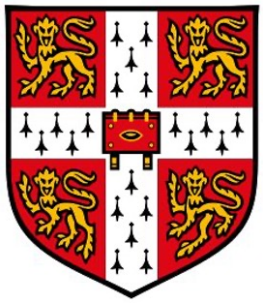


Unveiling the Milky Way's Formation History:

A Comparison of APOGEE and GALAH in Resolving Chemo-Dynamical Substructure

---

Author

Jacob Tutt

Degree

MPhil Data Intensive Science

Departments

Department of Applied Mathematics and Theoretical Physics

Cavendish Laboratory, Department of Physics

Institute of Astronomy

Supervisor

Dr GyuChul Myeong

Word Count

6,979*

June 26th, 2025

*Excluding Figures, Tables and Bibliography

Abstract

We present a comparative analysis of the Milky Way's stellar halo's decomposition using high-eccentricity samples from APOGEE DR17 and GALAH DR3, with the primary aim of reproducing the results and extending the unsupervised presented by Myeong et al. [1]. Applying Extreme Deconvolution to both chemo-dynamical datasets, we resolve three in-situ populations, Splash, Aurora, and Eos, and one accreted component, corresponding to the debris of the GS/E merger. By implementing an improved scaling scheme to increase stability, we recover distributions consistent with the original study, supporting its claims of being unbiased, objective and reproducible. Through UMAP embeddings, we find clear evidence that GALAH's higher dimensionality facilitates greater isolation of independent nucleosynthetic pathways and thus more stable clustering. Surprisingly, even in a comparable six-dimensional feature space, GALAH outperforms APOGEE, potentially indicating underlying advantages in its data, contradicting literature. We not only demonstrate the feasibility of achieving comparable results by clustering directly in embedding space, we show it requires $\sim 0.04\%$ of the computational runtime. This low-dimensional approach also proves more sensitive to resolving substructures within the GALAH dataset, successfully splitting the GS/E merger and revealing a potentially unseen decomposition of Splash. We conclude by proposing a hybrid approach for future surveys such as 4MOST and WEAVE, leveraging both the speed and stability of low-dimensional clustering with accuracy and uncertainty awareness of high-dimensional methods.

Contents

1 Introduction

1.1 The ΛCDM's Hierarchical Assembly

Although only a singular example from a diverse population, the unparalleled detail in which we can observe the Milky Way allows it to serve as a crucial benchmark for refining theories of galaxy formation and evolution [2], a central theme of modern astrophysics. The currently favored Λ-cold dark matter (ΛCDM) model proposes structural formation as a hierarchical process, in which galaxies like our own undergo a gradual series of merger events over cosmic time [3]. This results in stellar halos that are predominantly composed of debris from a few massive dwarf galaxies accreted early in the galaxy's history [4].

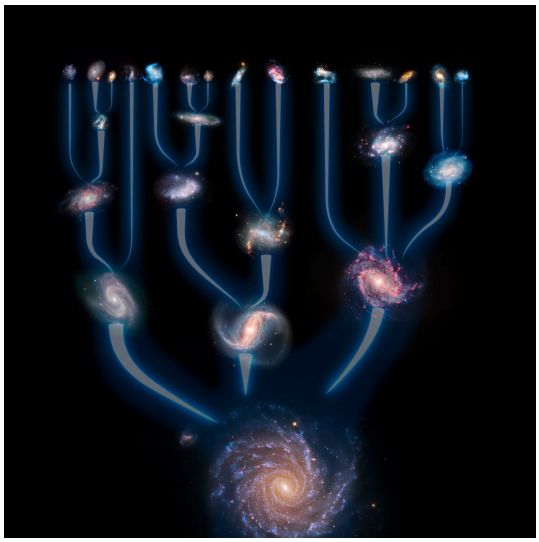


Figure 1: Illustration of hierarchical galaxy formation.
Credit: ESO/L. Calçada. [5]

Alongside this, the Milky Way halo's vast spatial extent and long orbital periods allow it to preserve the coherence of a progenitor system's debris over gigayear timescales and act as a fossil record of its hierarchical assembly. Pioneered by Eggen, Lynden-Bell, and Sandage [6] in 1962, the field of Galactic Archaeology seeks to reconstruct the forgotten history of our Galaxy and infer its formation by decoding its stellar chemistry (chemical tagging) and six-dimensional phase-space (orbital properties).

Eggen, Lynden-Bell, and Sandage [6] were the first to identify the large population of metal-poor ([Fe/H]) stars in the Galactic halo on highly eccentric orbits. Although the strong correlation they observed in this e-[Fe/H] plane was later shown to result from proper-motion selection biases in their dataset [7], thereby disproving their proposed model of an early, rapid collapse, their work nonetheless catalysed interest in the field. Eventually, extensions that build on these early insights, led to the first observational evidence for the significant accretion of halo stars by Chiba and Beers [8].

1.2 A New Era for Galactic Archeology

As highlighted by Deason and Belokurov [9] the study of the galactic halo has long been hindered by limitations in the field's datasets, whether that be incomplete sky coverage, sparse chemo-dynamical information, or insufficient sample sizes. However, the advent of the next generation of astronomical surveys is rapidly accelerating us into a new era of 'big data astronomy'. Although Gaia observed its final star on 15 January 2025 [10], it has already delivered full astrometric solutions, including positions, parallaxes, and proper motions, for approximately 1.46 billion sources [11, 12]. With its fourth data release anticipated in 2026, the insights provided by Gaia will continue to transform our understanding of the Milky Way's structure and evolutionary history for years to come. Complementing this in chemical space, several wide-field spectroscopic surveys such as the Apache Point Observatory Galactic Evolution Experiment (APOGEE) [13] and Galactic Archaeology with HERMES (GALAH) [14] provide elemental abundances of $\sim$ 734,000 and $\sim$ 918,000 stars respectively. Although these surveys are already transforming the field from theories to quantitative reconstructions [9], they are simply the beginning. Upcoming projects such as WEAVE (WHT Enhanced Area Velocity Explorer) [15] and 4MOST (4-metre Multi-Object Spectroscopic Telescope) [16, 17] promise to provide deeper, more complete, and higher-dimensional data to extend this revolution.

In parallel, the field is also being transformed by exponential advances in computational power. Firstly, these developments have enabled increasingly detailed simulations of gravitational N-body dynamics and hydrodynamical processes [18], exemplified by large-scale projects such as EAGLE [19], AURIGA [20] and LATTE [21]. These simulations allow our theoretical models to be refined through cross-referencing with observational discoveries. Secondly, and of particular interest to this work, is the emergence of machine learning methodologies, especially unsupervised clustering algorithms. As summarised by Sante et al. [22], these approaches disentangle hidden, unlabeled structures from within high-dimensional datasets, enabling us to identify theorised populations and previously undiscovered stellar components.

1.3 The Emerging Picture of the Galactic Halo

The arrival of new surveys has already enabled a much more detailed dissection of both the accreted and in-situ populations that occupy our local stellar halo. While the remnants of numerous lower-mass accretion events have been identified [23, 24, 25, 26, 27], there remains only one widely accepted example of an ancient massive merger: the so-called 'Gaia–Sausage' [28] or 'Gaia–Enceladus' [29] event, hereafter referred to as the GS/E merger. Cosmological simulations, guided by the dark-matter halo's fundamental mass function ($dn/dM \propto M^{-1.9}$) [9, 30], suggest Milky Way–mass halos typically experience numerous minor mergers (N $\sim$ 30) and a smaller number of major ones (N $\sim$ 3) [31]. In comparison to this, the Milky Way appears to have historically undergone a relatively subdued and restrained accretion process.

It is important to caveat this with the recognition of other significant merger events, such as the more recent examples visibly unraveling in the distant halo. Notably, the Sagittarius (Sgr) Dwarf Galaxy [32] whose vast tidal arms wrap around the Galaxy [33], and the Magellanic system [34, 35] which is the most massive gas-rich system currently undergoing infall [36]. Both of these are estimated to reside

in dark-matter halos comparable to the GS/E's [37, 38]. In addition, several studies have proposed the presence of debris from early accretion events concealed within the densely populated inner Galaxy, such as Heracles [39], Kraken [40] and Koala [41]. However, Naidu et al. [42] postulated that these structures may all share a common origin, and Myeong et al. [1] further questioned whether an additional merger is required to explain their observed stellar properties.

With an estimated stellar mass of $M_\star \sim 5 \times 10^8 \mathrm{M}_\odot$ and dark matter halo of $M_{\mathrm{DM}} \sim 2 \times 10^{11} \mathrm{M}_\odot$ [43], the GS/E merger is thought to contribute up to 50% of the inner halo ($\lesssim 30$ kpc) [44]. It is therefore difficult to overstate its role in shaping the Milky Way. Even prior to the Gaia mission, Deason et al. [45] noted the correlation, in the Bullock and Johnston [4] simulation suite, between broken stellar halo density profiles and the accretion histories of galaxies. This led to the hypothesis, that the Milky Way's rapid density drop of at ~ 20 kpc was the result of the 'apocenter pile-up' [46] caused by a singular, relatively early, and massive accretion event which now forms the dominant progenitor of the inner halo. Subsequent observations from Gaia provided overwhelming support for this scenario [28, 25]. In particular, Belokurov et al. [28] reported the dominance of a 'more metal-rich portion ($-1.7 <$[Fe/H]< -1) of the halo, exhibiting extreme radial anisotropy', an orbital signature difficult to explain through multiple small mergers. This finding was interpreted by both Belokurov et al. [28] and Haywood et al. [47] as strong evidence for this singular, massive accretion event that occurred $\sim$ 8–10 Gyr ago.

As the field has developed a more holistic view of the downstream effects of the GS/E merger, it is increasingly accepted that its impact extends beyond the deposition of an accreted halo population and is responsible for a significant number of highly eccentric in-situ stars. A sizeable population of high [α/Fe] and [Fe/H] (> -0.7 dex) stars, chemically and kinematically similar to the thick disc, was detected by Di Matteo et al. [48]. However, our modern understanding of their origin was not clear until the observations by Gallart et al. [49] revealed these stars to be significantly older than the bulk of the thick disc, and more aligned with the accreted stars' age. First named the 'Splash' by Belokurov et al. [50], it is theorised these stars were formed within the high-[α/Fe] proto-disk of the Milky Way, and then 'kicked out' as a result of the gravitational perturbations from the GS/E merger. More recently Myeong et al. [1] proposed the identification of a new halo component referred to as 'Eos'. Through comparison with simulations presented by Grand et al. [51] and Renaud et al. [52], it was suggested that it originated from GS/E-polluted gas before evolving into the outer thin disk. This population is thus consistent with the two-infall model used to explain the observed α-dichotomy in the galactic disk.

Finally, the halo's kinematically hot, low-metallicity ([Fe/H] $\lesssim -1.3$) component was proposed by Belokurov and Kravtsov [53] to be formed in the galaxy's messy state before its 'spin-up' and was dubbed 'Aurora'.

1.4 Objectives of this work

The majority of the halo substructures identified to date have been discovered through manual exploration of high-dimensional datasets and subsequent attempts to decompose these components have typically relied on manually defined decision boundaries. As we enter an era of automated discovery driven by machine learning methodologies, Myeong et al. [1] introduced the first 'unbiased and self-consistent' approach to

identify the highly eccentric (halo) components in the solar neighborhood using Gaussian Mixture Models (GMMs). In leveraging this unsupervised, probabilistic framework, their work (hereafter referred to as the original work) was presented as both objective and reproducible, a claim which the primary aim of this paper is to verify. Additionally, the paper goes on to extend the original work by (i) applying dimensionality reduction to compare the stability of the clusters between the independent datasets and (ii) exploring alternative clustering approaches to improve the convergence and computational efficiency of the analysis in anticipation of significantly larger future datasets.

2 Data

To rigorously ensure reproducibility, preprocessing pipelines were created in line with the scheme outlined in the original study, to test whether equivalent samples could be reconstructed directly from the raw survey releases and associated Value-Added Catalogues (VACs). Although the APOGEE DR17 sample [13] was originally presented as the primary source of analysis, with GALAH DR3 [14] serving as a verification tool, this work treats both surveys with equal weighting, comparing their respective advantages throughout.

The APOGEE selection, drawn primarily from the chemical information provided in its `allstarlite` catalogue, is augmented with Gaia EDR3's [54] orbital information [54] and the Bayesian distance estimates from Bailer-Jones et al. [55]. This enables orbital parameters to be derived from the placement of each source within the Milky Way potential presented by McMillan [56]. Similarly, the GALAH sample is constructed by linking the `allstar` catalogue with the `GaiaEDR3` and `dynamics` VACs.

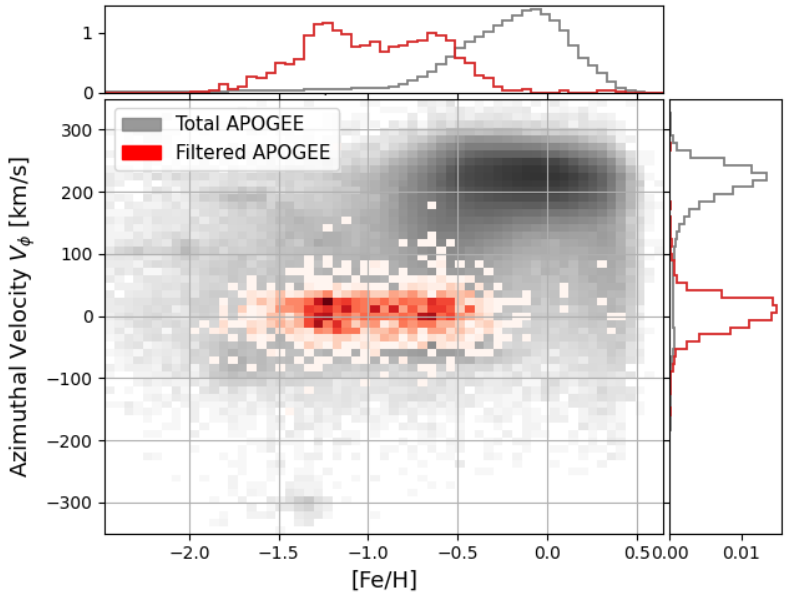


Figure 2: Distribution of Metallicity vs. Azimuthal Velocity V_ϕ in the resultant APOGEE Dataset relative to the Overall Survey

To reduce contamination from disk and bulge populations and better highlight the radial-orbit halo components, a strict eccentricity cut of $e > 0.85$ is applied, along with a minimum orbital apocenter of 5 kpc. Furthermore, stars of orbital energies $< 0 km^2 s^{-2}$ or distance uncertainties (from Bailer-Jones et al. [55]) >1.5 kpc are removed. As a result of these selection cuts, we must consider the weightings of

the recovered populations in relation to their radial anisotropy and thus inclusion in our sample. The effects of these cuts are illustrated by the azimuthal velocity distribution shown in Figure 2. Additionally, we introduce sharp, biased boundaries in 'dynamical space', reducing the dimensions that can be used accurately by the clustering algorithm.

2.1 APOGEE Dataset

Following the recommendations outlined in the original work and ASPCAP documentation [57] only high-quality red giant stars are retained. We exclude sources flagged as unreliable by the ASPCAPFLAG or associated with non-main survey fields (`magclouds` flag), and select only those with $\log g < 3.0$. To ensure reliable chemistry, we impose element-specific quality cuts by requiring no pipeline abundance flags and uncertainties < 0.1 dex for all elements used, except for [Ce/Fe], where a more relaxed threshold of 0.15 dex is applied. In addition to the original work, we perform individual flag and uncertainty checks for the six α-elements used to derive [α/Fe] (O, Mg, Si, S, Ca, and Ti) as no dedicated flag is provided. This is deemed a conservative approach since APOGEE's [α/M] abundance is derived from a global fit, rather than simply combining individual abundances.

Alongside orbital energy (E) and metallicity [Fe/H], the following four chemical dimensions were selected, balancing the power to distinguish between distinct nucleosynthetic pathways with reliability. Firstly, the [α/Fe] ratio due to its ability to trace the relative fractional enrichment from core-collapse (Type II) and thermonuclear (Type Ia) supernovae. Similarly, [Mg/Mn] is also included as Hawkins et al. [58] suggests Mg and Mn have stronger correlations with SNII and SNIa respectively, and is thus a more powerful tracer. To help differentiate between accreted and in-situ components, we add the odd-Z nucleosynthesis tracer of [Al/Fe]. Finally, as supported by Casali et al. [59], [Ce/Fe] is used to identify the slow neutron-capture (s-process), primarily associated with asymptotic giant branch (AGB) stars and later evolutionary enrichment.

Overall, we obtained a 6-D sample of 1612 stars, 5.5% smaller than that presented in the original work. The primary sources of this reduction were the failure of the [Ti/Fe] quality flag (29 of 94) and accounting for both lower and upper distance uncertainties (62 of 94). To assess the significance of this refinement, the distributions of the remaining and rejected samples of each dimension were compared. Although a slight bias toward excluding lower metallicity and [Al/Fe] stars was observed, it is not deemed significant enough to heavily affect the results of the analysis.

2.2 GALAH Dataset

A similar approach is taken with the GALAH data, in which we choose to exclude stars with `snr_c3_iraf` $>$ 30 or any flagged abundances. Additionally, a more lenient uncertainty threshold of 0.2 dex is used. Because GALAH surveys a broader range of chemical species, we are able to obtain a higher dimensional clustering space: Energy (E), [Fe/H], [α/Fe], [Na/Fe], [Al/Fe], [Mn/Fe], [Y/Fe], [Ba/Fe], [Eu/Fe], [Mg/Cu], [Mg/Mn], and [Ba/Eu]. Overall we obtain a 12-D sample of 1061 stars, 2.7% smaller than that presented by Myeong et al. [1].

2.3 Expected Performance of Datasets

Before considering the results of the analysis, the comparative strengths and limitations of the APOGEE and GALAH surveys are assessed a priori. APOGEE's operation in near-infrared wavelengths (H-band), its selection of predominantly red giant stars, and its broader focus on all major Galactic components provide a more spatially extensive dataset, as shown in Figure 3. In comparison, GALAH was primarily designed for an optical view of the nearby disk, targeting main-sequence and turnoff stars. This results in APOGEE offering a sample 51.9% larger than its GALAH counterpart, reaching further into the halo's population of dimmer stars. APOGEE therefore achieves a lower metallicity limit ([Fe/H] $\approx$ -2, compared to -1.3 for GALAH), and is thus theorised to better identify and trace the more metal-poor components of the halo, such as the debris from the GS/E merger. Additionally, APOGEE exhibits significantly smaller abundance uncertainties across all elements considered, suggesting a greater ability to resolve subtle chemo-dynamical populations, such as the proposed 'Aurora' component. However, it is important to consider warnings in the literature, notably by Boulet [60], on the potential underestimation of APOGEE's uncertainties, with particular caution to the inclusion of [Ce/Fe].

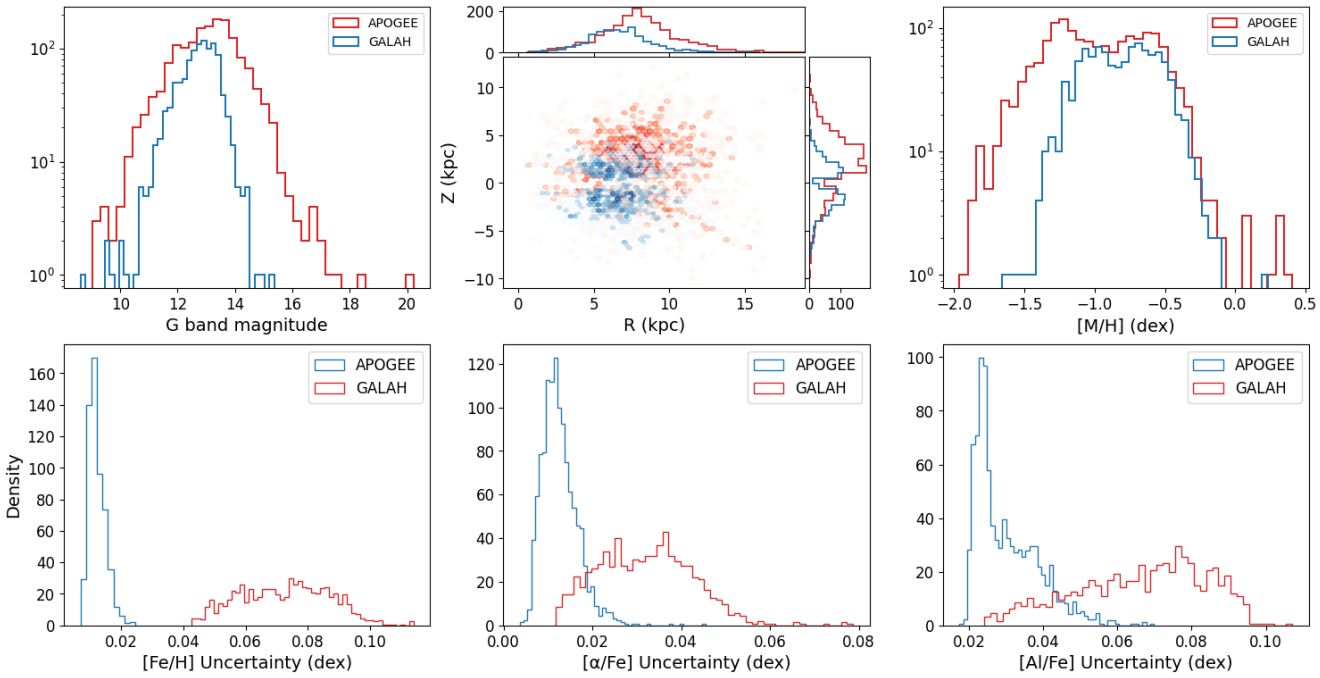


Figure 3: Comparison of the distribution of APOGEE and Galah Datasets

On the other hand, GALAH's broader chemical coverage allows the creation of a higher-dimensional space which may offer greater discriminatory power between populations with subtle differences in nucleosynthetic history. For example, the inclusion of the *r*-process tracer [Eu/Fe], which is unavailable in APOGEE. It is important to note, however, that several of these elements trace similar enrichment processes such as [Y/Fe] and [Ba/Fe] both probing the *s*-process, while [Na/Fe] and [Al/Fe] are associated with odd-Z element production. Therefore these overlaps may not necessarily offer improved insight but simply increase the attention that each process receives during clustering.

In summary, the relative advantages of each survey likely depend on whether the separation power offered by GALAH's higher dimensionality is sufficient to outweigh its smaller sample size, limited spatial reach, and larger measurement uncertainties.

3 Methods

3.1 Extreme Deconvolution (XD)

This study employs *Extreme Deconvolution* (XD) [61], a Gaussian Mixture Model (GMM) based algorithm, capable of accounting for heteroskedastic uncertainties. This is motivated by Figure 3 in which the observational uncertainties vary significantly across both dimensions and surveys. In doing so, XD allows the true distribution of the stellar populations to be inferred rather than that of the noisy observations. Bovy, Hogg, and Roweis [61] achieved this by reformulating the likelihood of each observed data point $\mathbf{w}_i$ as a convolution of the latent distribution with an uncertainty covariance $\mathbf{S}_i$. This allows the Expectation-Maximisation (EM) algorithm to optimise the likelihood of the deconvolved (underlying) distribution. Mathematically, with no incomplete data (i.e. an identity projection matrix, $R_i = \mathbf{I}$), this can be expressed as follows.

The latent distribution of true values $\mathbf{v}$ is modeled as a mixture of K Gaussians:

$$p(\mathbf{v}) = \sum_{j=1}^{K} \alpha_j \, \mathcal{N}(\mathbf{v} \,|\, \mathbf{m}_j, \mathbf{V}_j), \tag{1}$$

The likelihood of a noisy observation $\mathbf{w}_i$ given the model parameters $\theta = (\alpha_j, \mathbf{m}_j, \mathbf{V}_j)$ is computed by marginalising over the latent variable $\mathbf{v}$:

$$p(\mathbf{w}_i|\theta) = \sum_j \int d\mathbf{v} \, p(\mathbf{w}_i|\mathbf{v}) \, p(\mathbf{v}|j,\theta) \, p(j|\theta). \tag{2}$$

Where:

$$p(\mathbf{w}_i|\mathbf{v}) = \mathcal{N}(\mathbf{w}_i|\mathbf{v}, \mathbf{S}_i), \tag{3}$$

$$p(\mathbf{v}|j,\theta) = \mathcal{N}(\mathbf{v}|\mathbf{m}_j, \mathbf{V}_j), \tag{4}$$

$$p(j|\theta) = \alpha_j. \tag{5}$$

As a result, the total likelihood of $\mathbf{w}_i$ simplifies to another mixture of Gaussians:

$$p(\mathbf{w}_i|\theta) = \sum_j \alpha_j \, \mathcal{N}(\mathbf{w}_i \,|\, \mathbf{m}_j, \mathbf{T}_{ij}), \tag{6}$$

where the effective covariance $\mathbf{T}_{ij}$ accounts for both the Gaussian component and the measurement uncertainty:

$$\mathbf{T}_{ij} = \mathbf{V}_j + \mathbf{S}_i. \tag{7}$$

To determine the model parameters, we maximise the log-likelihood across all N data points:

$$\phi = \sum_i \ln p(\mathbf{w}_i|\theta) = \sum_i \ln \sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{w}_i \,|\, \mathbf{m}_j, \mathbf{T}_{ij}). \tag{8}$$

3.2 Pipeline

Although the XD package offers a fast implementation via its C backend, it lacks several features common in modern GMM libraries. Therefore this project develops a pipeline that builds upon it by automating multiple random initialisations, computing model selection criteria, including the Bayesian Information Criterion (BIC) [62] and Akaike Information Criterion (AIC) [63], and 'error-aware' cluster assignment.

In practice, the pipeline is run with the number of Gaussian components $K \in [1, 10]$, each with 100 random initialisations and a maximum of 10^9 Expectation-Maximisation (EM) iterations per run. To ensure robustness and reproducibility, the entire procedure is repeated three times.

$$\mathrm{AIC} = 2k - 2\ln\mathcal{L}, \tag{9}$$

$$\mathrm{BIC} = k\ln n - 2\ln\mathcal{L}, \tag{10}$$

It is noted that the AIC and BIC are commonly in competition and there is no universally preferred criterion. Shmueli [64] and Sober [65] present their subtle distinction as a 'philosophical' debate, with AIC oriented toward selecting the model with the best out-of-sample predictive accuracy, while BIC focuses more on the goodness of fit. In this work, we report both scores, typically preferring the criterion that yields a minimum in model comparison.

Additionally, the pipeline supports two preprocessing schemes: full standard scaling of all dimensions, or as used by the original work a simple rescaling of only the Energy (E) by a factor of 10^5 to approximately match the range of other features. Each approach is paired with a corresponding random initialisation strategy and applies automated rescaling of the covariance and mean. In principle, GMMs use of covariance matrices allows it to intrinsically account for the differences in scale, and avoiding scaling preserves the appreciation of physical units, which can be valuable. However, we find that standard scaling yields more stable results that show greater agreement with the original work and are therefore the results presented in this work.

3.3 Caveats

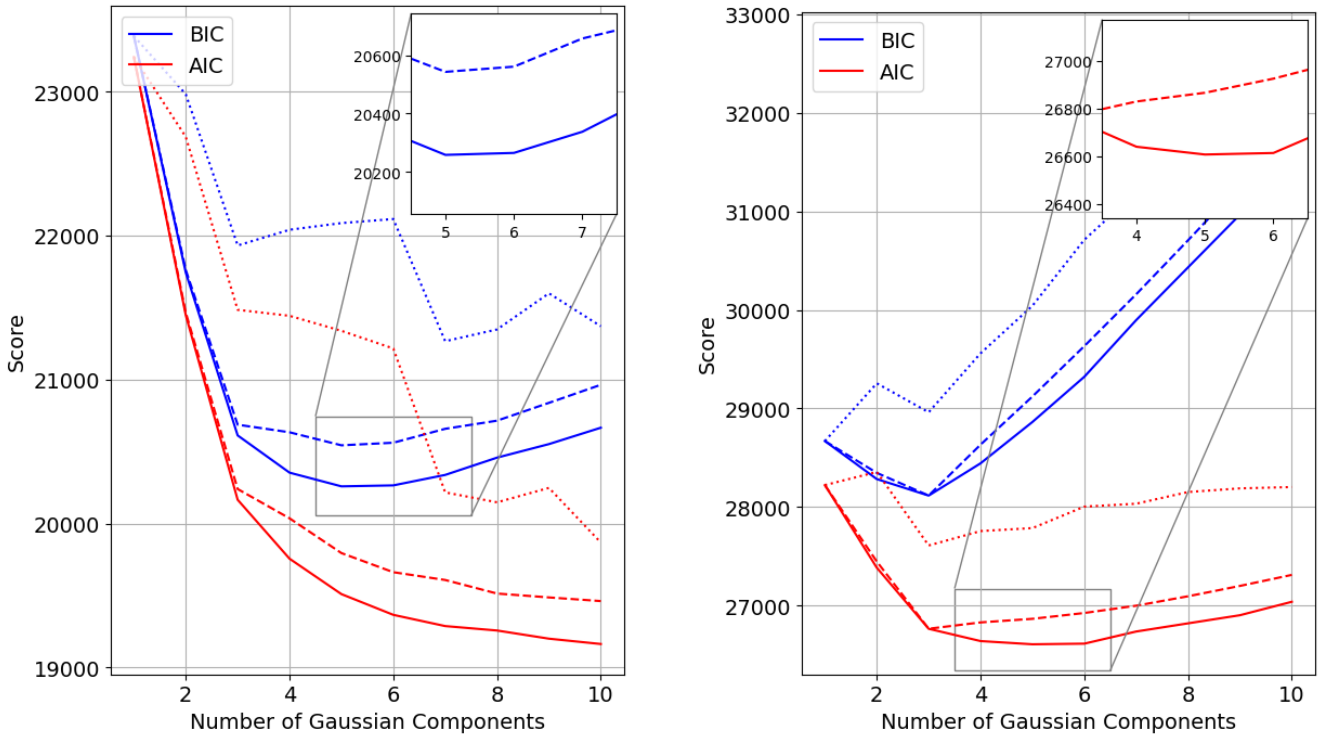
This work acknowledges the limitation that the adopted approach assumes that stellar populations trace multivariate Gaussian distributions. In reality, many stellar populations do not support this underlying assumption, particularly those formed over extended timescales or within rapidly evolving chemo-dynamical environments. A notable example is the distinct alpha plateau and knee, which form a clearly non-Gaussian structure in the [α/Fe]–[Fe/H] plane. That being said, while individual astrophysical populations may not be strictly Gaussian, GMMs can approximate dominant complex structures by representing them as a

combination of multiple Gaussian components. From these, their global properties can be approximated as:

$$\boldsymbol{\mu}_{\text{combined}} = \frac{\sum_{i=1}^{N} w_i \boldsymbol{\mu}_i}{\sum_{i=1}^{N} w_i}, \tag{11}$$

$$\boldsymbol{\Sigma}_{\text{combined}} = \frac{\sum_{i=1}^{N} w_i \left[\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{\text{combined}})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{\text{combined}})^\top\right]}{\sum_{i=1}^{N} w_i}. \tag{12}$$

4 Results



(a) APOGEE-Gaia Sample **(b)** GALAH-Gaia Sample

Figure 4: Comparison of AIC and BIC scores as a function of the number of Gaussian components. Solid lines show the best (lowest) score across all 100 initialisations, dashed lines indicate the median and dotted lines represent the worst fit.

4.1 APOGEE - High Dimensional Clustering

A comparison of the AIC and BIC trends with those presented by Myeong et al. [1] reveals partial agreement alongside subtle differences. As shown in Figure 4a, the overall trend is consistent with the original work: the AIC continues to decrease, failing to reach a minimum, while the BIC does and is thus preferred for model selection in this case. We note that although the use of scaled input data causes the absolute values of the information criteria (and likelihoods) to differ, this shift is arbitrary as values are compared relatively. Consistent with the original work, similar BIC scores are achieved for 4 to 8 components; however, our analysis shows a true minimum at 5 Gaussian components compared to the

original 7. This discrepancy is likely a result of the scaling scheme and marginal differences in the dataset used and highlights the importance of visually cross-referencing the results which achieve numerically comparable scores. In particular, the transition from 7 to 5 components appears to first correspond to the elimination of an additional 'background' component followed by the loss of the component associated with the 'Aurora' population.

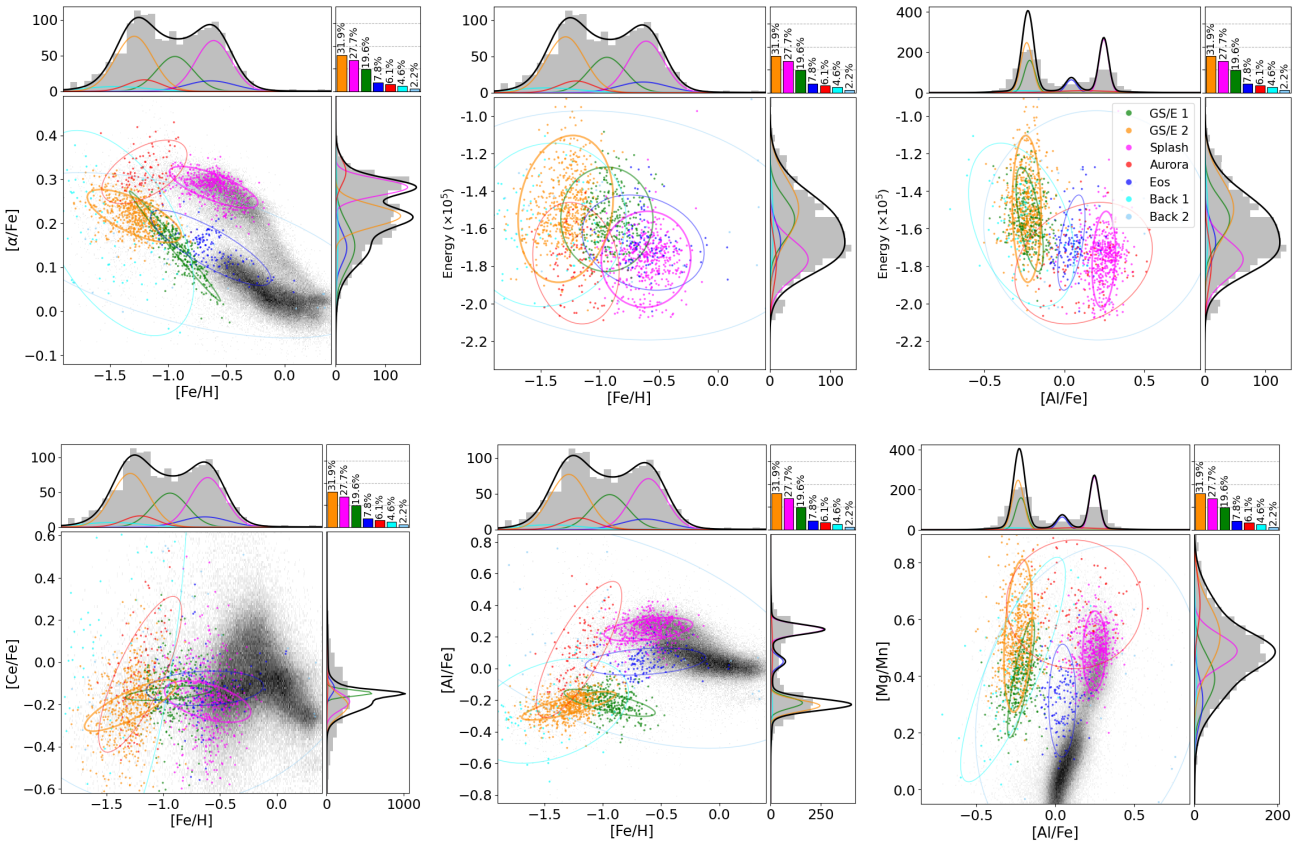


Figure 5: Chemo-dynamical projections of seven GMM-fitted components to the high-eccentricity APOGEE–Gaia sample. Shown are 2σ ellipses, marginal histograms, and fractional weights, with a greyscale density map of all of APOGEE's reliable data.

The loss of the 'Aurora' component is not entirely unexpected, given the population's reported isotropic velocity dispersion [53] and therefore weak weighting in our high-eccentricity sample. However, this is the first indication within this work of the fragility of Aurora's detection from within APOGEE's six-dimensional sample. To support the analysis of the Aurora population and allow a fairer comparison with the original study, we present the results from the 7-component model in the subsequent analysis.

In comparing the best 7-component fit with the original work, we find strong support that the model successfully recovers the same key stellar populations. Firstly, the mean values across all dimensions of the Splash, Aurora, and Eos populations, presented in Table 1 and Figure 5, show excellent agreement. The most notable difference is this work presents consistently smaller intrinsic uncertainties across the [Al/Fe] and [Ce/Fe] dimensions. The variations in these dimensions may reflect their potential instabilities and be a result of APOGEE's underestimation of elemental abundance uncertainties, particularly for [Ce/Fe] [59].

More significant discrepancies are seen in comparing the individual components of the GS/E merger,

Table 1: Summary of results from the 7 component XD fit on the APOGEE-Gaia sample.

Component	Weight (%)	Count	[Fe/H]	[α/M]	Energy ($10^5 \mathrm{km}^2 s^{-2}$)	[Ce/Fe]	[Al/Fe]	[Mg/Mn]
GS/E 1	19.6	299	-0.94 ± 0.16	0.15 ± 0.05	-1.55 ± 0.11	-0.15 ± 0.01	-0.22 ± 0.03	0.39 ± 0.08
GS/E 2	31.9	541	-1.29 ± 0.16	0.22 ± 0.02	-1.50 ± 0.15	-0.20 ± 0.05	-0.24 ± 0.04	0.54 ± 0.11
Splash	27.7	447	-0.61 ± 0.15	0.28 ± 0.02	-1.76 ± 0.10	-0.19 ± 0.04	0.25 ± 0.03	0.49 ± 0.06
Aurora	6.1	95	-1.20 ± 0.15	0.32 ± 0.03	-1.78 ± 0.13	-0.06 ± 0.15	0.12 ± 0.17	0.65 ± 0.09
Eos	7.8	126	-0.63 ± 0.21	0.14 ± 0.03	-1.67 ± 0.10	-0.12 ± 0.03	0.04 ± 0.03	0.31 ± 0.08
Back 1	4.6	70	-1.45 ± 0.27	0.18 ± 0.09	-1.58 ± 0.18	0.40 ± 0.60	-0.27 ± 0.13	0.42 ± 0.16
Back 2	2.2	34	-0.85 ± 0.66	0.13 ± 0.08	-1.58 ± 0.25	0.08 ± 0.31	0.20 ± 0.29	0.30 ± 0.23
$\mathrm{GS/E}_{tot}$	51.6	840	-1.16 ± 0.23	0.19 ± 0.05	-1.52 ± 0.14	-0.18 ± 0.05	-0.23 ± 0.04	0.48 ± 0.12
Back_{tot}	6.8	104	-1.25 ± 0.52	0.16 ± 0.09	-1.58 ± 0.20	0.29 ± 0.54	-0.11 ± 0.30	0.38 ± 0.19

where the fractional contribution of the GS/E_1 component is reduced relative to the original analysis (19.6% from 27.6%), while that of GS/E_2 increased (31.9% from 23.0%). However, this difference is simply a result of the way in which the GMM has decomposed the complex, non-Gaussian structure of the high-α plateau and chemical knee. This inconsistency in distinguishing these clusters is unsurprising given their shared origin. This claim is verified by the aggregated GS/E population (GS/E_{tot}) showing very strong agreement with the original analysis in both weight ($\approx 51\%$) and distribution in chemo-dynamical space. Finally, two 'implausible' low-weight background populations are also recovered.

4.2 GALAH - High Dimensional Clustering

In contrast to the APOGEE results, we find that the AIC provides the most informative score for model comparison, supporting this work's non-prescriptive consideration of both metrics. The analysis suggests 5 Gaussian components as optimal, consistent with the results presented by Myeong et al. [1]. The best 5 component solution obtains near-perfect agreement across all stellar components and chemo-dynamical dimensions, when compared to the original study, illustrated in Table 2. In Figure 6 we present the equivalent six dimensional grid to that used in APOGEE, with plots of the additional chemical abundances shown in Figure 7. Furthermore, by examining fits with a higher number of components (e.g. six and seven) it is found that additional Gaussians only capture extra background populations rather than meaningful astrophysical substructures.

4.3 Agreements with the Original Work

While simply reproducing the discussion of the original study is not the focus of this work, we briefly highlight the agreement between our results and several of the key findings, despite subtle differences in the shapes of the APOGEE data's fitted ellipses.

Our analysis supports the conclusion that the GS/E merger is the only accreted component, primarily based on its significantly lower [Al/Fe] abundances, suggestive of its origin in a less efficient star-forming progenitor. Furthermore, the chemical placement of the Eos component between the GS/E population and the thin disk (as illustrated by the background distribution) supports Myeong et al. [1]'s proposal that it formed from gas enriched by the GS/E merger and later developed into the thin disk. Finally, despite the slight differences in other populations' uncertainty ellipses in APOGEE's [Fe/H]–[Al/Fe] plane, we robustly recover Aurora's previously reported rapid chemical evolution (shown by the steep correlation).

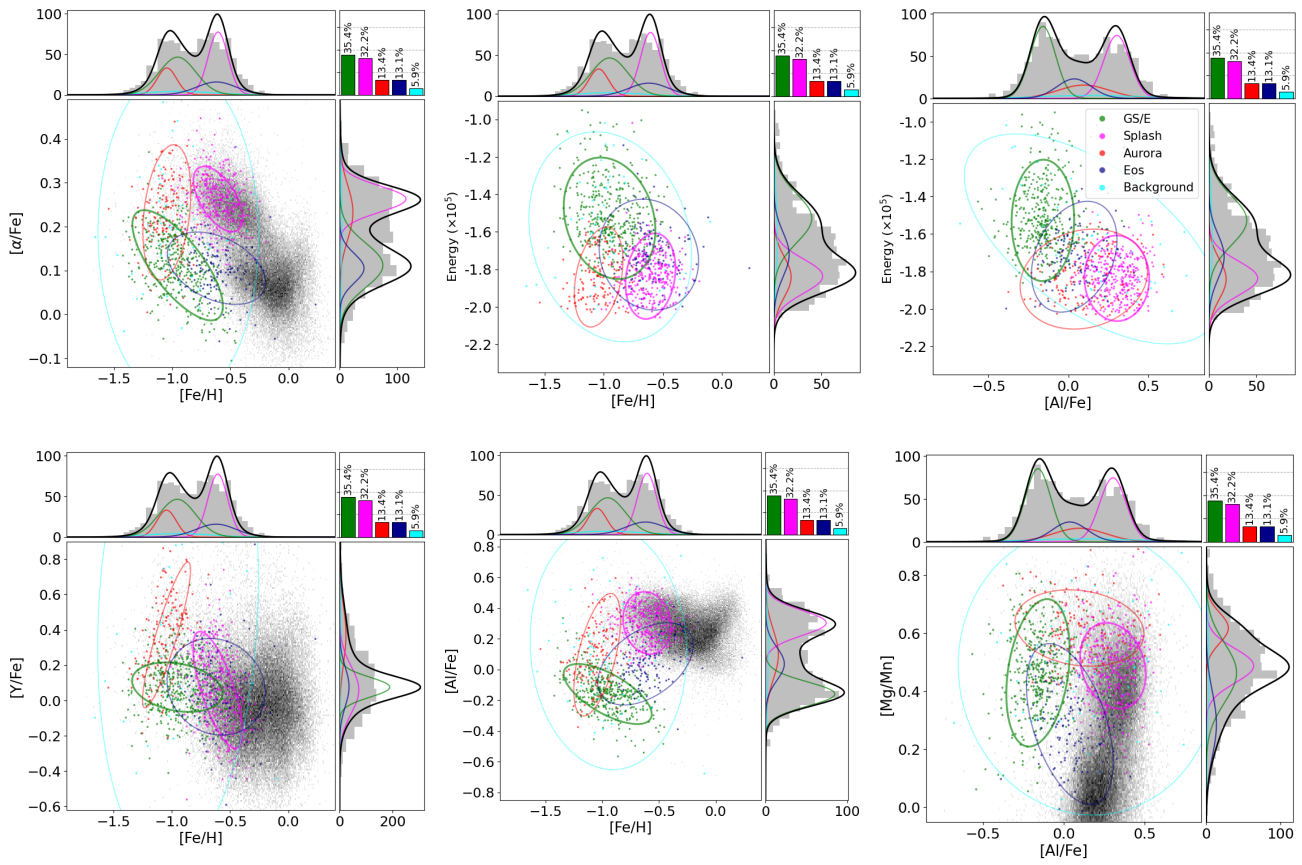


Figure 6: Chemo-dynamical projections of five GMM-fitted components to the high-eccentricity GALAH–Gaia sample. Shown are 2σ ellipses, marginal histograms, and fractional weights, with a greyscale density map of all of GALAH's reliable data.

Taken in parallel with its striking similarity to the proposed Hercules population [42], this is what led to the claim that Hercules may not necessarily correspond to an additional accretion event.

In summary, despite differences in dataset selection and the application of a new scaling scheme, these results recover the same structures identified in the original study. This provides strong support for the original work's claim of presenting an unbiased, objective, and reproducible decomposition of the local stellar halo in high-dimensional chemo-dynamical space.

4.4 Initial Dataset Comparison

As the focus of this paper turns to the comparison of the insights offered by the different datasets, there are some key initial observations that shed light on the potential advantages offered by each survey. Firstly, APOGEE's broader metallicity coverage ([Fe/H] $\gtrsim -2$) provides a more complete view of the GS/E merger's α plateau. This is shown by its ability to decompose the structure into two distinct components, a result not possible using high-dimensional clustering on the GALAH data, due to its metallicity limit of [Fe/H] $\simeq -1.5$. This broader view allows for a more accurate assessment of the relative contribution of accreted stars in the local solar neighborhood. Specifically, APOGEE suggests the GS/E population accounts for approximately 51.6% of the sample, consistent with the findings of Iorio and Belokurov [44]. On the other hand, GALAH's view leads to an underestimated fraction of around 35.4%.

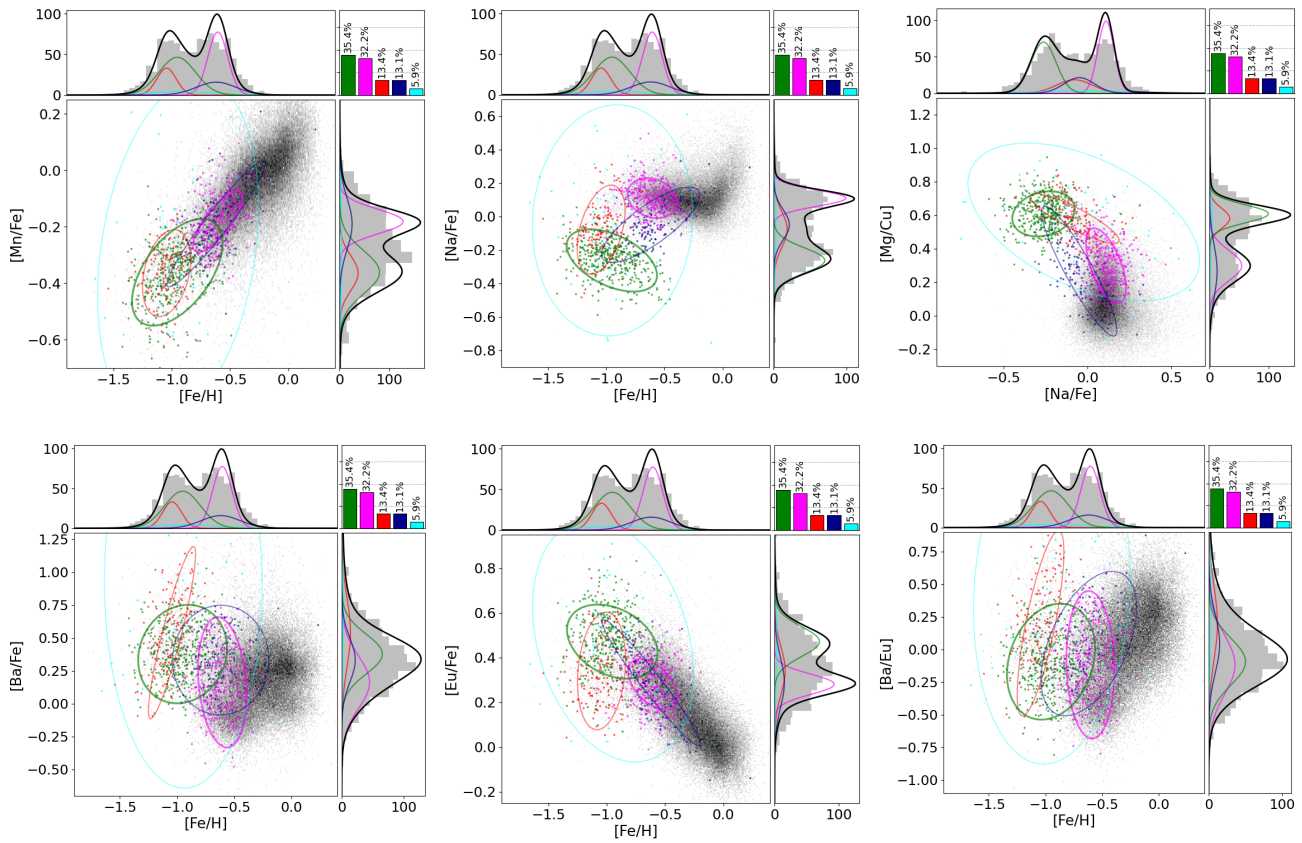


Figure 7: Additional chemo-dynamical projections of five GMM-fitted components to the high-eccentricity GALAH–Gaia sample. Here an insight into the dimensions not available in APOGEE is shown. Included are 2σ ellipses, marginal histograms, and fractional weights, with a greyscale density map of all of GALAH's reliable data.

However, despite this additional coverage, we observe the greater fragility in resolving the low-metallicity Aurora population within the APOGEE dataset. This is shown by the BIC score indicating that the five-component model, excluding Aurora, provides a better statistical fit to the data. On the other hand, the GALAH dataset demonstrates relatively stronger agreement with the original work, even for the weakly populated background component. Additionally, when comparing the Aurora, Eos and Splash population's intrinsic uncertainties recovered from both GALAH and APOGEE, we find very similar values. This demonstrates XD's effectiveness in deconvolving the underlying structures from the observational uncertainties and thus shows that GALAH's higher errors can be successfully accounted for and mitigated.

Collectively, this provides an early indication of GALAH's potential advantages in isolating halo substructures. As discussed in subsection 2.3, a likely explanation of this is its higher dimensionality providing clearer separation and increasing the robustness and stability of clustering. A more detailed investigation of this is provided throughout section 5.

Table 2: Summary of results from the 5 component Extreme Deconvolution fit on the GALAH-Gaia sample.

Component	Weight (%)	Count	[Fe/H]	[α/Fe]	[Na/Fe]	[Al/Fe]	[Mn/Fe]	[Y/Fe]	[Ba/Fe]	[Eu/Fe]	[Mg/Cu]	[Mg/Mn]	[Ba/Eu]	Energy (10^5 km^2 s^{-2})
GS/E	35.4	389	-0.95 ± 0.16	0.11 ± 0.05	-0.26 ± 0.07	-0.16 ± 0.08	-0.36 ± 0.08	0.08 ± 0.06	0.38 ± 0.15	0.47 ± 0.07	0.61 ± 0.05	0.46 ± 0.10	-0.09 ± 0.18	-1.53 ± 0.13
Splash	32.2	345	-0.60 ± 0.09	0.26 ± 0.03	0.11 ± 0.05	0.30 ± 0.08	-0.18 ± 0.04	0.06 ± 0.13	0.16 ± 0.20	0.28 ± 0.05	0.30 ± 0.09	0.49 ± 0.06	-0.12 ± 0.23	-1.84 ± 0.09
Aurora	13.4	141	-1.04 ± 0.08	0.23 ± 0.06	-0.08 ± 0.11	0.10 ± 0.16	-0.36 ± 0.06	0.37 ± 0.17	0.54 ± 0.26	0.34 ± 0.11	0.58 ± 0.06	0.62 ± 0.05	0.20 ± 0.29	-1.84 ± 0.10
Eos	13.1	129	-0.62 ± 0.17	0.11 ± 0.03	-0.05 ± 0.09	0.04 ± 0.11	-0.19 ± 0.09	0.08 ± 0.11	0.33 ± 0.17	0.29 ± 0.11	0.27 ± 0.16	0.29 ± 0.11	0.04 ± 0.22	-1.72 ± 0.12
Back	5.9	57	-0.95 ± 0.28	0.18 ± 0.15	-0.02 ± 0.28	0.14 ± 0.32	-0.29 ± 0.25	0.49 ± 0.57	0.92 ± 0.63	0.48 ± 0.22	0.56 ± 0.19	0.47 ± 0.20	0.44 ± 0.53	-1.63 ± 0.23

5 Dimensionality Reduction

This section analyses the cohesion and isolation of the substructures identified during high-dimensional clustering, with the aim of providing further insight into the comparisons made in subsection 4.4. To aid in the interpretation of these complex structures, non-linear dimensionality reduction techniques are employed, which aim to preserve relationships within a lower-dimensional projection. While the runtime performance is not a limiting factor for the datasets in this report, this paper aims to present scalable methods suitable for application to next-generation astronomical surveys. Accordingly, although several methods were explored, including 't-Distributed Stochastic Neighbor Embedding' (t-SNE) [66], this work ultimately favours 'Uniform Manifold Approximation and Projection' (UMAP) [67], due to its reported advantages in computational efficiency and its superior preservation of both local and global structures.

To optimise the projections, this section exploited UMAP's hyperparameters, namely `n_neighbors`, which controls the balance between preserving the local and global structure of the input data and `min_dist` which given the interpretation of the input, determines the spacing between points in the low-dimensional embedding. Additionally, we help to demonstrate the cohesion of these clusters through the success of `scikit-learn`'s GMM package [68] at re-identifying these populations from within the lower-dimensional embedding space, a concept explored more rigorously in section 6.

Below, we present the APOGEE and GALAH datasets in a two-dimensional UMAP embedding space, alongside the initial results of GMM clustering. These are shown across a range of `min_dist` values to explore the trade-off between visual separation and the intrinsic embedding projection.

5.1 APOGEE in Embedding Space

One of the most promising outcomes from Figure 8 is the presence of clear boundaries between the populations. An exception to this is the overlap between the GS/E_1 and GS/E_2 components, which is expected given their shared origin and the somewhat arbitrary nature of their separation. This observation helps validate two key processes of the analysis: firstly, the effectiveness of extreme deconvolution in identifying meaningful structures, and secondly, UMAP's success in preserving these structures in a lower-dimensional embedding.

When considering the cluster's isolation, one of the most interesting insights is the positioning of the Aurora population, which appears between the lower-metallicity GS/E component (GS/E_2) and the Splash population. As `min_dist` decreases (from right to left), Aurora fragments and becomes largely indistinguishable from each of the neighbouring components. This difficulty in resolving Aurora within the embedding space is consistent with its behavior in high-dimensional space, where its Gaussian ellipse regularly overlaps with both Splash and GS/E_2. This overlap is further supported by the assignment probabilities in the APOGEE data: of the 95 stars assigned to Aurora, 49 have Splash, and 22 have GS/E_2 as their second most likely component.

This overlap is also supported by astrophysical reasoning. At its lowest metallicities, Aurora traces some of the earliest star formation in a poorly enriched Milky Way, under conditions similar to those of the GS/E progenitor. However, due to the Milky Way's superior star formation efficiency, the Aurora

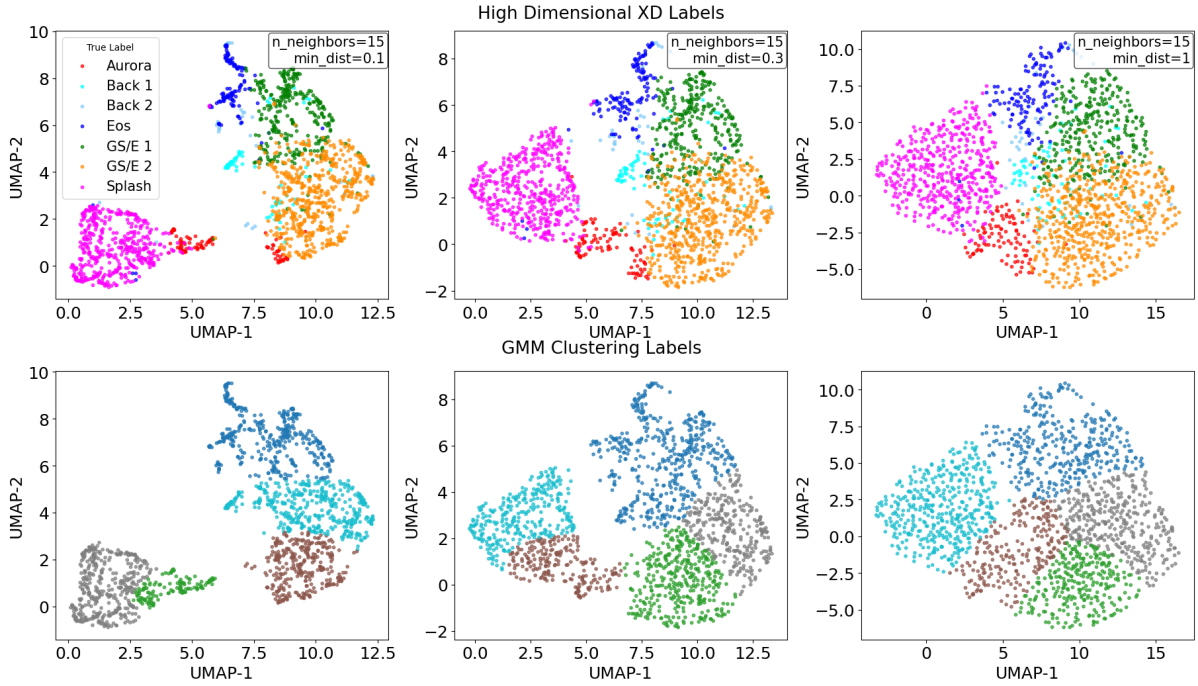


Figure 8: UMAP embeddings of the six-dimensional APOGEE dataset, shown for a range of `min_dist` values. The top row shows points coloured by their original high-dimensional XD cluster assignments. The bottom row shows the results of applying GMM clustering directly to the 2D UMAP-projected data.

population quickly enriches and begins to overlap with the Splash population at a time when the Galactic proto-disk is starting to spin up. Overall this helps explain the BIC score's favoring of a 5 component model, neglecting Aurora. It was also considered whether this fragmentation indicated significant incorrect assignment to Aurora. However, a comparison of the probabilities between the top two assignments shows little evidence to support this interpretation with only three stars have marginal certainty (i.e. a difference in probability $< 50\%$).

In summary, we see that APOGEE's dataset provides relatively poor separation between the clusters, suggesting it struggles to robustly distinguish between them, which is likely a cause of the less stable results achieved. This observation is further supported by the application of GMM to the embedded space, which struggles to reliably re-identify the original structures.

5.2 GALAH in Embedding Space

In comparison, the GALAH dataset demonstrates a much greater ability to isolate these structures, with Aurora consistently shown as a cohesive and independent component across all `min_dist` parameters in Figure 9. The most prominent result of this analysis is the near-perfect success of the GMM in recovering the original structures within all low-dimensional projections, a result that inspires the work presented in section 6. Overall, these findings present further evidence for the advantages offered by the GALAH dataset over APOGEE, with a more detailed discussion of the underlying reasons presented in subsubsection 5.2.1.

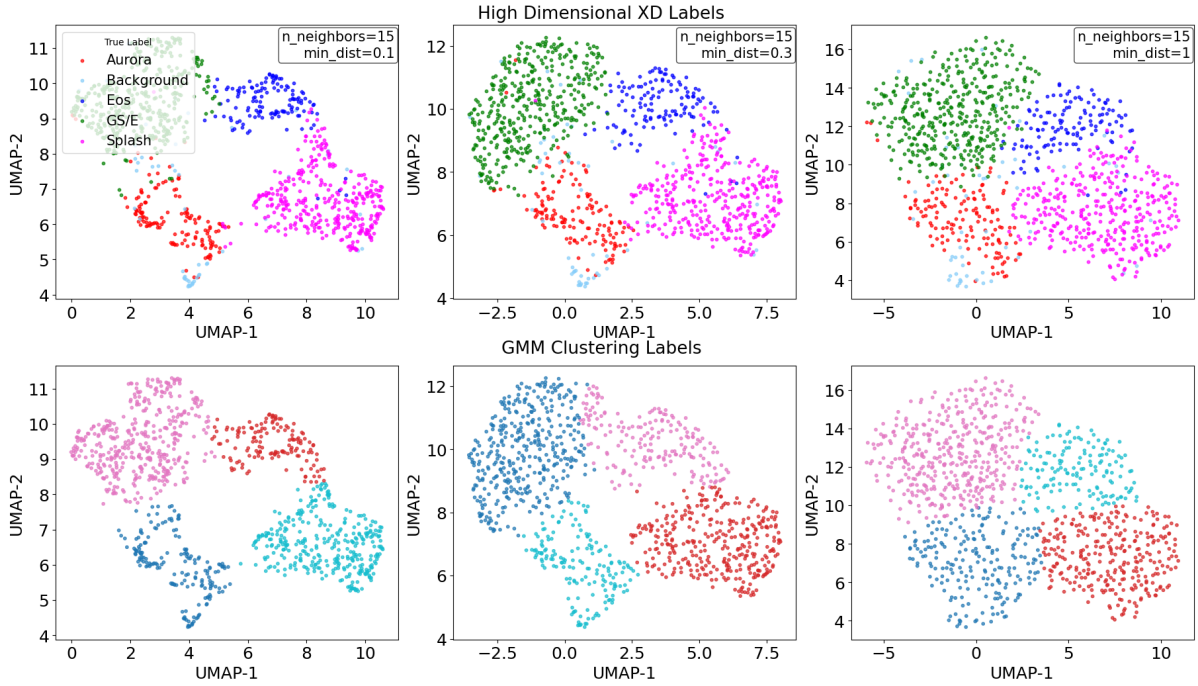


Figure 9: UMAP embeddings of the twelve-dimensional GALAH dataset, shown for a range of `min_dist` values. The top row shows points coloured by their original high-dimensional XD cluster assignments. The bottom row shows the results of applying GMM clustering directly to the 2D UMAP-projected data.

5.2.1 GALAH's Restricted View (6D)

This section aims to better understand the factors leading to GALAH's enhanced ability at isolating stellar populations. As previously suggested, a natural explanation is the increased dimensionality of the GALAH dataset, which enables more detailed nucleosynthetic distinction. However, as discussed in subsection 2.3, apart from the inclusion of the r-process tracer [Eu/Fe], many of the additional dimensions in GALAH trace identical evolutionary processes. Their inclusion therefore simply increases the relative weighting of the corresponding chemical channels in the clustering algorithm rather than introducing new information. If this is responsible for GALAH's improved performance, then a similar effect could be achieved in APOGEE by duplicating these key dimensions. For example, while GALAH incorporates both [Na/Fe] and [Al/Fe], each of which is an odd-Z element, a similar effect could be achieved in APOGEE by including [Al/Fe] twice.

This is investigated by analysing GALAH's view of the structures in an equivalent six-dimensional feature space to that used by APOGEE, comprising [Fe/H], [α/Fe], [Al/Fe], [Y/Fe], [Mg/Mn], and energy, E. Given GALAH's higher measurement uncertainties, smaller sample size, and restricted metallicity range, the UMAP projection of this space (Figure 10) was, a priori, expected to perform worse than that of APOGEE. However, in contradiction, while the lower-dimensional clusters are less distinct than those obtained from the full 12-dimensional GALAH dataset, they still show greater separation and cohesion than those from APOGEE. This is supported by the successful re-identification of the clusters (`min_dist` $= 0$), a result that was not achieved in Figure 8. Notably, five clusters were fitted, two of which fit the GS/E population (a result discussed in section 6). This counterintuitive result suggests that

GALAH's effectiveness may not only be a result of the number of features used but also potentially reflect advantages in the quality of GALAH's data despite previous discussions. Alternatively, a less dominant GS/E weighting may also advantage their identification. This work notes that further investigation is needed to explore this observation in detail.

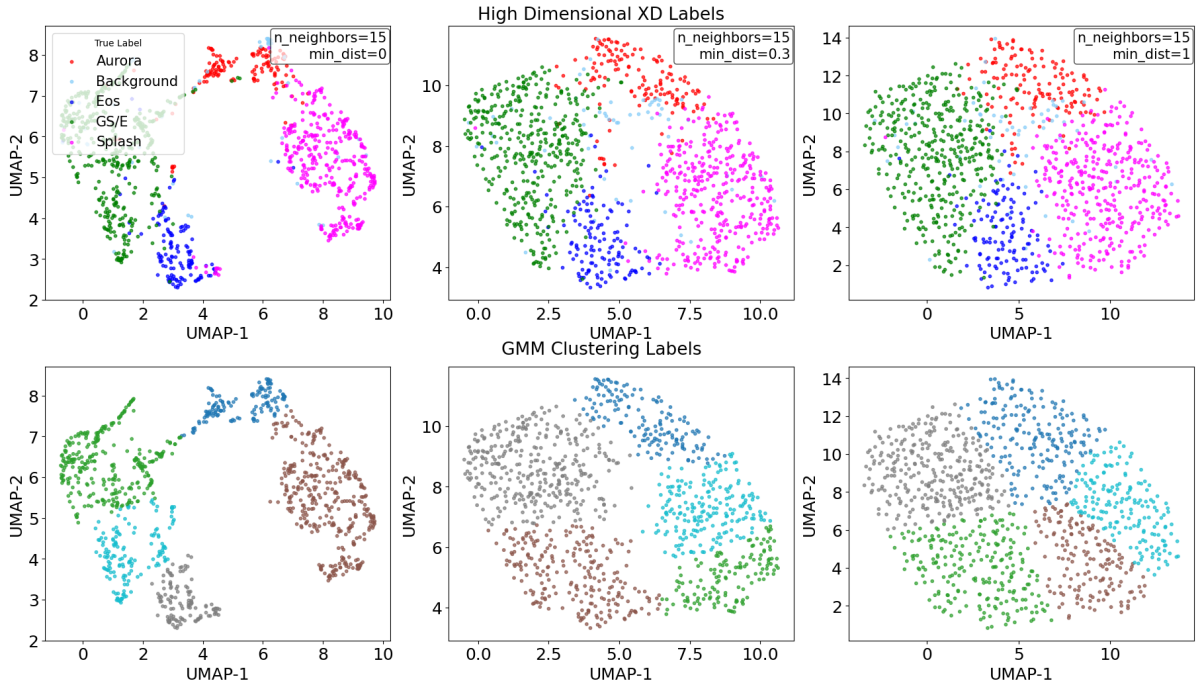


Figure 10: UMAP embeddings of the six-dimensional GALAH dataset, shown for a range of `min_dist` values. The top row shows points coloured by their original high-dimensional XD cluster assignments. The bottom row shows the results of applying GMM clustering directly to the 2D UMAP-projected data.

6 Clustering in Embedding Space

As datasets grow dramatically in both size and dimensionality, with 4MOST promising to observe over 25 million sources [16], the pipeline outlined in section 3 will likely face challenges. In particular, its high sensitivity to random initialisations will worsen and lead to greater computational costs and instability. Ultimately, this unreliable convergence could potentially limit the reproducibility of future studies.

Motivated by the initial success seen in Figure 9, this section explores the feasibility of applying clustering directly to the lower-dimensional embedding space to confront these limitations. Several clustering algorithms were considered, including HDBSCAN [69], due to its ability to automatically determine the number of clusters without prior specification. However, GMMs provided the most promising approach, and are therefore the focus of the results presented in this section.

6.1 Lower Dimensional Pipeline

This work presents an automated pipeline, similar in structure to that presented in subsection 3.2, which supports multiple initialisations, model comparison and cluster assignment to achieve robust results. Since UMAP's projection into embedding space does not support the propagation of measurement uncertainties, the use of `scikit-learn`'s GMM implementation [68] is sufficient. That said, the development of non-linear

dimensionality reduction methods which can incorporate observational errors is suggested as an interesting and valuable direction for future research.

We note that although the embedding space's fitted Gaussians are somewhat arbitrary, these clusters can be mapped back into physically meaningful chemo-dynamical structures through their assignments. For each cluster, a distribution in the original feature space can be recovered by fitting a multivariate Gaussian to its assigned stars. We caveat this with the fact there is no astrophysical justification that these populations will follow a probabilistic Gaussian in embedding space. Furthermore, due to UMAP being a non-linear projection, a Gaussian cluster in the embedding space does not necessarily resemble a Gaussian distribution in the original higher-dimensional space.

To enable greater comparability between the intrinsic structural covariances reported by Extreme Deconvolution in section 4 and the observational (uncertainty-unaware) covariances produced by this pipeline, we apply an approximate deconvolution of the cluster-level covariance matrices. For each structure, the intrinsic covariance matrix is estimated as:

$$\mathbf{\Sigma}_{\text{intr}} = \mathbf{\Sigma}_{\text{obs}} - \langle \mathbf{\Delta} \rangle, \tag{13}$$

where $\mathbf{\Sigma}_{\text{obs}}$ is the achieved covariance matrix, and $\langle \mathbf{\Delta} \rangle$ is the diagonal mean measurement error covariance matrix across all sources in the cluster.

6.2 4 Component Re-identification

To ensure robustness and allow comparability with the previous clustering methods and results, three runs were performed, each with 100 random initialisations for all models with 1 to 10 components. Across all runs, the solutions for each component model were similar, suggesting stable fitting. That being said, we find that neither AIC nor BIC scores reached a minimum, preventing an unbiased selection of the optimal number of components, a result which was possible in the high-dimensional case. In this section we present the results of fitting four components, attempting to recover the four astrophysical populations previously fitted by GALAH (excluding the background).

When comparing the chemo-dynamical distributions of the clusters assigned in Figure 11, shown in Figure 13 and Table 3, with those presented in section 4 and the original study, we find overwhelming consistency. The central advantage of this approach is the significant reduction in computational cost demonstrated by Figure 12. Overall, the pipeline's 100 initialisations of all 10 models are completed in ~ 7 seconds in contrast to the $\sim 1.7 \times 10^4$ second of the high-dimensional XD implementation, marking an overall speed-up factor of ~ 2500[1]. This remarkable decrease in computational cost, paired with the comparable results, not only verifies the validity of this low-dimensional approach but also presents it as a feasible solution to analysing the next generation of large-scale spectroscopic surveys.

Nevertheless, there are some limitations worth mentioning. Primarily, uncertainties in the fitted structures are marginally larger, typically $\sim 29.3\%$ across all dimensions and clusters. This reflects the method's lack of ability to rigorously incorporate measurement uncertainties, relying on a simplified, approximate approach to deconvolution in comparison to the expectation-maximisation algorithm utilised

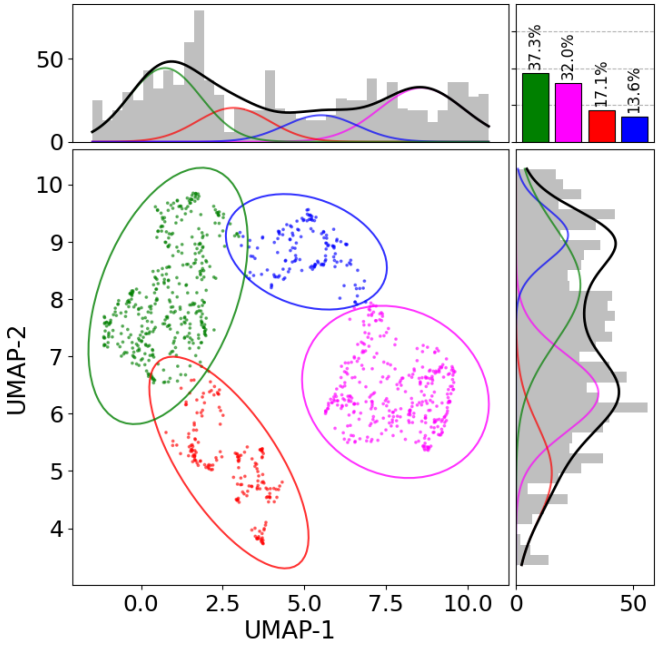


Figure 11: Four component GMM results when applied to UMAP embeddings of the twelve-dimensional GALAH dataset, with `n_neighbors` = 15, `min_dist` = 0.0

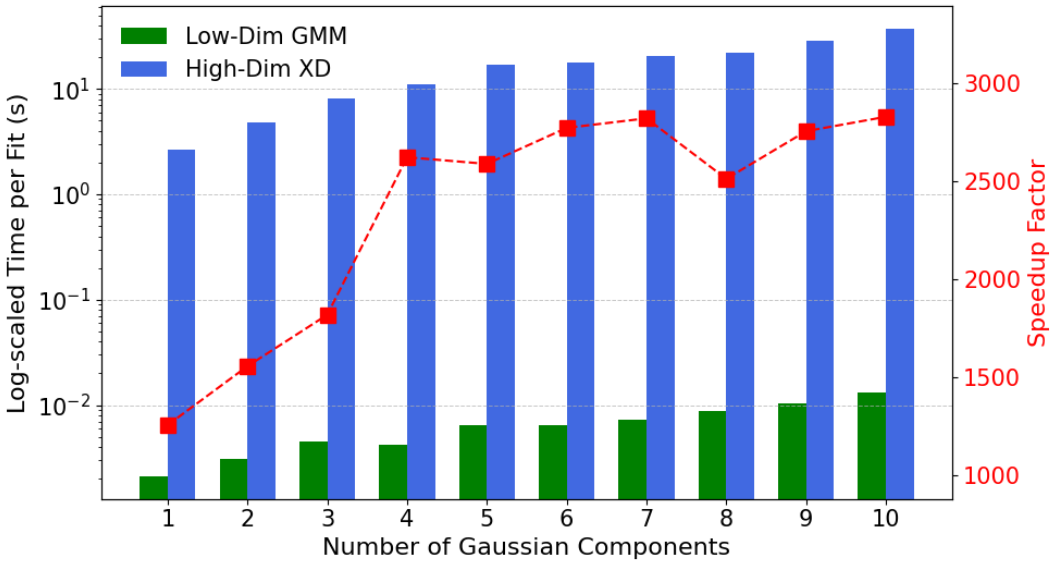


Figure 12: Comparison of runtimes for each model between the high-dimensional Extreme Deconvolution pipeline and the dimensionality-reduced GMM pipeline. Both pipelines used the same 12 chemo-dynamical features from GALAH, with results averaged over 100 initialisations.
[1] Benchmark timings were obtained using a 2020 MacBook Pro with a M1 chip and 16GB of memory.

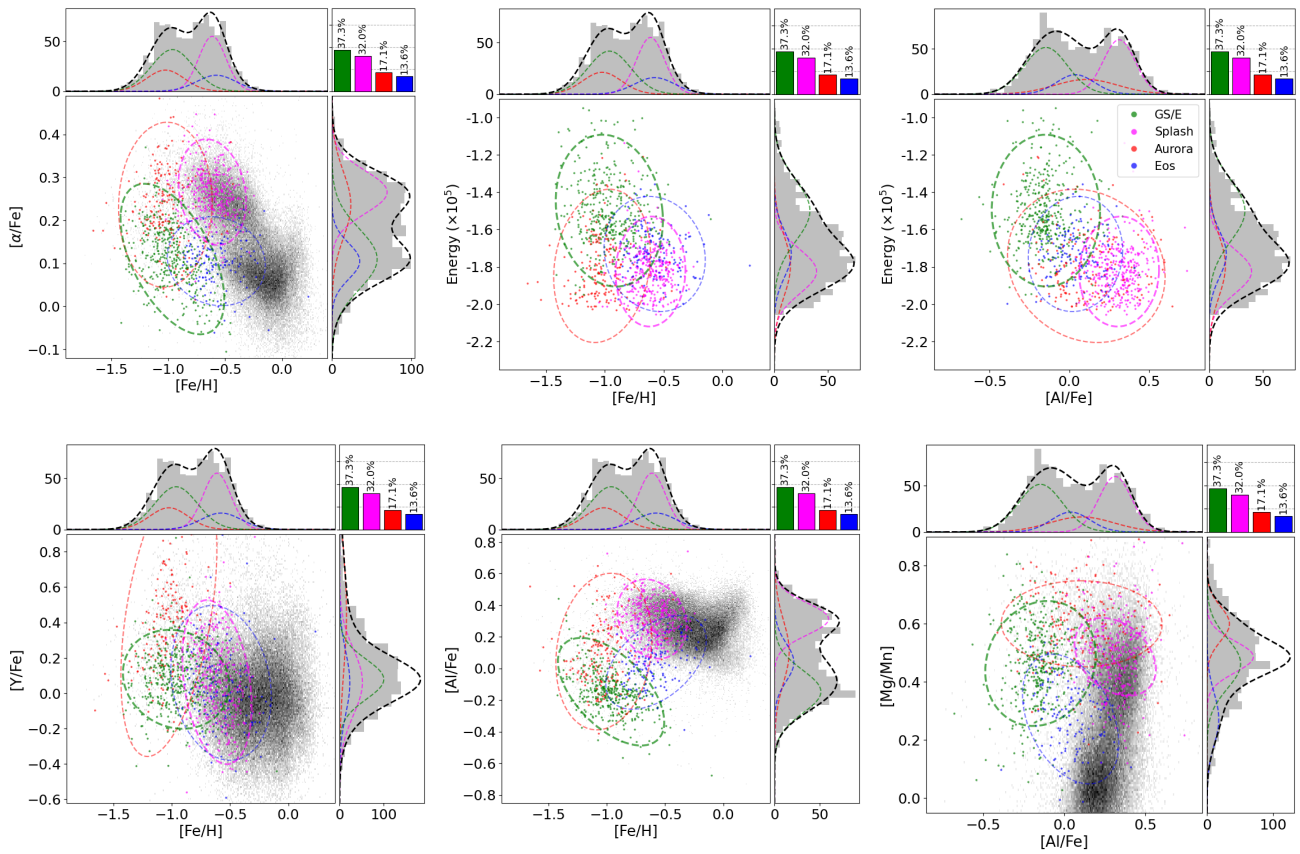


Figure 13: Chemo-dynamical projections of four GMM components to the GALAH–Gaia sample's UMAP embedding space. These projection are fitted using cluster membership. Shown are 2σ ellipses, marginal histograms, and fractional weights, with a greyscale GALAH background for context (only on axis it does not effect visualisation).

by XD. Additionally, in the low dimensional case, the chemo-dynamical distributions are fit using the hard assignments of each cluster rather than weighting all stars by membership probabilities. As a result, the understanding of the distribution is reduced compared to probabilistic assignments used in the direct high-dimensional GMM.

6.3 6 Component Identification

This work also provides evidence that the advantages of clustering in embedding space extends beyond reduced runtime. Specifically, it shows support for the method being more sensitive to subpopulations that Extreme Deconvolution was unable to resolve. As the number of fitted components is increased to six, the models split both the GS/E merger and the Splash cluster into two, with results presented in Table 4. By mapping these clusters back to the original feature space, we assess whether these splits correspond to meaningfully distinct chemo-dynamical phases.

Firstly, focusing on the GS/E merger, we find that the two subpopulations align with the lower-metallicity plateau and higher-metallicity 'knee' in the [α/Fe]–[Fe/H] plane (Figure 14), similar to that achieved in the APOGEE dataset. Although the data's metallicity limit at [Fe/H] $\sim$ -1.5 means these features are less well constrained than in Figure 5, this is a distinction which could not previously be

extracted from GALAH in section 4. As discussed earlier, high dimensional attempts to increase the number of components merely led to the assignment of more background populations.

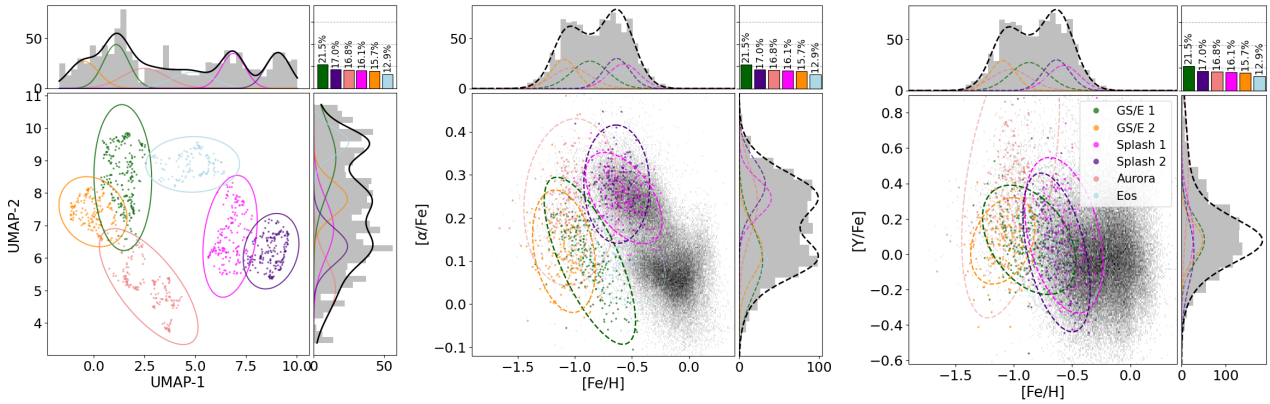


Figure 14: Chemo-dynamical and UMAP projections which highlight the seperation of $GS/E_{1/2}$ in the six GMM components from GALAH–Gaia sample's embedding space. These projection are fitted using cluster membership. Shown are 2σ ellipses, marginal histograms, and fractional weights, with a greyscale GALAH background. The colours of the Aurora and Eos have been dimmed to help increase visibility differences between subpopulations.

Secondly, the division of the Splash population is a distinction not identified elsewhere in this work or that presented by Myeong et al. [1]. While the two sub-clusters, labeled $Splash_1$ and $Splash_2$, are similar in many dimensions, supporting their common origin, subtle differences exist. As shown in Figure 15, $Splash_1$ displays higher [Al/Fe] and [Eu/Fe] abundances compared to $Splash_2$, but lower [Ba/Fe]. These unique chemical properties suggest potential subtle differences in their historical enrichment environments, such as contributions from core-collapse supernovae ([Al/Fe]), r-process events ([Eu/Fe]), and asymptotic giant branch stars ([Ba/Fe]). However, this is complicated by correlated dimensions not showing similar variations, such as [Y/Fe]. This raises the possibility that the separation is simply the result of a less informative structure in the embedding space. Overall, these potential evolutionary differences require further investigation in future work.

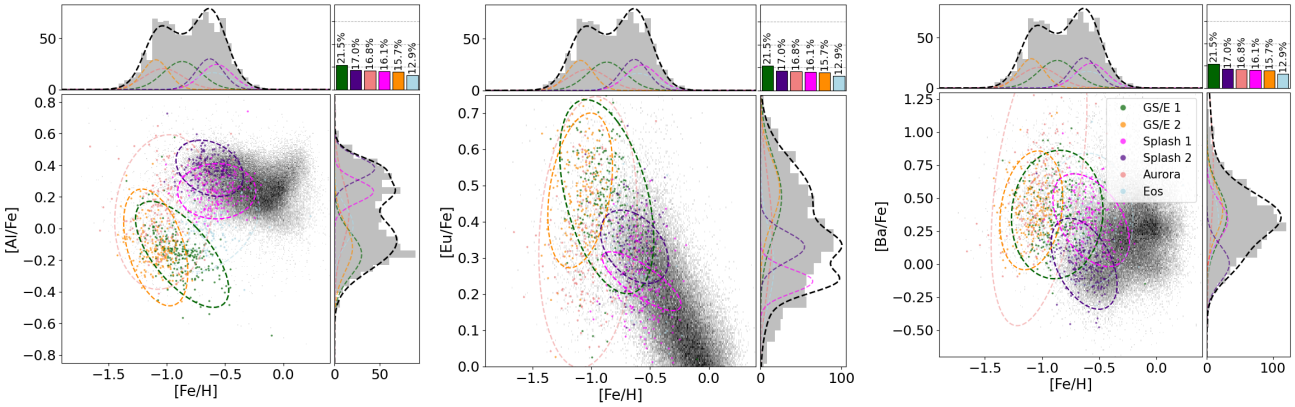


Figure 15: Chemo-dynamical projections which highlight the seperation of $Splash_{1/2}$ in the six GMM components from GALAH–Gaia sample's UMAP embedding space. These projection are fitted using cluster membership. Shown are 2σ ellipses, marginal histograms, and fractional weights, with a greyscale GALAH background. The colours of the Aurora and Eos have been dimmed to help increase visibility differences between subpopulations.

Table 3: Summary of Results from the 4 Component GMM Fit on GALAH-Gaia Sample's Low-Dimensional Embedding Space. The distribution uncertainties presented have been approximately deconvolved from the observational uncertainties.

Component	Weight (%)	Count	[Fe/H]	[α/Fe]	[Na/Fe]	[Al/Fe]	[Mn/Fe]	[Y/Fe]	[Ba/Fe]	[Eu/Fe]	[Mg/Cu]	[Mg/Mn]	[Ba/Eu]	Energy (10^5 km^2 s^{-2})
GS/E	37.3	399	-0.96 ± 0.18	0.11 ± 0.07	-0.27 ± 0.10	-0.15 ± 0.14	-0.36 ± 0.10	0.08 ± 0.11	0.39 ± 0.19	0.47 ± 0.10	0.61 ± 0.03	0.46 ± 0.09	-0.09 ± 0.21	-1.50 ± 0.16
Splash	32.0	340	-0.61 ± 0.12	0.27 ± 0.05	0.12 ± 0.08	0.31 ± 0.10	-0.18 ± 0.07	0.06 ± 0.19	0.16 ± 0.22	0.29 ± 0.06	0.29 ± 0.11	0.49 ± 0.05	-0.12 ± 0.23	-1.83 ± 0.12
Aurora	17.1	179	-1.02 ± 0.16	0.24 ± 0.08	-0.05 ± 0.15	0.11 ± 0.20	-0.34 ± 0.08	0.46 ± 0.33	0.68 ± 0.46	0.37 ± 0.15	0.59 ± 0.09	0.60 ± 0.06	0.31 ± 0.39	-1.79 ± 0.16
Eos	13.6	143	-0.58 ± 0.17	0.11 ± 0.04	-0.05 ± 0.10	0.04 ± 0.12	-0.17 ± 0.10	0.06 ± 0.18	0.33 ± 0.19	0.27 ± 0.10	0.26 ± 0.13	0.27 ± 0.09	0.06 ± 0.23	-1.73 ± 0.12

Table 4: Summary of Results from the 6 Component GMM Fit on GALAH-Gaia Sample's Low-Dimensional Embedding Space. The distribution uncertainties presented have been approximately deconvolved from the observational uncertainties.

Component	Weight (%)	Count	[Fe/H]	[α/Fe]	[Na/Fe]	[Al/Fe]	[Mn/Fe]	[Y/Fe]	[Ba/Fe]	[Eu/Fe]	[Mg/Cu]	[Mg/Mn]	[Ba/Eu]	Energy (10^5 km^2 s^{-2})
GS/E$_1$	21.5	226	-0.87 ± 0.16	0.10 ± 0.08	-0.26 ± 0.12	-0.16 ± 0.13	-0.30 ± 0.07	0.08 ± 0.12	0.38 ± 0.20	0.47 ± 0.11	0.58 ± 0.02	0.38 ± 0.03	-0.09 ± 0.24	-1.52 ± 0.16
GS/E$_2$	15.7	169	-1.09 ± 0.11	0.12 ± 0.06	-0.27 ± 0.09	-0.12 ± 0.15	-0.44 ± 0.05	0.08 ± 0.10	0.41 ± 0.18	0.49 ± 0.09	0.66 ± 0.01	0.57 ± 0.01	-0.08 ± 0.16	-1.46 ± 0.17
Splash$_1$	16.1	171	-0.57 ± 0.13	0.25 ± 0.04	0.09 ± 0.09	0.24 ± 0.07	-0.17 ± 0.05	0.11 ± 0.18	0.33 ± 0.14	0.24 ± 0.03	0.36 ± 0.09	0.49 ± 0.05	0.10 ± 0.09	-1.85 ± 0.11
Splash$_2$	17.0	180	-0.63 ± 0.12	0.28 ± 0.06	0.14 ± 0.06	0.38 ± 0.07	-0.19 ± 0.08	0.01 ± 0.18	0.00 ± 0.14	0.33 ± 0.04	0.22 ± 0.07	0.48 ± 0.06	-0.33 ± 0.09	-1.80 ± 0.13
Aurora	16.8	178	-1.03 ± 0.17	0.24 ± 0.08	-0.05 ± 0.15	0.10 ± 0.20	-0.34 ± 0.08	0.46 ± 0.33	0.68 ± 0.46	0.37 ± 0.15	0.59 ± 0.09	0.60 ± 0.05	0.31 ± 0.39	-1.80 ± 0.16
Eos	12.9	137	-0.60 ± 0.16	0.10 ± 0.04	-0.06 ± 0.11	0.02 ± 0.12	-0.18 ± 0.11	0.07 ± 0.17	0.34 ± 0.20	0.28 ± 0.10	0.28 ± 0.14	0.27 ± 0.09	0.06 ± 0.23	-1.73 ± 0.12
GS/E$_{tot}$	37.2	395	-0.96 ± 0.18	0.11 ± 0.07	-0.26 ± 0.11	-0.15 ± 0.14	-0.36 ± 0.10	0.08 ± 0.11	0.39 ± 0.19	0.48 ± 0.10	0.61 ± 0.03	0.46 ± 0.09	-0.09 ± 0.21	-1.50 ± 0.16
Splash$_{tot}$	33.1	351	-0.60 ± 0.13	0.26 ± 0.05	0.11 ± 0.08	0.31 ± 0.10	-0.18 ± 0.06	0.06 ± 0.18	0.16 ± 0.22	0.28 ± 0.06	0.29 ± 0.11	0.48 ± 0.06	-0.12 ± 0.23	-1.82 ± 0.12

6.4 Application in Future Work

Taken together, this work suggests a hybrid approach to future analyses. To reduce computational cost and instabilities, low-dimensional clustering is proposed as an initial step, in which approximate high-dimensional Gaussian distributions can be fit to the stellar populations. However, to more rigorously constrain the intrinsic uncertainty of the underlying structures, these Gaussians, with slight random perturbations, can then be used as initialisations for the Extreme Deconvolution algorithm. This would allow for a more robust consideration of measurement uncertainties. This combined strategy aims to offer a balance between computational efficiency, convergence stability, and accuracy, making it a promising pipeline for ensuring the reproducibility of results from increasingly large datasets.

7 Summary

This study presents a comparative analysis of the chemo-dynamical substructures in the Milky Way's halo revealed by APOGEE and GALAH's high eccentricity samples. Myeong et al. [1] was among the first to exploit Gaussian Mixture Models, specifically Extreme Deconvolution, to provide an unbiased decomposition of the stellar halo, which led to the discovery of the previously unresolved stellar population dubbed 'Eos' and provided evidence which disputed the suggested accreted origin of the Hercules 'merger'. Given these findings' significance, the present work aims to verify the reproducibility of Myeong et al. [1]'s results and thus the conclusions it drew. It further built on the original work's methodology to improve the efficiency and stability of the unsupervised clustering techniques in high-dimensional space. The core findings can be summarised as follows:

1. We find strong support for the original work's claim of being unbiased, objective and reproducible. This is achieved by recovering the same four stellar populations with near identical distributions for both APOGEE and GALAH to those presented in the original work. This study differs from the original work by employing stricter data cuts and the implementation of an improved feature scaling scheme to improve the EM algorithm's convergence.

2. Through the comparisons of UMAP projections we conclude that GALAH's high dimensionality enables greater isolation of halo substructures and hence improved clustering stability compared to APOGEE. This advantage persists, albeit less prominently, when the feature space is restricted to an equivalent six-dimensions ([Fe/H], [α/Fe], [Al/Fe], [Y/Fe], [Mg/Mn], E). This contradicts a priori discussions focused on GALAH's higher measurement uncertainties and restricted metallicity range and instead suggests GALAH may have advantages beyond its dimensionality. One possibility, supported by Casali et al. [59], is the potential underestimation of uncertainties in APOGEE.

3. We demonstrate the success of applying clustering directly within UMAP's embedding space from the GALAH dataset, achieving results comparable to that of the high dimensional analysis but in a fraction (0.04%) of the time. Additionally, further advantages are shown such as enhanced identification of chemically distinct sub-populations, including the distinction between the GS/E's [α/Fe]–[Fe/H] plateau and knee as well as the suggested decomposition of the Splash component. This

work's ability to propose a new implementation of the analysis which is more stable, computationally efficient, and sensitive highlights its promise for future applications.

There are two non-intuitive findings from this paper that require deeper exploration. Firstly, the suggestion that APOGEE may have some underlying disadvantage relative to GALAH for the subset of chemo-dynamical information used. Secondly, inquiring whether a clear astrophysical distinction exists between the two Splash subpopulations identified, requiring further cross-referencing with numerical simulations such as Auriga [20], VINTERGATAN [52], and ARTEMIS [70].

Looking ahead, this paper proposes multiple approaches for further developing the methodologies presented here for application in the upcoming large-scale spectroscopic surveys, such as 4MOST and WEAVE. Firstly, a hybrid pipeline is suggested. This would leverage low-dimensional clustering for its improved speed and convergence, followed by uncertainty-aware high-dimensional clustering to further refine the underlying substructures. Secondly, a more advanced direction would be the development of a non-linear dimensionality reduction technique capable of incorporating measurement uncertainties, enabling algorithms such as Extreme Deconvolution to be applied directly within the embedding space.

8 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 assisted in:

- 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.
- Correcting spelling, punctuation and grammar inconsistencies within the report.
- Offering more concise phrasing to help reduce word count.

References

[1] 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].

[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] 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.

[4] James S. Bullock and Kathryn V. Johnston. "Tracing Galaxy Formation with Stellar Halos. I. Methods". In: *The Astrophysical Journal* 635.2 (2005), pp. 931–949. DOI: 10.1086/497422.

[5] L. Calçada. *Hierarchical Galaxy Formation Illustration*. https://www.eso.org/public/images/1016-galaxy_formation_merger/. European Southern Observatory. 2023.

[6] O. J. Eggen, D. Lynden-Bell, and A. R. Sandage. "Evidence from the motions of old stars that the Galaxy collapsed." In: 136 (Nov. 1962), p. 748. DOI: 10.1086/147433.

[7] J. Norris, M. S. Bessell, and A. J. Pickles. "Population studies. I. The Bidelman-MacConnell "weak-metal" stars." In: 58 (July 1985), pp. 463–492. DOI: 10.1086/191049.

[8] Masashi Chiba and Timothy C. Beers. "Kinematics of Metal-poor Stars in the Galaxy. III. Formation of the Stellar Halo and Thick Disk as Revealed from a Large Sample of Nonkinematically Selected Stars". In: 119.6 (June 2000), pp. 2843–2865. DOI: 10.1086/301409. arXiv: astro-ph/0003087 [astro-ph].

[9] 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.

[10] Anthony G. A. Brown. *Gaia: Ten Years of Surveying the Milky Way and Beyond*. 2025. arXiv: 2503.01533 [astro-ph.GA]. URL: https://arxiv.org/abs/2503.01533.

[11] 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.

[12] Rimoldini, Lorenzo et al. "Gaia Data Release 3 - All-sky classification of 12.4 million variable sources into 25 classes". In: *AA* 674 (2023), A14. DOI: 10.1051/0004-6361/202245591. URL: https://doi.org/10.1051/0004-6361/202245591.

[13] 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.

[14] 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. URL: http://dx.doi.org/10.1093/mnras/stab1242.

[15] 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. URL: http://dx.doi.org/10.1093/mnras/stad557.

[16] 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. URL: http://dx.doi.org/10.1117/12.926239.

[17] Roelof S. De Jong et al. "4MOST: Project overview and information for the First Call for Proposals". In: *Published in The Messenger vol. 175* pp. 3-11 (2019), March 2019. DOI: 10.18727/0722-6691/5117. URL: https://doi.eso.org/10.18727/0722-6691/5117.

[18] E. L. House et al. "Disc heating: comparing the Milky Way with cosmological simulations: Disc heating". In: *Monthly Notices of the Royal Astronomical Society* 415.3 (June 2011), pp. 2652–2664. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2011.18891.x. URL: http://dx.doi.org/10.1111/j.1365-2966.2011.18891.x.

[19] Joop Schaye et al. "The EAGLE project: simulating the evolution and assembly of galaxies and their environments". In: *Monthly Notices of the Royal Astronomical Society* 446.1 (Nov. 2014), pp. 521–554. ISSN: 0035-8711. DOI: 10.1093/mnras/stu2058. URL: http://dx.doi.org/10.1093/mnras/stu2058.

[20] Robert J. J. Grand et al. *Overview and public data release of the augmented Auriga Project: cosmological simulations of dwarf and Milky Way-mass galaxies*. 2024. arXiv: 2401.08750 [astro-ph.GA]. URL: https://arxiv.org/abs/2401.08750.

[21] Andrew R. Wetzel et al. "Reconciling Dwarf Galaxies With CDM Cosmology: Simulating a Realistic Population of Satellites Around a Milky Way–Mass Galaxy". In: *The Astrophysical Journal Letters* 827.2 (Aug. 2016), p. L23. ISSN: 2041-8213. DOI: 10.3847/2041-8205/827/2/l23. URL: http://dx.doi.org/10.3847/2041-8205/827/2/L23.

[22] Andrea Sante et al. "Applying machine learning to Galactic Archaeology: how well can we recover the origin of stars in Milky Way-like galaxies?" In: *Monthly Notices of the Royal Astronomical Society* 531.4 (June 2024), pp. 4363–4382. ISSN: 1365-2966. DOI: 10.1093/mnras/stae1398. URL: http://dx.doi.org/10.1093/mnras/stae1398.

[23] G C Myeong et al. "Halo substructure in the SDSS–Gaia catalogue: streams and clumps". In: *Monthly Notices of the Royal Astronomical Society* 475.2 (Dec. 2017), pp. 1537–1548. ISSN: 1365-2966. DOI: 10.1093/mnras/stx3262. URL: http://dx.doi.org/10.1093/mnras/stx3262.

[24] G. C. Myeong et al. *The Shards of ω Centauri*. 2018. arXiv: 1804.07050 [astro-ph.GA]. URL: https://arxiv.org/abs/1804.07050.

[25] G. C. Myeong et al. "The Milky Way Halo in Action Space". In: *The Astrophysical Journal Letters* 856.2 (Mar. 2018), p. L26. ISSN: 2041-8213. DOI: 10.3847/2041-8213/aab613. URL: http://dx.doi.org/10.3847/2041-8213/aab613.

[26] Helmer Koppelman, Amina Helmi, and Jovan Veljanoski. "One Large Blob and Many Streams Frosting the nearby Stellar Halo in Gaia DR2". In: *The Astrophysical Journal Letters* 860.1 (June 2018), p. L11. ISSN: 2041-8213. DOI: 10.3847/2041-8213/aac882. URL: http://dx.doi.org/10.3847/2041-8213/aac882.

[27] Helmer H. Koppelman et al. "Multiple retrograde substructures in the Galactic halo: A shattered view of Galactic history". In: *Astronomy amp; Astrophysics* 631 (Nov. 2019), p. L9. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201936738. URL: http://dx.doi.org/10.1051/0004-6361/201936738.

[28] 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. URL: http://dx.doi.org/10.1093/mnras/sty982.

[29] 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. URL: http://dx.doi.org/10.1038/s41586-018-0625-x.

[30] Jeremy Tinker et al. "Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality". In: *The Astrophysical Journal* 688.2 (Dec. 2008), pp. 709–728. ISSN: 1538-4357. DOI: 10.1086/591439. URL: http://dx.doi.org/10.1086/591439.

[31] Onsi Fakhouri, Chung-Pei Ma, and Michael Boylan-Kolchin. "The merger rates and mass assembly histories of dark matter haloes in the two Millennium simulations: Merger rates". In: *Monthly Notices of the Royal Astronomical Society* 406.4 (June 2010), pp. 2267–2278. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2010.16859.x. URL: http://dx.doi.org/10.1111/j.1365-2966.2010.16859.x.

[32] R. A. Ibata, G. Gilmore, and M. J. Irwin. "Sagittarius: the nearest dwarf galaxy". In: *Monthly Notices of the Royal Astronomical Society* 277.3 (Dec. 1995), pp. 781–800. ISSN: 1365-2966. DOI: 10.1093/mnras/277.3.781. URL: http://dx.doi.org/10.1093/mnras/277.3.781.

[33] 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.

[34] P. Wannier and G. T. Wrixon. "An Unusual High-Velocity Hydrogen Feature". In: 173 (May 1972), p. L119. DOI: 10.1086/180930.

[35] D. S. Mathewson, M. N. Cleary, and J. D. Murray. "The Magellanic Stream." In: 190 (June 1974), pp. 291–296. DOI: 10.1086/152875.

[36] Thor Tepper-García et al. "The Magellanic System: the puzzle of the leading gas stream". In: *Monthly Notices of the Royal Astronomical Society* 488.1 (June 2019), pp. 918–938. ISSN: 1365-2966. DOI: 10.1093/mnras/stz1659. URL: http://dx.doi.org/10.1093/mnras/stz1659.

[37] S. L. J. Gibbons, V. Belokurov, and N. W. Evans. "A tail of two populations: chemo-dynamics of the Sagittarius stream and implications for its original mass". In: *Monthly Notices of the Royal Astronomical Society* 464.1 (Sept. 2016), pp. 794–809. ISSN: 1365-2966. DOI: 10.1093/mnras/stw2328. URL: http://dx.doi.org/10.1093/mnras/stw2328.

[38] D Erkal et al. "The total mass of the Large Magellanic Cloud from its perturbation on the Orphan stream". In: *Monthly Notices of the Royal Astronomical Society* 487.2 (May 2019), pp. 2685–2700. ISSN: 1365-2966. DOI: 10.1093/mnras/stz1371. URL: http://dx.doi.org/10.1093/mnras/stz1371.

[39] Danny Horta et al. "Evidence from APOGEE for the presence of a major building block of the halo buried in the inner Galaxy". In: 500.1 (Jan. 2021), pp. 1385–1403. DOI: 10.1093/mnras/staa2987. arXiv: 2007.10374 [astro-ph.GA].

[40] J M Diederik Kruijssen et al. "The formation and assembly history of the Milky Way revealed by its globular cluster population". In: *Monthly Notices of the Royal Astronomical Society* 486.3 (June 2018), pp. 3180–3202. ISSN: 1365-2966. DOI: 10.1093/mnras/sty1609. URL: http://dx.doi.org/10.1093/mnras/sty1609.

[41] Duncan A Forbes. "Reverse engineering the Milky Way". In: *Monthly Notices of the Royal Astronomical Society* 493.1 (Jan. 2020), pp. 847–854. ISSN: 1365-2966. DOI: 10.1093/mnras/staa245. URL: http://dx.doi.org/10.1093/mnras/staa245.

[42] Rohan P. Naidu et al. "Evidence from Disrupted Halo Dwarfs that r-process Enrichment via Neutron Star Mergers is Delayed by 500 Myr". In: *The Astrophysical Journal Letters* 926.2 (Feb. 2022), p. L36. ISSN: 2041-8213. DOI: 10.3847/2041-8213/ac5589. URL: http://dx.doi.org/10.3847/2041-8213/ac5589.

[43] Rohan P. Naidu et al. "Reconstructing the Last Major Merger of the Milky Way with the H3 Survey". In: *The Astrophysical Journal* 923.1 (Dec. 2021), p. 92. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ac2d2d. URL: http://dx.doi.org/10.3847/1538-4357/ac2d2d.

[44] Giuliano Iorio and Vasily Belokurov. "The shape of the Galactic halo withGaiaDR2 RR Lyrae. Anatomy of an ancient major merger". In: *Monthly Notices of the Royal Astronomical Society* 482.3 (Oct. 2018), pp. 3868–3879. ISSN: 1365-2966. DOI: 10.1093/mnras/sty2806. URL: http://dx.doi.org/10.1093/mnras/sty2806.

[45] A. J. Deason et al. "BROKEN AND UNBROKEN: THE MILKY WAY AND M31 STELLAR HALOS". In: *The Astrophysical Journal* 763.2 (Jan. 2013), p. 113. ISSN: 1538-4357. DOI: 10.1088/0004-637x/763/2/113. URL: http://dx.doi.org/10.1088/0004-637X/763/2/113.

[46] Alis J. Deason et al. "Apocenter Pile-up: Origin of the Stellar Halo Density Break". In: *The Astrophysical Journal Letters* 862.1 (July 2018), p. L1. DOI: 10.3847/2041-8213/aad0ee. URL: https://dx.doi.org/10.3847/2041-8213/aad0ee.

[47] M. Haywood et al. "In Disguise or Out of Reach: First Clues about In Situ and Accreted Stars in the Stellar Halo of the Milky Way from Gaia DR2". In: 863.2, 113 (Aug. 2018), p. 113. DOI: 10.3847/1538-4357/aad235. arXiv: 1805.02617 [astro-ph.GA].

[48] P. Di Matteo et al. "The Milky Way has no in-situ halo other than the heated thick disc: Composition of the stellar halo and age-dating the last significant merger with Gaia DR2 and APOGEE". In: *Astronomy amp; Astrophysics* 632 (Nov. 2019), A4. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201834929. URL: http://dx.doi.org/10.1051/0004-6361/201834929.

[49] Carme Gallart et al. "Uncovering the birth of the Milky Way through accurate stellar ages with Gaia". In: *Nature Astronomy* 3.10 (July 2019), pp. 932–939. ISSN: 2397-3366. DOI: 10.1038/s41550-019-0829-5. URL: http://dx.doi.org/10.1038/s41550-019-0829-5.

[50] 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].

[51] Robert J. J. Grand et al. "The dual origin of the Galactic thick disc and halo from the gas-rich Gaia-Enceladus Sausage merger". In: 497.2 (Sept. 2020), pp. 1603–1618. DOI: 10.1093/mnras/staa2057. arXiv: 2001.06009 [astro-ph.GA].

[52] Florent Renaud et al. "VINTERGATAN III: how to reset the metallicity of the Milky Way". In: 503.4 (June 2021), pp. 5868–5876. DOI: 10.1093/mnras/stab543. arXiv: 2006.06012 [astro-ph.GA].

[53] 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. URL: http://dx.doi.org/10.1093/mnras/stac1267.

[54] 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.

[55] 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.

[56] Paul J. McMillan. "The mass distribution and gravitational potential of the Milky Way". In: *Monthly Notices of the Royal Astronomical Society* 465.1 (Oct. 2016), pp. 76–94. ISSN: 1365-2966. DOI: 10.1093/mnras/stw2759. URL: http://dx.doi.org/10.1093/mnras/stw2759.

[57] Ana E. García Pérez et al. "ASPCAP: The APOGEE Stellar Parameter and Chemical Abundances Pipeline". In: 151.6, 144 (June 2016), p. 144. DOI: 10.3847/0004-6256/151/6/144. arXiv: 1510.07635 [astro-ph.SR].

[58] K. Hawkins et al. "Using chemical tagging to redefine the interface of the Galactic disc and halo". In: *Monthly Notices of the Royal Astronomical Society* 453.1 (Aug. 2015), pp. 758–774. ISSN: 1365-2966. DOI: 10.1093/mnras/stv1586. URL: http://dx.doi.org/10.1093/mnras/stv1586.

[59] G. Casali et al. "Time evolution of Ce as traced by APOGEE using giant stars observed with the Kepler, TESS and K2 missions". In: 677, A60 (Sept. 2023), A60. DOI: 10.1051/0004-6361/202346274. arXiv: 2305.06396 [astro-ph.GA].

[60] Thibault Boulet. "A catalogue of asteroseismically calibrated ages for APOGEE DR17: The predictions of a CatBoost machine learning model based on the [Mg/Ce] chemical clock and other stellar parameters". In: *Astronomy amp; Astrophysics* 685 (May 2024), A66. ISSN: 1432-0746. DOI: 10.1051/0004-6361/202348031. URL: http://dx.doi.org/10.1051/0004-6361/202348031.

[61] 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.

[62] Hirotugu Akaike. "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[63] Gideon Schwarz. "Estimating the Dimension of a Model". In: *Annals of Statistics* 6.2 (July 1978), pp. 461–464.

[64] Galit Shmueli. "To Explain or to Predict?" In: *Statistical Science* 25.3 (Aug. 2010), pp. 289–310. URL: https://www.jstor.org/stable/41058949.

[65] Elliott Sober. "Instrumentalism, Parsimony, and the Akaike Framework". In: *Philosophy of Science* 69.S3 (2002), pp. 112–123. DOI: 10.1086/341839.

[66] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[67] 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.

[68] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[69] 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.

[70] Andreea S Font et al. "The artemis simulations: stellar haloes of Milky Way-mass galaxies". In: *Monthly Notices of the Royal Astronomical Society* 498.2 (Aug. 2020), pp. 1765–1785. ISSN: 1365-2966. DOI: 10.1093/mnras/staa2463. URL: http://dx.doi.org/10.1093/mnras/staa2463.