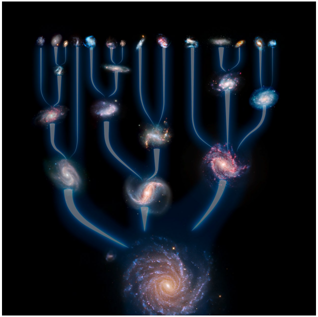
UNIVERSITY OF CAMBRIDGE

Gaia's Insight to the Milky Way's Hierarchical Assembly

Jacob Tutt[1]

[1] Department of Physics, University of Cambridge, UK

Galactic Archaeology's Motivation



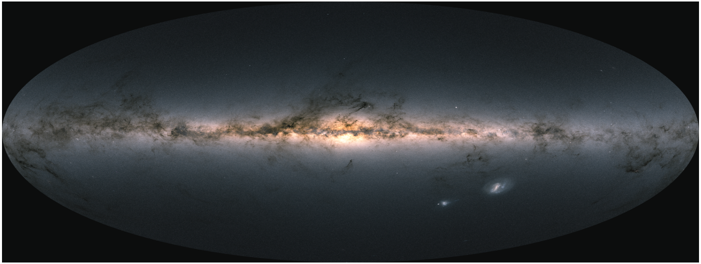
Credit: ESO/L. Calçada

- Uncovering hierarchical galaxy formation.
- Complements higher redshift galaxy formation surveys.
- Probe the ΛCDM model and dark matter distribution.

Merger Type	Number	Mass Ratio
Minor Mergers	$\sim$30	1:3 – 1:100
Major Mergers	$\sim$3	$>$1:3

N-Body Simulations from Fakhouri et al. (2010)

The ‘Big Data Era’ with Gaia



Credit: ESA/Gaia/DPAC, A. Moitinho.

- Long hampered by incomplete and insufficient data
- Gaia EDR3: Full astrometry for $> 1.4 \times 10^9$ sources (10 TB)
- Complementary surveys: APOGEE, GALAH, WEAVE, 4MOST

Automated Detection of Halo Substructures

Goal:

- Develop methods for detecting stellar substructures from the Galactic halo using the Gaia's Data

- **Approach:**
  - Optimising stellar halo (RGB) data selection
  - Construct all-sky maps for visual substructure identification
  - Developing automated detection algorithms for:
    - Globular clusters
    - Tidal streams
  - Cross-match structures with APOGEE spectroscopic data

Initial Data Acquisition

Criterion	Cut
Astrometric quality	RUWE < 1.4
Distance cut	Parallax $\varpi < 0.1$ mas
Galactic latitude	$\|b\| > 10^\circ$
Magnitude range	G-band $10 < G < 20.5$
Proper motion cut	$\sqrt{\mu_\alpha^2 + \mu_\delta^2} < 4$ or 12 mas/yr

Optimisation Criteria:

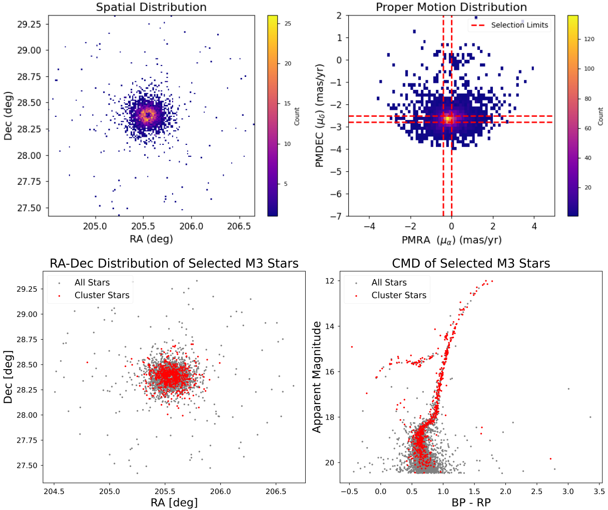
- **Red Giant Branch (RGB) stars**

Instead:

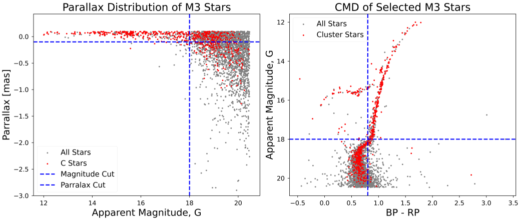
- 'Isolation purity' of test clusters **M3** and **NGC 1851**.

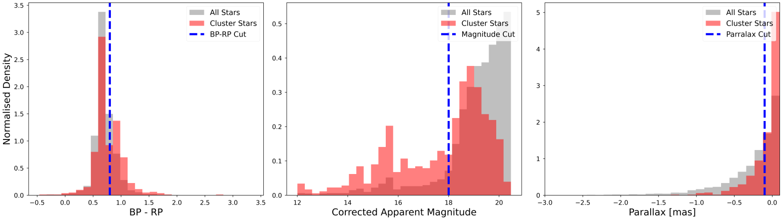
Extinction corrected using:
Dust Maps: Schlegel, Finkbeiner & Davis (1998) Calibration Coefficients: Casagrande et al. (2021)

M3 Isolation of Cluster Stars



M3 Isolation of Cluster Stars





Resultant (Optimised) Cuts

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

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%

Additional Investigations

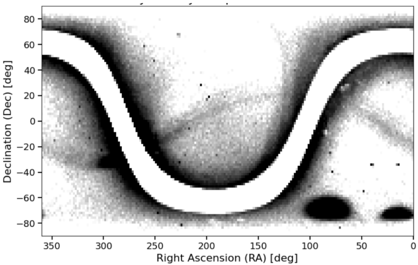
- Surface Gravity
- Effective Temperature
- Absolute Magnitude (exploiting Bailer-Jones dist)

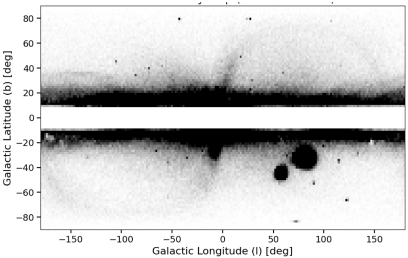
All-Sky Density Maps

Visualisation Techniques:

- Percentile Based Scaling
- Logarithmic Scaling
- Minimum Count Threshold
- False Colour Composites
  - Proper Motion
  - Apparent Magnitude

All-Sky Density Maps





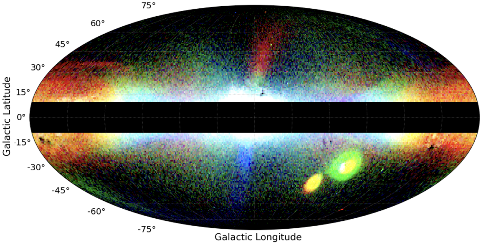
Visual Benchmark:

- Over-densities: 27 Over densities (+ LMC and SMC)
- Tidal Streams: Clear Sagittarius

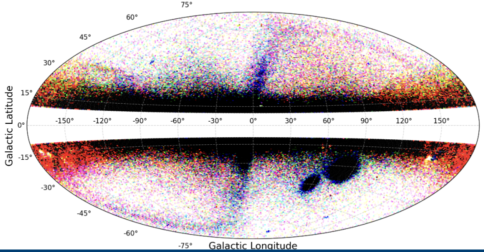
All-Sky Density Maps

Proper

Motion



Absolute Magnitude



Automated Over-Density Detection

Credit: ESA/Hubble & NASA, R. Cohen

Globular Clusters/ Dwarf Galaxies

- Localised
- Gravitationally Bound
- Heliocentric Positions

Traditional Methods

- Circular/Elliptical Projection
- Locally Distributed Perturbations
- Gaussian Mixture Models
  - Extreme Deconvolution (XD)

4D Clustering Algorithm

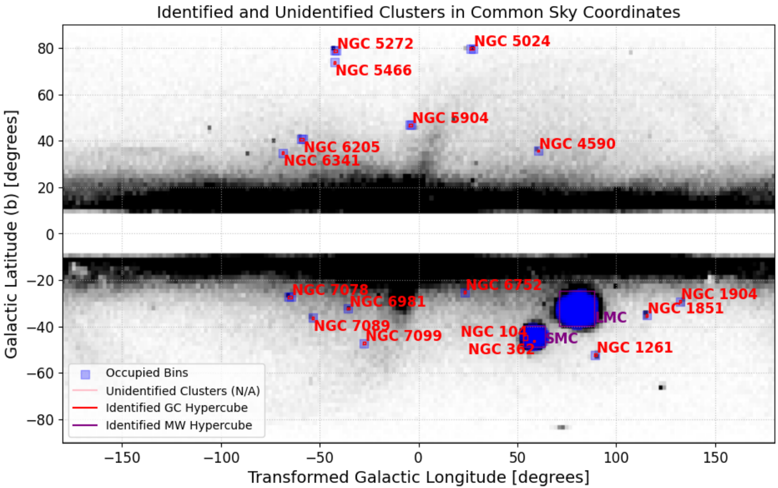
Algorithm:

- **Binning:** Data discretised into:
  - **4D:** Position and Proper motion
  - **2D:** Position
- **Relative Filtering**
- **Absolute Filtering**
- **Agglomerative Clustering**
  - 4D Adjacency Kernel
- **Nested Clusters Removal**
- **HyperCube Definition**

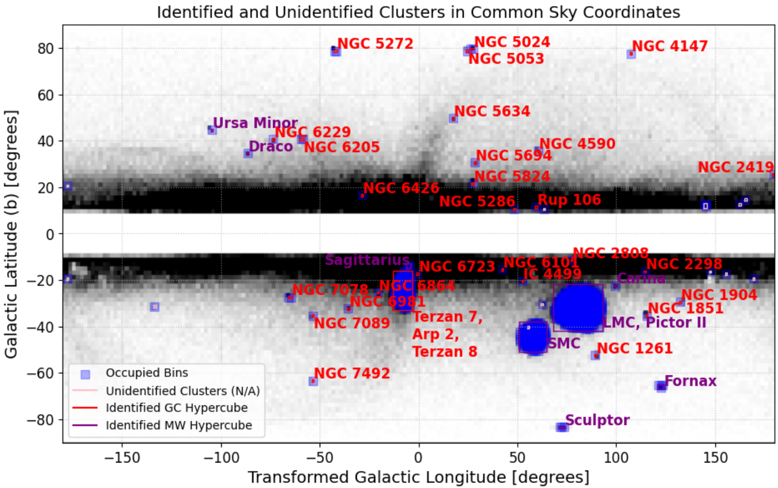
Parameters:

- Spatial bin width: $\Delta l = \Delta b = 1^\circ$
- Motion bin width: $\Delta\mu_\alpha = \Delta\mu_\delta = 0.65$ mas/yr
- Relative threshold: $f_{\rm thresh} = 0.28$
- Absolute count: $N_{\rm min} = 20$
- Connectivity: $C = 1$

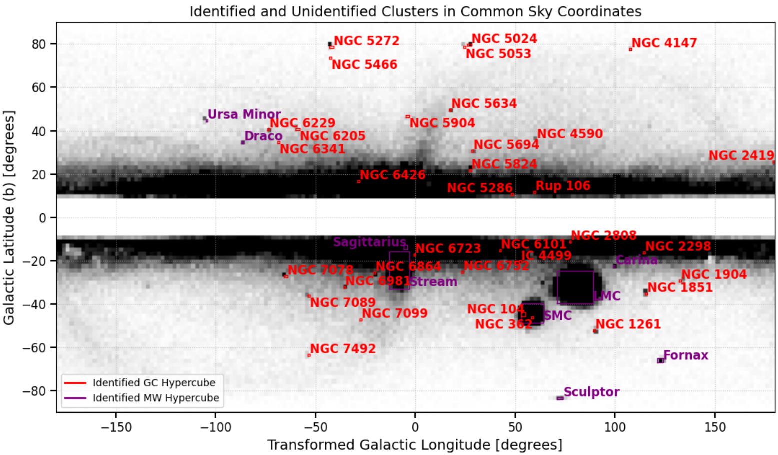
Higher PM Results



Lower PM Results



Overall Results



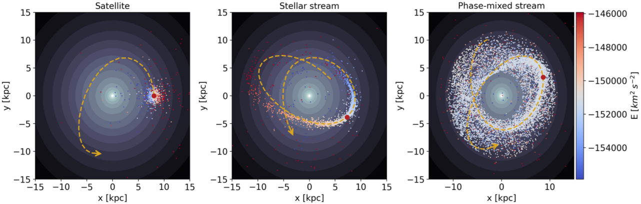
Stellar Streams

Properties

- Formed by tidal disruption
- Dispersed in 6D phase space
- Grouped in conserved quantities

Integral of Motion Space

- 301,642 stars with 6D data
- `Galpy`: L_z and E
- Requires:
  - Radial velocity
  - Distance



Credit: H.C. Woudenberg, 2023

HDBSCAN

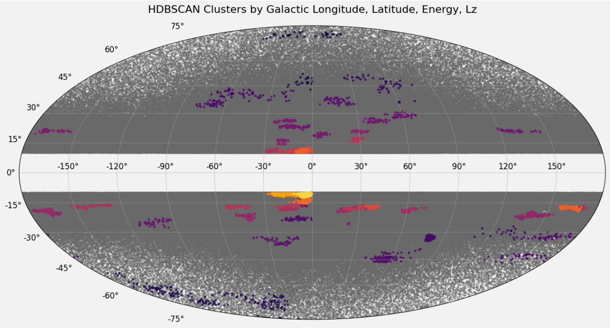
Properties:

- No prior assumptions
- Handle varying densities
- Hierarchical nature
- Robustness to noise

HDBSCAN Parameters

Parameter	Function	Value
`min_cluster_size`	Minimum cluster size	40
`min_samples`	Core point neighbourhood	15

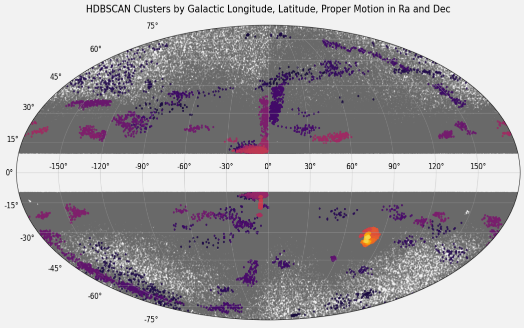
Integrals of Motion Results



$E - L_z$ **Clusters**
6 = Acheron-G09
42 = Gaia-8-I21
76 = NGC6397-I21
97 = New-3-I24
126 = Sagittarius-A20
135 = Tucana-III-S19

Credit: Mateu 2023, Galstreams

Heliocentric Results

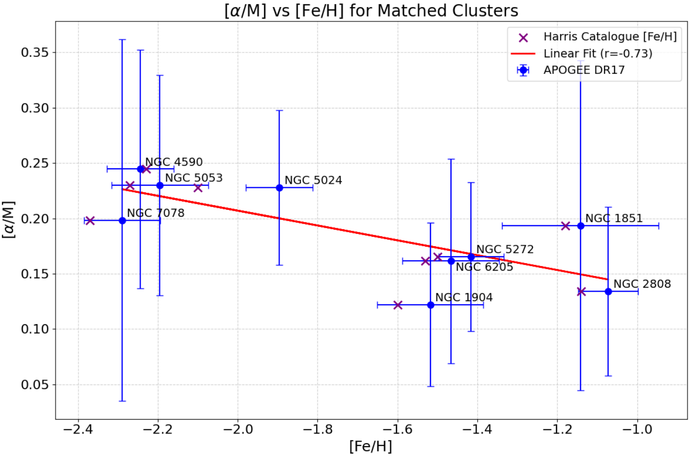


Proper Motion Clusters

5 = ACS-R21
20 = C-5-I24
26 = Cetus-Y13
28 = Corvus-M18
65 = Monoceros-R21
103 = New-8-I24
126 = Sagittarius-A20
128 = Scimitar-G17

Credit: Mateu 2023, Galstreams

Metallicities



Credit: Harris (2010)

Conclusions and Future Work

Localised Over-Densities

- Developed a 4D clustering algorithm for compact structures.
- Future: Primary data analysis tool in larger research.
    - GMM (Extreme Deconvolution)to hypercubes.

Stellar Streams

- Integral of motion space for identifying extended structures.
- Accuracy limited by distance and radial velocity uncertainties.
- Future: use GALAH/APOGEE and chemical abundances for chemo-dynamical tagging.

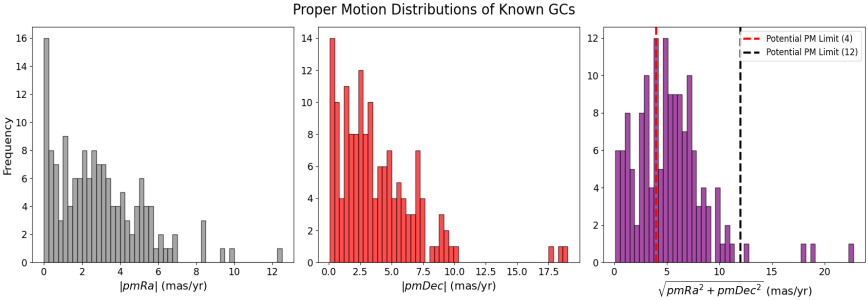
Questions

Thank you for your attention!

Jacob Tutt
Department of Physics, University of Cambridge
jlt67@cam.ac.uk
https://github.com/jacobtutt

Contextualising Data Preprocessing

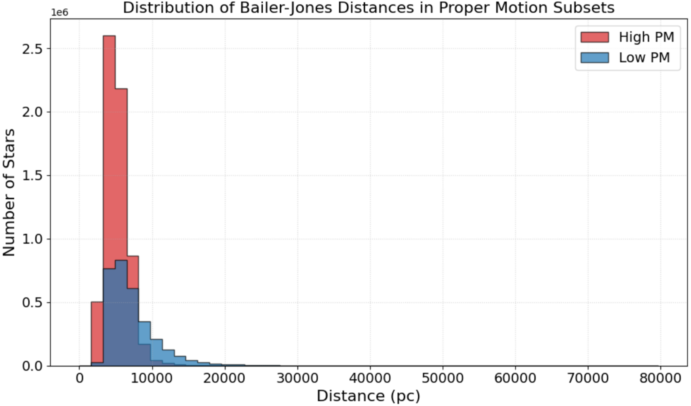
Known Globular Clusters in Proper Motion Space (μ_ℓ, μ_b)



Data sources: Vasileiv (2019)

Questions

Bailer Jones Distances



Questions

HDBSCAN - Step-by-step:

1. **Core Distance Calculation:**
   - For each point a, compute its **core distance**:

   $$\text{core}(a) = \text{distance to the } k^{\text{th}} \text{ nearest neighbor}$$

   - Here, $k = \texttt{min\_samples}$
2. **Mutual Reachability Distance:**
   - For each pair of points a and b, define:

   $$d_{\text{reach}}(a, b) = \max\left(\text{core}(a),\ \text{core}(b),\ d(a, b)\right)$$

   - This penalises sparse/outlier points by inflating their distances (density aware).

(Campello et al., 2013; McInnes et al., 2017)

Questions

HDBSCAN - Step-by-step:

3. **Minimum Spanning Tree (MST):**
   - Build a graph where edges connect all points using d_{reach}
   - Construct a minimum spanning tree (MST) using **single linkage**
4. **Hierarchy Construction:**
   - Progressively remove edges from the MST (starting with longest)
   - This creates a hierarchy of clusters at different density levels (dendrogram)
5. **Cluster Selection:**
   - Evaluate **stability** of clusters (how long they persist across scales)
   - Select most stable clusters and label others as noise or border points

(Campello et al., 2013; McInnes et al., 2017)

Questions

HDBSCAN vs DBSCAN

- **DBSCAN** uses a fixed distance threshold (`eps`) and struggles with clusters of varying densities.
- **HDBSCAN** builds a hierarchy of clusters across all density levels and selects the most **stable** ones, making it more robust to noise and variable-density regions.

Based on McInnes et al. (2017)