Unveiling the Milky Way's Formation History:
Resolving Chemo-Dynamical Substructures in APOGEE and GALAH

Jacob Tutt | MPhil Data Intensive Science | University of Cambridge

June 30, 2025

1 Motivation and Scientific Justification

Galaxy formation is proposed by the Λ-cold dark matter (ΛCDM) model to be a hierarchical process involving a series of merger events over cosmic time [1]. The debris of the Milky Way's accreted progenitors is therefore dispersed throughout the stellar halo, forming a fossil record of these events. As a result, the halo serves as a crucial benchmark against which simulations can be compared, providing valuable constraints for refining current cosmological models [2]. The field of Galactic Archaeology seeks to infer the Galaxy's past states by decoding the chemo-dynamical information hidden within datasets and isolating stellar populations with unique evolutionary origins.

The new generation of astronomical surveys including Gaia [3], APOGEE [4] and GALAH [5] have already driven major advances in our understanding of the accreted and in-situ components that occupy our stellar neighbourhood. One of the most dominant is the 'Gaia–Sausage/Enceladus' (GS/E) [6, 7], a massive ancient merger that deposited substantial debris into the halo. Through gravitational disruption and gas pollution this event also induced the formation of high-eccentricity in-situ populations, such as the 'Splash' [8] and the more recently discovered 'Eos' component [9]. Additionally, these surveys have enabled the isolation of an even older pre-disk population, known as 'Aurora' [10].

Modern advancements in unsupervised machine learning methodologies are being increasingly leveraged by the field due to their ability to identify components both objectively and autonomously. In this work, we use high eccentricity samples from APOGEE and GALAH to replicate the unbiased decomposition of the stellar halo presented by Myeong et al. [9] and assess the reproducibility of their findings. We then extend the analysis by comparing the datasets' respective advantages and providing an new approach to clustering that improves both computational efficiency and stability.

2 Methodology

This work employs *Extreme Deconvolution* (XD) [11], a GMM based algorithm due to its capability of accounting for heteroskedastic uncertainties. By allowing the Expectation-Maximisation process to robustly account for the different observational uncertainties present in APOGEE and GALAH, XD is able to infer the underlying intrinsic structures. It therefore enables a fairer comparison of the resultant distributions traced by each survey. To account for the features which XD lacks in comparison to modern GMM libraries we present a pipeline that enables automated optimisation via random initialisations, model selection using AIC or BIC and 'error-aware' cluster assignment. This pipeline also supports multiple scaling schemes with corresponding initialisation and transformations strategies.

Secondly, this project presents a new approach which aims to decrease computational cost and sensitivity to random initialisations by exploiting UMAP [12] for dimensionality reduction. This pipeline preforms clustering in a lower-dimensional embedding space before mapping the resultant structures back to distributions in the original feature space. Finally, this approach supports an approximate method for deconvolving the clusters' covariance matrices, with the aim of producing results comparable to those obtained by XD.

Overall, these pipelines allow analysis to be robust and unbiased through quantitative model comparison and repeat initialisations.

3 Key Findings

Firstly, by identifying four independent stellar populations from both the APOGEE and GALAH datasets, with near identical distributions to those presented in the original work, we find strong support for the work's claim of being unbiased, objective and reproducible. Additionally, despite slight instabilities in APOGEE's clustering, we are able to robustly recover the trends that led to the paper's key conclusions. Namely, the GS/E's significantly lower [Al/Fe] abundances suggests it is the only accreted component identified. Secondly, the Aurora population's rapid chemical evolution, coupled with its striking similarity to the proposed Hercules population, helps verify the claim an additional merger event is not necessarily required to explain the Hercules 'debris'.

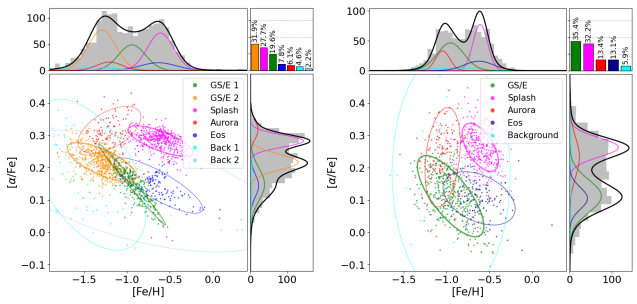

Figure 1. The chemo-dynamical distributions resolved by high dimensinal XD in APOGEE (left) and GALAH (right), with 2 σ uncertainties shown.

From both the increased clustering stability and comparisons of UMAP projections across various hyperparameters (`min_dist` and `n_neighbors`), we conclude that GALAH's higher dimensionality enables greater isolation of the halo's substructures. This observation persists, albeit less strongly when the GALAH dataset is restricted to an equivalent six-dimensional feature space as APOGEE thus suggesting that clustering with GALAH benefits from advantages beyond its dimensionality. This is concluded to be a result of the dataset's stunted metallicity range excluding regions of significant chemical overlap between Aurora and the GS/E debris.

Finally, this work demonstrates that by applying clustering to a two-dimensional embedding of the GALAH data we are able to achieve results comparable to that of the high-dimensional analysis. Although the underlying structures are slightly less well-constrained, with uncertainties on average 29.3% higher across all dimensions and populations, this approach achieves these results in just 0.04% of the runtime (across 100 initialisation of all models with 0 to 10 Gaussian components). Beyond this substantial speed up, the approach also shows greater sensitivity to identifying chemically distinct subpopulations, successfully making distinctions that could not be achieved in the original GALAH analysis. This includes the separation of the GS/E's [α/Fe]–[Fe/H] plateau and knee as well as a potential split in the Splash population that is previously unseen.

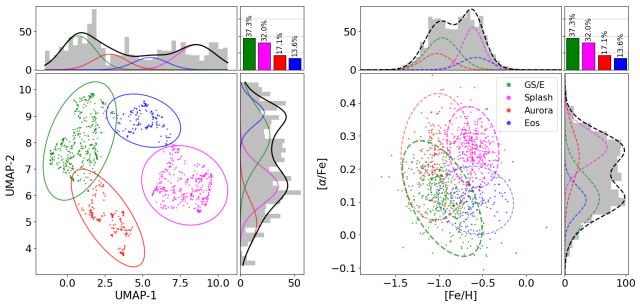

Figure 2. The clusters identified within GALAH using reduced dimensional pipeline showing distributions both in embedding space (left) and feature space (right).

4 Recommended Next Steps

One of the results this paper presents which merits further consideration in future work is the potential identification of two Splash 'sub-populations'. To determine whether a genuine physical distinction exists between them, will likely require comparisons to be made with simulations suites. Additionally, this work suggests the future development of non-linear dimensionality reduction techniques capable of propagating uncertainties into the embedding space as particularly promising.

5 Research Impact

With next-generation surveys such as WEAVE [13] and 4MOST [14] promising even larger and higher-dimensional datasets, this study presents an updated and more scalable methodology for implementing the analysis presented in the original work. A hybrid approach is suggested for future applications, balancing computational efficiency and stability of clustering in reduced-dimensional space with the accuracy and robust consideration of uncertainties offered by Extreme Decovolution. This can be achieved by initialising the high-dimensional XD pipeline with the results of the low-dimensional clustering.

References

[1] Gabriella De Lucia and Amina Helmi. "The Galaxy and its Stellar Halo: Insights on Their Formation from a Hybrid Cosmological Approach". In: *Monthly Notices of the Royal Astronomical Society* 391.1 (Nov. 2008), pp. 14–31. DOI: 10.1111/j.1365-2966.2008.13862.x.

[2] Ken Freeman and Joss Bland-Hawthorn. "The New Galaxy: Signatures of Its Formation". In: *Annual Review of Astronomy and Astrophysics* 40 (2002), pp. 487–537. DOI: 10.1146/annurev.astro.40.060401.093840.

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

[4] 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.

[5] Sven Buder et al. "The GALAH+ survey: Third data release". In: *Monthly Notices of the Royal Astronomical Society* 506.1 (May 2021), pp. 150–201. ISSN: 1365-2966. DOI: 10.1093/mnras/stab1242.

[6] V Belokurov et al. "Co-formation of the disc and the stellar halo". In: *Monthly Notices of the Royal Astronomical Society* 478.1 (June 2018), pp. 611–619. ISSN: 1365-2966. DOI: 10.1093/mnras/sty982.

[7] Amina Helmi et al. "The merger that led to the formation of the Milky Way's inner stellar halo and thick disk". In: *Nature* 563.7729 (Oct. 2018), pp. 85–88. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0625-x.

[8] Vasily Belokurov et al. "The biggest splash". In: 494.3 (May 2020), pp. 3880–3898. DOI: 10.1093/mnras/staa876. arXiv: 1909.04679 [astro-ph.GA].

[9] 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.

[10] Vasily Belokurov and Andrey Kravtsov. "From dawn till disc: Milky Way's turbulent youth revealed by the APOGEE+Gaia data". In: *Monthly Notices of the Royal Astronomical Society* 514.1 (May 2022), pp. 689–714. ISSN: 1365-2966. DOI: 10.1093/mnras/stac1267.

[11] 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.

[12] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML]. URL: https://arxiv.org/abs/1802.03426.

[13] Shoko Jin et al. "The wide-field, multiplexed, spectroscopic facility WEAVE: Survey design, overview, and simulated implementation". In: *Monthly Notices of the Royal Astronomical Society* 530.3 (Mar. 2023), pp. 2688–2730. ISSN: 1365-2966. DOI: 10.1093/mnras/stad557.

[14] Roelof S. de Jong et al. "4MOST: 4-metre multi-object spectroscopic telescope". In: *Ground-based and Airborne Instrumentation for Astronomy IV*. Ed. by Ian S. McLean, Suzanne K. Ramsay, and Hideki Takami. Vol. 8446. SPIE, Oct. 2012, 84460T. DOI: 10.1117/12.926239.

*** word count: 991 words**