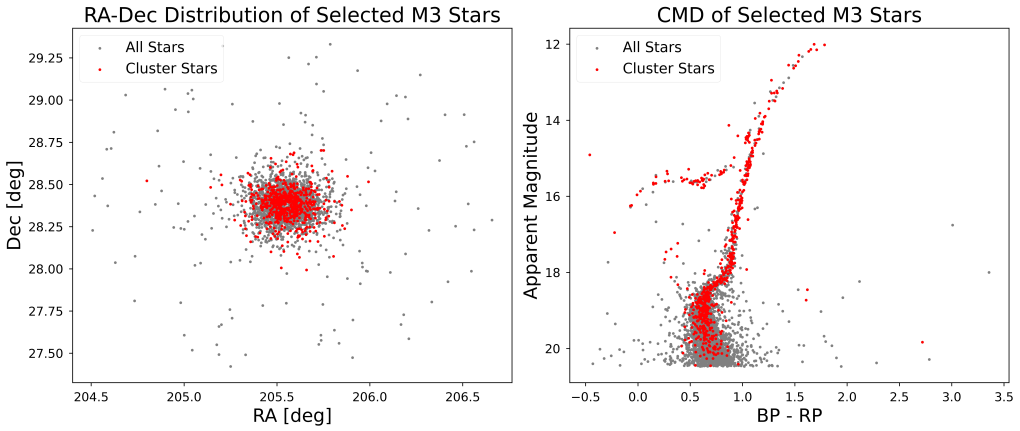
Given Gaia's astrometric focus, more advanced cuts (i.e. photometric) required both validation and optimisation. Furthermore, due to Gaia's negative parallaxes and systematic errors [8], a lower parallax bound was investigated, though applied conservatively due to their spatial dependence.

To evaluate the completeness and purity of filtering thresholds, globular clusters M3 (NGC 5272) and NGC 1851 were used as benchmarks. Stars passing the initial cuts (section 3) and within a 1° radius were selected, and likely cluster members isolated via proper motion.

As is the overarching aim of the Red Giant Branch Selection, we aimed to maximise the ratio of cluster to background stars. Although illustrative cuts for M3 are shown, final selection criteria were based on performance across both clusters (Table 2) and are summarised in Table 1.



(a) Cluster Star Identification



(b) Cluster Stars Results

Figure 3: Initial selection of cluster stars

Parameter	Value
BP–RP Cut	0.8
Magnitude Cut	18
Lower Parallax Cut	−0.1

Table 1: Filters for RGB selection.

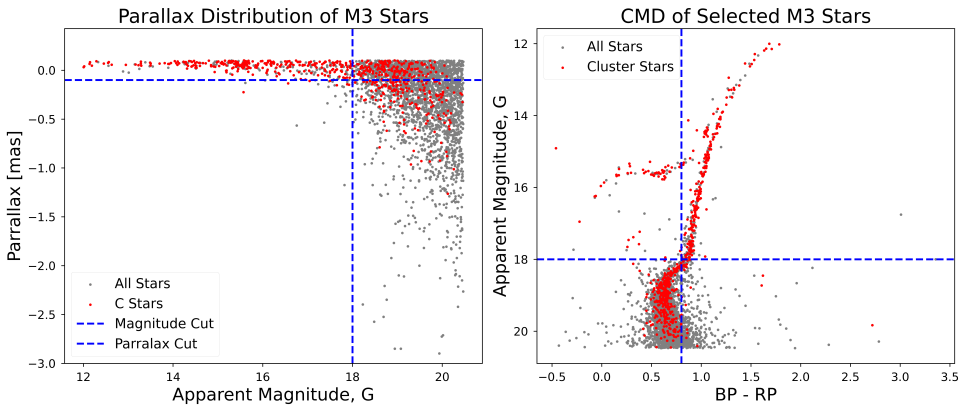
Cut Criteria	Total Stars	Cluster Stars	% Cluster	% Improvement
No Cut	3256	609	18.7%	N/A
Magnitude Cut	410	262	63.9%	45.2%
BP–RP Cut	769	235	30.6%	11.9%
Parallax Cut	1479	431	29.1%	10.4%
All Cuts	305	202	66.2%	47.5%

(a) M3 (NGC 5272)

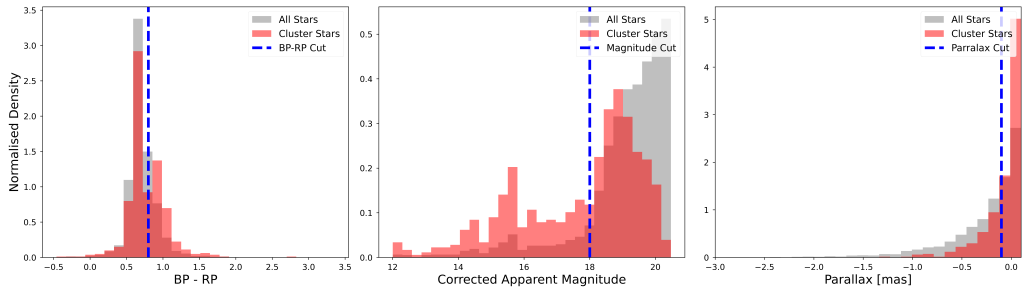
Cut Criteria	Total Stars	Cluster Stars	% Cluster	% Improvement
No Cut	1556	257	16.5%	N/A
Magnitude Cut	239	143	59.8%	43.3%
BP–RP Cut	562	160	28.5%	12.0%
Parallax Cut	829	211	25.5%	8.9%
All Cuts	193	123	63.3%	46.7%

(b) NGC 1851

Table 2: Filtering effects across cut criteria



(a) Optimised Cuts' Positions



(b) Distributions of populations

Figure 4: Optimised Cuts

The adopted cuts are qualitatively justified, as they select cooler, larger stars, typical of the RGB, and exclude highly uncertain or nonphysical parallax solutions. For both test examples, the fraction of cluster stars exceeded 63%, thus significantly improving the halo substructure selection.

Additional cuts were explored (Figure 5), including using the photogeometric distances from Bailer-Jones et al. [2] to compute absolute magnitudes, M_G, a more robust Red Giant Stars measure. Although a $M_G < 4$ threshold resulted in modest improvements for both clusters, the distance-dependent uncertainties and biases in the estimates limited their broader reliability. Furthermore, Gaia's surface gravities and effective temperatures were too sparsely reported and insufficiently accurate to offer practical benefit.

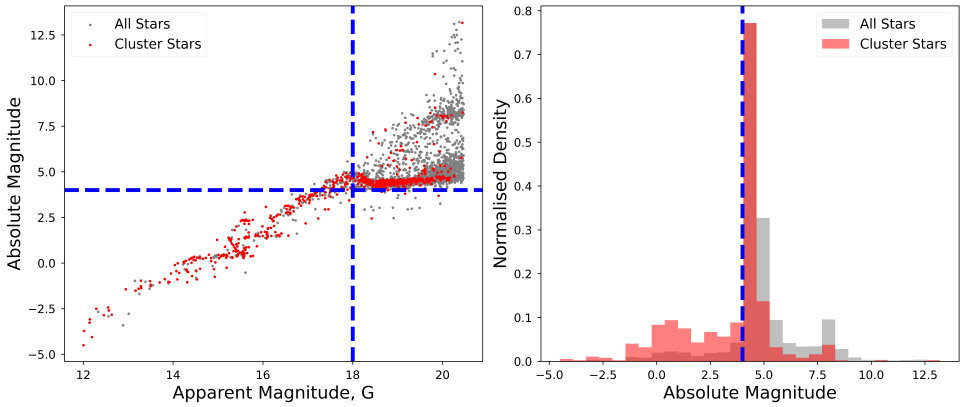
3.4 Resultant Samples

We obtain two final samples: a lower proper motion cut ($\mathrm{PM_{tot}} < 4$ mas/yr) and a higher proper motion cut ($\mathrm{PM_{tot}} < 12$ mas/yr), containing 3,105,304 and 2,452,277 stars, respectively. Although counter-intuitive, the lower proper motion sample is larger due to the initial query being fixed in size, and a higher proportion of stars passing the filters (7.31% vs 5.38%). This confirms the $\mathrm{PM_{tot}} < 4$ cut is more appropriate for isolating halo-like candidates.

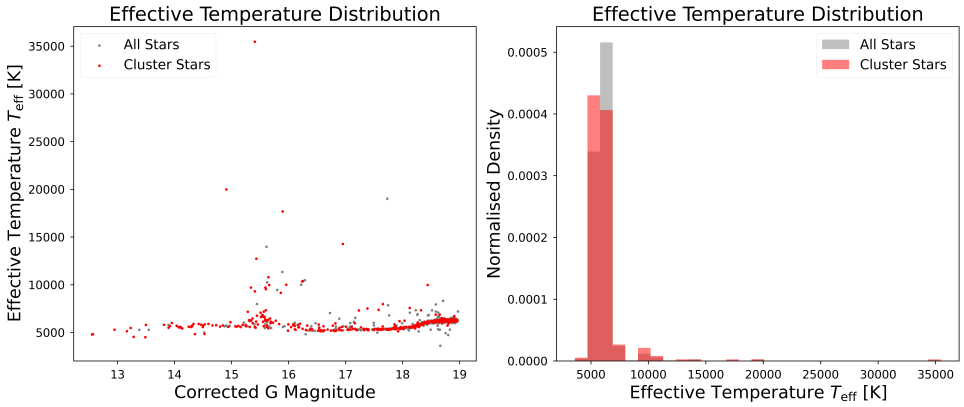
3.5 Radial Velocity Samples

A further subsample of 301,642 stars was extracted from the $\mathrm{PM_{tot}} < 4$ proper motion set by requiring radial velocity measurements. This enabled transformation into a Galactocentric frame and hypothetical reconstruction of the full 6D phase-space. This aids the identification of dynamically coherent tidal debris through conserved orbital quantities (section 6).

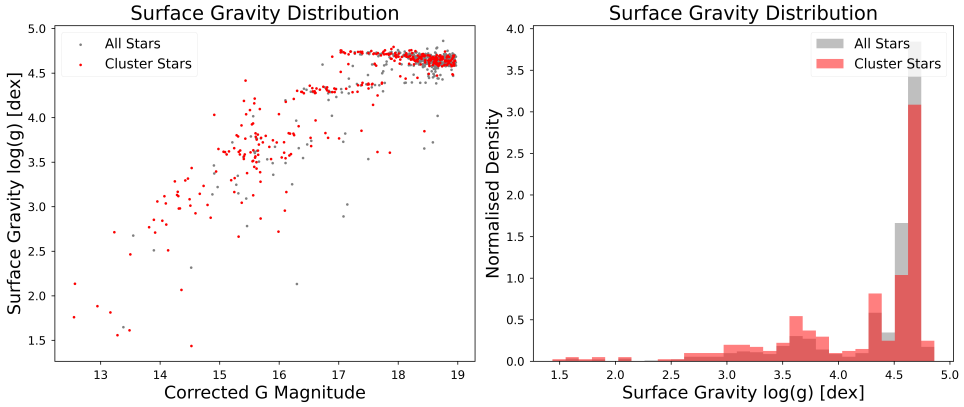
Using `SkyCoord` and `galpy` to model the Milky Way's potential, we compute the energy (E) and the z-component of angular momentum (L_z) for each star. These were selected as they could be calculated without orbital integration, which was computationally unfeasible at this scale. We also obtain Galactocentric radii (R_{gal}), which while not conserved, may help spatially isolate structures.



(a) M3: Absolute Magnitude, M_G



(b) M3: Effective Tempreture



(c) M3: Surface Gravity

Figure 5: Parameter Distributions from Gaia

3.6 Distances Probed

Gaia's documentation [18] warns distances it provides are only reliable within ≈2kpc, and become systematically underestimated from large fractional parallax errors beyond this. Although similarly limited and still unreliable, we analyse the photogeometric distances [2], which probabilistically combines parallaxes, colours, apparent magnitudes, and prior information. Overall, they estimate the two samples have 0.1772% (low PM) and 3.1976% (high PM) of stars lie beyond 15 kpc, with a shared maximum of 79.66 kpc.

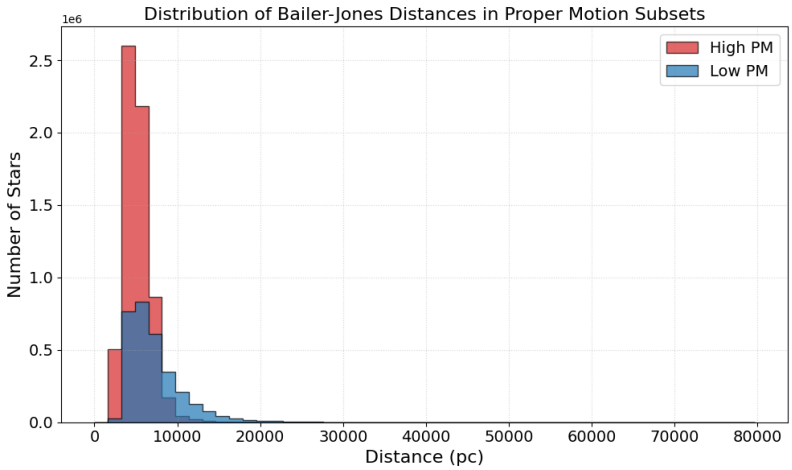


Figure 6: Photogeometric Distance Estimates Bailer-Jones et al. [2]

4 Density Maps

To visualise coherent structures we construct all-sky stellar density maps which aim to balance resolution, contrast, and colour mapping to suppress noise while enhancing meaningful features. Each map applies a percentile-based scaling to clip extreme values, followed by a logarithmic scaling to enhance contrast. To further minimise noise, pixels below a minimum count threshold are discarded. False-colour RGB composites are produced by binning stars into three intervals, allowing for the independent normalisation of distinct subgroups.

We treat the scaling bounds, colour binning schemes, and noise thresholds as hyperparameters to optimise and present plotting pipelines within `GA_Analysis` that allow easily adjustment.

4.1 Field of Streams

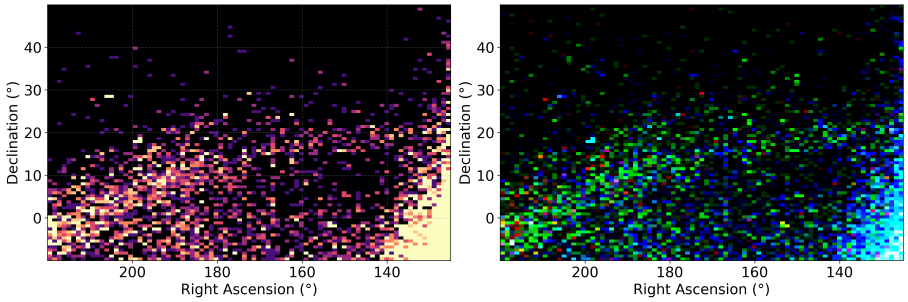
We begin by reproducing the Field of Streams (FoS) image from Belokurov et al. [3] to highlight the Sagittarius stream. Optimal contrast was achieved using a rather extreme 60–95 percentile scaling and three apparent magnitude channels with boundaries at $G = 16.1$ and 17.1 (Figure 8b).

4.2 All Sky View

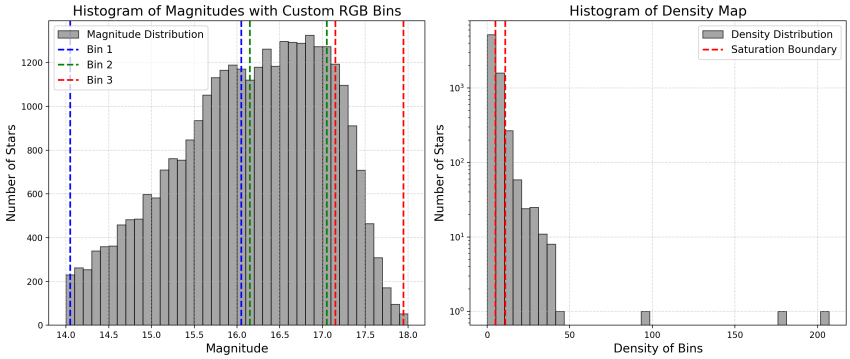
Extending this to an all-sky map, a simple 'binary' color map was found to best visualized over densities and the sagitarius stream can be clearly seen. The scaling percentiles of 5 and 95 were found to be optimal followed by additionally removing all bins with fewer than 11 samples post log-scaling to further reduce noise. Additional mollewide projections are included within the repository.

4.3 RGB Colour Maps Scheme

To improve structure identification, RGB composites used two binning strategies: (1) apparent magnitude bins (same as in Figure 7), and (2) linear proper motion bins between 0.5–3.5 mas/yr. Each 'colour channel' used the same normalisation strategies as 4.2. These composites are effective at isolating coherent tidal features with both showing the full (360) Sagittarius stream whereas single channel binning are concluded to better show over densities.

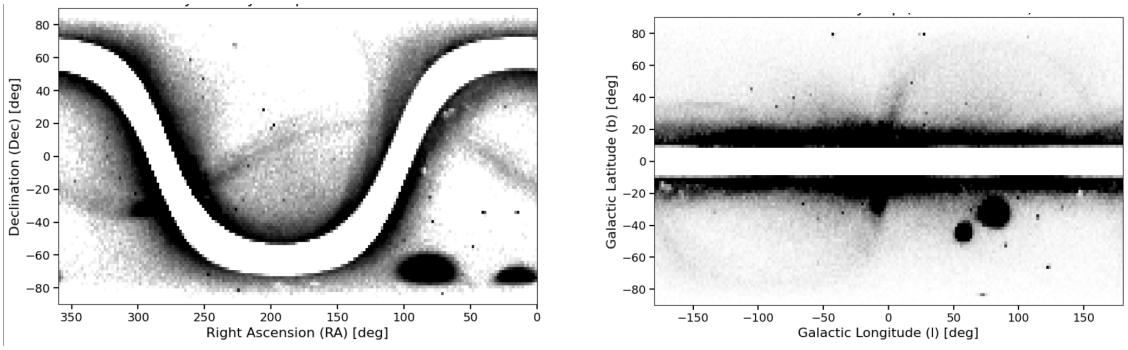


(a) Achieved Visualisation



(b) Hyperparameters

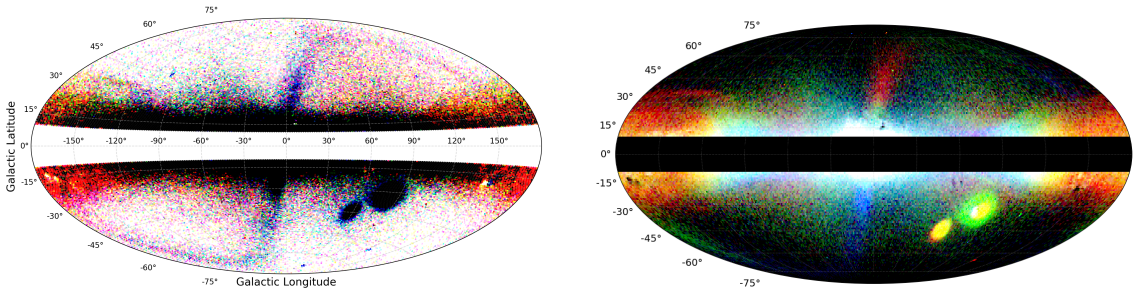
Figure 7: Field of Streams



(a) Equatorial Coordinates

(b) Galactic Coordinates

Figure 8: All-Sky Stellar Density Maps



(a) Apparent Magnitude Composites

(b) Proper Motion Composites

Figure 9: False RGB Composites

4.4 Visual Results

From the maps, we visually identify 27 over-densities, Sagittarius (with its stream) as well as the LMC and SMC. We therefore aim to devise automated detection strategies that are able to surpass this level of identification thus offering viable benefits.

5 Substructure Identification

We now aim to automatically detect over densities corresponding to globular clusters and dwarf galaxies using the Gaia astrometry. Due to their compact scales and gravitationally bound structure, such systems are identifiable via clustering in heliocentric positions (RA/Dec or l, b) and corresponding proper motions (unlike extended systems in section 6).

5.1 Traditional Approaches

These compact systems commonly have a spherical or elliptical projection on our sky and exhibit individual stellar movement that can be modeled as random perturbations to the systems global motion, making Gaussian Mixture Modelling (GMM) a natural choice.

Early tests used Extreme Deconvolution (XD) [5] due to its ability to incorporate Gaia's 'heteroskedastic uncertainties' and thus suitability for 'noisy, heterogeneous and incomplete observations'. The result showed far worse preformance than visual inspection, due to the true substructures occupying extremely small sections of each scaled dimension and the overwhelming noise. It was concluded a custom approach was necessary.

5.2 Novel Algorithm

A custom algorithm which combines simple binning with multi-dimensional clustering framework is defined. It requires no prior assumptions about the number of clusters (unlike GMM) and can be expanded for any dimensionality. The general structure of the algorithm is described below. The relative hyperparameters are the width of each dimensions bins, the absolute threshold and the relative threshold.

- **Binning:** The dataset is discretised into a 4D histogram using two spatial and two apparent motion dimensions.
- **Local Density Normalisation:** To account for variations in sky density, a 2D histogram is constructed by marginalising over proper motion.
- **Relative Filtering:** Each 4D bin is evaluated relative to the count of its associated 2D spatial bin. Bins that don't exceed a fixed fractional threshold are discarded. This reduces bias from crowded regions by using relative concentration rather than raw counts.
- **Absolute Filtering:** Additionally, 4D bins with insufficient absolute counts are discarded, removing bins in very sparsely populated areas that may by default pass the relative threshold.
- **Agglomerative Clustering:** Remaining bins are agglomerated by merging those that are direct neighbors in all dimensions using a strict 4D adjacency kernel. This condition distinguishes between true clusters, nearby but kinematically isolated structures, or spatially disjoint structures with similar motion.
- **Redundancy Removal:** To avoid fully nested clusters, those entirely contained within the 4D boundaries of another are removed. This still allows for partial nested/ overlapping clusters.
- **Results:** Each final structure is defined by the box's boundaries in position–motion space, forming a 4D hypercube.

The resultant hypercube boundaries, along with the Gaia source IDs of the stars contained within them, are stored in `.json` files within the `data` directory. For comparison of results, we use the strict condition that any recorded structure's values must fit fully within the 4D hypercube of a given cluster. For dwarf galaxies from Nadler et al. [14], where proper motions are unavailable, only spatial boundaries are used. Optimised bin widths of $\Delta l = \Delta b = 1^\circ$ and $\Delta\mu_\alpha, \Delta\mu_\delta = 0.65$

mas/yr are used. A fractional threshold of $f_{\text{thresh}} = 0.28$ and an absolute minimum bin count of $N_{\text{min}} = 20$ are used. An immediate benefit of this algorithm is both sets of results are achieved in less than 7 seconds.

5.3 Higher PM Results

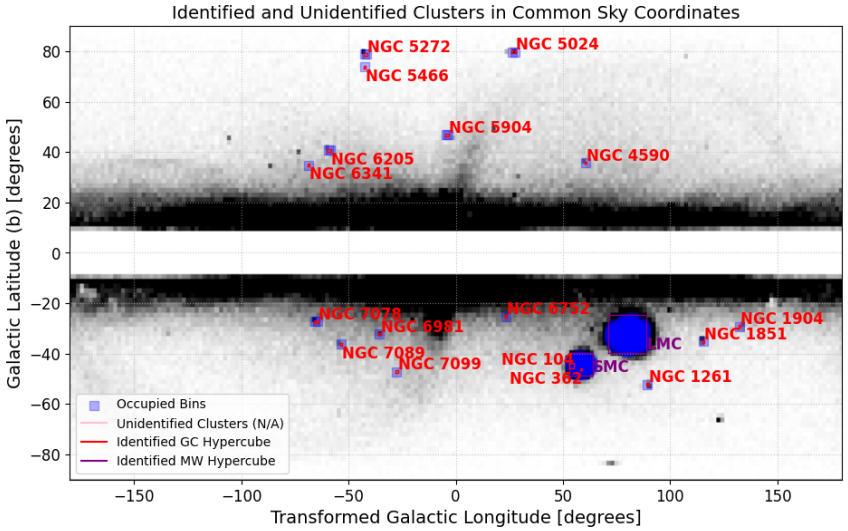


Figure 10: Higher PM: Custom Algorithm Results

For the higher PM dataset, we find 19 distinct clusters all of which have matches to the reference catalogs (section 2). Two promising results are the 0% false positive rate and the isolation of NGC 362 and NGC 104 from within the backdrop of the Small Magellanic Cloud, thus proving the success of the strict 4D adjacency kernel by providing benefits beyond visual inspection. Despite this, it does not identify all clusters than can be identified by eye.

5.4 Lower PM Results

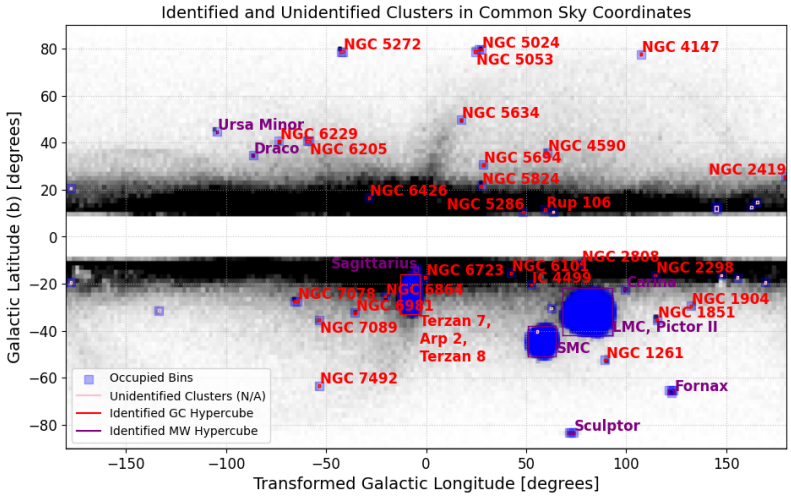


Figure 11: Lower PM: Custom Algorithm Results

The lower proper motion dataset produced significantly better results, indicating that the increased purity cut outweighs the risk of excluding some substructures. In total, 36 known objects were automatically identified, along with 12 additional unmatched structures. As this is greater than the number detected by visual inspection its supports the method's effectiveness. The cluster marked as a potential match to Terzan 7, Arp 2, or Terzan 8, exhibited boundary ranges too broad for this to be reliable and the algorithm is concluded to in fact be identifying the Sagittarius

stream. Finally, the split match with Pictor II is disregarded and it is concluded to be the LMC alone.

5.5 Overall Results

Across the two datasets, we automatically identify 43 unique known substructures, demonstrating the effectiveness of the custom algorithm in denoising Gaia's data and detecting local overdensities. The recovered structures include Sagittarius (and its tidal stream), the LMC, SMC, M3 (NGC 5272), and NGC 1851. For broader research, we present this algorithm as a potential preprocessing tool for isolating candidate regions of interest prior to applying more sophisticated methods. A prototype pipeline is outlined in Notebook 6.3, where XD is applied to the stars within each detected hypercube. It aims to fit a two-component Gaussian model in an attempt to further separate the cluster and background populations, though this was not finalised. A full breakdown of all resultant clusters is provided in the Appendices. 9.

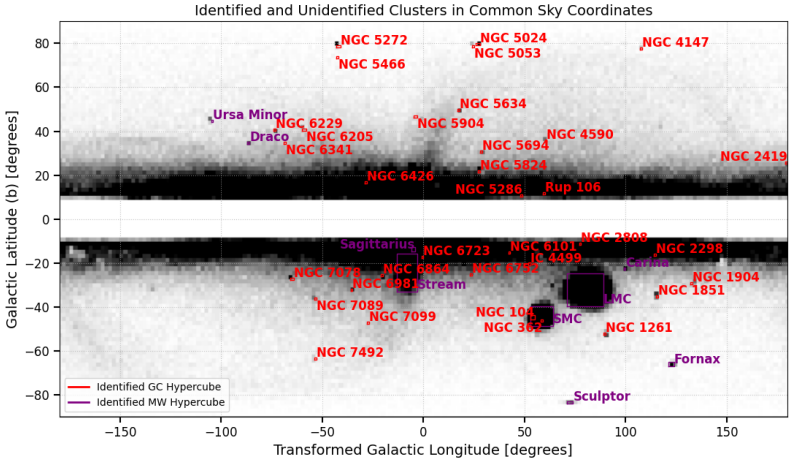


Figure 12: Overall Matched Structures

6 Tidal Stream Identification

Tidal streams are diffuse structures formed as progenitors are stretched along their orbit by the Milky Way's potential. Due to their extended and disrupted nature, proper motion is an unreliable clustering parameter due to its distortion in our heliocentric frame, as shown by the colour differences between the leading and trailing arms of the Sagittarius stream in Figure 9b. To overcome this, the `galpy` subset data generated in subsection 3.5 is used, which provides Galactocentric radius R_{gal} and integrals of motion such as Energy and L_z.

6.1 HBDSCAN

We employ the density-based hierarchical clustering algorithm, HDBSCAN [11], due to its ability to robustly handle noise and its hierarchical nature not requiring a prior on the number of clusters. Throughout we assign a minimium cluster size of 40 and a local density threshold (min_samples) of 15. For each run, all clustering parameters are scaled to be standard normal distirbutions. These were run over a range of permutations with two presented below. Additionally we have excluded clusters found within $|b| < 15$ and $|l| > 30$ to reduce disk biased results without removing Sagittarius.

6.2 Resultant Maps

Here we present the results for two runs. Firstly, we use l, b, Energy and Lz from the dataset defined in subsection 3.5 and secondly for a baseline the use of the l, b, $\Delta\mu_\alpha$ and $\Delta\mu_\delta$ from the full lower PM dataset.

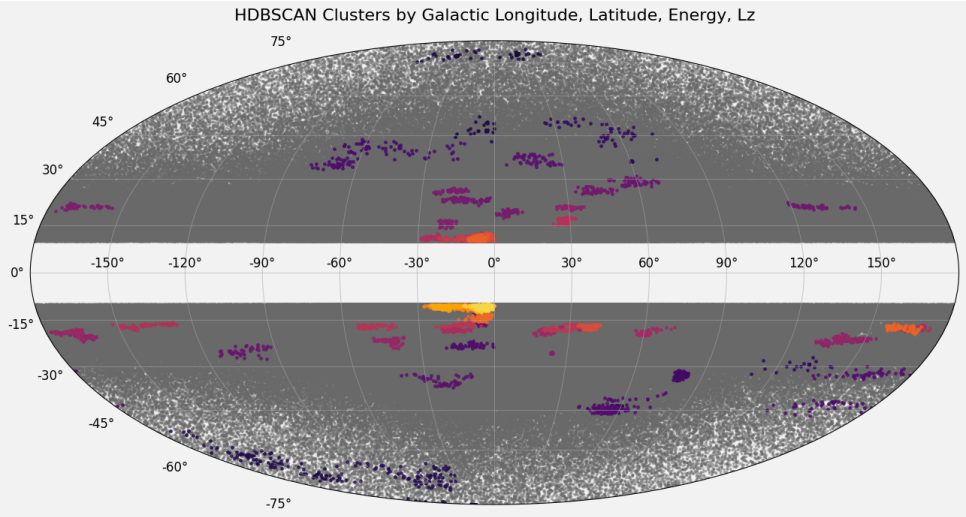


Figure 13: HDBSCAN using Integral of Motion (E/Lz)

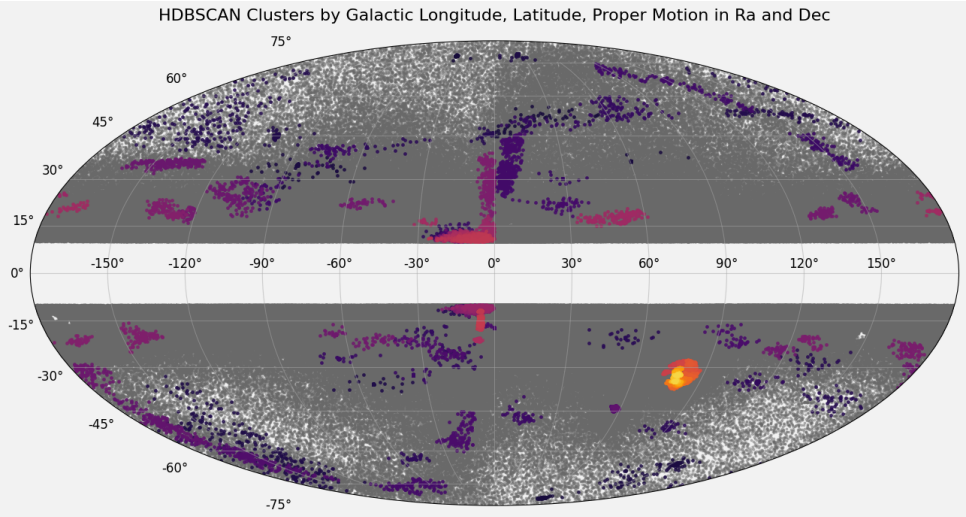


Figure 14: HDBSCAN using Proper Motion

6.3 Discussion of Results

The maps are compared against the `galstreams` library introduced in Mateu [12] (Figure 16). While both reveal structures compatible with stellar streams, the proper motion–based clustering appears to perform better which is unexpected given earlier discussion. This is attributed to two factors: the integral-of-motion dataset being substantially smaller due to the radial velocity requirement and the dependence on distance estimates in computing orbital parameters. As discussed in subsection 3.6, their systematic underestimation thus likely produces unreliable parameters and limits performance in energy–L_z space. Below we provide a table of `potential` identified streams through simple visual comparison however due to the high false positive rate these identifications should be treated with caution.

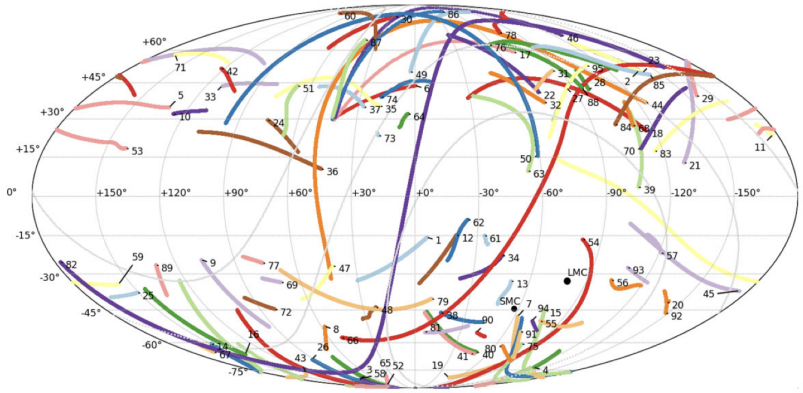


Figure 15: Galstream projections (Mateu [12])

Table 3: `Potential` Identifications (by visual inspection)

Energy-Lz Clusters	**Proper Motion Clusters**
6 = Acheron-G09	5 = ACS-R21
42 = Gaia-8-I21	20 = C-5-I24
76 = NGC6397-I21	26 = Cetus-Y13
97 = New-3-I24	28 = Corvus-M18
126 = Sagittarius-A20	65 = Monoceros-R21
135 = TucanIII-S19	103 = New-8-I24
	126 = Sagittarius-A20
	128 = Scimitar-G17

7 Metallicities Investigation

Finally, we compare our identified structures metallicities to those reported in Harris [9], using data from the APOGEE DR17 catalog [1]. For each cluster identified in section 5, we extract the GAIA source IDs of all stars that fall within the hypercube's position-motion space. These IDs were used to query APOGEE DR17 for available stellar abundance measurements.

For each cluster, we use the matched data to create global statisitcs, specifically for [Fe/H], [M/H], and [α/M]. For globular clusters, the [Fe/H] can be cross-referenced with those listed in Harris [9]. Below, result are shown for clusters with at least ten APOGEE-matched stars and an available Harris comparison, however all results are provided in the repository.

Notably, all calculated values for [Fe/H] (except 2 - NGC 3201 and NGC 5024) fall within 1σ of Harris' catalog values. This further supports the developed algorithm's ability to isolate globular clusters stars in 4D space. However, these values are systematically overestimated suggesting there is still contamination from younger, more metal-rich stars, likely from the bulges foreground. It is suggested the XD pipeline provided in Notebook 6.3 may be able to further improve this.

Overall, the halo clusters identified are shown to have low iron abundances ([Fe/H]), consistent with their origin within a poorly enriched interstellar medium in the early stages of the Milky Way. Their higher [α/M] ratios further support this by showing that despite lower overall metallically, they have a high relative enrichment of alpha elements (e.g., O, Mg, Si, Ca, Ti). This suggests

Table 4: Higher PM Metallicities (APOGEE DR17 vs. Harris)

Name	No. Stars	[Fe/H]	[M/H]	[α/M]	Harris [Fe/H]
NGC 5904	23	-1.198 ± 0.096	-1.216 ± 0.108	0.197 ± 0.076	-1.29
NGC 6752	10	-1.505 ± 0.107	-1.483 ± 0.124	0.231 ± 0.068	-1.54
NGC 6205	10	-1.517 ± 0.084	-1.509 ± 0.143	0.171 ± 0.090	-1.53
NGC 5272	25	-1.416 ± 0.079	-1.427 ± 0.091	0.163 ± 0.064	-1.50

Table 5: Lower PM Metallicities (APOGEE DR17 vs. Harris)

Name	No. Stars	[Fe/H]	[M/H]	[α/M]	Harris [Fe/H]
NGC 5272	90	-1.415 ± 0.082	-1.433 ± 0.092	0.165 ± 0.067	-1.50
NGC 6205	30	-1.467 ± 0.121	-1.442 ± 0.130	0.161 ± 0.092	-1.53
NGC 7078	84	-2.290 ± 0.096	-2.213 ± 0.143	0.198 ± 0.163	-2.37
NGC 1904	36	-1.518 ± 0.134	-1.514 ± 0.159	0.122 ± 0.074	-1.60
NGC 1851	38	-1.142 ± 0.196	-1.160 ± 0.224	0.194 ± 0.150	-1.18
NGC 2808	76	-1.072 ± 0.074	-1.072 ± 0.090	0.134 ± 0.076	-1.14
NGC 4590	28	-2.243 ± 0.084	-2.181 ± 0.120	0.245 ± 0.108	-2.23
NGC 5024	22	-1.896 ± 0.088	-1.887 ± 0.099	0.228 ± 0.070	-2.10
NGC 5053	13	-2.195 ± 0.121	-2.163 ± 0.139	0.230 ± 0.100	-2.27

they formed in a period dominated by Type II (core-collapse) supernovae before the onset of iron-rich Type Ia supernovae. This is characteristic of the early universe where the lack of heavy elements, led to inefficient gas cooling and higher Jeans masses for collapsing clouds. Therefore, the universe was dominated by the more massive stars, with short lifetimes and extreme deaths (Type II supernovae).

From the above discussion, we expect a negative correlation between [Fe/H] and [α/M]. This is verified using the data from Table 5, in which we find a Pearson correlation of -0.731 and a p-value of 0.0251, showing strong support for our results and analysis. Overall, those with lower metallicities (eg. NGC 7078) are likely older.

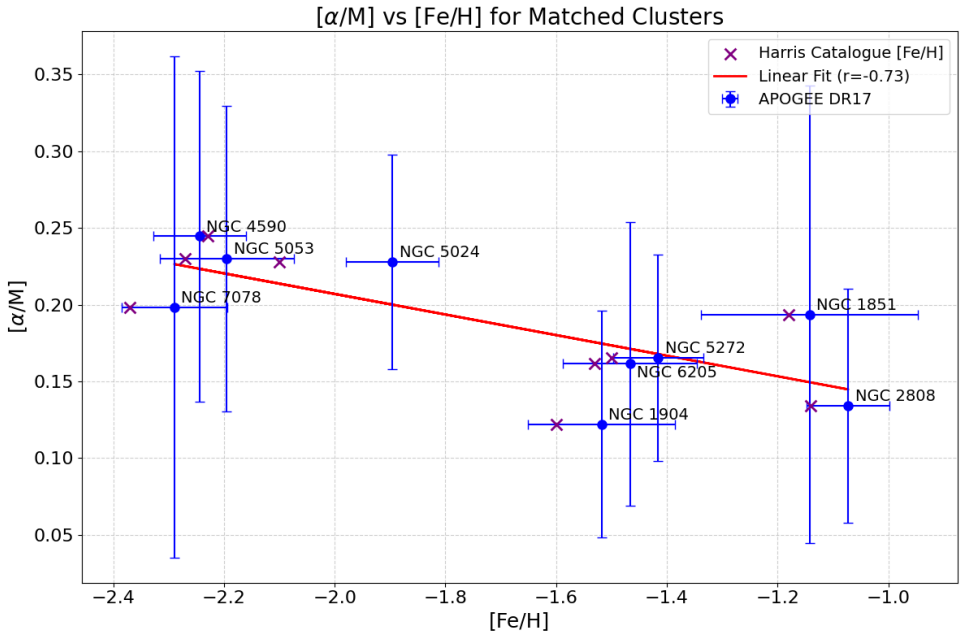


Figure 16: [Fe/H] and [α/M] relation (Table 5)

8 Conclusion

Through effective cuts for halo RGB stars, this project visualised local overdensities and stellar streams through all-sky maps. We then presented two clustering approaches, a novel 4D binning method and HDBSCAN, both of which outperformed visual inspection and identified 44 confirmed

substructures and up to 13 potential streams. The success of the isolation of cluster stars was further validated through strong agreement with Harris [9]. While this project was limited by the use of Gaia astrometry alone, future work aims to integrate complementary spectroscopic surveys, such as APOGEE and GALAH, to enable full chemo-dynamical analyses, as in Myeong et al. [13].

9 Appendices

Table 6: Detected Substructure Parameter Ranges from Higher PM Dataset

Name	l **Range [deg]**	b **Range [deg]**	μ_α **Range [mas/yr]**	μ_δ **Range [mas/yr]**
	True l	True b	True μ_α	True μ_δ
NGC 5904	[3.00, 5.00]	[46.27, 47.27]	[3.60, 4.25]	[-10.05, -9.40]
	3.86	46.80	4.08	-9.85
NGC 7099	[27.00, 28.00]	[-47.73, -46.73]	[-0.95, -0.30]	[-7.45, -6.80]
	27.18	-46.84	-0.69	-7.27
NGC 6981	[35.00, 36.00]	[-32.73, -31.73]	[-1.60, -0.95]	[-3.55, -2.90]
	35.16	-32.68	-1.23	-3.29
NGC 5272	[41.00, 43.00]	[78.27, 79.27]	[-0.30, 0.35]	[-2.90, -2.25]
	42.22	78.71	-0.14	-2.65
NGC 5466	[42.00, 43.00]	[73.27, 74.27]	[-5.50, -4.85]	[-0.95, -0.30]
	42.15	73.59	-5.41	-0.81
NGC 7089	[53.00, 54.00]	[-33.73, -35.73]	[2.95, 3.60]	[-2.25, -1.60]
	53.37	-33.77	3.52	-2.14
NGC 6205	[58.00, 60.00]	[40.27, 41.27]	[-3.55, -2.90]	[-2.90, -2.25]
	59.01	40.91	-3.16	-2.59
NGC 7078	[64.00, 66.00]	[-27.73, -26.73]	[-0.95, -0.30]	[-4.20, -3.55]
	65.01	-27.31	-0.64	-3.76
NGC 6341	[68.00, 69.00]	[34.27, 35.27]	[-5.50, -4.85]	[-0.95, -0.30]
	68.34	34.86	-4.93	-0.54
NGC 1904	[227.00, 228.00]	[-29.73, -28.73]	[2.30, 2.95]	[-1.60, -0.95]
	227.23	-29.35	2.47	-1.57
NGC 1851	[244.00, 245.00]	[-35.73, -34.73]	[1.65, 2.30]	[-0.95, -0.30]
	244.51	-35.04	2.12	-0.59
NGC 1261	[270.00, 271.00]	[-52.73, -51.73]	[1.00, 1.65]	[-2.25, -1.60]
	270.54	-52.12	1.63	-2.04
LMC	[271.00, 289.00]	[-39.73, -24.73]	[1.00, 2.30]	[-0.95, 1.65]
	280.47	-32.89	N/A	N/A
SMC	[296.00, 307.00]	[-48.73, -39.73]	[-0.30, 1.65]	[-1.60, -0.95]
	302.80	-44.30	N/A	N/A
NGC 4590	[299.00, 300.00]	[35.27, 36.27]	[-2.90, -2.25]	[1.65, 2.30]
	299.63	36.05	-2.75	1.76
NGC 362	[301.00, 302.00]	[-46.73, -45.73]	[6.20, 6.85]	[-2.90, -2.25]
	301.53	-46.25	6.73	-2.54
NGC 104	[305.00, 307.00]	[-45.73, -43.73]	[4.90, 5.55]	[-2.90, -2.25]
	305.89	-44.89	5.24	-2.52
NGC 5024	[332.00, 334.00]	[79.27, 80.27]	[-0.30, 0.35]	[-1.60, -0.95]
	332.96	79.76	-0.15	-1.36
NGC 6752	[336.00, 337.00]	[-25.73, -24.73]	[-3.55, -2.90]	[-4.20, -3.55]
	336.49	-25.63	-3.17	-4.04

Table 7: Detected Substructure Parameter Ranges from Lower PM Dataset

Name	l **Range [deg]**	b **Range [deg]**	μ_α **Range [mas/yr]**	μ_δ **Range [mas/yr]**
	True l	True b	True μ_α	True μ_δ
NGC 6723	[0.00, 1.00]	[-17.87, -16.87]	[0.55, 1.20]	[-2.70, -2.05]
	0.07	-17.30	1.03	-2.44
Sagittarius	[4.00, 6.00]	[-14.87, -12.87]	[-2.70, -2.05]	[-1.40, -0.75]
	5.57	-14.17	N/A	N/A
Sag Stream	[3.00, 13.00]	[-32.87, -15.87]	[-3.35, -2.05]	[-2.05, -1.40]
	N/A	N/A	N/A	N/A
NGC 6864	[20.00, 21.00]	[-25.87, -24.87]	[-0.75, -0.10]	[-3.35, -2.70]
	20.30	-25.75	-0.56	-2.80
NGC 6426	[28.00, 29.00]	[16.13, 17.13]	[-2.05, -1.40]	[-3.35, -2.70]
	28.09	16.23	-1.86	-2.99
NGC 6981	[35.00, 36.00]	[-32.87, -31.87]	[-1.40, -0.75]	[-4.00, -2.70]
	35.16	-32.68	-1.23	-3.29
NGC 5272	[41.00, 43.00]	[78.13, 79.13]	[-0.75, -0.10]	[-2.70, -2.05]
	42.22	78.71	-0.14	-2.65
NGC 7492	[53.00, 54.00]	[-63.87, -62.87]	[0.55, 1.20]	[-2.70, -2.05]
	53.39	-63.48	0.80	-2.27
NGC 7089	[53.00, 54.00]	[-35.87, -34.87]	[3.15, 3.80]	[-2.70, -2.05]
	53.37	-35.77	3.52	-2.14
NGC 6205	[58.00, 60.00]	[40.13, 41.13]	[-3.35, -2.70]	[-2.70, -2.05]
	59.01	40.91	-3.16	-2.59
NGC 7078	[64.00, 66.00]	[-27.87, -26.87]	[-0.75, -0.10]	[-4.00, -3.35]
	65.01	-27.31	-0.64	-3.76
NGC 6229	[73.00, 74.00]	[40.13, 41.13]	[-1.40, -0.75]	[-0.75, -0.10]
	73.64	40.31	-1.19	-0.44
Draco	[86.00, 87.00]	[34.13, 35.13]	[-0.10, 0.55]	[-0.75, -0.10]
	86.37	34.71	N/A	N/A
Ursa Minor	[104.00, 105.00]	[44.13, 45.13]	[-0.75, 0.55]	[-0.10, 0.55]
	104.98	44.81	N/A	N/A
NGC 2419	[180.00, 181.00]	[25.13, 26.13]	[-0.10, 0.55]	[-0.75, -0.10]
	180.37	25.24	-0.01	-0.56
NGC 1904	[227.00, 228.00]	[-29.87, -28.87]	[1.85, 3.15]	[-2.05, -1.40]
	227.23	-29.35	2.47	-1.57
Fornax	[236.00, 239.00]	[-66.87, -64.87]	[-0.10, 0.55]	[-0.75, -0.10]
	237.24	-65.67	N/A	N/A
NGC 1851	[244.00, 245.00]	[-35.87, -34.87]	[1.85, 2.50]	[-0.75, -0.10]
	244.51	-35.04	2.12	-0.59
NGC 2298	[245.00, 246.00]	[-16.87, -15.87]	[3.15, 3.80]	[-2.70, -2.05]
	245.63	-16.01	3.32	-2.19
NGC 4147	[252.00, 253.00]	[77.13, 78.13]	[-2.05, -1.40]	[-2.70, -2.05]
	252.85	77.19	-1.71	-2.11
Carina	[260.00, 261.00]	[-22.87, -21.87]	[0.55, 1.20]	[-0.10, 0.55]
	260.11	-22.22	N/A	N/A
LMC	[267.00, 292.00]	[-41.87, -21.87]	[1.20, 2.50]	[-0.75, 1.85]
	280.47	-32.89	N/A	N/A
NGC 1261	[270.00, 271.00]	[-52.87, -51.87]	[1.20, 1.85]	[-2.70, -1.40]
	270.54	-52.12	1.63	-2.04
NGC 2808	[282.00, 283.00]	[-11.87, -10.87]	[0.55, 1.20]	[-0.10, 0.55]
	282.19	-11.25	1.00	0.27
Sculptor	[286.00, 289.00]	[-83.87, -82.87]	[-0.10, 0.55]	[-0.75, -0.10]
	287.70	-83.15	N/A	N/A
SMC	[295.00, 309.00]	[-50.87, -37.87]	[-0.10, 1.85]	[-1.40, -0.75]
	302.80	-44.30	N/A	N/A
NGC 4590	[299.00, 300.00]	[35.13, 36.13]	[-3.35, -2.05]	[1.20, 1.85]
	299.63	36.05	-2.75	1.76
Rup 106	[300.00, 301.00]	[11.13, 12.13]	[-1.40, -0.75]	[-0.10, 0.55]
	300.89	11.67	-1.26	0.40
IC 4499	[307.00, 308.00]	[-20.87, -19.87]	[-0.10, 0.55]	[-0.75, -0.10]

Name	l **Range [deg]**	b **Range [deg]**	μ_α **Range [mas/yr]**	μ_δ **Range [mas/yr]**
	True l	True b	True μ_α	True μ_δ
	307.35	-20.47	0.49	-0.49
NGC 5286	[311.00, 312.00]	[10.13, 11.13]	[-0.10, 0.55]	[-0.75, -0.10]
	311.61	10.57	0.21	-0.11
NGC 6101	[317.00, 318.00]	[-15.87, -14.87]	[1.20, 1.85]	[-0.75, -0.10]
	317.75	-15.82	1.76	-0.22
NGC 5694	[331.00, 332.00]	[30.13, 31.13]	[-0.75, -0.10]	[-1.40, -0.75]
	331.06	30.36	-0.49	-1.07
NGC 5824	[332.00, 333.00]	[21.13, 22.13]	[-1.40, -0.75]	[-2.70, -2.05]
	332.55	22.07	-1.17	-2.23
NGC 5024	[332.00, 334.00]	[79.13, 80.13]	[-0.75, -0.10]	[-1.40, -0.75]
	332.96	79.76	-0.15	-1.36
NGC 5053	[335.00, 336.00]	[78.13, 79.13]	[-0.75, -0.10]	[-1.40, -0.75]
	335.70	78.95	-0.37	-1.25
NGC 5634	[342.00, 343.00]	[49.13, 50.13]	[-2.05, -1.40]	[-2.05, -1.40]
	342.21	49.26	-1.72	-1.51
Unknown 1	[133.00, 134.00]	[-31.87, -30.87]	[-0.10, 0.55]	[-0.10, 0.55]
	N/A	N/A	N/A	N/A
Unknown 2	[177.00, 178.00]	[-19.87, -18.87]	[-0.10, 0.55]	[-1.40, -0.75]
	N/A	N/A	N/A	N/A
Unknown 3	[177.00, 178.00]	[20.13, 21.13]	[-0.10, 0.55]	[-1.40, -0.75]
	N/A	N/A	N/A	N/A
Unknown 4	[190.00, 191.00]	[-19.87, -18.87]	[-0.10, 0.55]	[-1.40, -0.75]
	N/A	N/A	N/A	N/A
Unknown 5	[194.00, 195.00]	[14.13, 15.13]	[-0.10, 0.55]	[-1.40, -0.75]
	N/A	N/A	N/A	N/A
Unknown 6	[197.00, 198.00]	[12.13, 13.13]	[-0.75, -0.10]	[-1.40, -0.75]
	N/A	N/A	N/A	N/A
Unknown 7	[204.00, 205.00]	[-17.87, -16.87]	[-0.10, 0.55]	[-0.75, -0.10]
	N/A	N/A	N/A	N/A
Unknown 8	[212.00, 213.00]	[-16.87, -15.87]	[-0.10, 0.55]	[-0.75, -0.10]
	N/A	N/A	N/A	N/A
Unknown 9	[214.00, 216.00]	[11.13, 13.13]	[-0.75, -0.10]	[-0.75, -0.10]
	N/A	N/A	N/A	N/A
Unknown 10	[296.00, 297.00]	[10.13, 11.13]	[-4.00, -3.35]	[-0.10, 0.55]
	N/A	N/A	N/A	N/A
Unknown 11	[297.00, 298.00]	[-30.87, -29.87]	[1.85, 2.50]	[-0.10, 0.55]
	N/A	N/A	N/A	N/A
Unknown 12	[304.00, 305.00]	[-40.87, -39.87]	[0.55, 1.20]	[-2.05, -1.40]
	N/A	N/A	N/A	N/A

10 Declaration of Use of Autogeneration Tools

This report made use of Large Language Models (LLMs), to assist in the development of the project. These tools have been employed for assisting:

- Formatting plots to enhance presentation quality.
- Generating docstrings for the repository's documentation.
- Performing iterative changes to already defined code.
- Debugging code and identifying issues in implementation.
- Latex formatting for the report.
- Identifying spelling and punctuation inconsistencies within the report.
- Suggesting more concise phrasing to reduce the word count.

References

[1] Abdurro'uf et al. "The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data". In: *The Astrophysical Journal Supplement Series* 259.2 (Mar. 2022), p. 35. ISSN: 1538-4365. DOI: 10.3847/1538-4365/ac4414. URL: http://dx.doi.org/10.3847/1538-4365/ac4414.

[2] 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". In: *The Astronomical Journal* 161.3 (Feb. 2021), p. 147. ISSN: 1538-3881. DOI: 10.3847/1538-3881/abd806. URL: http://dx.doi.org/10.3847/1538-3881/abd806.

[3] V. Belokurov et al. "The Field of Streams: Sagittarius and Its Siblings". In: *The Astrophysical Journal* 642.2 (Apr. 2006), pp. L137–L140. ISSN: 1538-4357. DOI: 10.1086/504797. URL: http://dx.doi.org/10.1086/504797.

[4] E. Bica et al. "A census of new globular clusters in the Galactic bulge". In: *Astronomy amp; Astrophysics* 687 (July 2024), A201. ISSN: 1432-0746. DOI: 10.1051/0004-6361/202346377. URL: http://dx.doi.org/10.1051/0004-6361/202346377.

[5] Jo Bovy, David W. Hogg, and Sam T. Roweis. "Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations". In: *The Annals of Applied Statistics* 5.2B (June 2011). ISSN: 1932-6157. DOI: 10.1214/10-aoas439. URL: http://dx.doi.org/10.1214/10-AOAS439.

[6] Luca Casagrande et al. "The GALAH survey: effective temperature calibration from the InfraRed Flux Method in the Gaia system". In: *Monthly Notices of the Royal Astronomical Society* 507.2 (Aug. 2021), pp. 2684–2696. ISSN: 1365-2966. DOI: 10.1093/mnras/stab2304. URL: http://dx.doi.org/10.1093/mnras/stab2304.

[7] Alis J. Deason and Vasily Belokurov. "Galactic Archaeology with Gaia". In: *New Astronomy Reviews* 99 (Dec. 2024), p. 101706. ISSN: 1387-6473. DOI: 10.1016/j.newar.2024.101706. URL: http://dx.doi.org/10.1016/j.newar.2024.101706.

[8] Mark A. Fardal et al. "Mapping Gaia Parallax Systematic Errors over the Sky with Faint Milky Way Stars". In: *The Astronomical Journal* 161.2 (Jan. 2021), p. 58. ISSN: 1538-3881. DOI: 10.3847/1538-3881/abcccf. URL: http://dx.doi.org/10.3847/1538-3881/abcccf.

[9] William E. Harris. *A New Catalog of Globular Clusters in the Milky Way*. 2010. arXiv: 1012.3224 [astro-ph.GA]. URL: https://arxiv.org/abs/1012.3224.

[10] Helmi, Amina et al. "A box full of chocolates: The rich structure of the nearby stellar halo revealed by Gaia and RAVE". In: *AA* 598 (2017), A58. DOI: 10.1051/0004-6361/201629990. URL: https://doi.org/10.1051/0004-6361/201629990.

[11] Claudia Malzer and Marcus Baum. "A Hybrid Approach To Hierarchical Density-based Cluster Selection". In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, Sept. 2020, pp. 223–228. DOI: 10.1109/mfi49285.2020.9235263. URL: http://dx.doi.org/10.1109/MFI49285.2020.9235263.

[12] Cecilia Mateu. "galstreams: A library of Milky Way stellar stream footprints and tracks". In: *Monthly Notices of the Royal Astronomical Society* 520.4 (Jan. 2023), pp. 5225–5258. ISSN: 1365-2966. DOI: 10.1093/mnras/stad321. URL: http://dx.doi.org/10.1093/mnras/stad321.

[13] G. C. Myeong et al. "Milky Way's Eccentric Constituents with Gaia, APOGEE, and GALAH". In: *Astrophys. J.* 938.1 (2022), p. 21. DOI: 10.3847/1538-4357/ac8d68. arXiv: 2206.07744 [astro-ph.GA].

[14] E. O. Nadler et al. "Milky Way Satellite Census. II. Galaxy–Halo Connection Constraints Including the Impact of the Large Magellanic Cloud". In: *The Astrophysical Journal* 893.1 (Apr. 2020), p. 48. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ab846a. URL: http://dx.doi.org/10.3847/1538-4357/ab846a.

[15] Ciaran A. J. O'Hare et al. "Velocity substructure from Gaia and direct searches for dark matter". In: *Phys. Rev. D* 101.2 (2020), p. 023006. DOI: 10.1103/PhysRevD.101.023006. arXiv: 1909.04684 [astro-ph.GA].

[16] T. Prusti et al. "TheGaiamission". In: *Astronomy amp; Astrophysics* 595 (Nov. 2016), A1. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201629272. URL: http://dx.doi.org/10.1051/0004-6361/201629272.

[17] David J. Schlegel, Douglas P. Finkbeiner, and Marc Davis. "Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds". In: *The Astrophysical Journal* 500.2 (June 1998), p. 525. DOI: 10.1086/305772. URL: https://dx.doi.org/10.1086/305772.

[18] A. Ulla et al. *Gaia DR3 documentation Chapter 11: Astrophysical parameters*. Gaia DR3 documentation, European Space Agency; Gaia Data Processing and Analysis Consortium. June 2022.

[19] A. Vallenari et al. "GaiaData Release 3: Summary of the content and survey properties". In: *Astronomy amp; Astrophysics* 674 (June 2023), A1. ISSN: 1432-0746. DOI: 10.1051/0004-6361/202243940. URL: http://dx.doi.org/10.1051/0004-6361/202243940.

[20] Eugene Vasiliev. "Proper motions and dynamics of the Milky Way globular cluster system fromGaiaDR2". In: *Monthly Notices of the Royal Astronomical Society* 484.2 (Jan. 2019), pp. 2832–2850. ISSN: 1365-2966. DOI: 10.1093/mnras/stz171. URL: http://dx.doi.org/10.1093/mnras/stz171.