Clustering Methods for Studying Galactic Components

Anna Parul, Jeremy Quijano (team f)

April 29, 2021

1 Introduction

One of the ways to unveil the history of a galaxy is to study the chemical composition of stars. Hydrogen and Helium, two of the most abundant elements in the Universe, along with small amount of Lithium and Deuterium originate from primordial nucleosynthesis whilst metals (all elements heavier than He are called *metals* in astrophysics) are produced in stellar interiors and are deposited into the interstellar medium (ISM) through supernova explosions and stellar winds. The chemical information about the ISM is preserved in the atmospheres of the stars that were born from it therefore stars formed from the same patch of ISM should have a similar set of the chemical signatures and should be clustered in chemical space.

The idea of chemical tagging — grouping together stars with similar chemical properties was introduced in [1]. "Strong chemical tagging" seeks to identify stars born from the same gas cloud. Since most of the open clusters dissolve on timescales of ∼100 Myr studying of present-day spatial distribution of stars born together should tell us a history of radial migrations and other dynamical processes and could be used to adjust dynamical models. Weak limit of chemical tagging works with larger groups of stars and can be used, for example, to decompose disk into thick and thin part ([2]) or, with the use of kinematic data, to study structure of stellar halo ([3]) and identify stellar streams ([4])

Modern spectroscopic surveys like APOGEE, RAVE, LAMOST and GALAH provide an enormous amount of high-precision measurements of chemical composition of stars in the Milky Way. Dissecting the multidimensional chemical space and identifying chemically similar groups of stars is the task well suitable for machine learning clustering algorithms. Price-Jones et al. in [5] applied density-based clustering algorithm (DBSCAN) to 8D-chemical space to detect star cluster candidates that dissolved across the Milky Way disc, Anders et al. in [6] with the help of t-SNE separated disk components and identified chemically peculiar stars that might have originated from a dwarf satellite galaxy.

In this project we plan to use several clustering algorithms (t-SNE, DBSCAN, UMAP) on galaxies from cosmological simulation to compare their performance and study limits of clustering in chemical space

2 Dataset

In our work we used galaxy **m12f** from Latte suite. The Latte suite of FIRE-2 cosmological zoom-in baryonic simulations of Milky Way-mass galaxies ([7]), part of the Feedback In Realistic Environments (FIRE) simulation project, were run using the Gizmo gravity plus hydrodynamics code in meshless finite-mass (MFM) mode ([8]) and the FIRE-2 physics model ([9]). FIRE-2 reproduces realistic gas physics through including free-free emission, photoionization/recombination, Compton scattering, metal-line, molecular, fine structure, dust-collisional processes, and cosmic-ray heating. It implements detailed simulation of stellar feedback from supernovae of Type Ia and II and mass loss from OB/AGB stars. Star formation occurs in dense (n > 1000), self-gravitating, Jeans-unstable molecular gas with 100% efficiency per free-fall time. Every star particle is considered as a simple stellar population with Kroupa IMF which evolves along stellar population models from STARBURST99. The simulation traces 11 elements (H, He, C, N, O, Ne, Mg, Si, S, Ca, Fe), despite the fact that this number is big for simulation this set of elements is smaller than the set used in works on chemical tagging performed on real data. Also the fact that each star particle represent the whole population means that we cannot aim to perform a strong version of chemical tagging.

We consider two samples:

A) *stars born inside 20 kpc region around center of host galaxy ($R_{birth} < 20$ kpc).* This is smaller sample which focuses on the host galaxy only, here we expect to recover thick and thin disk of galaxy, stellar halo, bulge and, hopefully, smaller subpopulations.

B) *stars born in the spherical shell between 30 kpc and 500 kpc from the host center.* This sample contains satellites, stellar streams and accreted stars brought in by minor mergers.

3 Parameter space

Common technique used to separate galactic components in chemical space is to study stellar distribution on the plane [Fe/H]-[X/Fe] [1]. Such diagrams illustrate many element behaviour patterns, for example, the decrease of [α/Fe] for higher values of [Fe/H] which is associated with increasing deposit of iron from supernovae type Ia at later stages of galaxy evolution. However Ting in [10] showed with Principal Component Analysis that dimensionality of chemical space in a Solar neighbourhood is about 8 to 9. Anders et al. ([6]) used combinations of 13 elements (Mg, Al, Si, Ca, TiI, Fe, Cu, Zn, Sr, T, ZrII, Ce, Ba) for clustering with t-SNE, Price-Jones et al. ([5]) explored parameter spaces formed by combinations of 5 to 15 elements and showed that 8 dimensions (Mg, Al, Si, K, Ti, Mn, Fe, Ni) is sufficient to recover birth clusters of stars, FIRE simulation does not track all elements that were used in mentioned papers (we have mostly α-elements) therefore discriminating power of clustering algorithm might be limited. We proceed with analysis of chemical space which includes following dimensions: [Fe/H], [O/Fe], [Mg/Fe], [Ca/Fe], [C/Fe], [N/Fe], [Ne/Fe], [Si/Fe], [S/Fe].

[1] $[\mathrm{X/Y}] \equiv \log_{10}(\mathrm{N_X/N_Y})_{\mathrm{star}} - \log_{10}(\mathrm{N_X/N_Y})_{\mathrm{Sun}}$, where N_X and N_Y are the abundances of elements X and Y

4 Overview of algorithms

t-SNE

T-SNE, t-distributed stochastic neighbor embedding, converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. We used t-SNE with 2-dimensional embedding space, default early exaggeration parameter (*early_exaggeration=12*) and *learning_rate* = 1000 as it is close to 10% from sample size.

One important parameter of t-SNE is perplexity (p) which is related to the number of close neighbours for each data point. First we studied how the overall appearance of t-SNE map depends on p. We performed t-SNE for p in 50, 100, 150, 500, 750, 1000 (Figure 1). Larger values of p tend to flatten the map while smaller values decrease the distance between clusters. In our case the best result was obtained with $p = 100$ so we use this value for our analysis of both sample A and B.

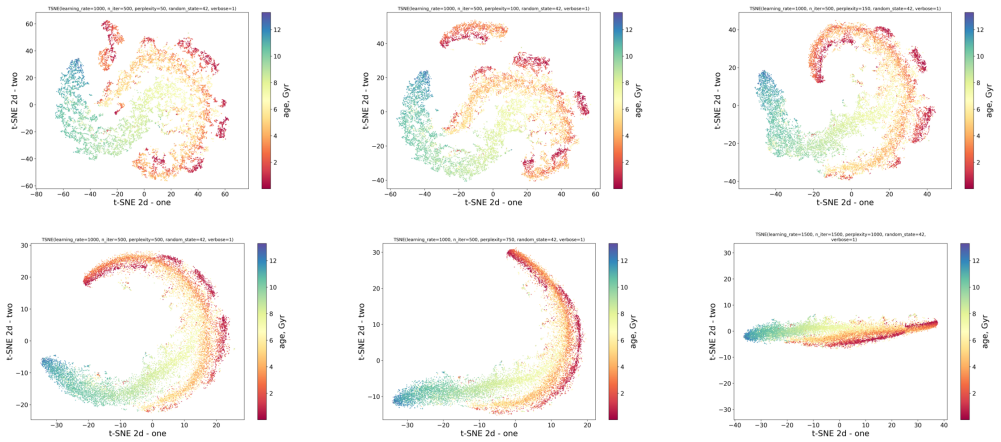

Figure 1: T-SNE maps for different values of *perplexity* (p is increasing from top left to bottom right)

UMAP

UMAP (uniform manifold approximation and projection) is an algorithm for dimension reduction based on manifold learning techniques and ideas from topological data analysis. It is similar to t-SNE, but offers a number of advantages such as increased speed and better preservation of the data's global structure.

We started from a search for the optimal parameters for the algorithm as we did with t-SNE. Figure 2 shows compilation of results of performing UMAP with different values of *n_neighbours*, which constrains the size of the local neighborhood UMAP will look at when attempting to learn the manifold structure of the data, and *min_dist*, which controls how tightly UMAP is allowed to pack points together in low-dimensional representation. As opposed to t-SNE, UMAP does not show a lot of variations in the final results, so for our analysis we chose *n_neighbours* to be 20 and *min_dist* 0.01. As with t-SNE we used these parameters for our analysis of both sample A and B.

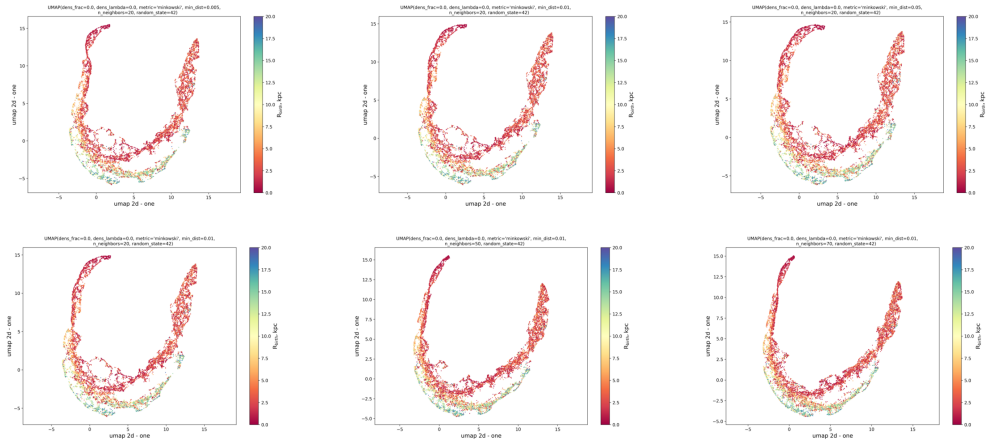

Figure 2: UMAP maps for different values of *min_dist* and *n_neighbors*

DBSCAN

DBSCAN, Density-Based Spatial Clustering of Applications with Noise, finds core samples of high density and expands clusters from them. One of the differences of DBSCAN compared to UMAP and t-SNE is that it provides labels for each data point instead of representation in space of lower dimension.

We tried several parameters but we were not able to utilize this method on our data in the given time. We were able to get various number of clusters (from 5 to 60), but majority of data points were attributed to the same label, so we could not find any clusters that had physical meaning. Figure 3 depicts the clusters we were getting.

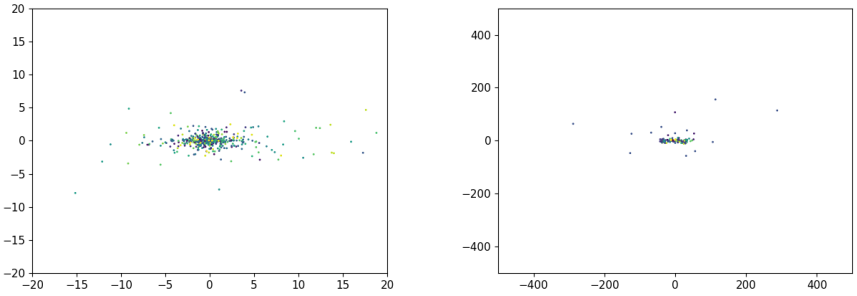

Figure 3: Example plots for DBSCAN. Left plot is for sample A and right plot sample B. We did not continue working on this due to the time constraints.

5 Results

Sample A

Figures 4 and 5 show distribution of stars on XY and XZ plane. The left plot on both panels shows the distribution of stars colored by density of points per pixel in logarithmic scale, the middle plot is the weighted age of the stars in the pixel, and the right plot represents the birth radius. The left plot indicates the presence of thick and thin disk with bulge in the center and halo. Right plot demonstrates the so called inside-out growth when star formation proceeded from the center to outskirts, left plot indicates that majority of the stars stays near their birth place during their evolution, however there is a hint of outward migration in the disk plane.

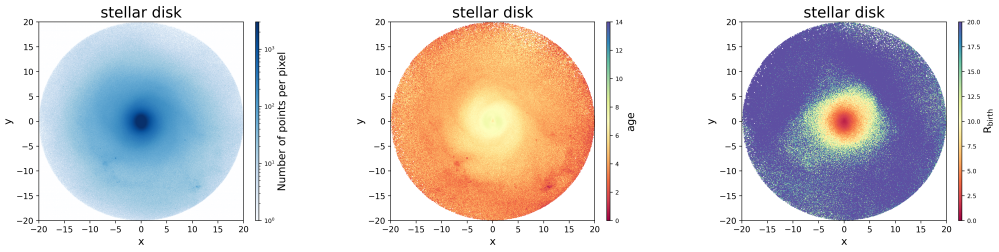

Figure 4: Distribution of stars from sample A on XY plane. Color shows density of points in pixel on the left plot, age on the middle plots, birth radius on the right plot.

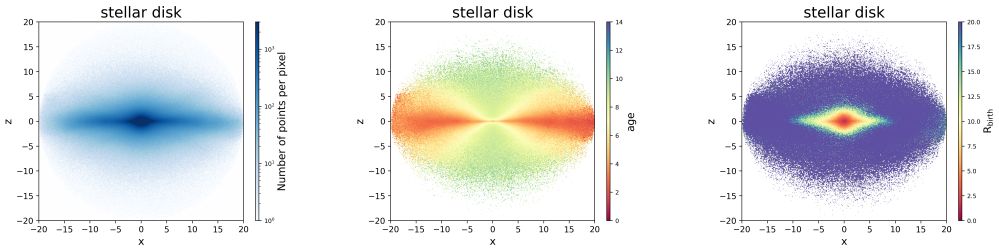

Figure 5: Distribution of stars from sample A on XZ plane. Color shows density of points in pixel on the left plot, age on the middle plots, birth radius on the right plot.

t-SNE

To explore the meaning of different groups seen on t-SNE map we color-coded the results of mapping by various properties (Figure 6). Top left plot which is colored by age also contains groups that were visually identified.

Group I contains old stars, born in the inner parts of the galaxy, rich in most of the alpha-elements, but with moderate abundance of iron. As for kinematics, these stars consitute spherical component of the galaxy which could be seen from near-zero values of J_z/J_{circ}.

Group II contains very young disk ($J_z/J_{circ} \sim 1$) stars formed near the center of galaxy which is in contradiction with inside-out growth. These stars have the highest amount of iron and low

amount of alpha-elements, which is in agreement with their young age. In terms of oxygen content they almost do not show any deviations from the values in group III, while in other alpha-elements their abundances has extreme values (lowest for Mg, Ca, Si, S and Ne and highest for N and C).

Group III contains 2 smaller subclusters (IIIb and IIIc) which are characterized mainly by their age. Overall group III represents thick disk component of the galaxy.

To confirm our guess about the physical nature of groups on t-SNE plane we map the identified clusters back on XZ plane (Figure 7). Indeed we see the spherical, thick disk and thin disk components identified purely from clustering in chemical abundance space.

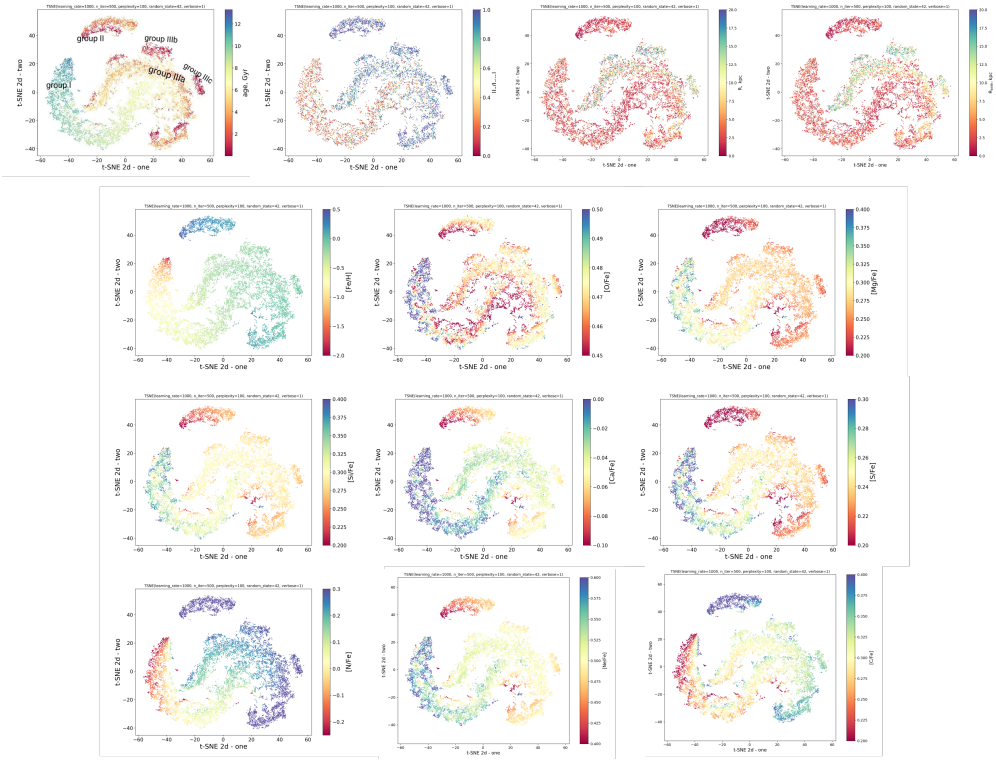

Figure 6: T-SNE projection of sample A color-coded by various parameters (elemental abundances, age, birth radius, current radius and J_z/J_{circ})

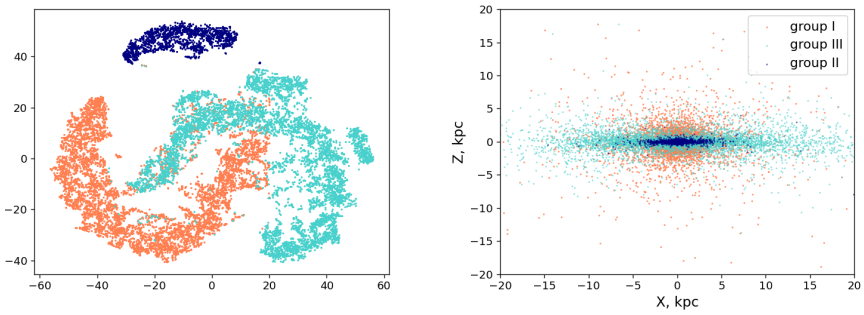

Figure 7: Comparison of clusters in embedding space of t-SNE with stellar distribution in the coordinate space of the galaxy.

UMAP

Figure 8 shows UMAP results color-coded by chemical abundances, age, birth and current radii and angular momentum ratio. The separation of clusters is less clear than in case of t-SNE, however it is still possible to identify a few groups, corresponding to disk and spherical component, mainly due to J_z/J_{circ}. As expected, the disk component contains younger and more metal-rich stars.

The chemical space was also similar to t-SNE in that we found group I (older stars) to contains almost no Fe and a higher abundance of alpha elements (Mg, Ca, S, and Ne).

Group II (younger stars) was the opposite. We found a larger abundance of Fe, N, and C, which supports the ISM having a higher metallicity as it evolves.

Figure 9 demonstrates how stars from these subgroups are located in the galaxy. As compared to t-SNE we were not able to identify the thin disk component as easily. We go into detail later, but we found the thin disk by comparing different portions of t-SNE to UMAP and finding the thin disk this way. Much like t-SNE, we find the expected results for the spherical and disk components.

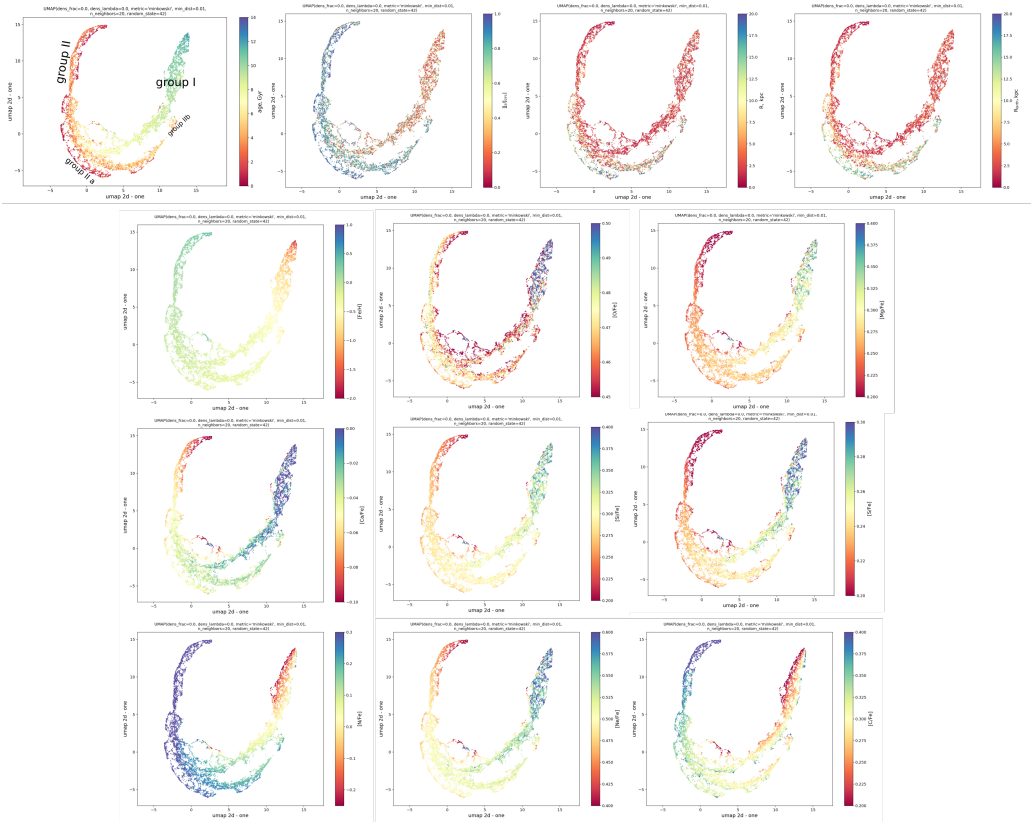

Figure 8: UMAP projection of sample A color-coded by various parameters (elemental abundances, age, birth radius, J_z/J_{circ})

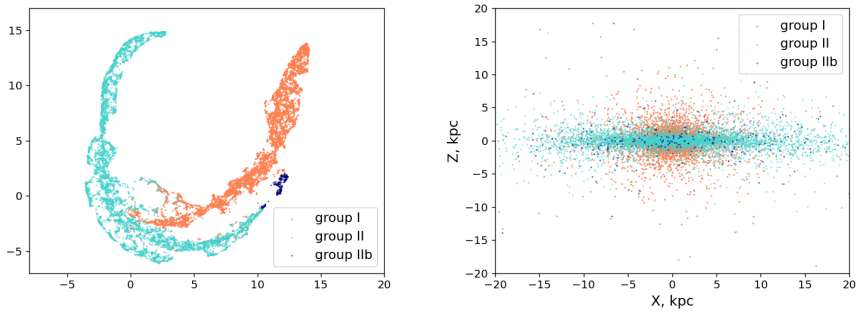

Figure 9: Comparison of clusters in embedding space of UMAP with stellar distribution in the coordinate space of the galaxy.

Sample B

Similar to sample A Figures 10 and 11 show distribution of stars on XY and XZ plane. The left plot on both panels shows the distribution of stars colored by density of points per pixel in logarithmic scale, the middle plot is the weighted age of the stars in the pixel, and the right plot represents the birth radius. All plots show the signs of a few tidally stripped satellites with slightly different age and birth radii.

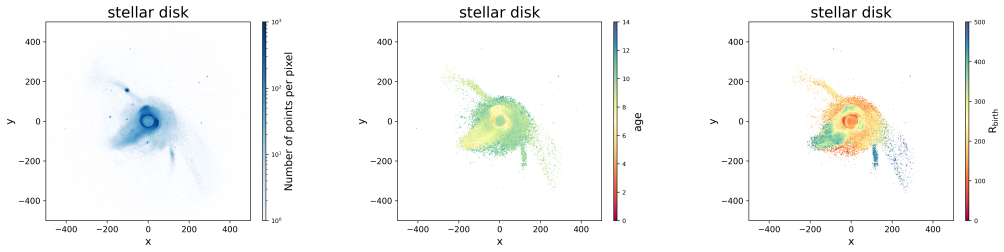

Figure 10: Distribution of stars from sample B on XY plane. Color shows density of points in pixel on the left plot, age on the middle plots, birth radius on the right plot.

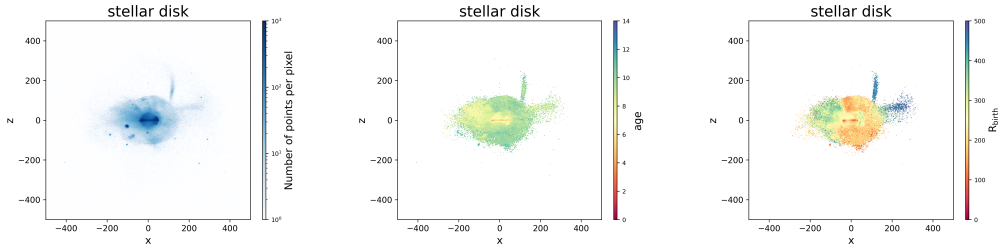

Figure 11: Distribution of stars from sample B on XZ plane. Color shows density of points in pixel on the left plot, age on the middle plot, birth radius on the right plot.

t-SNE

Similar with sample A, we use the color-coded plots to show the results for the various mappings. Figure 12 shows these various properties including age, birth and current radius and all nine elemental abundances.

The most clearly separated groups are labeled as in-situ and ex-situ stars as they correspond to stars that were born inside and outside host galaxy respectively. We have in-situ stars in the sample B, because filtering host stars based on their birth radius is not enough to completely clean up the sample.

Beside these two large groups the t-SNE projection color-coded by elemental abundances other than [Fe/H] and [O/Fe] (two bottom rows on Figure 12) reveal some complicated substructure in the ex-situ cluster. This fact points out that different satellites are chemically distinct enough to be identified by means of chemical tagging, however identifying indices of stars from substructure in low-dimensional space was complicated and we were not able to proceed with further analysis

of the origin of these smaller groups. It's interesting to note that projection color-coded by age and birth radius does not show the same type of structure, that is because star formation in the satellites involved in minor mergers continued throughout their fall onto the host galaxy.

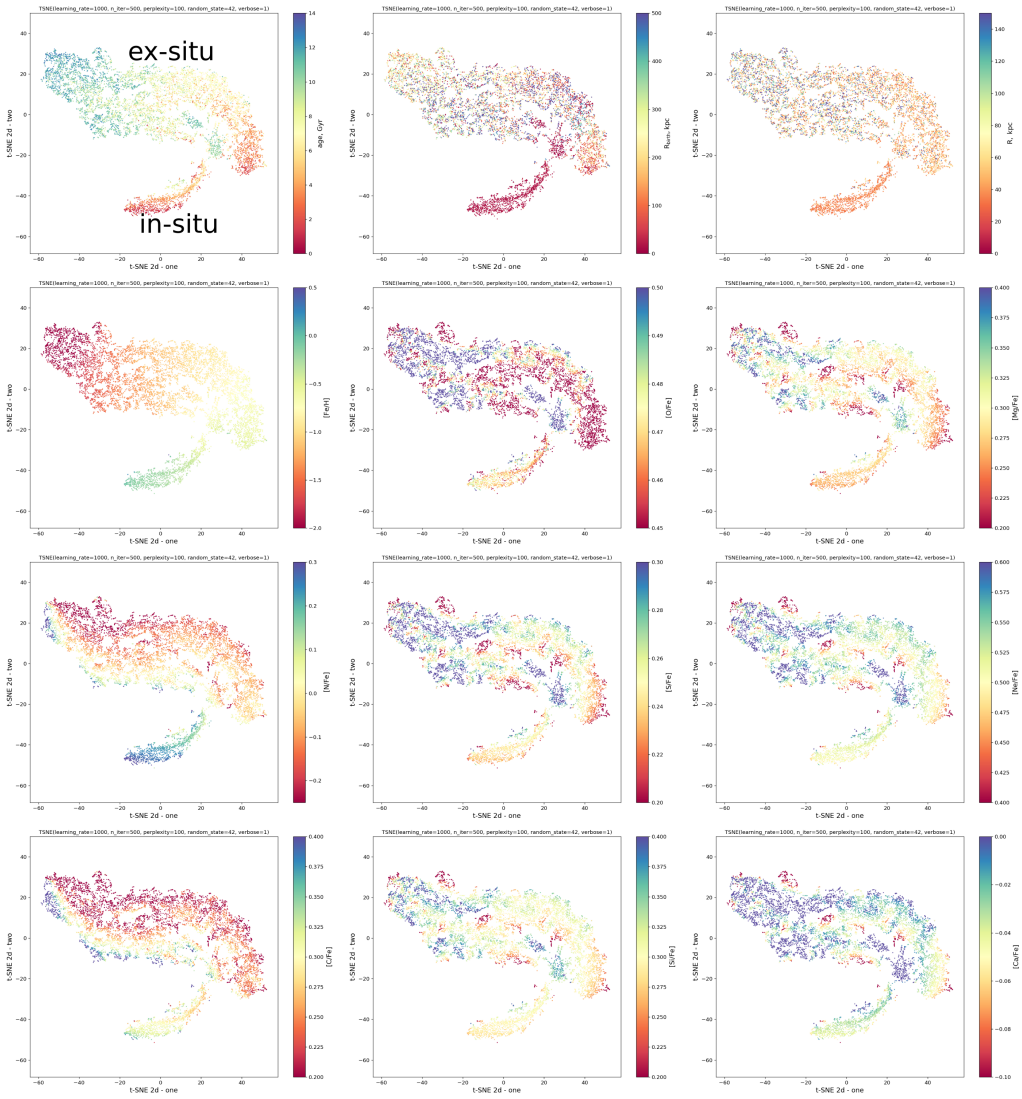

Figure 12: T-SNE projection of sample B color-coded by various parameters (elemental abundances, age, birth radius and current radius.)

UMAP

We found that UMAP followed similar trends as t-SNE for sample B. Figure 13 shows the results for the various parameter spaces we used. We also were able to define in-situ and ex-situ stars as in

the case of t-SNE projection, we also can see signs of smaller scale substructure, however on t-SNE projection it seemed to be more well-defined.

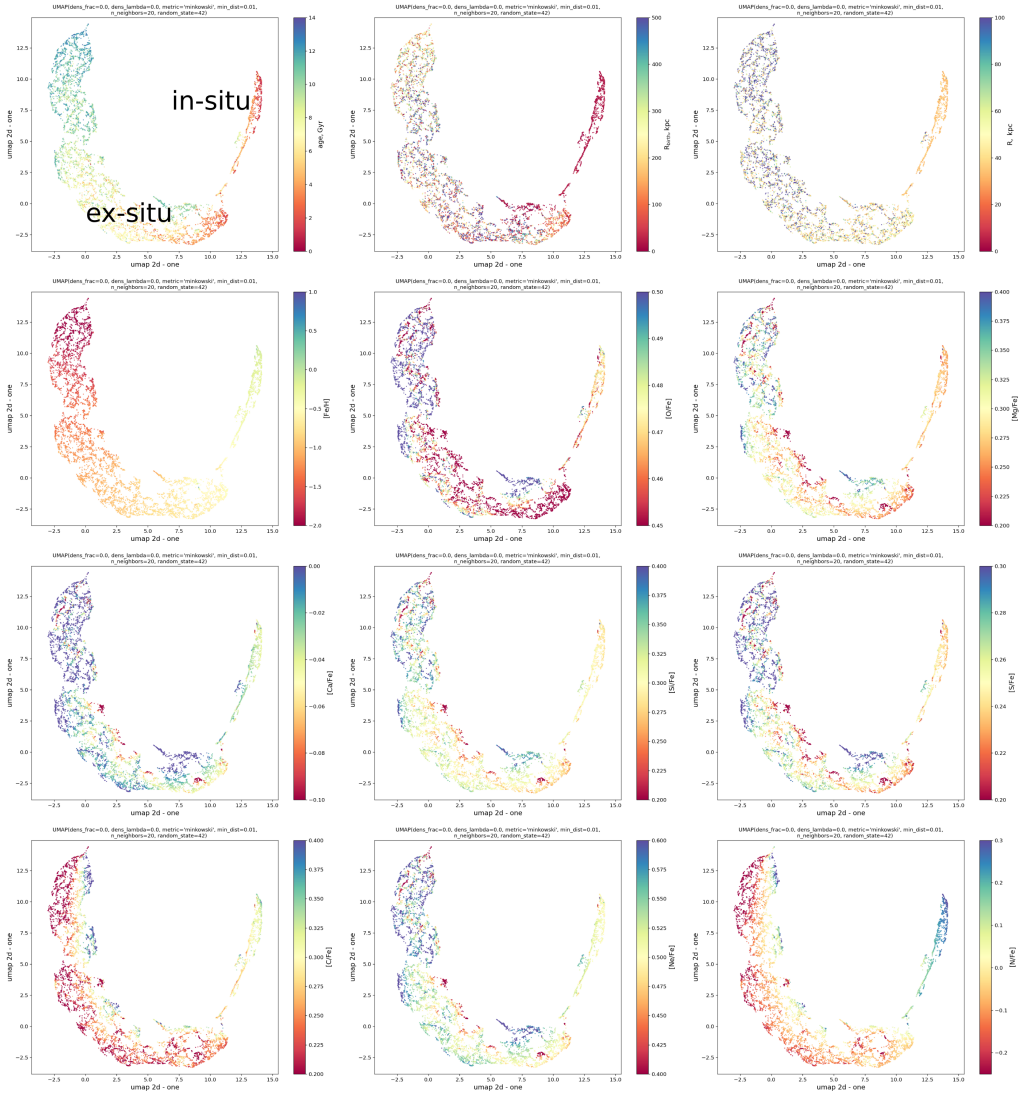

Figure 13: UMAP projection of sample B color-coded by various parameters (elemental abundances, age, birth radius and current radius.)

Comparison between methods

One method of comparison is to measure the overlap between groups corresponding to different components defined by different methods and different samples. We are referring to "overlap" as the amount that t-SNE and UMAP have in common for each group and each sample.

Sample A overlap was measured corresponding to thick/thin disk and spheroidal component. The spherical component we see that 90% of t-SNE and 94% of UMAP overlap. For our thick disk 97% of t-SNE and 78% of UMAP overlap. Unfortunately, UMAP was not able to easily cluster the thin disk and we had to find it another way.

We chose this as a metric to see whether the clustering power of each algorithm was consistent. This provides an easy way to see how each algorithm is handling the data and if they are able to accurately distinguish the chemical space. By this metric we believe that t-SNE performed slightly better because it was able to easily pick out the three components we were expecting to see.

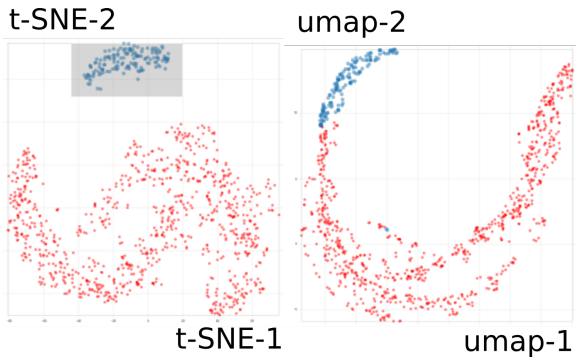

Figure 14: Comparison of Group II from t-SNE and UMAP projection for sample A.

To visualize the location of the same stars on plots of different projections we used brushable plots which allow to select data points on one plot and automatically highlight them on other plots, therefore we could select area on, for example, t-SNE projection and find out the location of the included stars on other projection (UMAP, X-Z, $j_z - E$). We used publicly available Observable notebook with scatterplot matrix (`https://observablehq.com/@d3/brushable-scatterplot-matrix`) written on D3.js by Mike Bostock and uploaded our data.

With brushable plot we finally were able to locate the thin disk on UMAP projection for sample A (Figure 14). In contrast to t-SNE, thin disk in UMAP did not appear as a separate cluster but only as a part of disk component.

For sample B Figure 15 shows that in-situ group, indeed, corresponds to inner stars, while Figure 16 demonstrates the difficulties we have faced while analysing the ex-situ groups, mainly, that cores of tidally stripped satellites that could be located in phase-space appear to be scattered throughout the low-dimensional projection.

These results lead us to believe that the stellar streams and satellite chemical space is more complex than the main galaxy. This difference causes the algorithms to break down and is hard to find the structure.

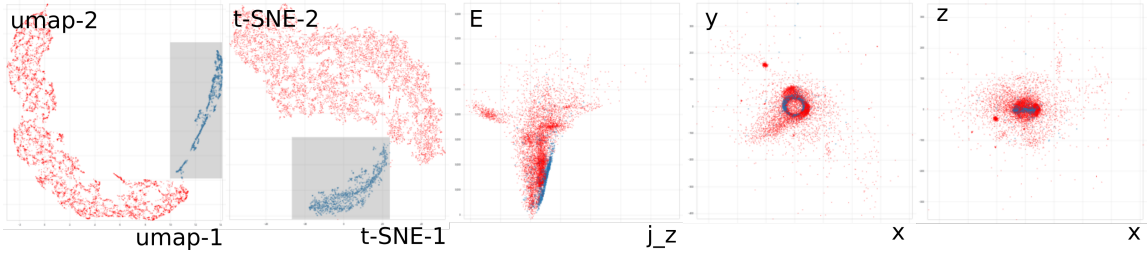

Figure 15: Here we compare the clusters between UMAP and t-SNE and where these stars are located in coordinate space. Smaller cluster on both projections correspond to stars from the host galaxy.

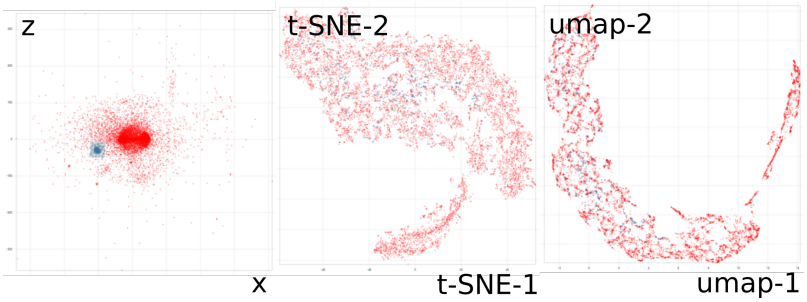

Figure 16: The left plot shows the star distribution in red and the area of interest in blue. We correlate the blue stars from the left plot to the t-SNE and UMAP plots to see if the algorithms found chemically similar clusters. Even though stars in blue are the part of tidally stripped satellites and are located in the dense core of this satellite, they are not considered to be close in chemical space by t-SNE and UMAP algorithm.

6 Conclusion

Through our testing we found that t-SNE and UMAP produce the results we were expecting to see for sample A but only partially for sample B. Sample A stars were expected to be clustered into three main groups: spherical, thin disk, and thick disk. We were able to obtain these groups with t-SNE and partially (only disk and spheroidal component without further separation into thin and thick disk) for UMAP. Sample B, however, appeared to be more challenging. The chemical space for both samples do seem to have some structure, but it is hard to fully discern for sample B.

We were not able to produce any meaningful results with DBSCAN in the timeframe of this project, but we believe it can be used as an effective chemical tagging tool for this data.

However, we have shown that both t-SNE and UMAP are effective tools for chemical tagging for this data and may be improved upon for further analysis. Again, with the time restriction of the project we believe that sample B can be improved by optimizing the parameters for both algorithms. We chose to use the same parameters for both samples and found that the results were far better with sample A than sample B.

References

[1] K. Freeman and J. Bland-Hawthorn, "The New Galaxy: Signatures of Its Formation," , vol. 40, pp. 487–537, Jan. 2002.

[2] J. Wojno, G. Kordopatis, M. Steinmetz, P. McMillan, G. Matijevič, J. Binney, R. F. G. Wyse, C. Boeche, A. Just, E. K. Grebel, A. Siebert, O. Bienaymé, B. K. Gibson, T. Zwitter, J. Bland-Hawthorn, J. F. Navarro, Q. A. Parker, W. Reid, G. Seabroke, and F. Watson, "Chemical separation of disc components using RAVE," , vol. 461, no. 4, pp. 4246–4255, Oct. 2016.

[3] R. P. Naidu, C. Conroy, A. Bonaca, B. D. Johnson, Y.-S. Ting, N. Caldwell, D. Zaritsky, and P. A. Cargile, "Evidence from the H3 Survey That the Stellar Halo Is Entirely Comprised of Substructure," , vol. 901, no. 1, p. 48, Sep. 2020.

[4] L. Necib, B. Ostdiek, M. Lisanti, T. Cohen, M. Freytsis, S. Garrison-Kimmel, P. F. Hopkins, A. Wetzel, and R. Sanderson, "Evidence for a vast prograde stellar stream in the solar vicinity," *Nature Astronomy*, vol. 4, pp. 1078–1083, Jul. 2020.

[5] N. Price-Jones, J. Bovy, J. J. Webb, C. Allende Prieto, R. Beaton, J. R. Brownstein, R. E. Cohen, K. Cunha, J. Donor, P. M. Frinchaboy, D. A. García-Hernández, R. R. Lane, S. R. Majewski, D. L. Nidever, and A. Roman-Lopes, "Strong chemical tagging with APOGEE: 21 candidate star clusters that have dissolved across the Milky Way disc," , vol. 496, no. 4, pp. 5101–5115, Aug. 2020.

[6] F. Anders, C. Chiappini, B. X. Santiago, G. Matijevič, A. B. Queiroz, M. Steinmetz, and G. Guiglion, "Dissecting stellar chemical abundance space with t-SNE," , vol. 619, p. A125, Nov. 2018.

[7] A. R. Wetzel, P. F. Hopkins, J.-h. Kim, C.-A. Faucher-Giguère, D. Kereš, and E. Quataert, "Reconciling Dwarf Galaxies with ΛCDM Cosmology: Simulating a Realistic Population of Satellites around a Milky Way-mass Galaxy," , vol. 827, no. 2, p. L23, Aug. 2016.

[8] P. F. Hopkins, "A new class of accurate, mesh-free hydrodynamic simulation methods," , vol. 450, no. 1, pp. 53–110, Jun. 2015.

[9] P. F. Hopkins, A. Wetzel, D. Kereš, C.-A. Faucher-Giguère, E. Quataert, M. Boylan-Kolchin, N. Murray, C. C. Hayward, S. Garrison-Kimmel, C. Hummels, R. Feldmann, P. Torrey, X. Ma, D. Anglés-Alcázar, K.-Y. Su, M. Orr, D. Schmitz, I. Escala, R. Sanderson, M. Y. Grudić, Z. Hafen, J.-H. Kim, A. Fitts, J. S. Bullock, C. Wheeler, T. K. Chan, O. D. Elbert, and D. Narayanan, "FIRE-2 simulations: physics versus numerics in galaxy formation," , vol. 480, no. 1, pp. 800–863, Oct. 2018.

[10] Y.-S. Ting, K. C. Freeman, C. Kobayashi, G. M. De Silva, and J. Bland-Hawthorn, "Principal component analysis on chemical abundances spaces," , vol. 421, no. 2, pp. 1231–1255, Apr. 2012.