Galaxy Formation and Evolution via Phase-temporal Clustering with FuzzyCat ∘ AstroLink

William H. Oliver and **Tobias Buck**
Interdisciplinary Center for Scientific Computing, University of Heidelberg
Im Neuenheimer Feld 205, D-69120 Heidelberg, Germany
`william.oliver@iwr.uni-heidelberg.de` & `tobias.buck@iwr.uni-heidelberg.de`

Abstract

We demonstrate how the composition of two unsupervised clustering algorithms, `AstroLink` and `FuzzyCat`, makes for a powerful tool when studying galaxy formation and evolution. `AstroLink` is a general-purpose astrophysical clustering algorithm built for extracting meaningful hierarchical structure from point-cloud data defined over any feature space, while `FuzzyCat` is a generalised soft-clustering algorithm that propagates the dynamical effects of underlying data processes into a fuzzy hierarchy of stable fuzzy clusters. Their composition, `FuzzyCat` ∘ `AstroLink`, can therefore identify a fuzzy hierarchy of astrophysically- and statistically-significant fuzzy clusters within any point-based data set whose representation is subject to changes caused by some underlying process. Furthermore, the pipeline achieves this without relying upon strong assumptions about the data, the change process, the number/importance of specific structure types, or much user input – thereby making itself applicable to a wide range of fields in the physical sciences. We find that for the task of structurally decomposing simulated galaxies into their constituents, our context-agnostic approach has a substantial impact on the diversity and completeness of the structures extracted as well as on their relationship within the broader galactic structural hierarchy – revealing dwarf galaxies, infalling groups, stellar streams (and their progenitors), stellar shells, galactic bulges, and star-forming regions.

1 Motivation and related work

A pressing and continually evolving sub-field of astrophysics is the study of galaxy formation and evolution, which according to ΛCDM cosmology, assemble hierarchically with time through mergers and the accretion of smaller galaxies [44]. To understand how and why a galaxy and its substructure develop within the context of the surrounding environment and of the underlying cosmological model, astrophysicists and cosmologists will look to both observational and simulation data. With observations, we may learn from a very large number of galaxies that are each observed at a unique snapshot in time and that arise from the *ground-truth* cosmology of our Universe. While with simulations, we may learn from many snapshots of a comparatively small number of galaxies that depend on a pre-specified cosmological model. By comparing these two data types, we can hope to constrain our cosmological models as well as our understanding of galaxy formation and evolution.

A typical approach towards studying galaxy formation and evolution in the context of simulated data is to use a halo finder (+ merger tree) code [e.g. 4, 18, 22, 43] to find a catalogue of (sub-)haloes and their merger history (and then to analyse the physical properties of their outputs). These codes routinely perform similarly and robustly [2, 3, 14, 20, 21, 23, 38, 39], however they only consider mostly or completely self-bound groups that satisfy a minimum overdensity threshold. If this threshold is too high then some haloes may not be detected and if it is too low then some haloes

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

can be disregarded in the unbinding procedure. As such, these codes will tend not to capture tidally disrupted groups or fleeting structures resulting from density waves, hydrodynamical effects, or star-formation events. Not including these kinds of structures in any subsequent analysis ensures that cosmological models used in simulations are never constrained against their existence – even though their analogues are observed to be present in our Universe [26, 27, 29]. It is for these reasons that we investigate the usefulness of a more generalised and context-agnostic approach.

2 Finding stable clusters from evolving data: FuzzyCat ∘ AstroLink

The `FuzzyCat` ∘ `AstroLink` pipeline operates on a data set whose representation is subject to changes from an underlying process (such as stochastic resampling, temporal-evolution, etc.). The approach composes the two clustering algorithms by first applying `AstroLink` [31, 37] to the various realisations of the data and then applying `FuzzyCat` [33] to the various `AstroLink` outputs. The result is an unsupervised machine learning pipeline that produces a fuzzy hierarchy of astrophysically- and statistically-significant fuzzy clusters that encapsulate the effects of the underlying process(es) implicit within an evolving input data set. To our knowledge, such a pipeline has not yet existed.

2.1 The AstroLink algorithm

`AstroLink` is an unsupervised astrophysical clustering algorithm that extracts arbitrarily-shaped hierarchical clusters from an arbitrarily-shaped point-based data set such that the clusters found are statistical outliers from noisy density fluctuations. It is an improvement to its predecessors, `CluSTAR-ND` [36] and `Halo-OPTICS` [35], and by comparison boasts increased clustering power in shorter run-times. It also shares algorithmic ties to, but is more statistically robust than, `OPTICS` [1] and `HDBSCAN` [12, 28]. These can be thought of hierarchical extensions of `DBSCAN` [16], which itself can be thought of as a more-robust-to-noise version of the Friends-Of-Friends (`FOF`) algorithm [13] – an algorithm commonly used to identify galaxies/haloes from cosmological simulations.

The `AstroLink` algorithm performs five steps; (1) data rescaling, (2) local-density estimation, (3) data aggregation, (4) model-fitting and structure identification, (5) hierarchy correction. If $\texttt{adaptive} = 1$ (default), step 1 rescales the data to have unit variance – so as to remove the effect of differing units in the feature space. Step 2 calculates the local-density of each data point by applying a multivariate Epanechnikov kernel [15] and a balloon estimator [41] to its $k_{\rm den}$-neighbourhood ($k_{\rm den} = 20$, default) – the logarithm of this estimate is taken before all values are then normalised, i.e. $\log\hat{\rho} \in [0, 1]$. Step 3 tracks and records the connected components of data points that form as the edges of a local-density-weighted $k_{\rm link}$-nearest-neighbour graph ($k_{\rm link}$ is data-driven by default) are traversed in descending order – these components define a hierarchy of feature-space overdensities. In step 4, a model is fit to the *clusteredness* of these connected components and is used to identify the $\geq S\sigma$-outlier overdensities (S is data-driven by default) from the noisy local-density fluctuations inherent within the data. If $\texttt{h\_style} = 1$ (default), step 5 corrects the final hierarchy by incorporating some additional outlier overdensities – producing the final hierarchy of clusters. `AstroLink` does not require the user to make any hyperparameter choices as the performance of the entirely data-driven version of this process is near-optimal in nearly all cases. When applied to simulated galaxies, `AstroLink` does extraordinarily well at finding the remnants of infalling-satellites within the data. The implementation is described in more detail in the original science paper [37] as well as in the `AstroLink` ReadTheDocs page [32].

2.2 The FuzzyCat algorithm

`FuzzyCat` is an unsupervised general-purpose soft-clustering algorithm that, given a series of clusterings on object-based data, produces data-driven fuzzy clusters whose membership functions encapsulate the effects of changes in the clusters due to changes in the feature space representation of the objects themselves. The different input clusterings may be governed by any underlying process that affects the clustering structure (e.g. stochasticity, temporal evolution, model hyperparameter variation, etc.). In effect, `FuzzyCat` propagates these effects into a soft-clustering which has had these effects abstracted away into the membership functions of the original object-based data.

At its core, `FuzzyCat` is very similar to `AstroLink` – procedures mimicking steps 2, 3, and 4 are analogously performed – except that it takes a data set of clusters as input as opposed to one of data points. It is in this sense that FuzzyCat is actually a clustering algorithm that operates on the Jaccard-space of a catalogue of clusterings in order to produce *clusters* of clusters (i.e. fuzzy clusters). As such the Jaccard index, which is calculated for every pair of clusters in the input, is analogous to the AstroLink $\log \hat{\rho}$ calculated in step 2 – although in FuzzyCat there is no equivalent to k_{den} and k_{link} is effectively fixed at the size of the data set. The final fuzzy clusters, found after a process equivalent to the AstroLink step 3, must also meet thresholds (measured by the Jaccard index) of internal similarity ($J_{\mathrm{min\_intra}} = 0.5$, default) and external dissimilarity ($J_{\mathrm{max\_inter}} = 0.5$, default), as well as remain stable over at least a minimum number of data set realisations. The latter condition is governed by the minStability hyperparameter, which we change from the default value of 0.5 for the applications in Sec. 4. These conditions ensure statistical robustness with the corresponding hyperparameters effectively playing the role of the AstroLink S-parameter. The final fuzzy clusters are then translated into membership functions with respect to the underlying object-based data by counting the number of data realisations for which each object appears within each fuzzy cluster.

It is worth clarifying that FuzzyCat is never provided any knowledge of the feature space representation of the object-based input data set nor of the underlying change process that acts upon it – i.e. it makes no assumptions about *why* clusters of these objects exist nor about *how* they may change between any two clusterings. It is easier to see why FuzzyCat can be applied to a stochastically changing clustering (as in Sec. 3) as opposed to a temporally evolving one (as in Sec. 2) – however the strong temporal correlation between consecutive snapshots produces a statistically significant clustering signal resulting in physically meaningful cluster tracking. We refer the reader to the FuzzyCat ReadTheDocs page [34] for more details.

3 A simple use case

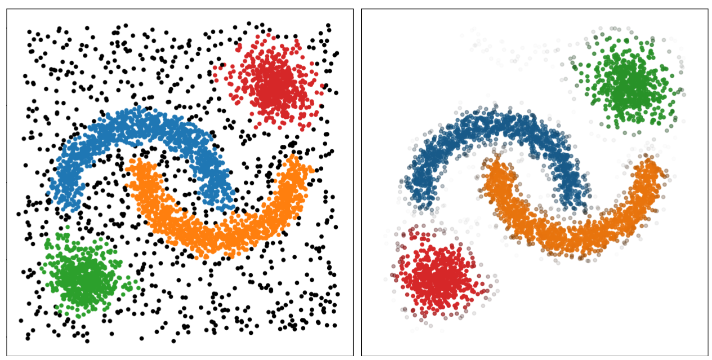

Figure 1: The FuzzyCat $\circ$ AstroLink pipeline applied to uncertain toy data (code & results found **here**). *Left*: A random sample of the data with coloured points belonging to AstroLink clusters. *Right*: The resultant fuzzy clusters found by FuzzyCat after 100 resamplings/clusterings of the data with the opacity and colour of points representing the membership function (in this case a probability) of a data point to belong to a particular fuzzy cluster.

We first take a detour from the focus of this work and demonstrate the versatility of this pipeline by applying it to a 2D toy data set whereby the underlying process that changes the data is stochastic resampling due to the effect of uncertainties. In this exercise, each data point's uncertainty profile is a 2D Gaussian distribution with identical covariance matrices equal to $\sigma^2 I$ where $\sigma = 0.05$. The data has been resampled and clustered 100 times by AstroLink and then FuzzyCat is applied to these resampled clusterings (with both algorithms using their default settings). Fig. 1 depicts the results, where we see that the effect of stochastically resampling from the uncertainty profiles of the data points is to give fuzzy boundaries to the AstroLink clusters. Although we don't explore this function further in this work, it shows that the FuzzyCat $\circ$ AstroLink pipeline is capable of propagating uncertainties into clusters. Such a function would also be highly beneficial for studying galaxy formation and evolution in the context of observational data, as well as for other areas of the physical sciences where uncertainties are inherent to the data used.

4 Applications to NIHAO-UHD galaxies and comparison with AHF

We now apply the FuzzyCat $\circ$ AstroLink pipeline to a set of six simulated galaxies from the NIHAO-UHD suite [9] – detailed in Sec. 4.1. We first apply AstroLink with its default settings to the 6D position-velocity feature space of the stellar particles in each snapshot of each galaxy. We then provide the resultant clusterings to FuzzyCat where we choose the minStability hyperparameter such that the resultant fuzzy clusters exist for ≥ 230 Myr (approximately the period of the Sun's orbit within the Milky Way). So as to draw a comparison to current methods, we also apply Amiga's Halo Finder (AHF) to the same set of galaxies – see Sec. 4.2 for details on this code.

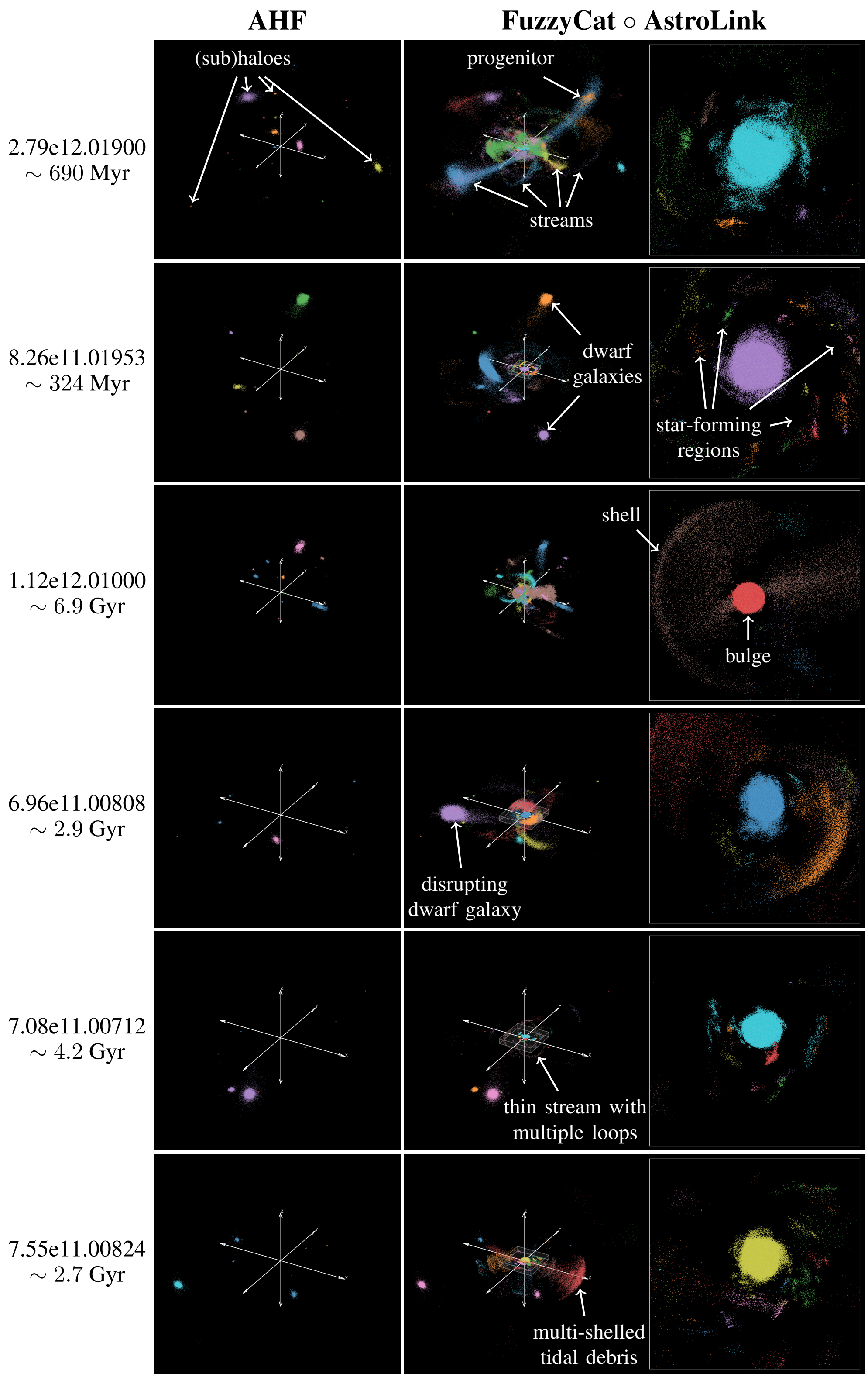

Figure 2: Frames from animations (code & results found **here**) of clusterings produced by `AHF` (left) and the `FuzzyCat` ∘ `AstroLink` pipeline (right). Each row shows the snapshot's file name (i.e. `galaxy_mass.snapshot_number`) and look-back time. Each panel depicts a full 3D projection (axes extend 100 kpc from the galactic centre), and for our approach, a zoomed-in top-down view of the galactic disk (i.e. everything within the central white prism of width 50 kpc and height 10 kpc).

Fig. 2 depicts various frames from the animations generated via both AHF and the FuzzyCat ∘ AstroLink pipeline. Among the structures extracted by our approach are; dwarf galaxies, infalling groups, stellar streams (and their progenitors), stellar shells, galactic bulges, and star-forming regions. By comparison, traditional approaches are not able to find most of this structure beyond a subset of which is (or is mostly) self-bound – this can be seen with the corresponding results from AHF. Our pipeline therefore reveals much more of the information content of these galaxies, and as such, lends itself as a powerful tool for analysing galaxy formation and evolution in a modern setting. We refer the reader to the FuzzyCat ReadTheDocs page [34] for a coded tutorial of this work as well as animations of the phase-temporal clusterings for each of the six galaxies by each of the codes.

4.1 The NIHAO-UHD suite of cosmological hydrodynamical simulations

The NIHAO-UHD suite [9] is an ultra-high resolution subset of the Numerical Investigation of a Hundred Astronomical Objects (NIHAO) simulation suite [45] which assumes cosmological parameters from the 2015 Planck Collaboration et al. [40]. These galaxies are chosen to reflect the most MW-like galaxies in terms of mass, size and disk properties. Parts of the simulation suite have previously been used to study the build-up of MW's peanut-shaped bulge [5, 7], investigate the stellar bar properties [19], infer the MW's dark halo spin [30], study the dwarf galaxy inventory of MW mass galaxies [8, 11], and investigate the age-metallicity relation of MW disk stars [24] including the chemical bimodality of disk stars [6, 10], their abundances [25] and the origin of very metal-poor stars inside the stellar disk [42]. Because of their realistic cosmology, complex hydrodynamical nature, and advanced physical realism, the NIHAO-UHD galaxies serve as an excellent probe for studying galaxy formation and evolution – as well as to showcase the novel capabilities of our pipeline.

4.2 Amiga's Halo Finder

As a comparison to our approach, we apply the halo finder AHF2 [17, 22] to each snapshot of each simulated galaxy. It works by recursively refining a grid in a top-down manner to identify spatial regions within the simulation box that meet a certain overdensity threshold (200 times the critical density of the Universe, in this case). Once particles are identified as belonging to such regions, an iterative unbinding procedure is performed to remove all particles whose velocity exceeds the escape velocity at that particle's position within the given halo (this assumes a spherically symmetric density profile for the halo). The recursive and iterative nature of this algorithm yields a hierarchy of (sub)haloes – bounded groups with an overdensity above a threshold. With the haloes of each snapshot now found, temporal merger trees are calculated using the AHF analysis tool, MergerTree, which traces the particle IDs of all particles throughout the snapshots and identifies all progenitors of a given halo. The animations, from which the frames in Fig. 2 are taken, plot the star particles from each AHF halo with a colour that is passed down from its 'father' halo (its most similar progenitor from the previous snapshot). In this sense, these animations are the AHF analogue of those produced using the results of the FuzzyCat ∘ AstroLink pipeline.

5 Conclusion and outlook

In this work, we have demonstrated the effectiveness of the FuzzyCat ∘ AstroLink pipeline as a novel unsupervised machine learning approach – particularly as a tool for analysing simulated galaxies in the context of galaxy formation and evolution. By applying our pipeline to the NIHAO-UHD suite, we have shown that it can successfully identify a diverse range of astrophysical structures that traditional halo finder (+ merger tree) methods do not – capturing transient and tidally disrupted structures that are often overlooked in conventional analyses. As such, it provides the means to a more comprehensive understanding of galaxy formation and evolution.

The ability of the FuzzyCat ∘ AstroLink pipeline to adapt to data with underlying processes, such as stochastic variations and temporal evolution, positions it as a powerful tool for future studies in astrophysics and in other fields where data is fuzzy, dynamic, and complex. By overcoming the limitations of existing methods, our approach offers its user a more flexible and detailed examination of the hierarchical and multifaceted nature of astrophysical structures. We now intend to improve upon our pipeline through massive parallel (re-)implementations of AstroLink and FuzzyCat – opening application avenues to a broader range of cosmological simulations and observational data sets, thereby enhancing our understanding of our Universe's structure at multiple scales.

Broader impact statement

The authors are not aware of any immediate ethical or societal implications of this work. This work purely aims to aid scientific research and proposes a method of using a pipeline of clustering algorithms to learn about galaxy formation and evolution.

Acknowledgments and Disclosure of Funding

This work is funded by the Carl-Zeiss-Stiftung through the NEXUS program.

References

[1] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. In *ACM Sigmod record*, volume 28, pages 49–60. ACM, 1999. ISBN 1581130848. doi: 10.1145/304182.304187.

[2] Santiago Avila, Alexander Knebe, Frazer R. Pearce, Aurel Schneider, Chaichalit Srisawat, Peter A. Thomas, Peter Behroozi, Pascal J. Elahi, Jiaxin Han, Yao-Yuan Mao, Julian Onions, Vicente Rodriguez-Gomez, and Dylan Tweed. SUSSING MERGER TREES: the influence of the halo finder. *Monthly Notices of the Royal Astronomical Society*, 441(4):3488–3501, 05 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu799. URL `https://doi.org/10.1093/mnras/stu799`.

[3] Peter Behroozi, Alexander Knebe, Frazer R. Pearce, Pascal Elahi, Jiaxin Han, Hanni Lux, Yao-Yuan Mao, Stuart I. Muldrew, Doug Potter, and Chaichalit Srisawat. Major mergers going Notts: challenges for modern halo finders. *Monthly Notices of the Royal Astronomical Society*, 454(3):3020–3029, 10 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv2046. URL `https://doi.org/10.1093/mnras/stv2046`.

[4] Peter S. Behroozi, Risa H. Wechsler, and Hao-Yi Wu. The rockstar phase-space temporal halo finder and the velocity offsets of cluster cores. *The Astrophysical Journal*, 762(2):109, 2012. ISSN 0004-637X 1538-4357. doi: 10.1088/0004-637x/762/2/109. URL `http://dx.doi.org/10.1088/0004-637X/762/2/109`.

[5] T. Buck, M. Ness, A. Obreja, A. V. Macciò, and A. A. Dutton. Stars behind Bars II: A Cosmological Formation Scenario for the Milky Way's Central Stellar Structure. *The Astrophysical Journal*, 874:67, March 2019. doi: 10.3847/1538-4357/aaffd0.

[6] Tobias Buck. On the origin of the chemical bimodality of disc stars: a tale of merger and migration. *Monthly Notices of the Royal Astronomical Society*, 491(4):5435–5446, February 2020. doi: 10.1093/mnras/stz3289.

[7] Tobias Buck, Melissa K. Ness, Andrea V. Macciò, Aura Obreja, and Aaron A. Dutton. Stars Behind Bars. I. The Milky Way's Central Stellar Populations. *The Astrophyscial Journal*, 861 (2):88, July 2018. doi: 10.3847/1538-4357/aac890.

[8] Tobias Buck, Andrea V. Macciò, Aaron A. Dutton, Aura Obreja, and Jonas Frings. NIHAO XV: the environmental impact of the host galaxy on galactic satellite and field dwarf galaxies. *Monthly Notices of the Royal Astronomical Society*, 483(1):1314–1341, February 2019. doi: 10.1093/mnras/sty2913.

[9] Tobias Buck, Aura Obreja, Andrea V. Macciò, Ivan Minchev, Aaron A. Dutton, and Jeremiah P. Ostriker. NIHAO-UHD: the properties of MW-like stellar discs in high-resolution cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 491(3):3461–3478, January 2020. doi: 10.1093/mnras/stz3241.

[10] Tobias Buck, Jan Rybizki, Sven Buder, Aura Obreja, Andrea V. Macciò, Christoph Pfrommer, Matthias Steinmetz, and Melissa Ness. The challenge of simultaneously matching the observed diversity of chemical abundance patterns in cosmological hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 508(3):3365–3387, December 2021. doi: 10.1093/mnras/stab2736.

[11] Tobias Buck, Aura Obreja, Bridget Ratcliffe, Yuxi(Lucy) Lu, Ivan Minchev, and Andrea V. Macciò. The impact of early massive mergers on the chemical evolution of Milky Way-like galaxies: insights from NIHAO-UHD simulations. *Monthly Notices of the Royal Astronomical Society*, 523(1):1565–1576, July 2023. doi: 10.1093/mnras/stad1503.

[12] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data*, 10(1), July 2015. ISSN 1556-4681. doi: 10.1145/2733381. URL `https://doi.org/10.1145/2733381`.

[13] Marc Davis, George Efstathiou, Carlos S Frenk, and Simon DM White. The evolution of large-scale structure in a universe dominated by cold dark matter. *The Astrophysical Journal*, 292:371–394, 1985. doi: 10.1086/163168.

[14] Pascal J. Elahi, Jiaxin Han, Hanni Lux, Yago Ascasibar, Peter Behroozi, Alexander Knebe, Stuart I. Muldrew, Julian Onions, and Frazer Pearce. Streams going Notts: the tidal debris finder comparison project. *Monthly Notices of the Royal Astronomical Society*, 433(2):1537–1555, 06 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt825. URL `https://doi.org/10.1093/mnras/stt825`.

[15] Vassiliy A Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[17] Stuart P. D. Gill, Alexander Knebe, and Brad K. Gibson. The evolution of substructure - I. A new identification method. *Monthly Notices of the Royal Astronomical Society*, 351(2):399–409, June 2004. doi: 10.1111/j.1365-2966.2004.07786.x.

[18] Jiaxin Han, Shaun Cole, Carlos S. Frenk, Alejandro Benitez-Llambay, and John Helly. hbt+: an improved code for finding subhaloes and building merger trees in cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 474(1):604–617, 10 2017. ISSN 0035-8711. doi: 10.1093/mnras/stx2792. URL `https://doi.org/10.1093/mnras/stx2792`.

[19] T. Hilmi, I. Minchev, T. Buck, M. Martig, A. C. Quillen, G. Monari, B. Famaey, R. S. de Jong, C. F. P. Laporte, J. Read, J. L. Sanders, M. Steinmetz, and C. Wegg. Fluctuations in galactic bar parameters due to bar-spiral interaction. *Monthly Notices of the Royal Astronomical Society*, 497(1):933–955, September 2020. doi: 10.1093/mnras/staa1934.

[20] Alexander Knebe, Steffen R. Knollmann, Stuart I. Muldrew, Frazer R. Pearce, Miguel Angel Aragon-Calvo, Yago Ascasibar, Peter S. Behroozi, Daniel Ceverino, Stephane Colombi, Juerg Diemand, Klaus Dolag, Bridget L. Falck, Patricia Fasel, Jeff Gardner, Stefan Gottlöber, Chung-Hsing Hsu, Francesca Iannuzzi, Anatoly Klypin, Zarija Lukić, Michal Maciejewski, Cameron McBride, Mark C. Neyrinck, Susana Planelles, Doug Potter, Vicent Quilis, Yann Rasera, Justin I. Read, Paul M. Ricker, Fabrice Roy, Volker Springel, Joachim Stadel, Greg Stinson, P. M. Sutter, Victor Turchaninov, Dylan Tweed, Gustavo Yepes, and Marcel Zemp. Haloes gone mad14: The halo-finder comparison project. *Monthly Notices of the Royal Astronomical Society*, 415(3):2293–2318, 2011. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2011.18858.x. URL `https://doi.org/10.1111/j.1365-2966.2011.18858.x`.

[21] Alexander Knebe, Noam I Libeskind, Frazer Pearce, Peter Behroozi, Javier Casado, Klaus Dolag, Rosa Dominguez-Tenreiro, Pascal Elahi, Hanni Lux, Stuart I Muldrew, et al. Galaxies going mad: the galaxy-finder comparison project. *Monthly Notices of the Royal Astronomical Society*, 428(3):2039–2052, 2013.

[22] Steffen R. Knollmann and Alexander Knebe. AHF: Amiga's Halo Finder. *The Astrophysical Journal Supplement Series*, 182(2):608–624, June 2009. doi: 10.1088/0067-0049/182/2/608.

[23] Jaehyun Lee, Sukyoung K. Yi, Pascal J. Elahi, Peter A. Thomas, Frazer R. Pearce, Peter Behroozi, Jiaxin Han, John Helly, Intae Jung, Alexander Knebe, Yao-Yuan Mao, Julian Onions, Vicente Rodriguez-Gomez, Aurel Schneider, Chaichalit Srisawat, and Dylan Tweed. Sussing

merger trees: the impact of halo merger trees on galaxy properties in a semi-analytic model. *Monthly Notices of the Royal Astronomical Society*, 445(4):4197–4210, 11 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu2039. URL `https://doi.org/10.1093/mnras/stu2039`.

[24] Yuxi Lucy Lu, Melissa K. Ness, Tobias Buck, and Christopher Carr. Turning points in the age-metallicity relations - created by late satellite infall and enhanced by radial migration. *Monthly Notices of the Royal Astronomical Society*, 512(4):4697–4714, June 2022. doi: 10.1093/mnras/stac780.

[25] Yuxi (Lucy) Lu, Melissa K. Ness, Tobias Buck, Joel C. Zinn, and Kathryn V. Johnston. Similarities behind the high- and low-α disc: small intrinsic abundance scatter and migrating stars. *Monthly Notices of the Royal Astronomical Society*, 512(2):2890–2910, May 2022. doi: 10.1093/mnras/stac610.

[26] Khyati Malhan, Rodrigo A. Ibata, Sanjib Sharma, Benoit Famaey, Michele Bellazzini, Raymond G. Carlberg, Richard D'Souza, Zhen Yuan, Nicolas F. Martin, and Guillaume F. Thomas. The Global Dynamical Atlas of the Milky Way Mergers: Constraints from Gaia EDR3-based Orbits of Globular Clusters, Stellar Streams, and Satellite Galaxies. *The Astrophysical Journal*, 926(2):107, February 2022. doi: 10.3847/1538-4357/ac4d2a.

[27] Alan W. McConnachie, Rodrigo Ibata, Nicolas Martin, Annette M. N. Ferguson, Michelle Collins, Stephen Gwyn, Mike Irwin, Geraint F. Lewis, A. Dougal Mackey, Tim Davidge, Veronica Arias, Anthony Conn, Patrick Côté, Denija Crnojevic, Avon Huxor, Jorge Penarrubia, Chelsea Spengler, Nial Tanvir, David Valls-Gabaud, Arif Babul, Pauline Barmby, Nicholas F. Bate, Edouard Bernard, Scott Chapman, Aaron Dotter, William Harris, Brendan McMonigal, Julio Navarro, Thomas H. Puzia, R. Michael Rich, Guillaume Thomas, and Lawrence M. Widrow. The large-scale structure of the halo of the andromeda galaxy. ii. hierarchical structure in the pan-andromeda archaeological survey. *The Astrophysical Journal*, 868(1):55, 2018. ISSN 1538-4357. doi: 10.3847/1538-4357/aae8e7. URL `http://dx.doi.org/10.3847/1538-4357/aae8e7https://iopscience.iop.org/article/10.3847/1538-4357/aae8e7/pdf`.

[28] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205. URL `https://doi.org/10.21105/joss.00205`.

[29] Juan Miro-Carretero, Maria A. Gomez-Flechoso, David Martinez-Delgado, Andrew P. Cooper, Santi Roca-Fabrega, Mohammad Akhlaghi, Annalisa Pillepich, Konrad Kuijken, Denis Erkal, Tobias Buck, Wojciech A. Hellwing, and Sownak Bose. Extragalactic Stellar Tidal Streams: Observations meet Simulation. *arXiv e-prints*, art. arXiv:2409.03585, September 2024. doi: 10.48550/arXiv.2409.03585.

[30] Aura Obreja, Tobias Buck, and Andrea V. Macciò. A first estimate of the Milky Way dark matter halo spin. *Astronomy & Astrophysics*, 657:A15, January 2022. doi: 10.1051/0004-6361/202140983.

[31] William H Oliver. Astrolink, 09 2024. URL `https://github.com/william-h-oliver/astrolink`.

[32] William H Oliver. Astrolink, 09 2024. URL `https://astrolink.readthedocs.io/en/latest/`.

[33] William H Oliver. Fuzzycat, 09 2024. URL `https://github.com/william-h-oliver/fuzzycat`.

[34] William H Oliver. Fuzzycat, 09 2024. URL `https://fuzzycat.readthedocs.io/en/latest/`.

[35] William H Oliver, Pascal J Elahi, Geraint F Lewis, and Chris Power. The Hierarchical Structure of Galactic Haloes: Classification and Characterization with Halo-OPTICS. *Monthly Notices of the Royal Astronomical Society*, 501(3):4420–4437, 12 2021. ISSN 0035-8711. doi: 10.1093/mnras/staa3879. URL `https://doi.org/10.1093/mnras/staa3879`.

[36] William H Oliver, Pascal J Elahi, and Geraint F Lewis. The hierarchical structure of galactic haloes: generalized N-dimensional clustering with CluSTAR-ND. *Monthly Notices of the Royal Astronomical Society*, 514(4):5767–5785, 06 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac1701. URL `https://doi.org/10.1093/mnras/stac1701`.

[37] William H. Oliver, Pascal J. Elahi, Geraint F. Lewis, and Tobias Buck. The hierarchical structure of galactic haloes: differentiating clusters from stochastic clumping with ASTROLINK. *Monthly Notices of the Royal Astronomical Society*, 530(3):2637–2647, May 2024. doi: 10.1093/mnras/stae1029.

[38] Julian Onions, Alexander Knebe, Frazer R. Pearce, Stuart I. Muldrew, Hanni Lux, Steffen R. Knollmann, Yago Ascasibar, Peter Behroozi, Pascal Elahi, Jiaxin Han, Michal Maciejewski, Manuel E. Merchán, Mark Neyrinck, Andrés N. Ruiz, Mario A. Sgró, Volker Springel, and Dylan Tweed. Subhaloes going Notts: the subhalo-finder comparison project. *Monthly Notices of the Royal Astronomical Society*, 423(2):1200–1214, 06 2012. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2012.20947.x. URL `https://doi.org/10.1111/j.1365-2966.2012.20947.x`.

[39] Julian Onions, Yago Ascasibar, Peter Behroozi, Javier Casado, Pascal Elahi, Jiaxin Han, Alexander Knebe, Hanni Lux, Manuel E. Merchán, Stuart I. Muldrew, Mark Neyrinck, Lyndsay Old, Frazer R. Pearce, Doug Potter, Andrés N. Ruiz, Mario A. Sgró, Dylan Tweed, and Thomas Yue. Subhaloes gone Notts: spin across subhaloes and finders. *Monthly Notices of the Royal Astronomical Society*, 429(3):2739–2747, 01 2013. ISSN 0035-8711. doi: 10.1093/mnras/sts549. URL `https://doi.org/10.1093/mnras/sts549`.

[40] Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. XVI. Cosmological parameters. *Astronomy & Astrophysics*, 571:A16, November 2014. doi: 10.1051/0004-6361/201321591.

[41] Stephan R. Sain. Multivariate locally adaptive density estimation. *Computational Statistics & Data Analysis*, 39(2):165–186, 2002. ISSN 0167-9473. doi: https://doi.org/10.1016/S0167-9473(01)00053-6. URL `https://www.sciencedirect.com/science/article/pii/S0167947301000536`.

[42] Federico Sestito, Tobias Buck, Else Starkenburg, Nicolas F. Martin, Julio F. Navarro, Kim A. Venn, Aura Obreja, Pascale Jablonka, and Andrea V. Macciò. Exploring the origin of low-metallicity stars in Milky-Way-like galaxies with the NIHAO-UHD simulations. *Monthly Notices of the Royal Astronomical Society*, 500(3):3750–3762, January 2021. doi: 10.1093/mnras/staa3479.

[43] Volker Springel, Simon D. M. White, Giuseppe Tormen, and Guinevere Kauffmann. Populating a cluster of galaxies – i. results at z = 0. *Monthly Notices of the Royal Astronomical Society*, 328(3):726–750, 2001. ISSN 0035-8711. doi: 10.1046/j.1365-8711.2001.04912.x. URL `https://doi.org/10.1046/j.1365-8711.2001.04912.x`.

[44] Volker Springel, Simon D. M. White, Adrian Jenkins, Carlos S. Frenk, Naoki Yoshida, Liang Gao, Julio Navarro, Robert Thacker, Darren Croton, John Helly, John A. Peacock, Shaun Cole, Peter Thomas, Hugh Couchman, August Evrard, Jörg Colberg, and Frazer Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435(7042):629–636, 06 2005. doi: 10.1038/nature03597.

[45] L. Wang, A. A. Dutton, G. S. Stinson, A. V. Macciò, C. Penzo, X. Kang, B. W. Keller, and J. Wadsley. NIHAO project - I. Reproducing the inefficiency of galaxy formation across cosmic time with a large sample of cosmological hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 454:83–94, November 2015. doi: 10.1093/mnras/stv1937.