



# Revealing the Hierarchical Structure of Galactic Haloes with Novel Data Mining Algorithms

William H. Oliver

*This thesis is presented as part of the requirements for the conferral of the degree:*

Doctor of Philosophy (PhD)

Supervisor:  
Prof. Geraint F. Lewis

The University of Sydney  
School of Physics

2022

# Statement of Originality

I, *William H. Oliver*, declare that this thesis is submitted in partial fulfilment of the requirements for the conferral of the degree *Doctor of Philosophy (PhD)*, from the University of Sydney, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

---

**William H. Oliver**

August 19, 2022

# Abstract

The research within this thesis is directed developing, training, and testing unsupervised astrophysical clustering algorithms that extract meaningful structures from their input data. It is a well-studied consequence of the  $\Lambda$ CDM cosmological model of the Universe that these structures form hierarchically through the continual merging of smaller structures. Once a merger has occurred however, the mergers are not entirely lost but can instead remain detectable as coherent groups for some time – dependent on the ongoing conditions of the surrounding environment. As such, galaxies are expected to contain a myriad of substructure that act as fossil records of the galaxies themselves. As larger and more advanced surveys continue to be conducted, we are faced with the task of unearthing these galaxies and their substructures over a vast range of ever-more-complicated data sets. To tackle this issue, it is necessary to prepare ourselves with appropriate tools that can sift through these data sets and discover new structures. This is the goal that motivates the works within this thesis. First I developed **HALO-OPTICS**, a new algorithm designed to hierarchically classify astrophysical clusters within N-body particle simulations. I showed that it performs well against a current state-of-the-art code (e.g. **VELOCIRAPTOR**) even though it uses comparatively less of the available information within the simulation data. Next I developed **CLUSTAR-ND** and in doing so I make various algorithmic improvements upon its predecessor **HALO-OPTICS**. These upgrades dramatically improved **CLUSTAR-ND**'s computational footprint, its sensitivity to relevant clusters, and its capacity to operate over any size data set containing any number of dimensions. Finally, I developed **CLUSTARR-ND** which boasts all the operational virtues of **CLUSTAR-ND** while also providing an **OPTICS**-style representation of clustering structure and identifying clusters as statistically distinct overdensities (when compared to the noisy density fluctuations) of the input data. As the culmination of the sequential development of state-of-the-art unsupervised clustering algorithms, **CLUSTARR-ND** opens up the opportunity for adaptively providing a meaningful hierarchical astrophysical clustering of any n-point d-dimensional data set with an extremely modest computational demand, resulting in rapid run times.

# Acknowledgements

I want to thank my supervisor Prof. Geraint F. Lewis for his support, dedication, and guidance throughout my candidature. Without his mentorship I would not have discovered the passion I now have for the development and application of unsupervised machine learning algorithms, nor would I have gained so many key insights into the inner workings of a career in research.

I thank Dr. Pascal J. Elahi for his technical wisdom and selfless devotion to understanding and improving my research. I thank Zhen Wan for showing me the ropes on paper writing, modelling, and observing at the AAT.

I would also like to thank my PhD siblings, office mates, and SifA friends – Florian, Lawrence, Zhen, Joseph, Angela, Rasel, Sonia, Di, Gurashish, and Becky – for the many lunches well spent and thought-provoking discussions had.

To my wonderful partner Jorja, thank you for your sincere and much needed encouragement, for being a sounding board for all my ramblings, for making the coffee, for ensuring that I still went outside, and for being the love of my life. I couldn't have gotten this far without you.

To my incredible Mum and Grandma, thank you for nurturing my love of maths and science from a young age. Thank you to my Dad and sister Ruby for inspiring me to pursue the things I enjoy. Also to Mum, Rodney, Grandma, Rodd, and Sharon for being so supportive of my (seemingly perpetual) student life.

To my friends from home and Sydney, thanks for making the last three and half years enjoyable and for keeping me sane.

I am grateful to have had the financial support of the Hunstead Student Support Scholarship, the Paulette Isabel Jones PhD Completion Scholarship, and the Post-graduate Research Support Scheme during my candidature. I also wish to thank the University of Sydney HPC for providing their services and computational resources such as the virtual desktop environment Argus which has contributed to the research results within this thesis.

# List of Publications

The following refereed journal papers are reproduced with permission in this thesis:

1. *The Hierarchical Structure of Galactic Haloes: Classification and Characterisation with Halo-OPTICS*. **W. H. Oliver**, P. J. Elahi, G. F. Lewis, & C. Power. *MNRAS* 501, 4420, 2021. [[arXiv:2012.04823](#)].
2. *The Hierarchical Structure of Galactic Haloes: Generalised N-Dimensional Clustering with CluSTAR-ND*. **W. H. Oliver**, P. J. Elahi, & G. F. Lewis. *MNRAS* 514, 5767, 2022. [[arXiv:2201.10694](#)].

At the time of writing this thesis, the following paper has not yet undergone peer review. It will be submitted shortly to a refereed journal.

3. *The Hierarchical Structure of Galactic Haloes: Differentiating Clusters from Non-Poissonian Noise with CluSTARR-ND*. **W. H. Oliver**, P. J. Elahi, & G. F. Lewis. *in preparation*.

I am a co-author of the following refereed journal papers, which are included with permission in App. A of this thesis for completeness.

- A1. *On the Origin of the Asymmetric Dwarf Galaxy Distribution around Andromeda*. Z. Wan, **W. H. Oliver**, G. F. Lewis, J. I. Read, & M. L. M. Collins. *MNRAS* 492, 456, 2020. [[arXiv:1912.02393](#)].
- A2. *The Dynamics of the Globular Cluster NGC 3201 out to the Jacobi Radius*. Z. Wan, **W. H. Oliver**, H. Baumgardt, G. F. Lewis, M. Gieles, V. Hénault-Brunet, T. de Boer, E. Balbinot, G. Da Costa, & D. Mackey. *MNRAS* 502, 4513, 2021. [[arXiv:2102.01472](#)].

The copyright for papers 1, 2, A1, and A2 is held by Oxford University Press (OUP) and the articles are reproduced here with permission from OUP:

*"Rights retained by ALL Oxford Journal authors: [...] The right to include the article in full or in part in a thesis or dissertation, provided that this is not published commercially."* [[OUP](#), 06/2022].

## Authorship Attribution Statements

The author contribution statements for each paper are correct and are specified in their respective chapters.

---

**William H. Oliver**

August 19, 2022

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements throughout this thesis are correct.

---

**Geraint F. Lewis**

August 19, 2022

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>List of Publications</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Structure Formation within the Universe . . . . .	1
1.1.1 The $\Lambda$ Cold Dark Matter Cosmological Model . . . . .	1
1.1.2 Galactic Building Blocks . . . . .	3
1.1.3 Galaxies . . . . .	7
1.1.4 Large-Scale Structure . . . . .	11
1.2 Approach . . . . .	13
1.2.1 Outline . . . . .	14
<b>2 Clustering Algorithms</b>	<b>16</b>
2.1 Measuring Similarity . . . . .	17
2.1.1 Hamming Distance . . . . .	17
2.1.2 Minkowski Distance . . . . .	17
2.1.3 Mahalanobis Distance . . . . .	18
2.1.4 Wasserstein Distance . . . . .	18
2.2 Clustering Models . . . . .	18
2.2.1 Connectivity-based Clustering . . . . .	19
2.2.2 Centroid-based Clustering . . . . .	20
2.2.3 Distribution-based Clustering . . . . .	21
2.2.4 Density-based Clustering . . . . .	21
2.3 Computational Techniques . . . . .	22
2.3.1 Pre-Process Methods . . . . .	23
2.3.2 Mid-Process Methods . . . . .	24
2.3.3 Post-Process Methods . . . . .	25

2.4	Statistical Evaluation . . . . .	26
2.4.1	Internal Evaluation . . . . .	26
2.4.2	External Evaluation . . . . .	27
2.5	The State of Astrophysical Structure Finding . . . . .	28
2.5.1	Simulation Specific Finders . . . . .	28
2.5.2	Observation Specific Finders . . . . .	32
2.5.3	Generalised Structure Finders . . . . .	33
<b>3</b>	<b>A Novel Approach to Astrophysical Clustering</b>	<b>35</b>
3.1	Structure Finding with Halo-OPTICS . . . . .	36
<b>4</b>	<b>A More Suitable Algorithm for Big Data</b>	<b>55</b>
4.1	Structure Finding with CluSTAR-ND . . . . .	56
<b>5</b>	<b>An Entirely Data-Driven Method</b>	<b>76</b>
5.1	Structure Finding with CluSTARR-ND . . . . .	77
<b>6</b>	<b>Conclusions</b>	<b>90</b>
6.1	Future Outlook . . . . .	91
6.1.1	Additional Machine Learning Approaches . . . . .	92
6.1.2	Applications to Observational Data Sets . . . . .	94
<b>A</b>	<b>Contributing Publications</b>	<b>99</b>
A.1	The Asymmetric Dwarf Galaxy Distribution around Andromeda . . . . .	99
A.2	The Dynamics of NGC3201 . . . . .	112
<b>B</b>	<b>Supplementary Codes</b>	<b>126</b>
B.1	The Halo-OPTICS algorithm . . . . .	126
B.2	The CluSTAR-ND algorithm . . . . .	137
B.3	The CluSTARR-ND algorithm . . . . .	146
	<b>Bibliography</b>	<b>157</b>



# Chapter 1

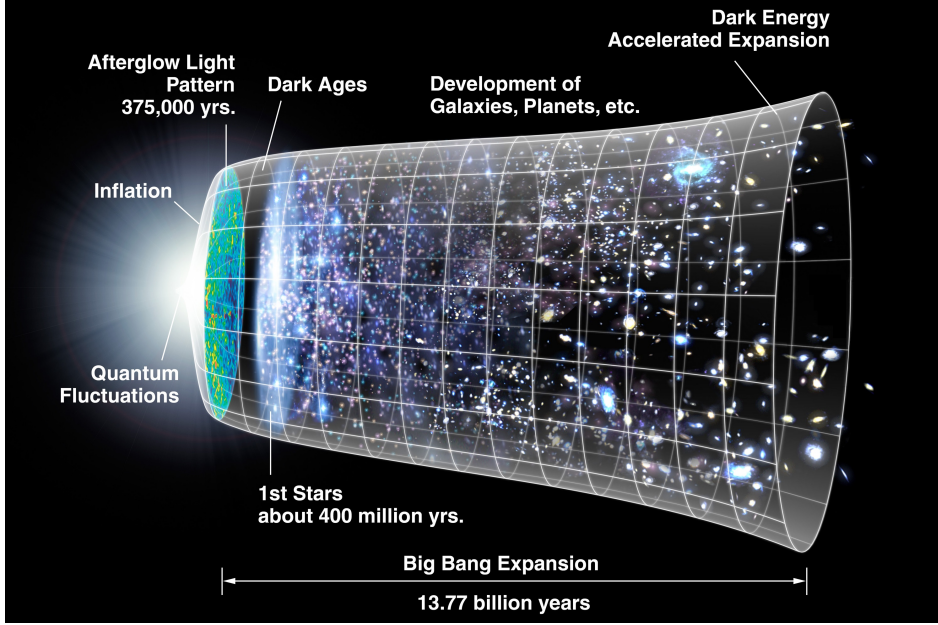
## Introduction

### 1.1 Structure Formation within the Universe

The observed matter content of the Universe is particularly clumpy with discernible astrophysical structures existing over a mass range spanning  $\gtrsim 20$  orders of magnitude, from stars (down to  $\sim 10^{-1}M_{\odot}$  for hydrogen-fusing main-sequence red dwarf stars [1–3] and down to  $\sim 10^{-2}M_{\odot}$  for deuterium-fusing brown dwarf stars [4]) to cluster of galaxies (up to  $\sim 10^{19}M_{\odot}$  as in the case of the largest known structure, the Hercules-Corona Borealis Great Wall [5–7]). The formation of these structures is guided by the evolution of the Universe – which under the widely accepted theory of the  $\Lambda$  Cold Dark Matter cosmological model ( $\Lambda$ CDM) [8–12] is predicted to have assembled hierarchically through the continual merging of smaller constituent substructures [13–16]. This process applies to structure formation on nearly all scales, from the formation of compact objects to the large scale structure of the Universe, and at nearly all times.

#### 1.1.1 The $\Lambda$ Cold Dark Matter Cosmological Model

The  $\Lambda$ CDM model of cosmology is a unifying theory that explains a wide range of observed and theorised phenomena within the Universe. Currently, it provides the most complete understanding of the Universe’s origin, evolution, and composition – as such it is often referred to as the standard model of cosmology. According to this standard model, the Universe; began with the Big Bang [17–22] and subsequently expanded exponentially for a brief period known as *inflation* [23]; is currently experiencing accelerated expansion due to *dark energy* making up  $\sim 68\%$  of its energy density (resulting in a positive cosmological constant,  $\Lambda$ ) [24–27]; is also composed of  $\sim 5\%$  baryonic matter and  $\sim 27\%$  cold dark matter [28, 29] such that



**Figure 1.1:** An artistic illustration of the history of the Universe and the formation of the structure within it as predicted by the  $\Lambda$ CDM model of cosmology. Depicted is the quantum fluctuations of the Big Bang, inflation, the CMB, the dark ages that preceded the birth of the first stars, and then the gradual structure formation that occurred within the expanding Universe that is accelerated by the presence of dark energy. This figure has been reproduced from [43].

its energy-density combines precisely to allow for it to be spatially flat [30–33]; and is gravitationally described by the General Theory of Relativity [34, 35], specifically by the Friedmann-Lemaître-Robertson-Walker metric [20, 36–42] on cosmological scales.

Accordingly, the  $\Lambda$ CDM model has 6 independent parameters: the physical baryon density ( $\Omega_b h^2 = 0.0224 \pm 0.0001$ ); the physical dark matter density ( $\Omega_d h^2 = 0.120 \pm 0.001$ ); the age of the Universe ( $t_0 = 13.787 \pm 0.020$  Gyr); the scalar spectral index ( $n_s =$ ); the curvature fluctuation amplitude ( $\Delta_R^2 = 2.441^{+0.088}_{-0.092} \times 10^{-9}$  [44]); and the reionization optical depth ( $\tau = 0.054 \pm 0.007$ ). Implicitly, there are a further 6 assumed-to-be-fixed parameters underlying this model: the total density ( $\Omega_{\text{tot}} = 1$ ); the equation of state of dark energy ( $w = -1$ ); the tensor/scalar ratio ( $r = 0$ ); the running of the spectral index ( $dn_s/d\ln k = 0$ ), the sum of the three neutrinos masses ( $\Sigma m_\nu = 0.06$  eV/ $c^2$  [45]); the effective number of relativistic degrees of freedom ( $N_{\text{eff}} = 3.046$  [45]).

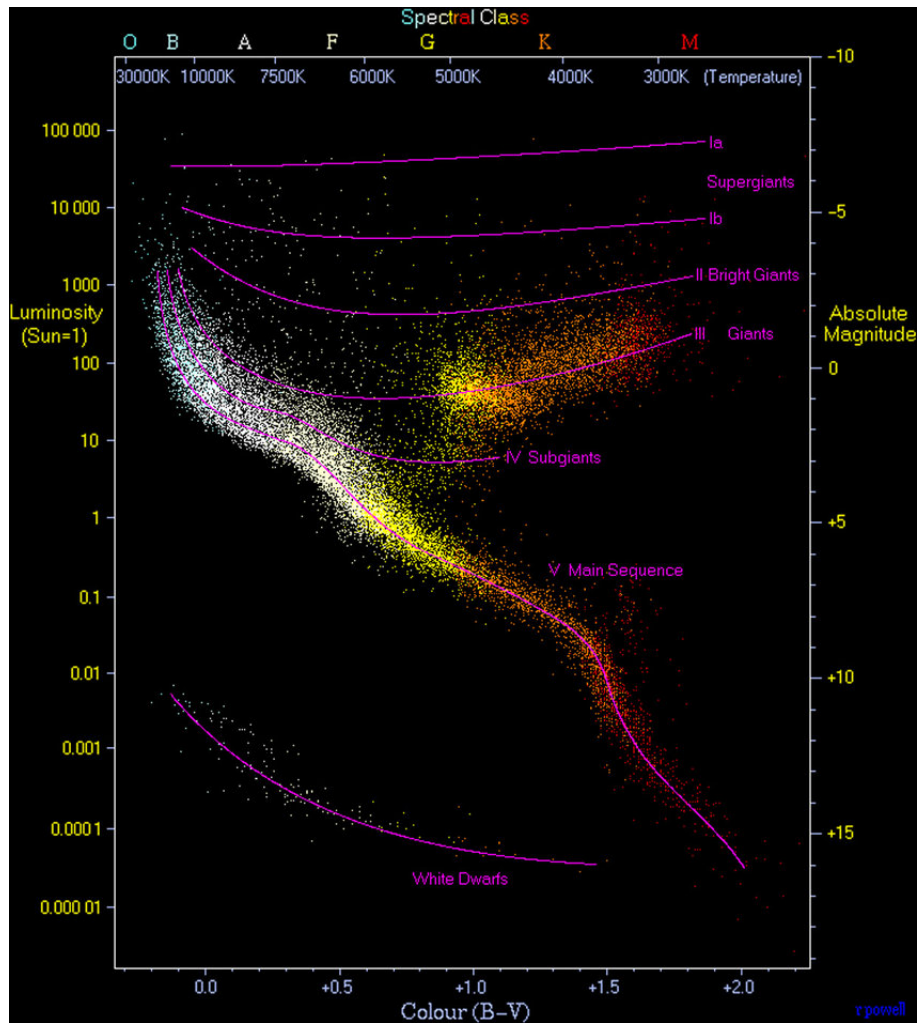
The values of these parameters have been confirmed by measuring them directly (from [33] unless otherwise cited). They can also be validated by matching observations with numerical simulations. Such simulations work by initialise a mock universe with particles constructed to represent small clusters of baryonic and dark matter. With these initial conditions, the simulation then exposes these particles to various

*test* laws of nature and evolves a universe whose structure and evolution can be compared with observation to credit or discredit those test laws.

As illustrated in Fig. 1.1, the  $\Lambda$ CDM model also predicts that following inflation the mass distribution of the Universe was described by a Gaussian distribution of adiabatic density fluctuations with a similar amplitude on all spatial scales [46] – arising from the quantum fluctuations inherent in the initial conditions of the Big Bang. This is confirmed through observations of the Cosmic Microwave Background (CMB) [30, 31, 47–50] and is the precursor for the evolution of all structure. It is the coupling of these small anisotropies with the effects of gravity that leads to localised gravitational instabilities which collapse over time to form structure [51]. The spatially flat and expanding Universe of  $\Lambda$ CDM, composed with  $\sim 5.4$  times as much cold dark matter as baryonic matter, dictates the way in which these instabilities collapse, the type of structures that form, as well as the manner in which this occurs – giving rise to the agglomerative (bottom-up) model of hierarchical structure formation.

### 1.1.2 Galactic Building Blocks

The first structures to form in the Universe were concentrations of dark matter and, attracted to these through their gravitational pull, clouds of baryonic particles. Initially, these clouds were made of  $\sim 75\%$  protons and  $\sim 25\%$  helium nuclei with small traces of the nuclei of deuterium, lithium, and beryllium – according to accepted models of the Big Bang Nucleosynthesis (BBN) that followed inflation [52]. However as these clouds cooled, these particles were able to recombine with electrons when the Universe was  $\sim 3.8 \times 10^5$  years old [53] making the clouds mostly neutral hydrogen – and later molecular hydrogen. Over time these gas clouds became more dense and developed their own substructure due to their own internal gravitational instabilities. Broadly speaking, these gas clouds can be categorised into Giant Molecular Clouds (GMCs), *clumps*, and *cores* [51] – with each of these being a substructure of the one before it. GMCs are large and typically have masses of  $\sim 10^5$ – $10^7 M_\odot$ , while clumps and cores – being smaller and denser – have masses of  $\sim 10^2$ – $10^4 M_\odot$  and  $\sim 10^{-1}$ – $10 M_\odot$  respectively. These gas cloud cores, often called protostellar cores, will eventually collapse to form a star [54]. Consequently, the GMCs and clumps will harbour the transformation of many cores into stars and as such form themselves into larger stellar structures.



**Figure 1.2:** An observational Hertzsprung-Russell diagram composed with 22000 stars plotted from the Hipparcos Catalogue [55] and 1000 stars from the Gliese Catalogue of nearby stars [56]. The diagram depicts the stars as described by their luminosity/absolute magnitude and the colour/spectral type/surface temperature. Multiple luminosity classes are also depicted here. This figure has been reproduced from [57].

## Stars

The characteristics of a star depend entirely upon the environment that it forms within. Though the star formation process is not fully understood, it is known that there are many factors that affect the type of the final star including the; protostellar core size and density; gas temperature and cooling rate; gas cloud angular momentum distribution; metallicity within the gas; radiation and magnetic field strengths; as well as the turbulence of the gas cloud [58–61]. The effects of these factors in the process of star formation give the resultant star its type which is typically described using the Morgan-Keenan (MK) classification system [62, 63]. The MK system classifies

stars based on their spectral type<sup>1</sup> and luminosity class<sup>2</sup> which can be visualised through the Hertzsprung-Russell diagram (a plot of absolute magnitude/luminosity vs spectral type/colour/temperature) in Fig. 1.2.

Stars can be categorised into additional physical classes – such as populations I, II [73], and III [13, 74, 75]. These populations correspond directly to age and metallicity (the abundance of elements heavier than hydrogen and helium), such that population III stars<sup>3</sup> were the first stars forming from only those light elements that were created during the BBN and population I stars are the most recently formed stars emerging from the stellar debris of the previous populations. Population I and II stars are observed directly in the Milky Way’s spiral arms and bulge/globular clusters respectively [80, 81]. Importantly, due to the differing chemical abundances between stars born from differing gas clouds and of different types, the abundances of each metal within each star serve as a record of its formation and evolution and can be used to associate the stars with their parent structure.

### Open and Globular Clusters

Among the structures defined by groupings of stars are open and globular clusters. Open clusters are the direct result of active star formation in a GMC, and as such, consist of population I stars and are located in dynamic regions of their parent structure<sup>4</sup> – where gas is condensed and star formation occurs [87–89]. The stars of an open cluster are only loosely bound by their mutual gravitational attraction and hence are easily disassociated from one another [90, 91]. For this reason, open clusters will often be observed to have an irregular shape and, as they age, will become spatially and kinematically indistinguishable as a coherent group when compared to their neighbouring stars [88, 89, 92–94]. Since the lifetime of the stars themselves is commonly much longer than the expected lifetime of the open cluster, once the cluster has disassociated, the stellar chemical abundances become the strongest predictor of its existence and former phase-space coherence.

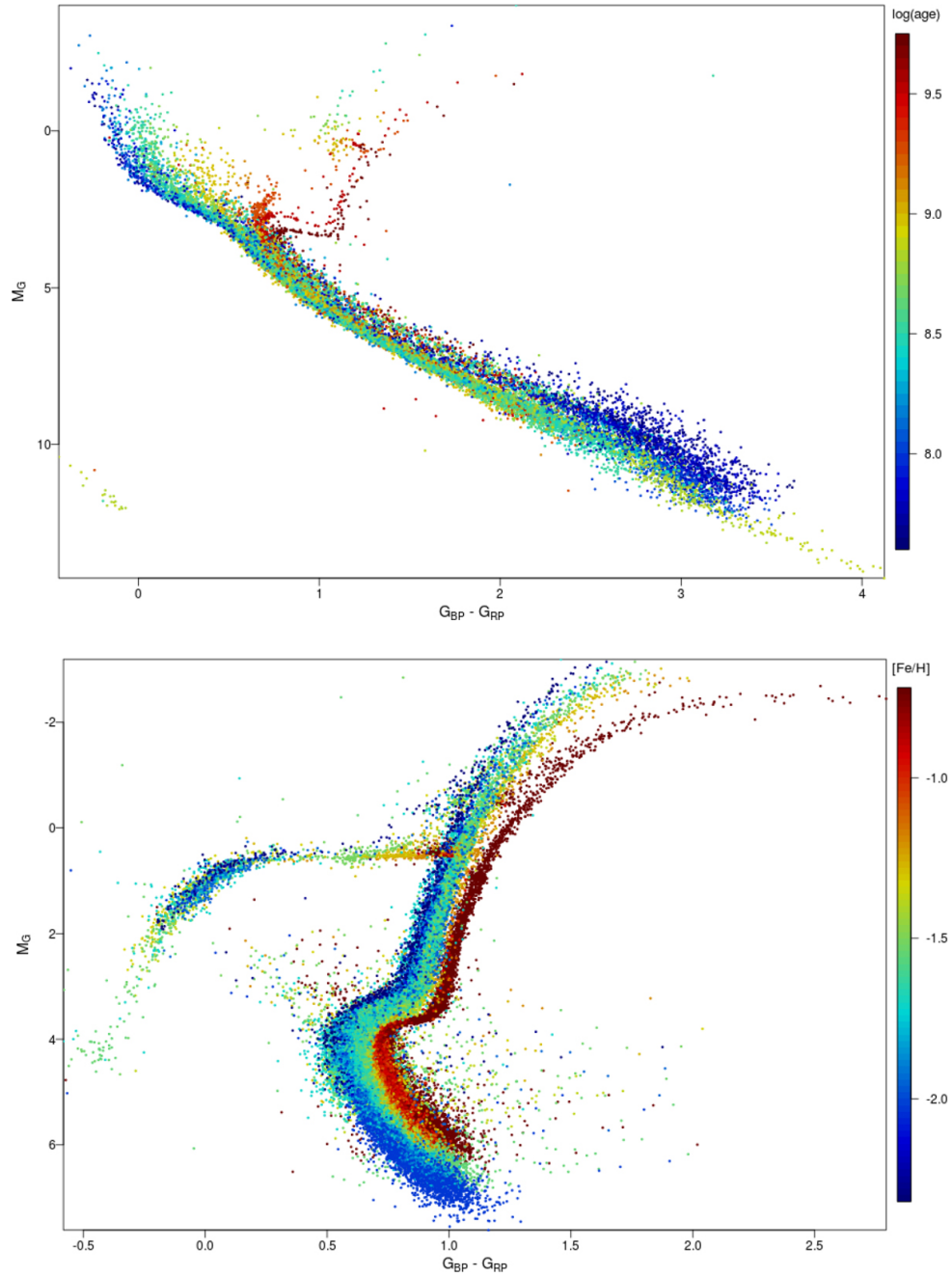
---

<sup>1</sup>Historically, these only included the O, B, A, F, G, K, and M types (in order of hot to cold), although now these types now also include the hotter W type Wolf-Rayet stars [64, 65] and the colder L, T, and Y type brown dwarf stars [66–72]. Each spectral type is also followed by a subdivision from 0 to 9 that provides a finer tuning to this classification.

<sup>2</sup>The luminosity classes are 0, I, II, III, IV, V, VI, and VII (or D) and classify whether a star is a hypergiant, supergiant, bright giant, giant, subgiant, dwarf (main-sequence), subdwarf, or white dwarf (or degenerate) respectively.

<sup>3</sup>At present, only indirect evidence of population III stars exists as these stars are hypothesised to have been very short-lived and only existed prior to the re-ionisation epoch [76–79].

<sup>4</sup>There are  $\gtrsim 3800$  known open clusters [82–85] within the Milky Way – most of which are found within the spiral arms [86] – with estimates predicting that the total number could be an order of magnitude larger than this.



**Figure 1.3:** Composite Hertzsprung-Russell diagrams consisting of 32 open clusters (top) and 14 globular clusters (bottom) determined using Gaia DR2 data [95]. The open cluster stellar members are coloured according to  $\log(\text{age})$  using extinction and distance moduli as determined from the Gaia data. The globular cluster stellar members are coloured according to metallicity,  $[\text{Fe}/\text{H}]$ . This figure has been reproduced from [96].

Contrary to open clusters, globular clusters are more massive ( $\sim 10^6 M_\odot$ ), much older (up to  $\sim 13 \times 10^9$  years old and contain predominantly population II stars), and are typically entirely devoid of dark matter, gas, and dust [97]. While the mechanisms behind their formation is still poorly understood, they are commonly found in the stellar haloes of galaxies. Here they only experience weak tidal forces, which – along with them being more massive – is thought to be the reason that they are able to remain self-bound for extended time periods [98]. Globular clusters are typically pressure supported and hence have large velocity dispersion. As such, globular clusters are predominantly identified by being strongly spatially and chemically coherent<sup>5</sup>. Fig. 1.3 depicts composite Hertzsprung-Russell diagrams for both open and globular clusters as observed with Gaia DR2 – indicating that members of these structures can be determined through the use of stellar isochrone models.

### 1.1.3 Galaxies

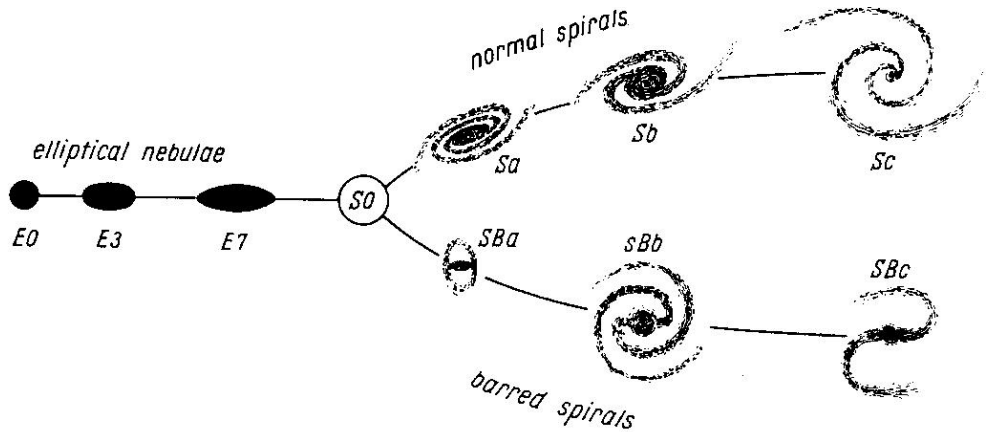
As is mentioned in Sec. 1.1, the  $\Lambda$ CDM cosmological model predicts that all baryonic structure forms through the continual merging and accretion of smaller stellar structures. The main driver of this formation for galaxies is the presence of large amounts of dark matter which had gradually formed into self-bound haloes following inflation [51, 104]. The presence of these dark haloes facilitates the condensing of baryonic gas within them which then coalesces into increasingly large observable structures. Recent observations of the most distant galaxies suggest that dark protogalaxies had formed before the first stars had even emitted any light [105–110].

#### Galactic Classification

Galaxies have been observed throughout the Universe in various shapes and sizes ( $\sim 10^8$ – $10^{14} M_\odot$ ) [51]. Historically, the method of categorising these is through the Hubble classification system (see the original schematic in Fig. 1.4) which separates them into elliptical, spiral, lenticular, and irregular [111, 112]. Elliptical galaxies (E) appear smooth, without molecular gas, and are, as their name suggests, ellipsoidal in shape [113]. Elliptical galaxies are therefore divided into subclasses, dependent upon how elliptical they are, i.e. E1, E2, ..., E7<sup>6</sup> such that this integer has been rounded

<sup>5</sup>Globular clusters have also been sub-categorised by the degree of concentration of stars toward their core in a system known as the Shapley-Sawyer Concentration Class [99–103]. This system classifies globular clusters on a scale from one to twelve such that a Class I cluster has a high density of stars in its centre while a Class XII cluster has essentially uniform density throughout.

<sup>6</sup>There are no recorded E8, E9, or E10 galaxies and even most E4 – E7 are mis-classified lenticular galaxies [114–116].



**Figure 1.4:** The original Hubble tuning fork diagram indicating schematics of elliptical, lenticular, and spiral (with and without a bar) galaxies as published in [112].

from the result of  $10(1 - b/a)$  where  $a$  and  $b$  are the lengths of the semi-major and semi-minor axes respectively [117].

Spiral galaxies (S) have thin disks and spiral arms. Roughly two-thirds of these also exhibit a barred structure in their centre (denoted SB) [118]. As such, they are not only categorised based upon the presence of this bar, but also upon the; fraction of their total light that comes from their central bulge; the tightness with which their spiral arms are wound; and the degree to which the spiral arms are resolved into stars, molecular hydrogen, and ordered dust lanes. The latter three criteria are correlated such that; spiral galaxies with a distinct bulge usually also display tightly wound spiral arms with poorly resolved stellar, molecular hydrogen, and dust lane components are labelled as *Sa* or *SBa*; spiral galaxies with weak or absent central bulges usually have open arms with clearly resolved material components are labelled as *Sc* or *SBc*; and with a spiral galaxy whose features belong in between these types being labelled as *Sb* or *SBb*.

Lenticular galaxies (*S0*) are a galaxy classification that is intermediary to both elliptical and spiral galaxies. Similarly to elliptical galaxies, they exhibit a smoothly varying light distribution without spiral arms. Similarly to spiral galaxies, they feature a thin disk and a bulge – although this is typically more dominant than in a spiral galaxy. Lenticular galaxies may also have a central bar, in which case they are denoted by *SB0*. Irregular galaxies (Irr) are those that do not fit the characteristics of the other galaxy types. They lack any obvious symmetry, do not have a disk or central bulge. Generally, their appearance is uneven and patchy – although they do often show regions dominated by molecular hydrogen. Originally, galaxies were thought to progress from elliptical, through the lenticular stage, and then on to

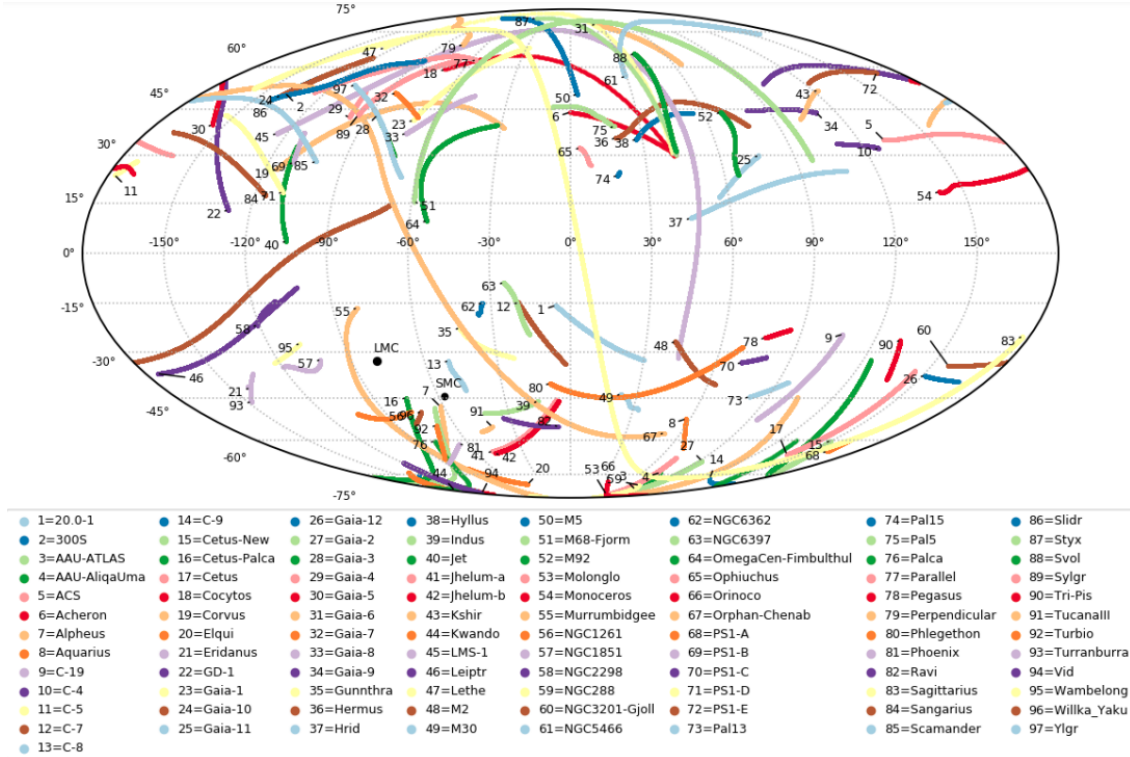


become a spiral galaxy – giving rise to the shape of the Hubble *tuning-fork* diagram shown in Fig. 1.4. While this is no longer an accepted theory of galactic formation and evolution, some research has suggested that the disk structure of spiral galaxies may be built around an existing elliptical galaxy [113, 119, 120]. The conditions that create each these galaxy types are still a topic of active research.

Further classifications of galaxies can be made to extend the Hubble classification scheme. Dwarf galaxies are denoted with the prefix of a lowercase 'd' to above types – although historically there have also been dwarf spheroidal galaxies (dSph) and there is now evidence for and against differentiating these from dwarf ellipticals. According to the system developed by de Vaucouleurs [121–123], galaxies can be classified using a finer gradation of Hubble's scheme e.g. S0a (between lenticular and a spiral with tightly wound spirals), Sbc (between Sb and Sc), and SA (for spirals without bars – and SAB for weakly barred spirals). In this system, there are new spiral types which appropriately extend the spiral galaxies towards the irregular galaxies (Scd, Sd, Sdm, Sm, Im, I0 – and their barred counter-types – where 'm' stands for Magellanic since the Magellanic Clouds are the model for the Im type) and there are also *peculiar*-typed galaxies (P) which do not fit the standard classification scheme. Peculiar galaxies show some coherent features but not others and are typically in the midst of transformation – having been strongly perturbed due to a recent major-merger such as the Antennae galaxy. An addition feature, rings, is also classified in this extended system – where '(r)' denotes galaxies possessing at least one ring-like structure, '(s)' for those without rings, and '(rs)' for transition galaxies.

## Galactic Substructure

During the hierarchical satellite merging process that guides the formation of galaxies, structures can become disrupted and the debris of this event can remain detectable as a distinct and coherent stellar structure for some time that is dependent on the ongoing conditions of the surrounding environment [125–128]. In such a disruption event, the tidal forces will begin to distort both structures involved in the merge along an axis pointing roughly towards and away from their perturber. This distortion will begin to shear the structures and introduce a velocity differential across the tidal force vector. The magnitude of this differential depends upon the mass and density profile of the structures as well as their positions and velocities, but predominantly, the smaller structure will be separated into its tidal debris and a progenitor. If both structures are of similar mass, then tidal debris will be ejected from both. As such, the presence of this debris is a strong indicator of past interactions between the



**Figure 1.5:** The positions on-the-sky of known tidal streams within the Milky Way found by various studies. This figure has been reproduced from [124].

host galaxy and other structures and therefore is effectively a record of its formation history [129–132]. The debris can form from the disruption of any globular cluster, dwarf or regular galaxy in the presence of another galaxy and as a result is not typically self-bound – although most are gravitationally bound to their host galaxy.

Categorically, the debris will either form an elongated curved structure or a thin membrane – commonly referred to as stellar (or tidal) streams and shells [133, 134] respectively. In observations, stellar streams are most commonly seen in spiral galaxies [135–139] while shells are more common within in elliptical galaxies [140–142] – although this discrepancy may be the result of our position with respect to these structures. In simulations, it has been shown that shells are produced following the disruption of satellites that have merged on near-radial orbits while streams are generated from disrupting satellites on more circular orbits [127, 128, 143, 144]. In the Milky Way, there are currently  $\sim 100$  known stellar streams as shown in Fig. 1.5 – the progenitors of which are disrupting or disrupted globular clusters and dwarf galaxies. Whereas in simulations, it is expected that many more streams and shells should exist [16, 145–149]. This is active point of research within the field of galactic archaeology.

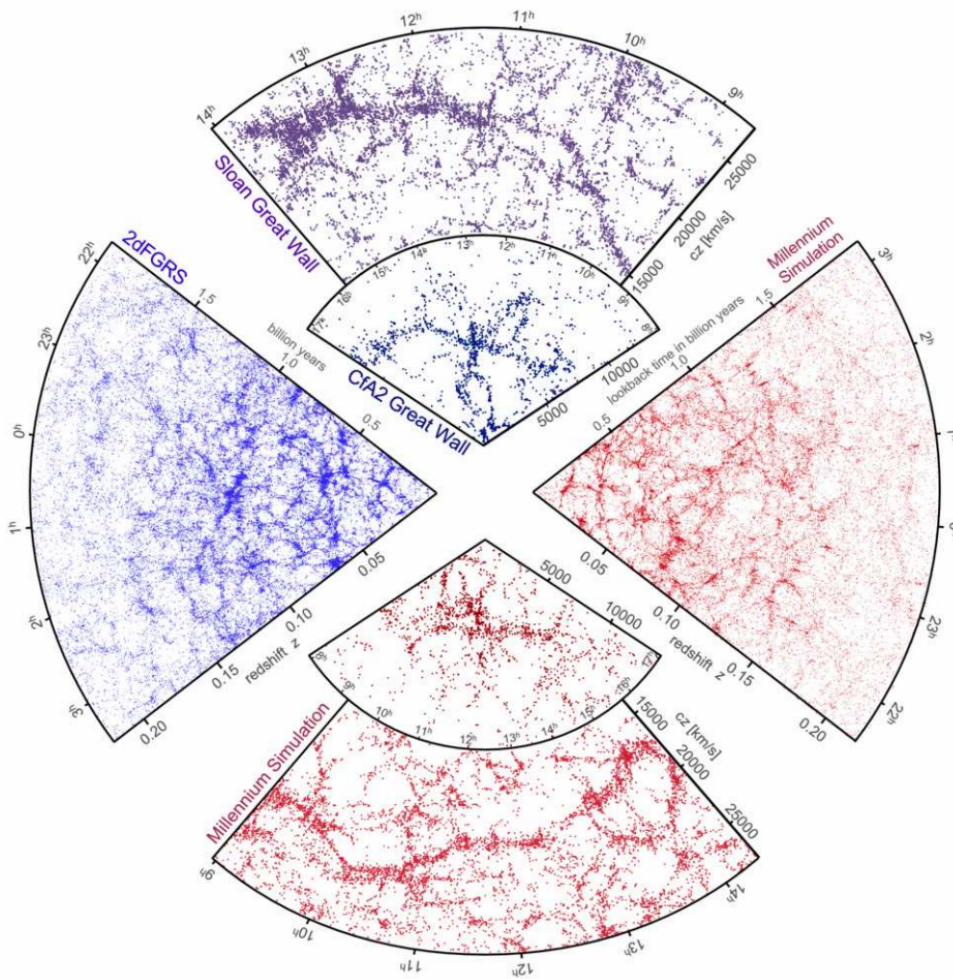
### 1.1.4 Large-Scale Structure

Just as gas and dust particles are grouped to form clouds and stars, and clouds and stars are grouped to form clusters and galaxies, galaxies are also grouped into parent structures on scales much larger than themselves. Referred to as large-scale structure, the mechanism behind the formation of these groupings of galaxies and their constituents is, at its core, largely similar to that of the formation of the galaxies themselves – the macroscopic stretching of Big Bang quantum fluctuations via inflation and the gravitational instability that resulted from this leading to the accretion of matter (particularly of dark matter). However, the configuration of these structures are distinct in shape [9].

Both  $\Lambda$ CDM-based simulations and observations show that the matter of the Universe organises itself on these into a cosmic web (as shown in Fig. 1.6). Ongoing and upcoming observational surveys such as DES [152], Euclid [153], and 4MOST [154] still require the identification of the cosmic web and its components within them. The cosmic web can be sub-categorised into voids, sheets, filaments, and knots – which can be thought of as corresponding to overdensities in 0, 1, 2, and 3 spatial dimensions respectively [155–158]. It is also in approximately this order that these structures will form chronologically. This can be understood by considering an ellipsoid-shaped Gaussian perturbation in the early Universe following inflation. The perturbation will collapse fastest along its shortest axis and create a *pancake*-like sheet. The intermediate axis will collapse and fragment next, turning the sheet into a filament (or perhaps multiple filaments). Finally, the longest axis will collapse into a knot.

#### Voids

Voids are the vast regions of space between sheets and filaments defined as containing very few or no galaxies and being dominated by accelerated expansion. In addition to emerging as the negative space adjacent to structure forming regions, the formation of voids is thought to be shaped by baryon acoustic oscillations in the early Universe – i.e. gravity driven oscillations occurring in collision-prone baryonic materials that are analogous to the compression and rarefaction of sound waves in air [159–162]. As inflation slowed, these oscillations resulted in both underdense and overdense regions. Where the latter have developed into baryonic structures, the former have grown into voids. Voids will typically be 10–200 Mpc in diameter with particularly large voids being labelled supervoids [163, 164].



**Figure 1.6:** A series of two-dimensional slices of the galaxy distribution found within spectroscopic redshift surveys (upper/left wedges) and mock catalogues from the Millennium N-body simulation [150] (lower/right wedges). The depiction not only shows the various types of large-scale structure but also that this has developed and become increasingly clustered over time – indicated by the increased smoothness in the left/right wedges at redshift increases. This figure has been reproduced from [151].

## Sheets

Sheets, also referred to as great walls, are large planar structures containing galaxies. As they age, the galaxies within them begin to align into linear filament-like structures, however the signature of the great wall remains intact as these linear structures will be locally co-planar. Great walls can be hundreds to thousands of Mpc in length and width while only being a few Mpc thick [5–7, 165–167].

## Filaments

Filaments are elongated structures that form from sheets and are observed to connect galaxy clusters linearly into the cosmic web [168, 169]. Their cross-section is roughly circular throughout the length of their principle axis. They are thought to play a major role in galaxy formation by directing the flow of extra-galactic material towards galaxies and clusters of galaxies. Some studies suggest that they are responsible for galaxy alignment within galaxy clusters [170, 171] and for statistically unusual alignments within the galaxies themselves such as hemispherical and planar asymmetries of galactic substructure [172–177].

## Knots

Knots, often referred to as galaxy clusters, are typically defined as having both a *sufficiently* high density and number of galaxies [178]. This leads to some ill-defined notions about exactly what a galaxy cluster is, however, typical galaxy clusters contain 50–10,000 member galaxies, have masses of  $10^{14}$ – $10^{15} M_{\odot}$ , and have diameters of 1–10 Mpc [179–181]. Galaxy clusters can be further classified into different morphological types using the Bautz-Morgan system [182, 183] – such that a type I galaxy cluster possesses a large bright central dominant galaxy, a type III galaxy cluster exhibits no dominant central galaxy<sup>7</sup>, and type II galaxy clusters contain central elliptical galaxies whose brightness is intermediate to those in types I and III. Lesser associations of galaxies that do not meet these requirements are typically labelled as galaxy groups [184–186] – such as the Local Group that the Milky Way belongs to – while associations of galaxy clusters and galaxy groups are called galaxy superclusters [187–190] (originally labelled second-order clusters [178]). Due to the bottom-up hierarchical formation of structure that occurs within the  $\Lambda$ CDM cosmological model, galaxy groups, clusters, and superclusters are the latest structures to form with their ongoing formation still occurring today.

## 1.2 Approach

In addition to providing a brief overview of the astrophysical structure within the Universe and the mechanisms under which it has formed, the secondary purpose of Sec. 1.1 is to highlight a vast network of classifications and classification systems that have been placed on the patterns that have emerged due to the nature of the Universe. These classification systems include those defined by; events that occur for

---

<sup>7</sup>Types IIIE and IIIS are often used to denote a higher proportion of elliptical or spiral galaxies respectively.

fixed periods of time (inflation, BBN, recombination, etc.); physical overdensities of matter that occur in space and time (gas clouds, stars, globular clusters, galaxies, filaments, etc.); subdivisions of object types that are labelled due to a value of a particular attribute (protostellar cores, population I stars, SBa galaxies, type III galaxy cluster, etc.); and the names of different models/theories ( $\Lambda$ CDM, General Theory of Relativity, etc.). Each of these systems of classification has come about through a process of pattern recognition known as clustering.

Clustering the objects of the Universe allows us to study the origin of these objects and make predictions about the nature of the Universe. Specifically, classifying stars separately from groupings of hydrogen gas allows us to study the conditions needed in order for the gas to undergo nuclear fusion. Similarly, classifying open clusters from stars simply belong to their host galaxy allows to study the conditions that lead to the expulsion of gas from the their interior. By classifying spiral galaxies from elliptical galaxies, tidal streams from shells, voids from knots all contributes to our ability to study the processes in play that govern their formation. However, while each of these structures can be defined as their constituent objects bounded by some region of overdensity (at least within some intrinsic feature space) that is in essence the limit of their similarities. Hence, in order to classify these structures and tell them apart from each other we need a systematic approach that robustly classifies them – preferably with the aid of as little prior knowledge of their existence as possible.

### 1.2.1 Outline

The focus of the research works within this thesis is the development of new astrophysical clustering algorithms that produce high quality clustering results and are applicable to any size data set with any feature space. In particular, the objective is to build these algorithms so that they can excel at uncovering galaxies and galactic substructure from both large-scale synthetic and observational survey data. By creating such algorithms, the ultimate goal of this thesis is provide the necessary tools that can be used to study the formation and evolution of galaxies such as our own MW and others in the Local Group.

To establish the existing and relevant work on the topic, I first review clustering algorithms in Chapter 2. I introduce their development and use throughout history as well as summarise their functionalities and the computational/statistical techniques used alongside them. I also discuss the current state of astrophysical clustering algorithms and outline the existing divide between simulation-specific and observation-specific astrophysical clustering algorithms – asserting that generalised structure

finding algorithms are sorely needed to bridge this gap.

In Chapter 3, I begin the task of developing such an algorithm by creating **HALO-OPTICS** – which extends the general-purpose density-based clustering algorithm **OPTICS** to be well-suited to find galaxies and their substructure from the spatial information of their constituents. I then develop the novel generalised astrophysical clustering algorithm **CLUSTAR-ND** in Chapter 4 which – when compared to **HALO-OPTICS** – boasts far more modest run-times and can be applied to data sets of any size and dimensionality while exhibiting a finer sensitivity to existing structure. As a work in progress and the last installation of this series of generalised astrophysical clustering algorithms, I then present **CLUSTARR-ND** in Chapter 5. **CLUSTARR-ND** reduces the complex cluster extraction process of its predecessors to one with a simple-to-interpret functionality, i.e. the returned clusters are statistically distinct from the implicit noisy density fluctuations that occur within the data. **CLUSTARR-ND** also produces an ordered-density plot that can be used to visually inspect the clustering structure – as is the case with **OPTICS** and **HALO-OPTICS**.

Finally in Chapter 6 I summarise these works and their findings, concluding that **CLUSTARR-ND** is ideally suited for generalised astrophysical structure finding in the contexts of both synthetic and observational data sets. I also provide my comments on future work including a series of upcoming applications of the **CLUSTARR-ND** algorithm to observational data sets and the additional machine learning techniques that can be employed to further maximise the clustering quality.

## Chapter 2

# Clustering Algorithms

Clustering is the meaningful grouping of similar objects. Methods that provide statistically motivated cluster analyses originated in the early 20<sup>th</sup> century within the field of anthropology [191, 192] and were later adapted for the classification of personality traits in psychology [193–196] in the 1930’s and 1940’s. These methods were mostly variations on what is today described as multiple group factor analysis, which in many of these cases was used to reduce the complex feature spaces involved in determining personality traits down to a small number of factors. However, Cattell [196] also laid the groundwork for a series of processes now known as single-linkage clustering. Despite this early work on cluster analysis within psychology, the field did not gain traction among the broader scientific community until the 1960’s and 70’s when computers became more widely available and the concepts broke out into various other fields including biology [197, 198], medicine [199–202], psychiatry [203–206], archaeology [207, 208], economics [209, 210], linguistics [211], legislation [212, 213], and of course astrophysics and cosmology (refer to Sec. 2.5).

Following this flourishing period for cluster analysis, research of the clustering algorithms themselves became widespread among generalist computing scientists. A wide range of more complex algorithms were developed in the following decades, however, finding structure from within a data set would still remain a difficult task to accomplish. This is due to the inherent subjectivity that surrounds the definition of such structures. Just as there are multiple contextual definitions of a cluster there are also multiple differing methods that can be used in order to find the various kinds of structure that may appear within any given data set. Hence, deciding upon which method will be most suitable in any given situation can be a challenge, requiring a thorough understanding of both the structures and the algorithms themselves to overcome [214]. Many clustering algorithms now exist within the machine learning field of data mining for the purpose of extracting statistically coherent groups from



data sets. As such, the present day field of clustering research is an expansive and prolific topic that is now summarised within the remainder of this chapter.

## 2.1 Measuring Similarity

The broadly defining notion of clustering is that the clusters are constructed such that the objects of one cluster are more similar to each other than they are to the objects of another cluster – thereby prescribing a statistical coherence to the clusters. There are many ways to assess this similarity and its exact definition must also depend on the data types that are being clustered over – sometimes requiring a custom metric. The typical definition relies on some distance metric ( $d$ ) defined over the input data ( $X$ ), the core axioms [215] of which are that for all  $x, y, z \in X$ ,

$$\begin{aligned} d(x, y) = 0 &\iff x = y, \\ d(x, y) &= d(y, x), \\ d(x, y) &\leq d(x, z) + d(y, z). \end{aligned} \tag{2.1}$$

Together these also imply that  $d(x, y) \geq 0$ . This is an intuitive way to construct the idea of similarity within a data set i.e. a small distance between two data points depicts them being more similar than if that distance had been larger. Some examples of distance metrics commonly used for clustering are as follows.

### 2.1.1 Hamming Distance

$$d(x, y) = \sum_{i=1}^n z_i, \quad z_i = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{otherwise} \end{cases} \tag{2.2}$$

The Hamming distance [216] is a simple discrete distance metric that can be used to define distances between two equal length sequences of symbols. It prescribes a distance equal to the number of positions in the sequences at which the corresponding symbols are different. The Hamming distance can be used within clustering algorithms for error detection and diagnosis as well as for genetic sequence clustering [217], among other purposes. Variants of this distance metric include the Lee distance [218] and the Levenshtein distance [219].

### 2.1.2 Minkowski Distance

$$d(x, y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}, \quad p \in \mathbb{Z}_{\geq 1} \tag{2.3}$$

The Minkowski distance is a generalisation of the Euclidean ( $p = 2$ ) and Manhattan ( $p = 1$ ) distance metrics to arbitrary integer exponents,  $p$ . Strictly speaking, this distance function is only a true metric for  $p \geq 1$  since it violates the triangle inequality for  $p < 1$  – however, a metric can still be obtained by removing the exponent of  $1/p$ . All  $p$ -Minkowski distances are translation-invariant and the Euclidean metric is also rotation-invariant making them a powerful choice when measuring similarity within data sets whose data points are defined relative to one another.

### 2.1.3 Mahalanobis Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})} \quad (2.4)$$

One extension of the Euclidean distance is that of the Mahalanobis distance [220] which is equivalent to defining Euclidean distances on a unit-scaled principal-component-analysis transformed data set. Here  $\boldsymbol{\Sigma}$  is the covariance matrix of the data set and constructing the distance metric in this way removes any global correlations and scaling dependencies in the data. As such, this distance metric is often used when to assess similarity within data sets that combine feature with different units of measurement.

### 2.1.4 Wasserstein Distance

$$W_p(\mu, \nu) = (\inf \mathbf{E} [d(X, Y)^p])^{1/p}, \quad p \in \mathbb{Z}_{\geq 1} \quad (2.5)$$

The Wasserstein distance [221] is a measure of the minimum cost required to transport the probability mass of one distribution to another – e.g. between  $\mu(x)$  and  $\nu(y)$ . Here, the inner distance function ( $d$ ) defines the distance (the *cost* of moving) between any two points ( $x \in X$  and  $y \in Y$ ) in the domain of the probability distributions. Hence given the optimal transportation plan, the  $p$ -Wasserstein distance generalises the  $p$ -Minkowski distance to measure the distance between probability distributions. As such, the  $p$ -Wasserstein distance can be used to assess the similarity in data sets consisting of fuzzy objects. Other metrics that provide similarity measures of between probability distributions include the Kullback-Leibler divergence [222, 223] and the Hellinger distance [224].

## 2.2 Clustering Models

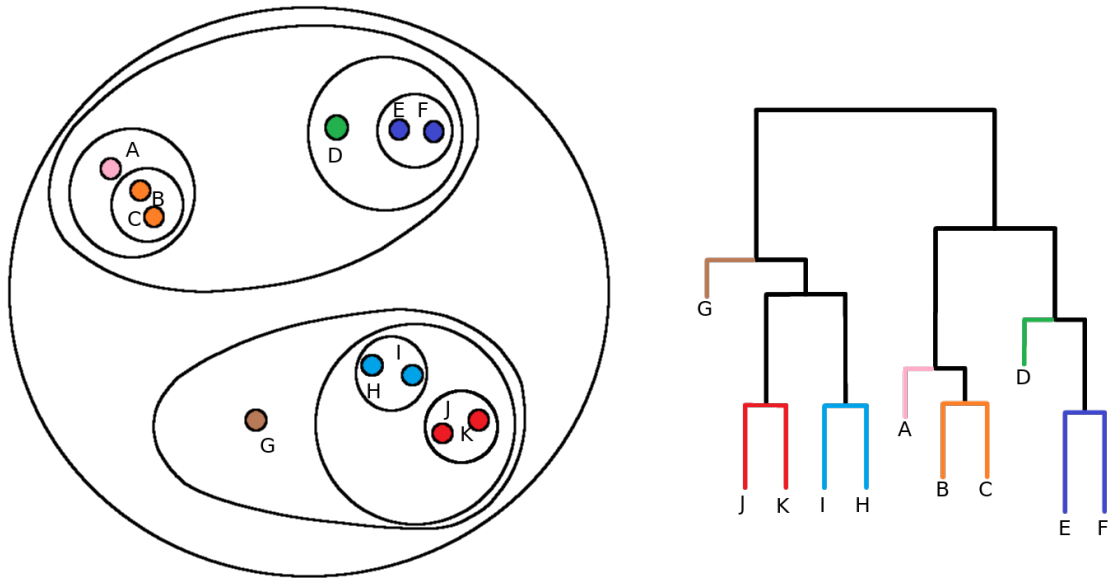
In general, the clusters that clustering algorithms produce will fall into one or more categories of various statistical models. The clustering model of a clustering

algorithm is the manner in which that algorithm constructs the resultant clusters. The clustering model of the algorithm is therefore a major contributor in how the algorithm prescribes the shape, size, and definition of the clusters it finds. There are various types of clustering models, some common and basic models being those that are; connectivity-based (often referred to as hierarchical); centroid-based; distribution-based; density-based; and graph-based. A brief overview is given for each of these below.

These models also couple with the cluster membership models, which can be broadly categorised hard or soft (also referred to as fuzzy) clustering. A hard clustering is produced if each data point either belongs to a cluster or not. A soft clustering is produced if each data point is associated to a cluster with some membership probability specific to that cluster. In this sense, a hard clustering can be thought of as a soft clustering where the only allowed probabilities are either 0 or 1. Generally speaking however, clustering algorithms can be further classified by the data-partitioning scheme that they use. Among these schemes, data points can belong to exactly one cluster such that clusters are mutually exclusive (strict partitioning) or data points can belong to multiple overlapping clusters such that clusters may be mutually inclusive (relaxed partitioning). In addition to these distinctions, clusters can also form a hierarchy such that each cluster can have a parent and/or child cluster(s). Furthermore, clusters may not completely partition the data and may leave some data points unclassified as outliers from all clusters or, in the case of hierarchical clusters, from each cluster within some level of the hierarchy.

### 2.2.1 Connectivity-based Clustering

Connectivity-based clustering algorithms create a data point hierarchy such that data points with a high similarity will be inter-connected through fewer connections than those with a lower similarity. The dendrogram – a trademark of connectivity-based clustering – can be found either; by merging groups of data points together in order of decreasing similarity (referred to as agglomerative clustering), or by splitting groups of data points in order of increasing similarity (referred to as divisive clustering). By keeping track of these similarities and the order in which these groups are merged or split, the resultant dendrogram (an example of which is shown in Fig. 2.1) represents the clustering structure within the data set which can be used to extract classifications of the data points. The standard variants of connectivity-based clustering algorithms are the single-linkage [225–227], complete-linkage [228], and weighted/unweighted average-linkage [229] which each boast different measures of



**Figure 2.1:** A toy data set and the dendrogram produced from using a connectivity-based clustering model with Euclidean distance as its distance metric. This figure has been reproduced from [230].

similarity between groups of data points.

Fig. 2.2 depicts the predicted clusters that various algorithms produce when applied to different toy data sets. The connectivity-based algorithms will have mixed as their results depend heavily upon the connectivity rules underlying their processes, however it tends to be true that any one algorithm will not be capable of finding clusters of various shapes, sizes, and densities.

## 2.2.2 Centroid-based Clustering

Centroid-based clustering algorithms create clusters on the basis that each object in a cluster is more similar to that cluster’s centroid than to the centroid of any other cluster. Typically this means that the number of clusters needs to be specified prior to clustering over the data set. The most well-known and, perhaps, simplest centroid-based clustering algorithm, **K-MEANS** [231, 232], finds  $k$  cluster centres and assigns data points to them by minimising the sum of squared distances between each of these data points and their closest centres. In **K-MEANS** the cluster centroids are chosen at random which only guarantees that the clustering solution is a locally optimal solution, however in **K-MEANS++** [233] the centroids are chosen less-randomly to ensure a more uniformly distributed initial set of  $k$ -centroids – which increases the probability of finding the globally optimal solution. Other variants of **K-MEANS** define their centroids and similarity measures differently e.g. **K-MEDIOIDS** [234] and

**K-MEDIANS** [235, 236]. Another adaption also defines fuzzy clusters, e.g. **FUZZY C-MEANS** [237].

By referring again to Fig. 2.2 it can be seen that centroid-based algorithms do not perform well when faced with clusters of arbitrary shape and size. While the definition of a centroid and its relation to the subsequently classified cluster members may change between each algorithm, their generally poor performance in these scenarios is largely due to them classifying clusters as the sets of point that are most similar to the centroids. As such, they will tend to lose their effectiveness when the clusters of the data set overlap, oddly shaped, or are difficult to separate.

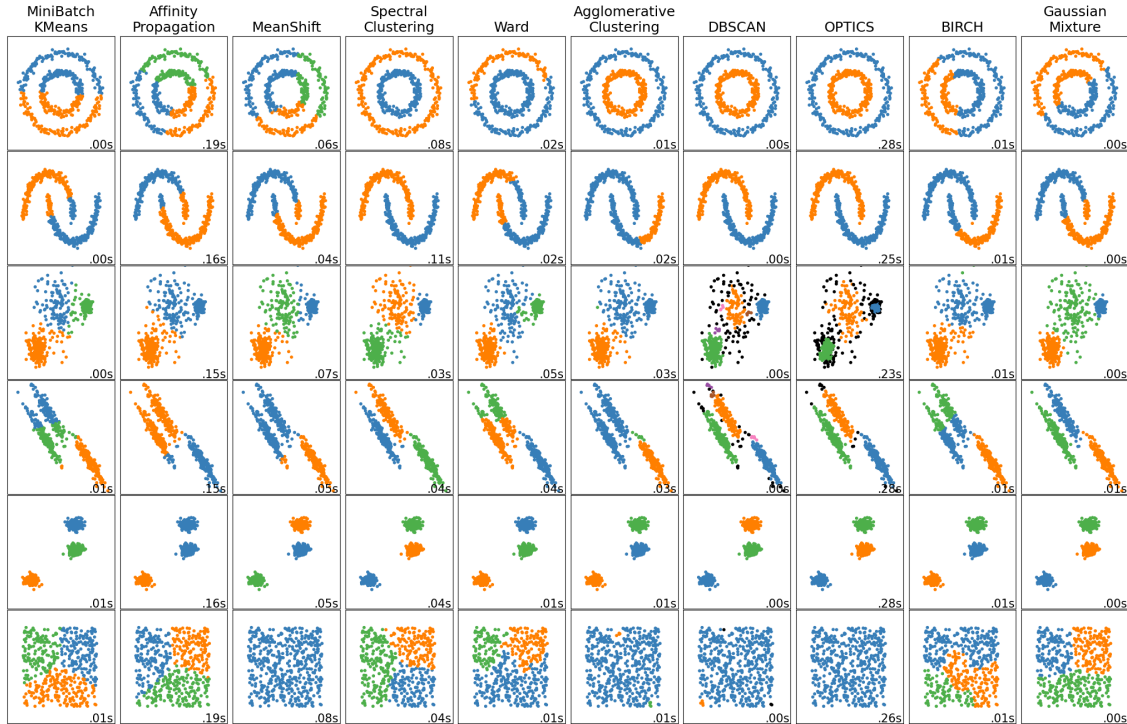
### 2.2.3 Distribution-based Clustering

Distribution-based clustering algorithms extract clusters from the data set such that the data points within each cluster belong to the same distribution model as each other. These algorithms can be subject to over-fitting, either because the distribution model of the clusters is overly complex for the data or because the cluster finding process prefers too many clusters. As such, these algorithms are usually restricted to find a fixed number of clusters using a simple distribution model. A common algorithm is the expectation-maximisation algorithm (**EM**) [238] which can fit a fixed number of distribution models from the exponential family of distributions to the data – although this is most commonly used to fit Gaussian distributions – also referred to as Gaussian mixture models. As with **K-MEANS** and the other common centroid-based clustering algorithms, the initial states of the Gaussian distributions are chosen at random, hence only guaranteeing a locally optimal solution and so is often computed multiple times in hopes of obtaining the globally optimal solution.

As shown in Fig. 2.2, Gaussian mixture models will perform well on Gaussian-like clusters even when those clusters are overlapping within the feature space. However, such a clustering model is not appropriate when the clusters deviate from this distribution. As seen with the concentric circles and the half moons, modelling the clusters as Gaussian distributions will not yield good quality results.

### 2.2.4 Density-based Clustering

Density-based clustering algorithms find clusters such that the data points within them are bounded by some contour surface of equal density. Unlike centroid- and distribution-based methods, the number of clusters does not need to be chosen prior to clustering and the clusters can be of arbitrary size and shape. A popular density-based clustering algorithm is **DBSCAN** [241] which extracts clusters defined by the



**Figure 2.2:** A visual comparison of the clustering output and run-times of various clustering algorithms available through the `SCIKIT-LEARN` software package [239]. Among the algorithms compared here there are connectivity-based (`WARD` and `AGGLOMERATIVE CLUSTERING`), centroid-based (`MINIBATCH K-MEANS`, `AFFINITY PROPAGATION`, `MEANSHIFT`, `SPECTRAL CLUSTERING`, `BIRCH`), distribution-based (`GAUSSIAN MIXTURE`), and density-based (`DBSCAN` and `OPTICS`) algorithms. The differences between these clustering models can be broadly understood by noticing the algorithm’s ability to match clusters of different sizes, shapes, and densities. This figure has been reproduced from [240].

same bounding density. Generalisations of `DBSCAN` such as `OPTICS` [242], `DELI-CLU` [243], and `HDBSCAN` [244, 245] extend this cluster definition to be hierarchical by using concepts from connectivity-based clustering methods. Other algorithms that do not rely on the connectivity-based methods, such as `MEAN-SHIFT` [246], also extract arbitrarily shaped density-based clusters.

Fig. 2.2 illustrates that density-based clustering algorithms perform well on a variety of cluster types. While `OPTICS` is more capable of appropriately dealing with clusters of varied densities, both codes can be seen outperforming the others in nearly all toy data set examples.

## 2.3 Computational Techniques

As the sizes and complexities of data sets have increased so to have the methods used for clustering over those data sets. These methods are an active topic of research in the

data mining field of machine learning are commonly used for; reducing computation times; appropriately reducing the amount of information available for clustering over; or highlighting the relevant information within a data set. Roughly speaking, these techniques can be thought of as pre-process, mid-process, and post-process methods.

### 2.3.1 Pre-Process Methods

To extract a *relevant* clustering of the input data it is essential to choose the appropriate similarity measure for the objects in that data as well as the appropriate clustering model. However, it is often also necessary to refine the data itself so that it is representative of the clustering structure that the user wishes to find. If, for example, a large multiple-choice-based questionnaire is conducted and the task at hand is to use the results of this to identify the set of most-common personality traits among people – then it is necessary to ask if the survey results are representative of the true distribution of the general populous. This is often a difficult query to confirm in the social sciences, however in the natural sciences this is usually motivated by enacting some ground truth concept about the data.

Most real-world data sets will contain some level of noise or contamination within them. Thus, before any clustering application it is worth attempting to *clean* the data set. In this case of the example above, this might be the removing of survey results for everyone who answered with all A's on their questionnaire or it may be the removing of foreground and background objects in the case of observational data of a galaxy. Simply put, there is no one-size-fits-all method for cleaning the data but there are some generalist techniques that can help. For instance, there are various local-outlier-detection methods, such as [247] and the many others that appear in the extensive review by Alghushairy et al. [248], that compute a local-outlier-factor (LOF) which can then be used to identify local-outliers within the data. In the case of [247], the LOF of a point,  $p$ , is computed by first finding the local-reachability-density (LRD) of each of  $p$ 's  $k$  nearest neighbours. The LRD of a point is the average reachability-distance (used in OPTICS [242]) from each of its own  $k$  nearest neighbours. The LOF of  $p$  is then the average of its neighbours LRDs divided by its own LRD. This gives a density-based LOF whose value is strictly positive and can be understood as indicating *how much less dense the neighbourhood of  $p$  is than the neighbourhoods of its neighbours are*. Typically, a local-outlier is considered to be any point with an  $\text{LOF} > 2$  – although this can and should be adjusted on a case-by-case basis.

Another common pre-process technique is the use of data transformation. There are many algorithms designed for this that can be used to remove unwanted correlations

in the data or remove over-dependencies on a particular feature of the data. The most common (and perhaps the most simple) of these is the principle-component-analysis (PCA) [249, 250] which can be used to reduce global correlations and over-dependencies on particular features. The PCA transformation successively finds orthogonal components so that the variances in these directions are maximised – which effectively imposes a Mahalanobis distance metric if a Euclidean distance metric is then used on the output. However, the PCA transformation can also then be used for dimensionality reduction, by choosing some number of dimensions (less than in the input data) in order of decreasing variance. Similarly to the PCA, the independent-component-analysis (ICA) [251–255] can be used to reduce additional non-Gaussian co-dependencies between the features of a data set – as well as reduce the data dimensionality in the same way as with PCA. While these techniques reduce the global correlations and feature over-dependencies, they can not do this locally within the data unless used iteratively (as is done within **ENBID** [256]). More complicated techniques, such as **UMAP** [257] and **t-SNE** [258], can achieve this while also performing a dimensionality reduction via a Riemannian manifold embedding.

### 2.3.2 Mid-Process Methods

Regardless of whether the choice similarity measure and clustering model are appropriate, and regardless of whether the data has been cleaned and transformed to highlight the relevant structure, the task of clustering can (and often does) still pose significant computational difficulties. This is particularly relevance if the clustering algorithm needs to perform a series of comparison between in order to operate. Of course, well-built clustering algorithms will make use of vectorised <sup>1</sup> and parallelised <sup>2</sup> operations wherever possible – but these techniques alone are often not enough to overcome the ordinarily extreme run-times of most complex algorithms.

If it is the mid-process task of a clustering algorithm to compare each object in the input data to other objects in the input data, then redefining the data into a tree structure can often reduce computational times significantly. For example, the  $k$  nearest neighbours or neighbours within some distance can be efficiently queried using the **KD-TREE** algorithm [259]. The **KD-TREE** algorithm iteratively partitions the data in two along cycling dimensions so as to create this tree structure. When querying the tree, entire regions of the data can be ignored by checking whether any objects in a branch will satisfy the query. This effectively reduces the computational time

---

<sup>1</sup>Vectorised operations are applied simultaneously to contiguous blocks of memory without the processor needing to retrieve any new instructions.

<sup>2</sup>Parallelised operations are a set of of instructions that are given to multiple processors to execute simultaneously.



complexity of querying the nearest neighbour of any object from  $O(n)$  to  $O(\log n)$ <sup>3</sup> (since there are now only  $\log n$  branches to check instead of naively checking every object in the input data,  $n$ ). Other tree-partitioning algorithms exist for this same purpose, such as **BALL-TREE** [260] and **R-TREE** [261]. Notably, approximate nearest neighbour search methods are also very powerful for reducing computational times. Implementations such as **FAISS** [262] can conduct exact and approximate nearest neighbour searches that are orders of magnitude faster than standard approaches – particularly for high dimensional data – and is capable of GPU acceleration.

If a clustering algorithm is required to minimisation of some cost/loss function, then it will likely be time-effective to implement an appropriate optimisation technique. The most appropriate technique will depend upon the complexity of the cost function landscape as well as the constraints and bounds of its variables. The restrictions on the cost functions themselves are minimal – i.e. any map between  $\mathbb{R}^n$  and  $\mathbb{R}$  – although it does need to be convex within the variable domain and appropriately represent the model that the minimisation provides a solution for [263]. Typical cost functions include the; sum of absolute errors between the model and the data ( $L_1$  norm); sum of squared errors between the model and data ( $L_2$  norm); negative log-likelihood of the model given the data; the Kullback-Leibler divergence [222, 223] between the desired function output and the model output given the data; the mutual information [264] between the desired function output and the model output given the data; among others. Simple minimisation techniques include the; downhill simplex method (zeroth order) [265]; steepest gradient descent (first order) [266, 267]; and the BFGS algorithm (second order) [268–271]. The order of these methods refers to the order of the function derivative that is used to approach the solution – which also gives an indication of the computational complexity of the algorithm and of the accuracy of the solution it provides. These methods however, do not guarantee a global minimum – only a local one. Guaranteeing the global solution is not always possible, but methods such as basin-hopping [272], simulated annealing [273–276], as well as genetic algorithms can do remarkably well at finding the global solution.

### 2.3.3 Post-Process Methods

In some cases, it may be necessary to apply further techniques to the already clustered data in order to extract additional information. It may also be the case that the use of a post-process method effectively achieves the same clustering results with the added benefit of reduced run times and/or computational resources. For example,

---

<sup>3</sup>Big-O notation denotes the asymptotic dependency of the computation time on any independent variable within the data or algorithm.

if a clustering is taken of a snapshot of a data stream, then it is likely useful to know how the predicted clusters evolve over time and which clusters new data points belong to without having to perform the entire clustering procedure over again.

A simple way that this can be achieved is through the use of a nearest neighbour classification algorithm [277, 278]. Nearest neighbour classification algorithms construct a tree structure that contains pre-labelled data (in this case the cluster labels of a previous clustering). The tree can then be queried with new unlabelled data to find which class (or set of classes) the new data point is most likely to belong to. The labels of the existing tree structure can then be updated and/or the new data can be added to the tree – effectively evolving the clusters in a modest  $O(\log n)$  time complexity. This technique is only reliable for so long as at some point the original clustering will not contain the relevant information for providing new and accurate classifications of novel data [279]. The nearest neighbours classification technique can be used to update the clustering results of any algorithm, however, there are stand-alone data stream clustering algorithms that do a better job at maintaining robust clustering results. The **STREAM** algorithm [280] works by creating a **K-MEDIANS**-like clustering of a flowing data stream. Similarly, the **BIRCH** algorithm [281] is a hierarchical clustering algorithm that can predict clusters from a single passing of a data stream.

## 2.4 Statistical Evaluation

Even with an appropriate clustering algorithm and a data set to apply it to, it is still pertinent to question the quality of the output in relation to the desired classification result. By assessing the quality of the clustering that is produced from a clustering algorithm and a data set, it is possible to; begin a process of hyperparameter optimisation; alter pre-, mid-, and post-process methods; or determine whether the procedure is even appropriate. In order to assess the quality of a clustering, an objective function that represents that quality needs to be constructed, the specifics of which will be entirely dependent on the clustering model and the desired classification. Broadly speaking, there are two modes of evaluation – internal and external.

### 2.4.1 Internal Evaluation

An internal evaluation measure assesses the clustering quality by considering only the data and resultant classifications. Typically, internal evaluation measures provide a balanced indication of the level to which the intra-cluster similarity and the inter-cluster dissimilarity are both maximised. Care must be taken when using any internal

evaluation as they are more suitable for quality assessment on some clustering models than others.

For example, using an internal evaluation measure to optimise the hyperparameters of a density-based clustering model with noise (such as **DBSCAN**) may incentivise the algorithm to return zero clusters. This is because intra- and inter-cluster similarities are simultaneously increased and decreased respectively when the density of clusters is increased. Without some care, the use of an unsuitable internal evaluation measure will motivate continued increasing of the threshold density of labelled clusters until the true clusters are no longer *dense enough* to be labelled as such. As a contrasting example, optimising the hyperparameters of a centroid-based clustering model without noise (such as **K-MEANS**) using an internal evaluation measure will appropriately incentivise self-similar and distinct clusters. Under the constraint that all points must be attributed to self-similar clusters there exists a configuration of  $k$  clusters that maximises both cluster self-similarity and distinctness [214].

The highest quality configuration will depend heavily upon the measure and so an understanding of the various measures is needed before any attempt at refining the process can be made. Some common measures include the Davies-Bouldin index [282], the Dunn index [283], and the Silhouette coefficient [284]. A comparison of these and more is performed by Hassani and Seidl [285] in the context of data stream clustering.

## 2.4.2 External Evaluation

An external evaluation measure assesses the quality of a clustering by considering how well the resultant classification system matches some predefined notion or label set that acts as a benchmark for the ideal clustering of the input data. As such, external evaluation measures quantify the similarity between the predicted clustering and the *ground truth* clustering benchmark. This means that the ground truth clustering must be heavily scrutinised and verified as such or else any optimisation that uses an external evaluation measure will inevitably produce sub-optimal results.

A few commonly used measures include recovery (often called recall), purity (often called precision) [286], and the Jaccard index [287]. These measures are simple to interpret and, given a true cluster and a predicted cluster, are equal to the size of the intersection between the true and predicted cluster as a ratio to the size of the; true cluster; predicted cluster; and union of the true and predicted clusters respectively. As such they provide a powerful indication of the match between two clusters. However, gaining an understanding of the match between a set of true clusters and a set of predicted clusters defined over the same data set can be difficult

as these measures do not take into account true negative predictions and are not straight forward to use when trying to optimise hyperparameters controlling the overlap of predicted clusters with multiple true clusters or vice versa. Other external evaluation measures, particularly those that are information-based (e.g. adjusted mutual information [288] and variation of information [289]), can demonstrate the quality of fit between two entire clusterings and typically do not suffer from the above downfalls.

## 2.5 The State of Astrophysical Structure Finding

Finding and classifying astrophysical structure is an important part of understanding the Universe as is detailed Sec. 1. The techniques used to do this do not always involve the use of a clustering algorithm as often the classification can be made via expert inspection. The planets of our solar system, stars, and the MW have been observable since antiquity. Galaxies and galaxy clusters [178] have been detected via inspection of photographic plates. Even more recently there are still studies that essentially make their discoveries via the inspection of data. For example, many of the MW's stellar streams have been detected simply by noticing overdensities in projections that surmount to comoving groups of stars [290–293]. While these structures have been identified robustly, inspection methods can not be used to search data sets exhaustively, and only with data mining algorithms can we hope to do so. Astro- and cosmo-related structure finding clustering algorithms have seen continued attention and growth over the last few decades. Functionally, these algorithms are often similar and will typically fall into one of a few common algorithm types.

### 2.5.1 Simulation Specific Finders

A common way to study the effect of cosmology and the structure that emerges due to this, is by conducting cosmological simulations that model the Universe (or a hypothetical one) from the Big Bang to the present time. Constrained by a cosmological model, the initial conditions of the Universe are created within the simulation. The spatial and kinematic information of the dark and baryonic particles are then evolved with time using equations of motion that arise from FLRW gravity. With ongoing structure formation taking place at every epoch of the simulation, it is necessary to be able to systematically classify this structure in order to understand the effect of the model.

**Table 2.1:** A chronological history of simulation-specific (sub)galactic structure finders. Among the algorithm’s year of development, the algorithm’s name, and the base code of the algorithm (a blank space indicating a first-of-its-kind algorithm) is the feature space (some combination of spatial, kinematical, metallicities and colour-magnitude diagram information, a blank space indicating that the algorithm can take any combination of this information) and a few indicators of the algorithm’s ability to find certain structure types. Namely, a tick or cross is given the Gal., Sub. and Tid. columns specifying whether the algorithm is capable of finding galaxies/haloes, subhaloes, and tidal debris respectively.

Year	Algorithm	Base	Space	Gal.	Sub.	Tid.
1974	SO [294]		x	✓	✗	✗
1985	FOF [295]		x	✓	✗	✗
1991	DENMAX [296, 297]	SO	x <sup>4</sup>	✓	✗	✗
1995	ADAPTIVE FOF [298]	FOF	x	✓	✗	✗
1996	ISO DEN [299]	SO <sup>5</sup>	x	✓	✗	✗
1997	BDM [300]	DENMAX <sup>5</sup>	x <sup>4</sup>	✓	✓	✗
1998	HOP [301]	DENMAX	x	✓	✓	✗
1999	HFOF [302]	FOF	x	✓	✓	✗
2001	SKID [303, 304]	DENMAX	x <sup>4</sup>	✓	✓	✗
2001	ENHANCED BDM [305]	BDM	x <sup>4</sup>	✓	✓	✗
2001	SUBFIND [306]	FOF	x <sup>4</sup>	✓	✓	✗
2004	MHF [307]	SO	x <sup>4</sup>	✓	✓	✗
2004	ADAPTAHOP [308]	HOP	x	✓	✓	✗
2004	DENMAX <sup>2</sup> [309]	DENMAX	x <sup>4</sup>	✓	✓	✗
2004	SURV [310, 311]	SO	{x, t} <sup>4</sup>	✓ <sup>6</sup>	✓	✗
2005	IMPROVED DENMAX [312]	DENMAX	x <sup>4</sup>	✓	✓	✗
2005	VoBoZ [313]	DENMAX	x <sup>4</sup>	✓	✓	✗
2006	PSB [314]	FOF	x <sup>4</sup>	✓	✓	✗
2006	6DFOF [315]	FOF	{x, v}	✓	✓	✗
2007	SUBHALO FINDER [316]	DENMAX <sup>5</sup>	x <sup>4</sup>	✓	✓	✗
2007	NTROPY-FOF [317, 318]	FOF	x	✓	✗	✗
2009	HSF [319]	SUBFIND	{x, v} <sup>4</sup>	✓ <sup>6</sup>	✓	✗
2009	LANL [320]	SO <sup>5</sup>	x	✓	✗	✗
2009	AHF [321]	MHF	x <sup>4</sup>	✓	✓	✗
2010	PHOP [322]	HOP	{x, v}	✓	✗	✗
2010	ASO HF [323]	SO	x <sup>4</sup>	✓	✓	✗

<sup>4</sup>An unbinding procedure is performed with both positions and velocities.

<sup>5</sup>This code also uses the FOF algorithm to first find field haloes.

<sup>6</sup>Can also be used to find large-scale-structure.

Year	Algorithm	Base	Space	Gal.	Sub.	Tid.
2010	P <sub>SO</sub> [324]	SO	$\mathbf{x}$	✓	✗	✗
2010	P <sub>FOF</sub> [325, 326]	FOF	$\mathbf{x}$	✓	✗	✗
2010	ORIGAMI [327]		$\mathbf{x}^4$	✓ <sup>6</sup>	✗	✗
2010	MENDIETA [328]	FOF	$\mathbf{x}^4$	✓	✓	✗
2010	ENHANCED SURV [329]	SURV	$\{\mathbf{x}, \mathbf{t}\}^4$	✓ <sup>6</sup>	✓	✗
2011	STF [330]	FOF	$\{\mathbf{x}, \mathbf{v}\}^4$	✓	✓	✓
2012	ROCKSTAR [331]	FOF	$\{\mathbf{x}, \mathbf{v}\}^4$	✓	✓	✓
2012	HBT [332]	FOF	$\{\mathbf{x}, \mathbf{t}\}^4$	✓	✓	✗
2013	S-TRACKER <sup>7</sup>	HBT	$\{\mathbf{x}, \mathbf{t}\}$	✓	✓	✓
2013	GRASSHOPPER <sup>7</sup>	HOP + SKID	$\mathbf{x}^4$	✓	✓	✗
2013	JUMP-D <sup>7</sup>	SO	$\mathbf{x}$	✓	✓	✗
2018	HBT+ [337]	HBT	$\{\mathbf{x}, \mathbf{t}\}^4$	✓	✓	✗
2019	VELOCIRAPTOR [338]	STF	$\{\mathbf{x}, \mathbf{v}\}^4$	✓	✓	✓
2020	HIKER [339]	MEAN-SHIFT	$\mathbf{x}^4$	✓	✓	✗
2021	FOF-HALO-FINDER [340]	FOF	$\mathbf{x}$	✓	✗	✗
2021	COMPASO [341]	SO <sup>5</sup>	$\mathbf{x}^4$	✓	✓	✗

In order to find galaxies, haloes, and subhaloes, clustering algorithms must identify some isolated region of space that has some pre-specified spatial overdensity and only contains self-bound particles. The earliest methods developed to achieve this are the Spherical Overdensity algorithm (SO; [294]) and the Friends-Of-Friends algorithm (FOF; [295]). The SO method finds density peaks within the data and then expands spherical surfaces out from each of these until the density within these regions (defined by the top-hat kernel) reaches the specified overdensity. The SO algorithm therefore uses a Euclidean distance-based similarity measure to define density and then employs a distribution-based clustering model to collect the particles that the structures are composed of. Contrarily, the FOF algorithm collects all particles that can be chained together by distances less than the FOF linking length. This linking length is typically chosen as  $0.2l_{\text{mean}}$  where  $l_{\text{mean}}$  is the mean particle separation within the simulation box. As such, the FOF algorithm also uses the Euclidean distance as its similarity measure except it uses a connectivity-based clustering model to assemble the structures.

Whether the clustering models of these algorithms perfectly suited the original intended definition of these structures or not is now barely relevant as most clustering algorithms designed for application to simulation data are built off of either the SO

<sup>7</sup>A paper for this algorithm was never published although it was used within one or more comparison papers, e.g. [333–336].

or the FOF algorithms – as depicted in Tab. 2.1. In this sense, these algorithms now effectively prescribe the definition of such structures. A series of comparison papers found that most modern galaxy/(sub)halo finders strongly agree on the; dark matter haloes [333]; galaxies [334]; subhaloes [342, 343]; tidal debris [335]; merger trees [344, 345]; and major mergers [346] that they find.

Among modern galaxy/(sub)halo finders there are three types; configuration space finders; phase space finders; and tracking finders. Configuration space finders (e.g. SUBFIND [306], AHF [321], and COMPASO [341]) use the 3D spatial positions of particles to find physical overdensities – which are then often reduced to bound haloes using kinematic information as well. Phase space finders (e.g. HSF [319], ROCKSTAR [331], and VELOCIRAPTOR [338]) use both the 3D spatial and 3D kinematical attributes of each particle to find structures. Tracking finders (e.g. SURV [310, 311], S-TRACKER<sup>7</sup>, and HBT+ [337]) use either configuration and/or phase space to find structures and keep track of them over time within a simulation.

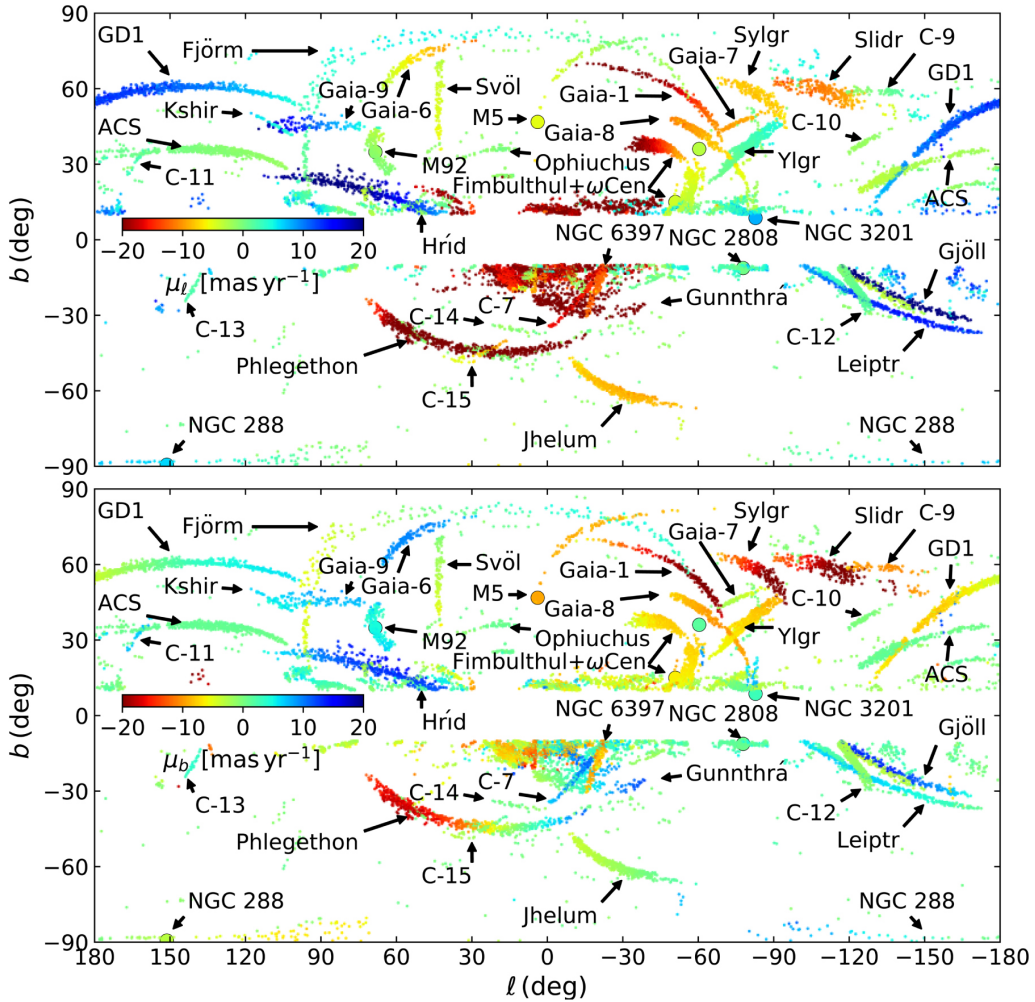
**Table 2.2:** A chronological history of observation-specific (sub)galactic structure finders. Among the algorithm’s year of development, the algorithm’s name, and the base code of the algorithm (a blank space indicating a first-of-its-kind algorithm) is the feature space (some combination of spatial, kinematical, metallicities and colour-magnitude diagram information, a blank space indicating that the algorithm can take any combination of this information) and a few indicators of the algorithm’s ability to find certain structure types. Namely, a tick or cross is given the Gal., Sub. and Tid. columns specifying whether the algorithm is capable of finding galaxies/haloes, subhaloes, and tidal debris respectively.

Year	Algorithm	Base	Space	Gal.	Sub.	Tid.
1996	GC3 [129]		$\mathbf{x}$	$\times$	$\times$	$\checkmark$
2002	MF [347]	[348]	$\{\mathbf{x}, \text{CMD}\}$	$\times$	$\times$	$\checkmark^8$
2011	MGC3 [351]	GC3	$\{\mathbf{x}, \mathbf{v}\}$	$\times$	$\times$	$\checkmark$
2011	IMPROVED MF [352]	MF	$\{\mathbf{x}, \text{CMD}\}$	$\times$	$\times$	$\checkmark^8$
2018	xGC3 [353]	MGC3	$\{\mathbf{x}, \mathbf{v}\}$	$\times$	$\times$	$\checkmark$
2018	STREAMFINDER [354]		$\{\mathbf{x}, \mathbf{v}\}$	$\times$	$\times$	$\checkmark$
2018	STARGO [355]	SOM [356]	$\{\mathbf{x}, \mathbf{v}\}$	$\times$	$\checkmark$	$\checkmark$
2022	HSS [357]		$\{\alpha, \delta\}$	$\times$	$\times$	$\checkmark$
2022	VIA MACHINAE [358]	HSS	$\{\mathbf{x}, \mathbf{v}\}$	$\times$	$\times$	$\checkmark$
2022	IOM [359]	SLINK [227]	$\{\mathbf{E}, \mathbf{J}\}$	$\times$	$\checkmark$	$\checkmark$

<sup>8</sup>This implementation is only used for stream finding, although matched filter techniques [348] have been used widely from the discovery of galaxy clusters [349] to exoplanets [350].

### 2.5.2 Observation Specific Finders

In the pursuit of structure discovery, many studies have conducted clustering analyses to find structure from large scale observational survey data sets such as Gaia [360], SDSS [361], PAndAS [362], and GALAH [363]. While there are many existing methods of observational structure finding that rely somewhat on an algorithmic-like process, there are far fewer named and stand-alone methods for this purpose than there are for simulation-based structure finding. Specifically, it is popular for studies to use ready-made clustering algorithms such as K-MEANS (e.g. [364, 365]), DBSCAN (e.g. [366, 367]), or HDBSCAN (e.g. [368, 369]) when searching observational data sets for structure. These are blind applications and without an understanding of the functionality of these algorithms it is not obvious whether these structures are statistically robust in an astrophysical setting.



**Figure 2.3:** The positions on-the-sky of a series of tidal streams recovered by Ibata et al. [139] using the `STREAMFINDER` algorithm with the Gaia EDR3 catalogue as its input. The stream members are coloured by their proper motions in the  $\mu_l$  (top) and  $\mu_b$  (bottom) directions.



The algorithms that have been built to find structure from observational data will typically use some physical model to motivate their findings. Many of these algorithms are shown in Tab. 2.2 and most commonly, they rely on some potential model of the host galaxy. Famously, the **STREAMFINDER** algorithm [354], which has discovered many new stellar streams within the Milky Way [139, 370–374], must use a potential model of the Milky Way and in recent applications has needed a set of stellar isochrone models as well. While this method is powerful and has been able to retrieve many streams from the Milky Way halo (as shown in Fig. 2.3), it struggles to find structures that are not well described by these models, i.e. those structures whose self-gravity is non-negligible and whose stars belong to multiple isochrones – hence why it only finds tidal streams.

For those observation specific substructure finders (e.g. **STARGO** [355], **HSS** [357], and **VIA MACHINAE** [358]) that do not require a model, certain projections of the data are often created (e.g. the Hough transform [375]) that restricts the information content of the data set. Again, this can work well for specific clustering situations but is not generally applicable to finding all relevant substructure.

### 2.5.3 Generalised Structure Finders

Contrary to simulation and observation specific structure finders, there exists another category of astrophysical clustering algorithm that align more closely with the generalised data mining methods used within observational studies such as **HDBSCAN**. However, instead of being applied blindly these methods are still purposefully created for astrophysical clustering. Unburdened by flat clustering methods (e.g. **FOF**) or galactic potential models, the clustering algorithms shown in Tab. 2.3 are able to retrieve all types of relevant structure and are usually applicable to any data set. While the **ENLINK** algorithm uses an entropy-based locally adaptive metric to improve the clustering power of what is otherwise essentially the **SUBFIND** algorithm without the unbinding procedure, and **HOT** is an adaptive extension of the **FOF** formalism using a minimum-spanning-tree, the **OPTICS** based codes are able to return a representation of the entire clustering structure. It is this adaptive measure of structure that lies at the centre of the scope of this thesis.

**Table 2.3:** A chronological history of generalised structure finders. Among the algorithm’s year of development, the algorithm’s name, and the base code of the algorithm is the feature space (some combination of spatial, kinematical, metallicities and colour-magnitude diagram information, a blank space indicating that the algorithm can take any combination of this information) and a few indicators of the algorithm’s ability to find certain structure types. Namely, a tick or cross is given the Gal., Sub. and Tid. columns specifying whether the algorithm is capable of finding galaxies/haloes, subhaloes, and tidal debris respectively.

Year	Algorithm	Base	Space	Gal.	Sub.	Tid.
2009	ENLINK [376]	SUBFIND		✓ <sup>9</sup>	✓	✓
2010	HOT <sup>10</sup>	MST [377]		✓ <sup>9</sup>	✓	✓
2017	FOPTICS [378]	OPTICS	{ <b>x</b> , <b>v</b> }	✓ <sup>9</sup>	✓	✓
2020	HALO-OPTICS (1)	OPTICS	<b>x</b>	✓ <sup>9</sup>	✓	✓
2022	CLUSTAR-ND (2)	HALO-OPTICS		✓ <sup>9</sup>	✓	✓
	CLUSTARR-ND (3)	CLUSTAR-ND		✓ <sup>9</sup>	✓	✓

<sup>9</sup>Can also be used to find large-scale-structure.

<sup>10</sup>A paper for this algorithm was never published although it was used within one or more comparison papers, e.g. [333–336].

# Chapter 3

## A Novel Approach to Astrophysical Clustering

As is outlined in Chapter 2, most astrophysical clustering algorithms are either designed for application to simulated data or observational data – and rarely both. Of the simulation specific finders, almost all are based off of either the **SO** [294] and/or the **FOF** [295] algorithms which alone can only provide flat clusterings of data unless applying they are applied iteratively or the method appeals to the physics of self-boundedness. Contrarily, most observation specific finders are either model dependent or create projections of the data that can limit the overall information content of the clustering. While these restrictions can prove powerful when targeting specific structure types, removing them can open up a clustering algorithm to be able to discover multiple cluster types simultaneously.

While only a small handful of purpose-built astrophysical clustering algorithms do not adhere to the formulae of the others, it is clear that many researchers in the field (particularly those working with observational data sets) have begun to notice the power that a generalised clustering algorithm can have in an astrophysical context. While the **K-MEANS** [231, 232], **DBSCAN** [241], and **HDBSCAN** [244, 245] algorithms have been used extensively in recent years, these applications are essentially blind – as without thorough tests the findings of these remains unclear in an astrophysical context. One such algorithm general-purpose hierarchical and density-based clustering algorithm that as seen extensive use outside of astrophysics and cosmology is **OPTICS** [242]. It has only been applied a handful times within astro- and cosmo-related fields [138, 378–385].

I create **HALO-OPTICS** from the **OPTICS** algorithm in order to overcome the restrictions of the simulation and observation specific structure finders and provide an adaptive measure of the spatial clustering structure within galactic haloes. **HALO-OPTICS**

is capable of choosing search radius parameter of OPTICS based on the more physically motivated overdensity factor,  $\Delta$ . HALO-OPTICS also comes with a cluster extraction method that derives astrophysical structure from the reachability plot that OPTICS produces. I optimise the HALO-OPTICS parameters and then compare it to the simulation specific algorithm, VELOCIRAPTOR [338]. In this comparison I find that although HALO-OPTICS only uses the 3D spatial positions of data points while VELOCIRAPTOR uses the entire phase-space (positions and velocities), HALO-OPTICS is able to provide a good match to the tidal debris that VELOCIRAPTOR finds. The following section presents the published paper in which this research is conducted and the code for the HALO-OPTICS algorithm can be found in App. B.1.

While the test cases used here illustrate the effectiveness of HALO-OPTICS when applied to simulated galactic haloes (and mock/toy data sets), it is important to recognise that at the core of the algorithm there is not any particular function that requires the input data set to be of this exact nature. Allowing for minor adjustments to the parameters of the algorithm, HALO-OPTICS can be meaningfully applied to any astrophysical data set whereby the data points represent objects in 3D space. Even though the latter does impose a limitation on the input data set’s feature space, the applicability of HALO-OPTICS to stars within galaxies, simulation particles, galaxies within simulation boxes, etc. serves as a critical step towards defining a robust generalised astrophysical structure finder.

### 3.1 Structure Finding with Halo-OPTICS

This section presents the published journal article:

1. *The Hierarchical Structure of Galactic Haloes: Classification and Characterisation with Halo-OPTICS*. **W. H. Oliver**, P. J. Elahi, G. F. Lewis, & C. Power. *MNRAS* 501, 4420, 2021. [[arXiv:2012.04823](https://arxiv.org/abs/2012.04823)].

*Author Contributions:* I developed and trained the HALO-OPTICS algorithm, produced the clustering outputs, drew comparisons between HALO-OPTICS and the state-of-the-art galaxy/(sub)halo finder VELOCIRAPTOR, and wrote the manuscript. Dr. Pascal J. Elahi assisted with the development of the algorithm training method and provided the clustering outputs from VELOCIRAPTOR – of which he is the creator. Prof. Geraint F. Lewis conceived the idea of using the OPTICS algorithm [242] for astrophysical clustering and supervised the project. Prof. Chris Power provided the data sets of the simulated MW-type galaxies. All authors reviewed and commented on the paper.

# The hierarchical structure of galactic haloes: classification and characterization with HALO-OPTICS

William H. Oliver,<sup>1</sup>★ Pascal J. Elahi<sup>1,2,3</sup>, Geraint F. Lewis<sup>1</sup> and Chris Power<sup>2,3</sup>

<sup>1</sup>*Sydney Institute for Astronomy, School of Physics A28, The University of Sydney, Sydney, NSW 2006, Australia*

<sup>2</sup>*International Centre for Radio Astronomy Research, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*

<sup>3</sup>*ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*

Accepted 2020 December 6. Received 2020 November 10; in original form 2020 August 20

## ABSTRACT

We build upon Ordering Points To Identify the Clustering Structure (OPTICS), a hierarchical clustering algorithm well known to be a robust data miner, in order to produce HALO-OPTICS, an algorithm designed for the automatic detection and extraction of all meaningful clusters between any two arbitrary sizes. We then apply HALO-OPTICS to the 3D spatial positions of halo particles within four separate synthetic Milky Way-type galaxies, classifying the stellar and dark matter structural hierarchies. Through visualization of the HALO-OPTICS output, we compare its structure identification to the state-of-the-art galaxy/(sub)halo finder VELOCIRAPTOR, finding excellent agreement even though HALO-OPTICS does not consider kinematic information in this current implementation. We conclude that HALO-OPTICS is a robust hierarchical halo finder, although its determination of lower spatial-density features such as the tails of streams could be improved with the inclusion of extra localized information such as particle kinematics and stellar metallicity into its distance metric.

**Key words:** galaxies: clusters: general – galaxies: structure – dark matter.

## 1 INTRODUCTION

A primary prediction of hierarchical galaxy formation in the cold dark matter (CDM) cosmological model is that galaxies should be surrounded by numerous low-mass satellites as a result of historical and ongoing accretion (White & Rees 1978; Kauffmann, White & Guiderdoni 1993; Ghigna et al. 1998). Simulations of galaxy formation under this regime have predicted that galaxies of a size similar to the Milky Way (MW) could harbour 300–500 satellites at least as massive as  $\sim 10^8 M_\odot$  at  $z = 0$  (Klypin et al. 1999; Moore et al. 1999; Reed et al. 2005; Springel et al. 2008; Tollerud et al. 2008; Ishiyama et al. 2013). Observations within and around the MW have identified  $\sim 60$  satellites of the same size; refer to tables A1 and A2 from Newton et al. (2018) and more recently Koposov et al. (2018), Homma et al. (2019), Mau et al. (2019), and Torrealba et al. (2019) for catalogues of these. This has become known as the missing satellite problem.

Studies have shown that by suppressing the small-scale power spectrum (Kamionkowski & Liddle 2000; Zentner & Bullock 2003), by considering warm dark matter as an alternative model (Colin, Avila-Reese & Valenzuela 2000; Bode, Ostriker & Turok 2001), or by enforcing that dark matter emerged from the late decay of a non-relativistic particle (Strigari, Kaplinghat & Bullock 2007), the inconsistency is brought to within a reasonable margin of error as a result of a reduced theoretical number of satellites (as well as the slowly increasing number of the observed). In conjunction

with these cosmological solutions, the observations of the faintest MW satellites have also suggested that the discrepancy is smaller than previously thought and that smaller dark matter haloes are far less efficient at forming stellar populations than their more massive counterparts (Bullock, Kravtsov & Weinberg 2000; Benson et al. 2002; Somerville 2002; Ricotti & Gnedin 2005; Moore et al. 2006). The implication is that the observed mass-scale of MW dwarfs may be an artefact of detection bias. By considering more complex baryonic physics such as reionization and supernovae feedback, the most recent cosmological  $N$ -body simulations suggest that the missing satellite problem may no longer be an obstacle for the CDM model (Brooks & Zolotov 2014; Sawala et al. 2015, 2016; Dutton et al. 2016; Wetzel et al. 2016; Zhu et al. 2016; Kim, Peter & Hargis 2018; Fielder et al. 2019).

Even though the tension of CDM with the number of satellites has relaxed, the search for satellites in the Local Group is ongoing. In simulations, the abundance of subhaloes has been shown to be dependent on the mass of the subhalo such that  $dn_{\text{sub}}/dM_{\text{sub}} \propto M_{\text{sub}}^{-\alpha}$  with  $\alpha \approx 1.9$  (Gao et al. 2004; Reed et al. 2005; Diemand, Kuhlen & Madau 2007; Springel et al. 2008; Angulo et al. 2009; Garrison-Kimmel et al. 2014; Xie & Gao 2015; Rodriguez-Puebla et al. 2016; Elahi et al. 2018), where  $n_{\text{sub}}$  is the number of subhaloes with mass greater than  $M_{\text{sub}}$  – the individual subhalo masses. This relation is appropriately descriptive of subhaloes whose dark matter hosts have masses within the range of  $10^{12}$ – $10^{15} M_\odot$ , although this range has a lower bound set by the each simulation’s particle mass resolution; however, it is expected to hold true for host halo masses much smaller than this. Observationally, this power law appears to hold for subhaloes with luminosity down to  $10^8 L_\odot$  – approximately the

\* E-mail: woli0618@uni.sydney.edu.au

luminosity of the brightest satellites of the MW and M31 – although it is not consistent for those satellites whose luminosity is below that limit (Tollerud, Boylan-Kolchin & Bullock 2014). This could again support the notion that currently our satellite detection capabilities fail to easily detect those ultrafaint satellites (Koposov et al. 2008; Tollerud et al. 2008; Walsh, Willman & Jerjen 2008; Bullock et al. 2010; Sesar et al. 2014) that are hosted by the aforementioned dark matter subhaloes responsible for suppressing the star formation within them (Wolf et al. 2010; Martinez et al. 2011).

Surveys have exposed a considerable amount of substructure surrounding the MW (e.g. Zhao et al. 2012; Blanton et al. 2017; Starkenburg et al. 2017; Gaia Collaboration et al. 2018) and other nearby galaxies such as M31 (McConnachie et al. 2009, 2018). We see that while this structure may not be directly associated with specific satellites of these galaxies, it is still intimately linked to them having arisen from the satellite–host tidal interactions. Characterizing the observed structure quantitatively will complement studies of satellite galaxies. As such an important component of subhalo analysis is the initial determination of such objects. There are many algorithms used to ascertain clustered data from data sets. Data miners such as K-MEANS (Lloyd 1982), MEAN-SHIFT (Fukunaga & Hostetler 1975), and DBSCAN (Ester et al. 1996) work well for certain cluster shapes, although they will not indicate the hierarchy of clusters present in a data set. Astrophysical clustering algorithms like SUBFIND (Springel et al. 2001), Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement (ROCKSTAR; Behroozi, Wechsler & Wu 2012), Amiga Halo Finder (AHF; Knollmann & Knebe 2009), VELOCIRAPTOR (Elahi et al. 2019), and others (refer to Knebe et al. 2011 for a comparison of these and 14 others) mostly rely on the Spherical Overdensity (SO) method (Press & Schechter 1974), Friends-Of-Friends (FOF; Davis et al. 1985), or some iterative combination of the two.

The SO method aims at identifying the density peaks enclosed within some dense region of  $N$  nearest neighbours. Following this, spherical surfaces expand about each peak until a specified overdensity is achieved within it whilst iteratively adapting the centre of the sphere to the new centroid of the enclosed particles. The biggest downfall of the SO method is that it fails to detect clusters below the specified overdensity threshold. Contrarily, the FOF algorithm endeavours to link together those particles that are physically close to each other and then subsequently computes the centroid of this particle composition. Neither the SO method nor the FOF algorithm is inherently hierarchical unless used iteratively on the findings of their previous applications with a larger overdensity or smaller linking length, respectively. Only then can these algorithms differentiate between clusters of two different densities within the same hierarchy. A novel and complementary algorithm to the traditional structure finders that is hierarchical in this sense is the Ordering Points To Identify the Clustering Structure (OPTICS) algorithm (Ankerst et al. 1999). OPTICS can be readily applied to observational data sets in ways that some traditional structure finders cannot since it does not require estimates of the gravitational potential.

We build upon OPTICS to develop HALO-OPTICS and apply it to the synthetic galactic haloes generated by Power & Robotham (2016) at redshift zero. We first summarize the details of this data set and outline our choice of physical quantities from these synthetic haloes in Section 2.1. We then summarize the OPTICS algorithm in detail in Section 2.2. Next, we present HALO-OPTICS; by first justifying our choice of the OPTICS hyperparameters  $\epsilon$  and  $N_{\min}$  in Section 3.1, then defining our automatic cluster extraction technique in Section 3.2. We conduct parameter optimization tests in

Section 3.3 and inform the reader about the nature of the HALO-OPTICS hierarchy in Section 3.4. We then present our findings; by visualizing the HALO-OPTICS output in Section 4.1, comparing HALO-OPTICS with VELOCIRAPTOR in Section 4.2, and by analysing the galactic hierarchy returned by HALO-OPTICS in Section 4.3. Later in Section 5, we discuss these results and the implications of providing HALO-OPTICS with extra localized information. Finally, we make our conclusions and express our intent for future works in Section 6.

## 2 BACKGROUND

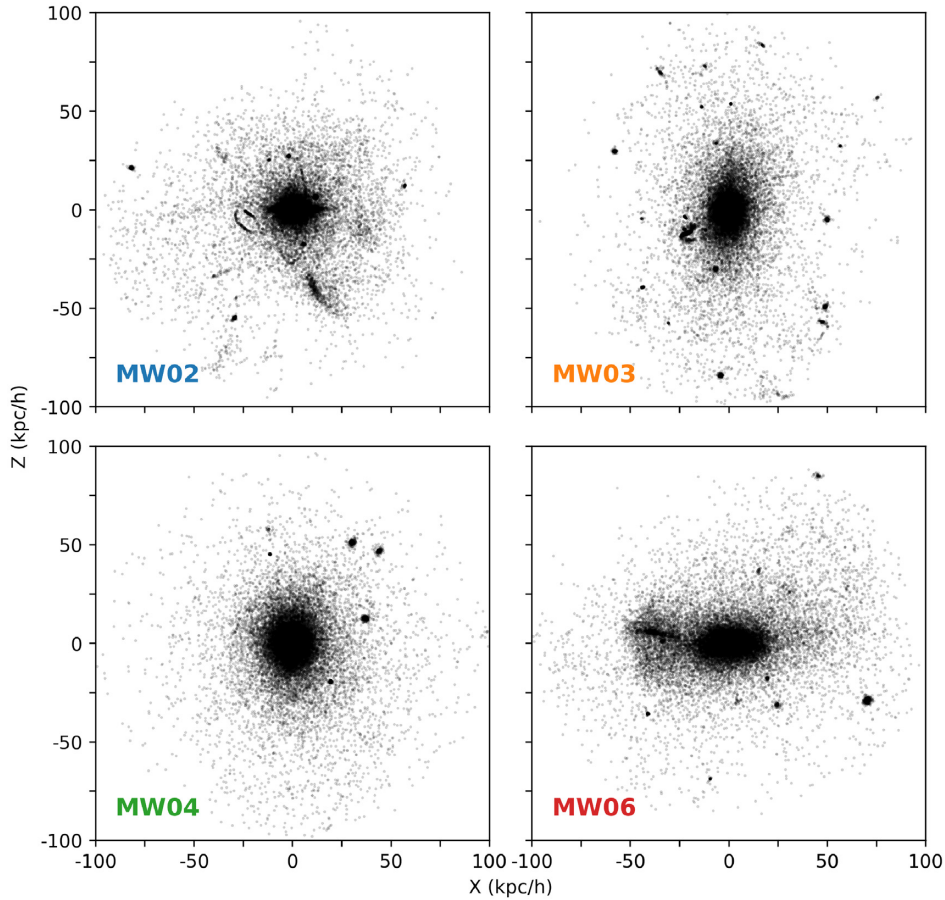
### 2.1 Synthetic haloes

For our synthetic halo data we use those produced by Power & Robotham (2016) at redshift zero. These haloes are drawn from a set of cosmological zoom simulations. The parent simulation (run with GADGET-2; Springel 2005) is a  $\Lambda$  cold dark matter ( $\Lambda$ CDM)  $N$ -body simulation conducted in a  $50 \text{ Mpc } h^{-1}$  cube with  $256^3$  particles. The total matter, baryon, and dark energy density parameters are  $\Omega_m = 0.275$ ,  $\Omega_b = 0.0458$ , and  $\Omega_\Lambda = 0.725$ , respectively, and the dimensionless Hubble constant is  $h = 0.702$ . The power spectrum normalization is  $\sigma_8 = 0.816$ , and the primordial spectral index is  $n_s = 0.968$ . At  $z = 0$ , the FOF algorithm was used to select MW-type haloes with  $M_{200} \approx 2 \times 10^{12} M_\odot h^{-1}$  that reside in low-density (void) regions, which were identified with the V-web algorithm of Hoffman et al. (2012). These galaxies were then resimulated with a version of GADGET-3, as discussed in Power & Robotham (2016), from  $z = 99$  to  $z = 0$  using all particles contained within a radius of  $5R_{200}$ . The resimulations include the baryonic physics of cooling, star formation, supernovae feedback, but do not include any chemical evolution. More details on these simulations are found in Power & Robotham (2016). The stellar particle mass and dark matter particle mass in these resimulated galaxies are  $M_s \approx 10^6 M_\odot h^{-1}$  and  $M_d \approx 5 \times 10^6 M_\odot h^{-1}$ , respectively.

To appropriately consider the structures within stellar haloes we use an open-source PYTHON package for data analysis, namely YT (Turk et al. 2011), to separately read the 3D positions and the masses for all stellar and dark matter particles present in each of the  $\Lambda$ CDM synthetic haloes at  $z = 0$ . To be unambiguous, we only use the 3D positions of these particles to identify the presence of clustering in the data sets. Fig. 1 provides a visualization of each of the synthetic haloes within  $100 \text{ kpc } h^{-1}$  of their barycentre. The barycentres are determined using the *shrinking spheres* method outlined in Power et al. (2003). Fig. 1 indicates that there is an abundance of hierarchical substructure present within each galactic halo.

### 2.2 Ordering points to identify clustering structure

The OPTICS algorithm is a robust tool for hierarchically identifying density-based structure in any  $n$ -dimensional data set for which a distance metric can be defined. It has been used across various fields to quantify human behaviour and mobility patterns (Zheng et al. 2008), characterize the genomic diversity in wheat (Wang et al. 2014), optimize the distribution of urban energy supply systems (Marquant et al. 2017), and more. For data sets containing variables with incompatible units, the distance metric can be difficult to construct. However, given a data set of spatial coordinates, the choice of a distance metric is obvious, namely the Euclidean distance. This certainty makes the application of OPTICS to the physical clustering of particles in 3D space very powerful and robust. Despite this, OPTICS



**Figure 1.** A 2D projection of each of the synthetic haloes within  $100 \text{ kpc } h^{-1}$  of their barycentre. Each of the panels is marked in their lower left according to the galaxy they contain. The colour scheme of these markings corresponds to those used to distinguish the galaxies in Fig. 12. It can be seen that each MW-type galaxy contains a dense central region with an abundance of substructure surrounding it.

has only been applied in an astrophysical context a handful of times (e.g. Fuentes, De Ridder & Debusscher 2017; McConnachie et al. 2018; Canovas et al. 2019; Massaro et al. 2019).<sup>1</sup>

For each point in a data set, OPTICS calculates a measure of the local density surrounding that point called a reachability distance – see equation (2) below. OPTICS also concurrently creates an ordered list of all points in the data set, such that any point with an ordered index of  $i$  is the most *reachable* previously unordered point to all points with an ordered index less than  $i$ . The visualization of the output of the OPTICS algorithm is the reachability plot – the reachability distances as a function of the ordered index for all points in the data set. Fig. 2 depicts the way in which OPTICS achieves this. The reachability distance is an inverse measure of the local density surrounding each point and as such, clusters in the data set present themselves as valleys in the reachability plot. Refer to Section 3.2 and Fig. 4 for

<sup>1</sup>Fuentes et al. (2017) examine the suitability of OPTICS to be applied to large-scale data sets by testing its performance when applied to a simulated astrophysical data set. McConnachie et al. (2018), Canovas et al. (2019), and Massaro et al. (2019) apply OPTICS to observational data sets to identify both new members of existing clusters and new clusters entirely. We build upon these works in order to produce HALO-OPTICS by standardizing the approach under which OPTICS should be applied to astrophysical data sets, establishing a cluster extraction method that appropriately identifies a full hierarchy of astrophysical clusters from the OPTICS output, and verifying its performance.

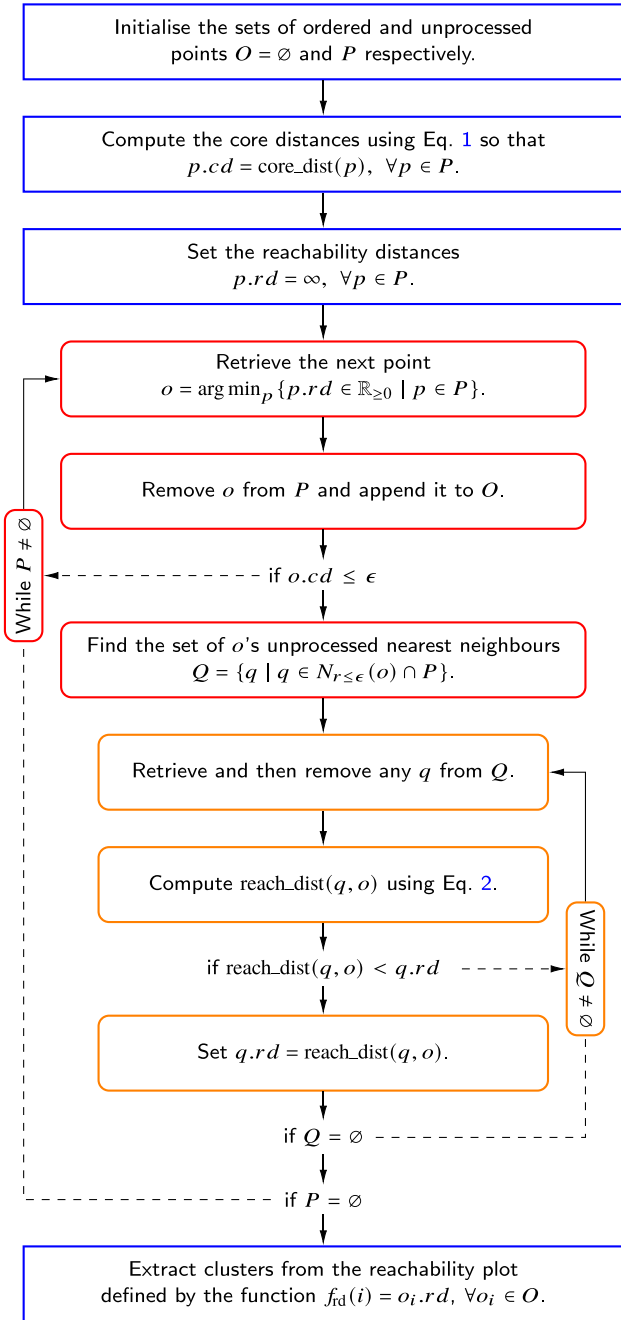
details on the algorithm we use to extract meaningful clusters from the reachability plot.

The power of OPTICS is in part owed to the minimal input required on the user’s behalf. OPTICS only requires the user to provide two parameters,  $\epsilon$  and  $N_{\min}$ . These parameters are chosen according to the data set and are robust enough that small changes in their choice do not strongly affect the reachability plot nor the determination of any clusters present in the data.<sup>2</sup> Refer to Section 3.1 for details regarding our choice of these parameters.

(i) The parameter  $\epsilon$  is the radius for which a nearest neighbour radius query is performed for each point in the data set, and consequently is also the largest possible reachability distance for any point. An appropriate choice of  $\epsilon$  is made through a consideration of the trade-off between the least dense structures that the user wishes to detect, as well as the runtime of the algorithm.

(ii) The parameter  $N_{\min}$  is the minimum number of points that a structure must contain such that it can be detected as a cluster. This parameter is also fundamental in the calculation of the reachability distance. Increasing  $N_{\min}$  reduces noise in the reachability plot, but limits the smallest possible structures determinable to clusters containing at least  $N_{\min}$  points.

<sup>2</sup>The OPTICS output can be greatly affected if the fractional change in  $N_{\min}$  is large even if the change in  $N_{\min}$  itself is small.



**Figure 2.** The OPTICS activity chart with nodes outlined in blue, red, and orange to indicate that they are one-time operations, part of the outer while loop, or part of the inner while loop, respectively. The solid and dashed lines indicate the paths to be taken if a condition is or is not satisfied, respectively. Paths that are not conditional are also shown as solid lines. Paths with arrow heads are unidirectional whereas those without arrow heads may be traversed in either direction depending on the current context, i.e. the process moves towards satisfied conditions. Traditionally the core distances are computed for each point individually after they are appended to  $O$ , however computing them collectively speeds up the process and also allows this step to be run in parallel without changing the reachability plot. The inner while loop may also be run in parallel for any given  $o$ , though the outer while loop cannot be parallelized due to the sequential data access order. It should be noted that there are no additional constraints when retrieving the next point to be ordered into  $O$ , if multiple points have equal reachability distances, then the next point is chosen randomly from them. It is also due to this that the first ordered point is simply a random element of  $P$ .

For a given point  $o$  in the data set, if at least  $N_{\min}$  points are returned from its nearest neighbour radius query, i.e.  $|N_{r \leq \epsilon}(o)| \geq N_{\min}$ ,<sup>3</sup> then that point is labelled as a core-point. Every point is assigned a core distance such that

$$\text{core\_dist}(o) = \|o - p\|_2, \quad (1)$$

where  $p$  is  $o$ 's  $N_{\min}$ th most nearest neighbour. It then follows that every core-point will have a core distance less than or equal to  $\epsilon$ . Given that  $o$  is a core-point, it will at some stage of the algorithm be used to ascertain the reachability distance for each of its currently (at that stage) unprocessed nearest neighbours,  $q$ , within a radius of  $\epsilon$  from  $o$ , i.e.  $q \in N_{r \leq \epsilon}(o) \cap P$ , where  $P$  are those currently unprocessed points. The reachability distance for each of these nearest neighbours with respect to  $o$  is

$$\text{reach\_dist}(q, o) = \max(\text{core\_dist}(o), \|q - o\|_2). \quad (2)$$

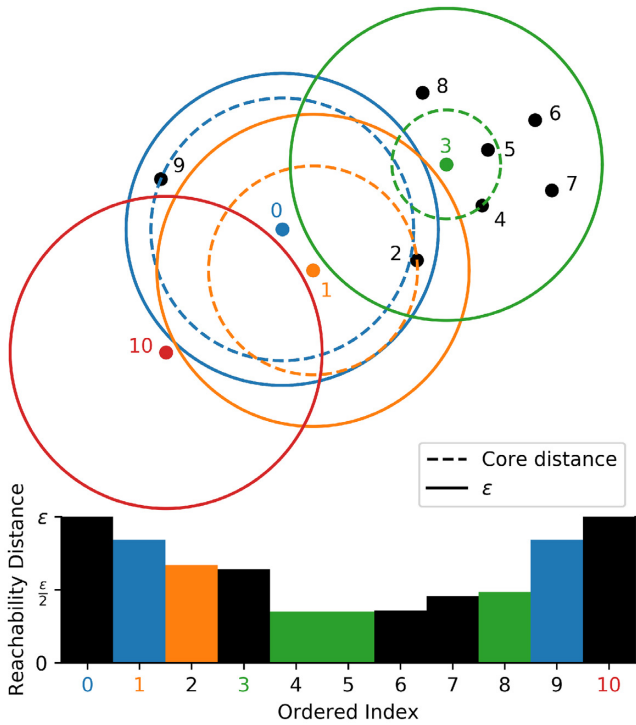
This ensures that the closest  $N_{\min}$  points to  $o$  have a reachability distance with respect to  $o$  equal to the core distance of  $o$ , while all other nearest neighbours of  $o$  have a reachability distance with respect to  $o$  equal to their Euclidean distance from  $o$ . The reachability distance with respect to  $o$  of any given unprocessed nearest neighbour  $q$  is assigned to  $q$  if it is smaller than  $q$ 's currently assigned reachability distance.

For clarity, each point in the data set is initialized with a reachability distance of infinity and the ordered list is determined by iteratively appending to it; the point with the smallest reachability distance. For each iteration, the above set of reachability distances with respect to the current point  $o$  is calculated and used to adjust the reachability distances in the data set as is described above and as in Fig. 2.

This process of adjusting the reachability distance ensures that the reachability plot remains smooth, contains less noise than it otherwise would, and ultimately gives a reliable representation of the density of any structures present within the data set. Furthermore, it is an effective process for limiting the algorithm's knowledge of local densities at any particular iteration to those points that have been ordered up until that iteration, whilst concurrently seeking out the regions of highest density from them. This process makes the reachability distance non-deterministic, as it depends upon the ordered list. As such, the final reachability distance of a point,  $q$ , is always  $\geq$  the smallest core distance of the points in the set of  $q$ 's  $N_{\min}$  reverse nearest neighbours – which is defined to be the set of all points in the data whose  $N_{\min}$  nearest neighbours contain  $q$ . Consequentially, and although not intended for this purpose, the reachability distance of a point (or rather the density within the volume of the  $n$ -sphere it encompasses) can only be interpreted as an approximate density estimator that has been found at the resolution of  $N_{\min}$  nearest neighbours. Density estimators commonly used in numerical cosmology such as the smoothed particle hydrodynamics (SPH; Monaghan 1992) and Voronoi (Voronoi 1908; van de Weygaert 1994; Okabe 2016) estimators are deterministic and do provide a unique measure of density for each point – qualities that the reachability distance does not possess. Ultimately, the reachability distance and the process under which it is created provides not only an approximate measure of local density but more importantly a means for ordering the points of a data set and thereby reducing the  $n$ -dimensionality of the clustering structure to a 2D representation of it. Fig. 3 shows a demonstration of the OPTICS process for a 2D toy data set. Here each point has been marked corresponding to its ordered index. Furthermore, the

<sup>3</sup>Note that by convention a point  $o$  is included amongst its own nearest neighbour search and hence the  $N_{\min}$  nearest neighbours include  $o$  as well.





**Figure 3.** A 2D toy example of the OPTICS algorithmic process and the corresponding reachability plot output. This example is conducted using  $N_{\min} = 3$  with an arbitrarily scaled  $\epsilon$  parameter, the corresponding core distances and nearest neighbour search radius are also shown for the ordered indices 0, 1, 3, and 10. The colour of each coloured bar in the reachability plot corresponds to the colour of the point that produced that bar’s reachability distance value. Notice that the points with ordered indices 0 and 10 actually have reachability distances of infinity (although it is not shown here) as a result of never having previously (at the time they are appended to the ordered list) been a part of any other core-point’s unprocessed nearest neighbourhood – refer to Fig. 2 and the main text of Section 2.2 for more details on the process.

core distances and commonly shared  $\epsilon$  parameter have been marked and uniquely coloured for the ordered indices 0, 1, 3, and 10. The figure illustrates that the more spatially clustered points have been consecutively ordered in the reachability plot and have been assigned smaller reachability distances than the other points.

One of the major drawbacks to OPTICS is that it is computationally demanding, particularly for large data sets. Ankerst et al. (1999) report a constant factor increase in runtime of 1.6 when compared to its predecessor DBSCAN (Ester et al. 1996). The main difference between the two being that DBSCAN returns one level of clustering, i.e. all lists of points that are densely connected through core-point neighbourhoods. Whereas OPTICS extends the rigidity of a point either being part of a cluster or not, to a measure of *how much is a point a part of a cluster* through the means of the reachability distance. The worst-case time complexity for OPTICS is  $\mathcal{O}(n^2)$ , although in general the time complexity is  $\mathcal{O}(n \times r_\epsilon)$ , where  $r_\epsilon$  is the average runtime of the nearest neighbour radius queries. Choosing fast nearest neighbour search algorithms, such as scikit-learn’s KDTree algorithm (Bentley 1975; Pedregosa et al. 2011) ( $\mathcal{O}(r_\epsilon) \rightarrow \mathcal{O}(\log(n))$ ), as well as taking small values for  $\epsilon$ , help to reduce the runtime of a nearest neighbour radius query. Other improvements of the total runtime can be made through seed list optimization, partial sorting, and parallelization (Fuentes et al. 2017), although the parallelization of OPTICS is notoriously difficult due to its strongly sequential data access order, and typically the algorithm must be altered (Patwary et al. 2013). Another property of OPTICS

is that it does not naturally identify the clusters present within the data since it only returns a measure of local density about each point. However as is explained below in Section 3.2, this is also its most advantageous quality as it allows the user to be specific about the density, hierarchy level distinction, and point inclusion criteria that they wish to use to define a cluster.

### 3 HALO-OPTICS: A HIERARCHICAL GALAXY/(SUB)HALO FINDER

#### 3.1 Choosing appropriate OPTICS hyperparameters

As is mentioned in Section 2.2, the choices of  $\epsilon$  and  $N_{\min}$  are determined through a consideration of the minimum size and density of structures the user wishes to detect, as well as the runtime constraints the user wishes to adhere to. In order to provide a substantially high degree of resolution for the identifiable clusters, we choose to set  $N_{\min} = 20$  – a common choice for the minimum size of meaningful clusters in substructure finders. The choice to set  $N_{\min}$  as a constant between each application to the galactic haloes is implemented so that the minimum possible mass of the clusters remains roughly equal between them (and only differs due to the small differences between particle masses in these simulations). The lower mass limit of clusters is  $\approx 2 \times 10^7 M_\odot h^{-1}$  for stellar clusters and  $\approx 1 \times 10^8 M_\odot h^{-1}$  for dark matter clusters. This ensures that we can meaningfully compare the clustering between different particle types and different haloes. Having such a small value of  $N_{\min}$  does however introduce more noise to the reachability plot than when compared to that of larger  $N_{\min}$  values. This extra noise makes meaningful cluster extraction more involved and less obvious, and so we have constructed our own algorithm for automatic cluster detection that we present in Section 3.2.

Making the choice for  $\epsilon$  is a little more difficult as this essentially specifies the size and extent of the least dense structures. The FOF analogue for this parameter is the linking length that, in cosmological simulations and given a halo with virial overdensity  $\Delta$  that contains  $N_\Delta$  particles within a radius of  $R_\Delta$ , may be chosen using  $l_x = (2\pi/N_\Delta)^{1/3} R_\Delta$  (Elahi, Thacker & Widrow 2011). To extrapolate this to the  $\epsilon$  parameter we need to account for the fundamental difference between the algorithms. A point will only be assigned a reachability distance if it has previously been included in a now-ordered core-point’s set of unprocessed nearest neighbours within a radius of  $\epsilon$ . Therefore, the least dense core such as a point can be a part of is the core that surrounds the core-point that has exactly  $N_{\min}$  nearest neighbours that extend out to exactly a radius of  $\epsilon$ . This can be leveraged to find the factor by which  $\epsilon$  must be larger than the FOF linking length to encompass the same overdensity.<sup>4</sup>

<sup>4</sup>Because of the fundamental differences between OPTICS and FOF there will not, in general, be a value for  $\epsilon$  that produces precisely the same grouping of points as FOF would by using a particular linking length  $l_x$ . This is because OPTICS does not care about intracore structure while FOF does, i.e. OPTICS is less susceptible to point–point noise than FOF is. It should also be noted that constructing  $\epsilon$  from  $l_x$  in this way is an approximation for finding haloes that enclose a specified overdensity. As detailed in the study by More et al. (2011), there does not exist a unique linking length that encloses a specified overdensity for all FOF haloes. It is shown therein that the resultant enclosed overdensity of FOF haloes is not only dependent on the number of particles in the halo but also the concentration of the density peak. As such, the enclosed overdensity of HALO-OPTICS haloes is also subject to this ambiguity. However, since the mapping between an FOF halo and a HALO-OPTICS halo is not exact either, we do not find it necessary to conduct a more thorough determination of  $l_x$  for the purpose of computing  $\epsilon$ .

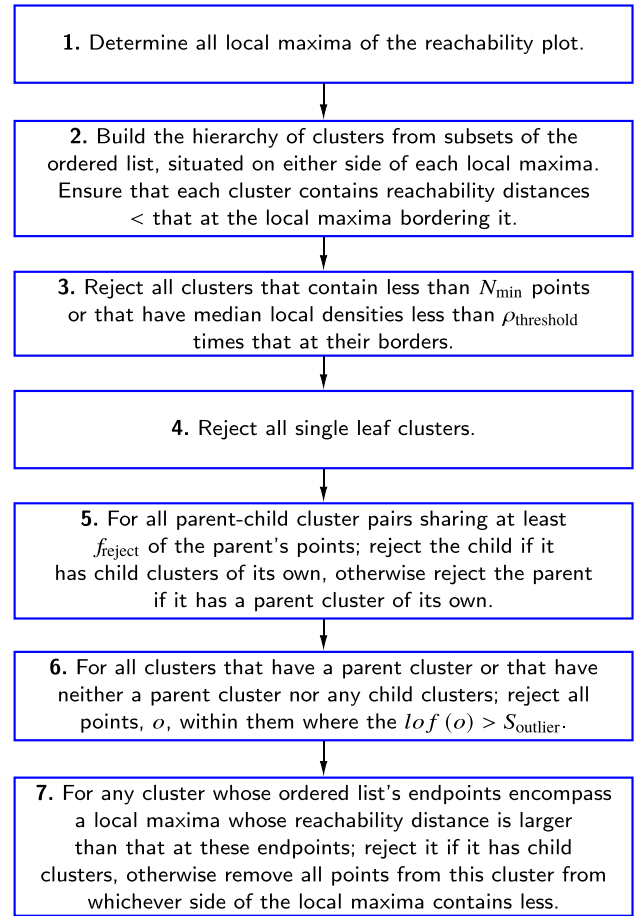
For each application of HALO-OPTICS, we first compute the corresponding FOF linking length (as detailed above) and then seek to multiply this by the mean of the end-to-end distance of a chain of  $N_{\min}$  points, each separated by unit length and each sampled from a directionally uniform probability distribution. Even though we only use  $N_{\min} = 20$  in this work, the following describes how we prepare the code for use with all possible values of  $N_{\min}$ . Since there is no analytical solution to the mean of this distance distribution in terms of elementary functions, we run a series of Monte Carlo simulations for  $N_{\min} = 3\text{--}20$  that compute it and store them as hard coded variables within the code. The case when  $N_{\min} = 2$  is trivial since this is identical to the FOF case and  $\epsilon = l_x$ . For larger values of  $N_{\min}$  we can base our approximation off the root-mean-square distance since it does have an analytical formula and is given by  $d_{\text{rms}}(N_{\min}) = \sqrt{N_{\min} - 1}$ . The root-mean-square of the end-to-end distance of the chain approaches a constant value of  $\sim 1.084$  times the Monte Carlo simulated mean of the end-to-end distance of the chain. Therefore, a good approximation of the mean of the end-to-end distance of the chain for  $N_{\min} > 20$  is given by  $d_{\text{rms}}(N_{\min})/1.084$ . For a set of 10 roughly log-spaced integers between  $N_{\min} = 21$  and 1000, this approximation is always within 0.2 per cent of the mean found from Monte Carlo simulations. All Monte Carlo simulations used here calculate  $10^7$  separate chains from which the end-to-end distances and resultant means are found. This calculation of  $\epsilon$  from  $N_{\min}$  and the FOF linking length,  $l_x$ , effectively switches the OPTICS input parameter  $\epsilon$ , to the physically motivated HALO-OPTICS input parameter  $\Delta$ , which we choose as  $\Delta = 200$  (times the critical density of our Universe). This also means that the root haloes (refer to Section 3.2 for a breakdown of this terminology) found by HALO-OPTICS encompass a similar overdensity to those that would be found by an FOF-based code.

### 3.2 Extracting clusters from OPTICS

Since the OPTICS algorithm itself does not return any clusters, the extraction of clusters from reachability plots is a separate and unique problem whose difficulties arise due to the innate subjectivity with regards to the definition of a cluster. Two commonly used automatic cluster extraction techniques are the  $\xi$ -step method, first proposed in the original OPTICS paper (Ankerst et al. 1999), and the DBSCAN method, which effectively returns those clusters that the OPTICS predecessor DBSCAN (Ester et al. 1996) would have. While the  $\xi$ -step method is able to extract a hierarchy and the DBSCAN method can do so after being applied iteratively, neither is robust enough to extract all necessary clusters at any overdensity. To combat these downfalls, we have developed our own extraction process based on the designs of Sander et al. (2003), Zhang et al. (2013), and McConnachie et al. (2018); Fig. 4 summarizes the steps involved. This extraction process produces a series of tree structures consisting of clusters for nodes that we refer to as the HALO-OPTICS hierarchy. In alignment with the standard terminology for this data structure, we refer to any pair of clusters separated by a single branch as a parent–child cluster pair, and we refer to a tree’s terminating clusters as the root and leaf clusters.

#### 3.2.1 Step 1

Given that dense regions of the data are described by valleys in the reachability plot, our first step to extracting the clusters present in the data is to define all local maxima in the reachability plot as the boundaries of clusters. For the purpose of finding these local maxima, we treat all undefined reachability distances as being equal to  $\epsilon$ , which



**Figure 4.** The cluster extraction activity chart summarizing the steps we take to determine clusters from the OPTICS output. Steps 1, 2, 3, and 6 can be parallelized, while steps 4, 5, and 7 can only be partially parallelized – due to the need to preserve the state of the hierarchy before these steps whilst the step is conducted, refer to step 3 in the main text for more details on this. However, we only perform these steps using a single core due to the entire extraction process only taking a small fraction of the time it takes to run OPTICS for any particular data set. It is important that these steps be performed exactly in this order else the extracted clusters may not all be meaningful.

occur when a point has never been included in a nearest neighbour radial query of another point that returns at least  $N_{\min}$  points.

#### 3.2.2 Step 2

We then construct clusters out of lists of consecutively ordered points contained within valleys of the reachability plot. Since each valley must be bordered by a local maximum, we link up contiguous sets of points on both sides of every local maximum in the reachability plot with reachability distance less than that at the corresponding local maximum. This creates two clusters for every local maximum. Following this step, every cluster will contain at least one local minimum, and be bordered by one (possibly two, if a valley is bordered by two local maxima with the same reachability distance) local maximum.

#### 3.2.3 Step 3

As expected, this process incurs many artefact clusters that are insignificant and so we now move through a rejection process that

removes these clusters from the list. Since the reachability distance of a particle is the radius of a particle encompassing sphere, we now approximate that the local density is inversely proportional to the volume of the  $n$ -sphere –  $n = 3$  in this paper – with a radius of the reachability distance. We use this concept in the first step of our rejection process that rejects any potential clusters that satisfy either of the following.

- (i) The cluster contains less than  $N_{\min}$  particles.
- (ii) The median local density of the cluster is less than  $\rho_{\text{threshold}}$  times that at the local maximum that was used to create it.

The first of these criteria ensures that clusters contain at least  $N_{\min}$  particles as required by the resolution of structures in OPTICS. For the second criterion, the chosen density contrast,  $\rho_{\text{threshold}}$ , guarantees that half of the points of a cluster must be at least  $\rho_{\text{threshold}}$  as dense as that cluster’s surroundings. Refer to Section 3.3 for our determination of a reasonable  $\rho_{\text{threshold}}$  value, ultimately we use  $\rho_{\text{threshold}} = 2$ . Some of the remaining steps are performed by first considering whether each cluster in the list satisfies a condition before then rejecting all such clusters at once. We conduct the process in this way due to some of these conditions depending on the state of the hierarchy and therefore whether or not a cluster satisfies such a condition is susceptible to change under a typical *reject mid iteration* type method.

### 3.2.4 Step 4

We now mark all clusters that are a single child of their parent cluster, before then rejecting each of them. We justify this as a necessary step to remove any clusters that are simply smaller versions of their parent cluster. Single child clusters occur in the hierarchy when one of the two clusters for each local maxima that was originally created in step 2, has been rejected during step 3.

### 3.2.5 Step 5

At this step, the list of clusters still typically contains many cascading parent–child clusters that share large numbers of points. For all parent–child cluster pairs sharing at least  $f_{\text{reject}}$  of the parent’s points; mark the child for rejection if it has child clusters of its own, otherwise mark the parent for rejection if it has a parent cluster of its own. We then reject those marked clusters after inspecting the entire list. This step further ensures the individuality of clusters in consecutive levels of the hierarchy. Refer to Section 3.3 for our determination of a reasonable  $f_{\text{reject}}$  value, ultimately we use  $f_{\text{reject}} = 90$  per cent.

### 3.2.6 Step 6

Another artefact of determining clusters in this way is that each cluster will likely contain outlier points that do not belong as part of the cluster. We now reject outlier particles from all clusters that either have a parent cluster, or that have neither a parent cluster nor any child clusters. By exempting all root clusters (with child clusters) from the outlier rejection, we maintain the lists of particles that give the best description of larger halo environments. For those clusters that we do apply the outlier detection to, we reject particles on the basis outlined by Breunig et al. (1999) who define the local-reachability-density of a point,  $o$ , as

$$\text{lrd}(o) = \frac{N_{\min}}{\sum_{q \in N_{r \leq o.cd}(o)} \text{reach\_dist}(o, q)}. \quad (3)$$

Here  $o.cd = \text{core\_dist}(o)$  from equation (1) and  $\text{reach\_dist}(o, q)$  is defined in equation (2). It should be noted that these values are found using only the points from within each cluster and in general will be different to those found during the OPTICS process, and will also differ from cluster to cluster for any point contained in multiple clusters. The local-outlier-factor of  $o$  is then defined as

$$\text{lof}(o) = \frac{\sum_{q \in N_{r \leq o.cd}(o)} \frac{\text{lrd}(q)}{\text{lrd}(o)}}{N_{\min}}. \quad (4)$$

We find the local-outlier-factor for all points of a cluster, for all clusters. For each cluster, we then reject all points from it that have a local-outlier-factor greater than  $S_{\text{outlier}}$ . Any point that is a part of  $n$ -many clusters will therefore have  $n$  individual local-outlier-factors that are respective to each. It then follows that such a point may be rejected from one cluster and not another. However, it also important to note that since a parent cluster contains a larger number of lower density points than its child cluster, the local-outlier-factor of a point contained in both parent and child clusters will always be larger than or equal with respect to the child cluster than it is with respect to the parent cluster. Therefore following this step, a child cluster will still never contain a point that a parent cluster does not. Refer to Section 3.3 for our determination of a reasonable  $S_{\text{outlier}}$  value, although ultimately we use the suggestion from Breunig et al. (1999) that  $S_{\text{outlier}} = 2$ .

### 3.2.7 Step 7

Following the rejection of outlier points from all clusters, a possible fringe case occurs where the ordered list for some clusters now encompasses one or more points (that are not necessarily contained in the cluster itself) whose local density is less than that of either of the points at the cluster’s ordered list bounds. This is essentially a discontinuity in the density field of the cluster as we have determined it thus far. So for any cluster that satisfies this condition, we reject it if it contains a child cluster, otherwise we remove all points from this cluster from whichever side (in the ordered list) of the local maximum contains less of them. If the removal of these points leaves the cluster with less than  $N_{\min}$  particles, then we reject the cluster.

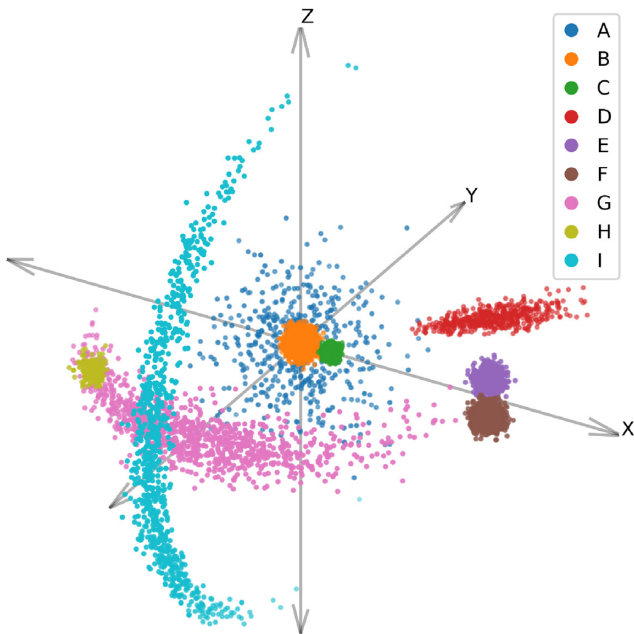
Following these steps, we are left with a list of clusters that we have determined to be significantly denser, distinct, and self-consistent when compared to their surroundings. Moreover, the process is completed with the addition of only three user-defined parameters –  $\rho_{\text{threshold}}$ ,  $f_{\text{reject}}$ , and  $S_{\text{outlier}}$ . It should be noted that the clusters and the hierarchy they are a part of is not necessarily the hierarchy that might be assigned based on physical reasons. It is particularly dependent upon  $\epsilon$  in the root level and similarly upon  $N_{\min}$  in the leaf level. Importantly, the detection of significant overdensities by HALO-OPTICS is still very informative even though it may not be physical, and further still there is the possibility for implementing changes to the extraction process such that it presents a more physical set of clusters.

## 3.3 Performance optimization

To justify our extraction process in Section 3.2 we now present some performance statistics of this process and how they are affected by the extraction parameters  $\rho_{\text{threshold}}$ ,  $f_{\text{reject}}$ , and  $S_{\text{outlier}}$ , type of structure present, as well as the level of non-structured background noise within the data set. To do this, we create a mock cluster set designed to mimic typical astrophysical structures. The data are contained within a unit cube centred on the origin and the total number of

**Table 1.** The descriptions and distributional parameters of each of the nine mock clusters used to investigate the purity and recovery statistics of HALO-OPTICS. The analysis is conducted inside a cube with a side length of 1 that is centred on the origin. The total number of points within each cluster is  $N \times (1 - f_b) \times f_c$  rounded to the nearest integer, where  $N$  is the total number of points inside the unit cube,  $f_b$  is the proportion of background noise, and  $f_c$  is the proportion of clustered points belonging to that particular cluster. Notice that  $\sum f_c = 1$ .

Cluster	Description	Centre coordinates	Spread	$f_c$
A	Sphere	$(x, y, z) = (0, 0, 0)$	$(\sigma_x, \sigma_y, \sigma_z) = (0.06, 0.06, 0.06)$	0.05
B	Sphere inside A	$(x, y, z) = (0, 0, 0)$	$(\sigma_x, \sigma_y, \sigma_z) = (0.01, 0.01, 0.01)$	0.25
C	Sphere at edge of B	$(x, y, z) = (0.05, 0, 0)$	$(\sigma_x, \sigma_y, \sigma_z) = (0.005, 0.005, 0.005)$	0.25
D	Cone extending radially from A	$(x, y, z) = (0.2, 0.2, 0)$	$(\sigma_r, \sigma_\theta, \sigma_\phi) = (0.05, 2^\circ, 2^\circ)$	0.05
E	Sphere nearby F	$(x, y, z) = (0.3, 0, 0.03)$	$(\sigma_x, \sigma_y, \sigma_z) = (0.01, 0.01, 0.01)$	0.1
F	Sphere nearby E	$(x, y, z) = (0.3, 0, -0.03)$	$(\sigma_x, \sigma_y, \sigma_z) = (0.01, 0.01, 0.01)$	0.1
G	Angular arc nearby H	$(x, y, z) = (0, -0.3, 0)$	$(\sigma_r, \sigma_\theta, \sigma_\phi) = (0.01, 5^\circ, 25^\circ)$	0.1
H	Sphere inside tail of G	$(r, \theta, \phi) = (0.3, 90^\circ, -135^\circ)$	$(\sigma_x, \sigma_y, \sigma_z) = (0.01, 0.01, 0.01)$	0.02
I	Angular arc nearby G	$(x, y, z) = (0, -0.4, 0)$	$(\sigma_r, \sigma_\theta, \sigma_\phi) = (0.01, 30^\circ, 2^\circ)$	0.08



**Figure 5.** Projection of the mock clusters listed in Table 1. The clusters, which are coloured by their true label, are designed to mimic a variety of typical astrophysical clusters that also provide many intricacies for OPTICS to interpret such as closely situated yet unique overdensities, elongated structures, and a (somewhat) arbitrarily multilevelled hierarchy.

points is kept at a constant  $N = 10^4$ . The OPTICS hyperparameters are chosen as  $N_{\min} = 20$  to mimic our application to the MW-type galaxies and  $\epsilon \rightarrow \infty$  so that the root cluster includes all points.

The clusters are created using 3D Gaussian distributions of various sizes, spreads and positions in both  $x, y, z$  and  $r, \theta, \phi$  coordinates.<sup>5</sup> The descriptions and distributional parameters of these clusters are presented in Table 1, and a 2D projection of one sampling is shown

<sup>5</sup>We do not use typical halo profiles here as the performance of HALO-OPTICS does not depend on the exact density profile of a cluster. This is due to OPTICS not using any information about the exact structure of a cluster in order to link the points within it, and is precisely why it excels at finding arbitrarily shaped clusters. We do, however, use a typical halo profile for clusters in Section 3.4 as the exact density profile will affect the conditions under which particle bridges are created – which in turn affects the shape of the HALO-OPTICS hierarchy.

in Fig. 5. The proportion of the total clustered points is given by (100 per cent  $- f_b$ ), where  $f_b$  is the percentage of background noise. We vary the background noise from  $f_b = 0$  per cent to 90 per cent in increments of 3 per cent and sample it using a uniform distribution of points throughout the space. For each level of  $f_b$  we run OPTICS 50 times, resampling from all distributions for each run.

We first assess the performance of HALO-OPTICS through measures of recovery and purity. We define the recovery to be a function of the true clusters and to be dependent on the levels of the predicted hierarchy such that

$$R(T|L) = \frac{\sum \{|T \cap C| \mid \forall C \in L\}}{|T|}, T \in M \wedge L \subset H. \quad (5)$$

Here  $T$  is a true (mock) cluster,  $M$  is the set of mock clusters,  $C$  is a HALO-OPTICS predicted cluster, and  $L$  is the set of predicted clusters that belong to the  $L$ th level of the predicted hierarchy,  $H$ . This way the recovery of a particular true cluster can be interpreted as the fraction of that cluster that is returned in the  $L$ th level of the predicted hierarchy. Since predicted clusters from HALO-OPTICS in the same hierarchical level cannot overlap, this value will always be contained to the interval  $[0, 1]$ .

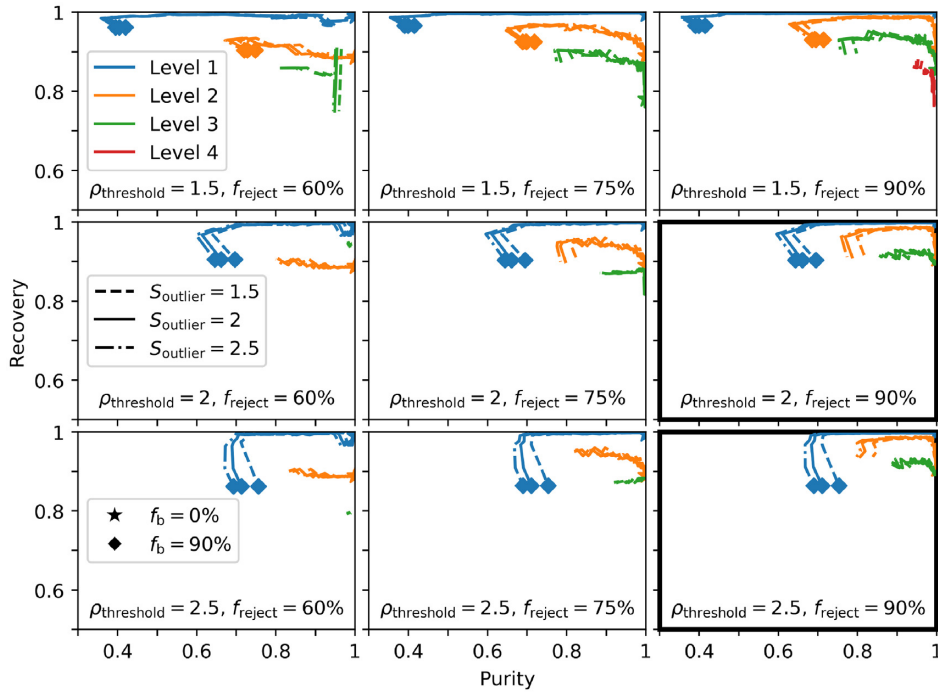
Similarly, we define the purity to be a function of the predicted clusters and to be dependent on the levels of the predicted hierarchy such that

$$P(C|L) = \frac{\sum \{|T \cap C| \mid \forall T \in M\}}{|C|}, C \in L \subset H. \quad (6)$$

Here  $T, M, C, L$ , and  $H$  are the same terms as in equation (5). This definition of the purity of a particular predicted cluster,  $C$ , can be interpreted as the fraction of  $C$  that is not background noise.

For each of the background noise/distribution resampling combinations we find the recovery versus purity relations for 27 HALO-OPTICS parameter combinations. The combinations draw from  $\rho_{\text{threshold}} \in \{1.5, 2, 2.5\}$ ,  $f_{\text{reject}} \in \{60 \text{ per cent}, 75 \text{ per cent}, 90 \text{ per cent}\}$ , and  $S_{\text{outlier}} \in \{1.5, 2, 2.5\}$ . Whilst masking all zero recovery and zero purity values, we then average over both the distributional resamplings and each level of the predicted hierarchy.

Fig. 6 depicts these recovery versus purity relations as dependent on the parameter combination, the level of the hierarchy, and the level of background noise. For all parameter combinations the recovery and purity decrease and increase, respectively, for a given level of background noise as the level of the hierarchy deepens. This should be expected, clusters in deeper levels of the hierarchy are denser and have fewer points than their parent clusters. As a result, they



**Figure 6.** The recovery and purity relations of HALO-OPTICS as dependent on various combinations of the extraction parameters –  $\rho_{\text{threshold}}$ ,  $f_{\text{reject}}$ , and  $S_{\text{outlier}}$  – after having been applied to the mock clusters listed in Table 1. Each panel displays the recovery versus purity as it is dependent on the various  $S_{\text{outlier}}$  values, hierarchy levels, and levels of background noise. We do not show level 0, the root level, as it contains all points in the data set and hence its recovery versus relation is a trivial line of constant recovery (=1) and decreasing purity ( $= 1 - f_b/100$  per cent). We find that the main contribution to the performance of HALO-OPTICS comes from the  $\rho_{\text{threshold}}$  and  $f_{\text{reject}}$  parameters with  $S_{\text{outlier}}$  making little difference. The bold bordered panels showcase the two best performing parameter combinations that we compare against each other by way of the maximum Jaccard index in Fig. 7.

are less affected by the increase in noise (higher purity per level of background noise) and are less likely to include more of the total clustered points in the data set (lower recovery in general).

Another noticeable feature in Fig. 6 is that some deep levels of the hierarchy reveal a drop in recovery at  $f_b = 0$  per cent. This characteristic is exaggerated for low  $\rho_{\text{threshold}}$  and low  $f_{\text{reject}}$  values. Following step 1 of the extraction process, the valley of the reachability plot that corresponds with any particular true cluster will typically have some non-zero number of cascading parent–child predicted clusters associated with it – these range from high density with a few points, to lower density with more points. In general, lowering  $\rho_{\text{threshold}}$  allows for the extraction of higher density child clusters within this valley, and lowering  $f_{\text{reject}}$  effectively removes the parent clusters above them. Increasing the level of background noise lowers the density contrast surrounding the each of the mock clusters, which in turn stops the low  $\rho_{\text{threshold}}$  from allowing the extraction of as many child clusters. Likewise, the resulting predicted cluster has a higher recovery and slightly lower purity for some  $f_b > 0$  per cent since it contains more points at a lower density. Eventually, as the background noise level increases, the prediction of the true clusters breaks down and the recovery drops dramatically. This can be seen at each hierarchy level for every parameter combination.

We see here that the choice of  $S_{\text{outlier}}$  makes little difference to the recovery versus purity relation and as such we choose to take  $S_{\text{outlier}} = 2$ , the suggestion by Breunig et al. (1999). Effectively,  $S_{\text{outlier}}$  is only responsible for removing a small number of points in comparison the size of the cluster, so it should be expected that this parameter has little effect on the recovery and purity. The bold bordered panels are the best performing  $\rho_{\text{threshold}}/f_{\text{reject}}$  parameter combinations. It is

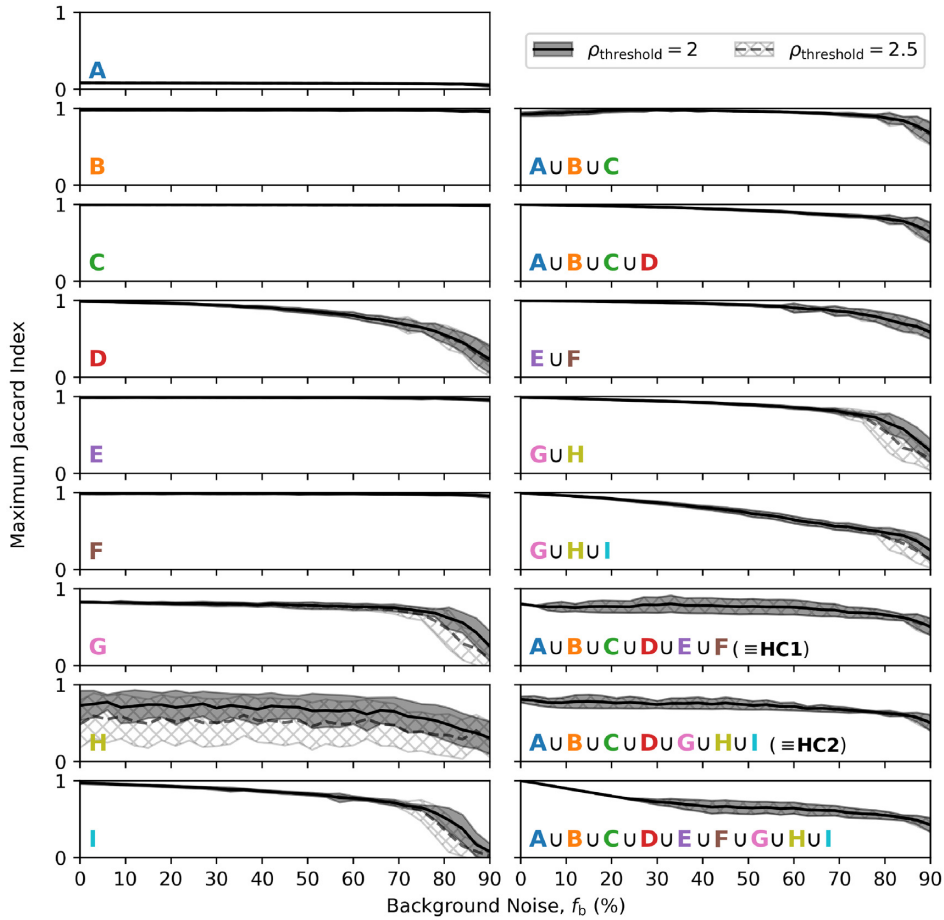
also clear here that  $f_{\text{reject}} = 90$  per cent is a good choice, however it is not immediately obvious as to whether  $\rho_{\text{threshold}} = 2$  or  $\rho_{\text{threshold}} = 2.5$  is a better choice.

To help distinguish between the two  $\rho_{\text{threshold}}$  choices, we also look at the maximum Jaccard index for each true cluster and some hierarchical combinations of them. The maximum Jaccard index is defined as

$$J_{\text{max}}(T) = \max \left\{ \frac{|T \cap C|}{|T \cup C|} \mid \forall C \in H \right\}, T \in M'. \quad (7)$$

Here  $M'$  is the set of true clusters and some typically predicted hierarchical combinations of them. Likewise,  $T$  is any element of  $M'$ , and  $C$  is any cluster in the predicted hierarchy,  $H$ , regardless of the hierarchy level it belongs to. The maximum Jaccard index provides a measure of *how close of a match does the best-fitting predicted cluster provide for any given true cluster* – a true cluster here being any element of  $M'$ . In this way, we can test the performance of HALO-OPTICS by examining how well it predicts  $M'$ . Fig. 7 depicts the mean and one standard deviation range of each  $T \in M'$  for both  $\rho_{\text{threshold}} = 2$  and  $\rho_{\text{threshold}} = 2.5$  when  $f_{\text{reject}} = 90$  per cent and  $S_{\text{outlier}} = 2$ . Here the mean and standard deviation are found from the series of HALO-OPTICS outputs over each of the distributional resamplings.

These also reveal a few features of the HALO-OPTICS predicted clusters in general. One such feature is that the stream-like structures – D, G, and I – are more affected by the increase in background noise. This could be expected since these structures are elongated and will have a lower spatial density than a Gaussian sphere, and therefore will become less distinguishable from the background noise more readily. Another feature is overencompassing clusters, such as cluster A, are



**Figure 7.** The maximum Jaccard index of all mock clusters, and some hierarchical combinations of them, as listed in Table 1. This measure indicates how close of a match the best-fitting predicted cluster from HALO-OPTICS is to any particular (union of) true cluster(s). Here we show the mean value and  $\mu \pm \sigma$  range of the maximum Jaccard index for both  $\rho_{\text{threshold}} = 2$  and  $\rho_{\text{threshold}} = 2.5$  when  $f_{\text{reject}} = 90$  per cent and  $S_{\text{outlier}} = 2$ . Here we see that the denser more spherical clusters are predicted almost perfectly, whereas the more stream-like cluster predictions made by HALO-OPTICS gradually suffer from the increase in background noise. However, it is obvious from this comparison that  $\rho_{\text{threshold}} = 2$  does perform better than  $\rho_{\text{threshold}} = 2.5$  in that the extraction process consistently produces better fitting predictions of the true clusters D, G, H, and I as well as some hierarchical combinations of these. This higher quality performance is particularly noticeable for the more stream-like structures as the background noise increases – a desired quality of structure finders.

not shown to be well matched here and have a lower maximum Jaccard index than other clusters. This is due to the algorithm not separating out the inner clusters, B and C, and in this way the best-fitting match for cluster A is most probably that provides the best match to the hierarchical combination of clusters A, B, and C. Of course, if the set difference of the best-fitting match to the latter was taken with the best-fitting matches to clusters B and C, we could provide a better match to cluster A alone. Interestingly, the maximum Jaccard index of the hierarchical combination of clusters A, B, and C has a slight peak around  $f_b \approx 30$  per cent. This occurs as a result of the reachability distance for the outer points of cluster A being reduced by the addition of the background noise points in these regions. This effect is small and, as the background noise increases, is outweighed by the inclusion of additional intracluster noise.

Fig. 7 also includes two panels of a pair of disjoint hierarchical combinations of clusters – namely the union of clusters A, B, C, D, E, and F, and the union of clusters A, B, C, D, G, H, and I. These have been included to express that the purposely arbitrary hierarchy we have constructed within our mock cluster set has been translated into an equally ambiguous predicted hierarchy. For the purpose of providing an explanation with regards to this, we will refer to the

former hierarchical combination as HC1 and the latter as HC2 for the remainder of this paragraph. The disjoint nature of HC1 and HC2 ensures that their respective best-fitting predicted clusters cannot be one and the same unless their shared best-fitting predicted cluster either contains only those leaf clusters that are common to both HC1 and HC2 (i.e. A, B, C, and D), or contains all leaf clusters (i.e. A–I). In the following we refer to these scenarios as case 1 and case 2, respectively.<sup>6</sup> For  $f_b = 0$  per cent, the Jaccard index for cases 1 and 2 is  $0.6/0.8 = 0.75$  and  $0.8/1 = 0.8$ , respectively, and hence the maximum Jaccard index for both HC1 and HC2 at this level of background noise is 0.8. However, as the background noise level increases the Jaccard index for each of these cases changes in a way that depends on the effective occupied volume with the unit cube of

<sup>6</sup>There are technically more cases that can occur, for example, where HC1 (or HC2) is best matched by a predicted cluster that includes only points from the leaf clusters within it (and some level of background noise), and then HC2 (or HC1) is best matched by either of the predicted clusters in cases 1 and 2. While this occurs for some other hyperparameter combinations that return a larger number of hierarchy levels, it does not occur within the hyperparameter combinations featured in Fig. 7 – and so we do not discuss these extra cases.

the true and predicted clusters proportionally – there is also some random component to this due to the randomized sampling of both the true clusters and the background noise. The occupied volume of the predicted clusters changes with the background noise as well. As a result, the best-fitting predicted cluster is not always provided by that from case 2, and the corresponding maximum Jaccard index to the best-fitting predicted cluster is not confined to less than 0.8 either. This is shown by the increase in the spread of the maximum Jaccard index for both HC1 and HC2.

From Fig. 7, we clearly see that with  $\rho_{\text{threshold}} = 2$ , HALO-OPTICS outperforms that with  $\rho_{\text{threshold}} = 2.5$  for true clusters D, G, H, and I – as well as some hierarchical combinations of the latter three – particularly when the background noise dominates the data. Not only is the mean value of the maximum Jaccard index for these clusters larger under the  $\rho_{\text{threshold}} = 2$  parameter scheme, but the spread is smaller too. It is now apparent that  $\rho_{\text{threshold}} = 2$  is the better parameter choice. So to summarize, we use  $\rho_{\text{threshold}} = 2$ ,  $f_{\text{reject}} = 90$  per cent, and  $S_{\text{outlier}} = 2$  as our (near) optimal HALO-OPTICS hyperparameters.

### 3.4 Understanding the HALO-OPTICS hierarchy

We now wish to inform the reader about the nature of the HALO-OPTICS hierarchy. To do this we construct another mock example of clusters, only now we intend for these clusters to constitute an easily understandable hierarchy. The mock example we use here consists of two distributions that are intended to represent a main halo and a satellite halo. Both of these haloes are modelled using the spherical Navarro–Frenk–White (NFW) profile (Navarro, Frenk & White 1996) that has a density profile of the form

$$\rho(r) = \frac{\rho_0}{\frac{r}{R_s} \left(1 + \frac{r}{R_s}\right)^2}, \quad (8)$$

where  $\rho_0$  and  $R_s$  are the characteristic density and radius, respectively. We use this to create a cumulative distribution function for the NFW profile from which to sample the main and satellite halo distributions from. We integrate the mass density profile in equation (8) over the volume and out to some variable radius  $r$ , and then renormalize this such that the integral out to some maximum radius,  $R_{\text{max}}$ , is unity. It then follows that an appropriate cumulative distribution function for the haloes is given by

$$F_{\text{NFW}}(r) = \frac{\left[\ln\left(1 + \frac{r}{R_s}\right) - \frac{r}{(r+R_s)}\right]}{\left[\ln\left(1 + \frac{R_{\text{max}}}{R_s}\right) - \frac{R_{\text{max}}}{(R_{\text{max}}+R_s)}\right]}, \quad (9)$$

where  $0 \leq r \leq R_{\text{max}}$ . For the main halo we choose  $R_s = 1$  and for the satellite halo we choose  $R_s = 0.2$ . For both haloes we use  $R_{\text{max}} = 10R_s$  and for the main halo we treat the effective  $R_{\Delta}$  as being equal to  $5R_s$  for the purposes of constructing  $\epsilon$  from the data in the way described in Section 3.1.

We now vary the resolution, mass fraction, and separation distance of the two-halo system and run HALO-OPTICS over each realization using the previously mentioned values of the HALO-OPTICS hyperparameters. The resolution is the total number of points in the data set that we sample from 13 logarithmically spaced values between 40 and  $10^4$ . The mass fraction is the ratio between the number of points in the satellite and the number of points in the main halo that we sample from 11 logarithmically spaced values between 0.002 and 1. These values correspond to the number of particles belonging to the satellite being equal to the lowest possible limit of detection by HALO-OPTICS when the resolution is  $10^4$  (when using

**Table 2.** Details of the number of clusters,  $|H|$ , in the HALO-OPTICS hierarchy and the dependency of this on the resolution, mass fraction, and separation distance variables of the two-profile system described in Section 3.4. Because of the non-linear dependencies between the four, discretely sampled, variables we only give a qualitative description of the generalized domain for which a particular number of clusters may be found in the input space. For the purposes of being succinct, we refer to the resolution, mass fraction, and separation distance as  $R$ ,  $M$ , and  $S$ , respectively.

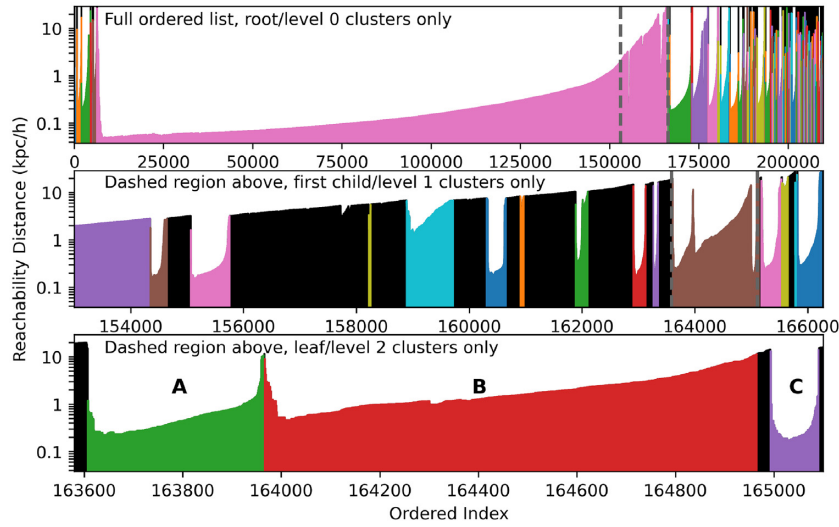
$ H $	Occurrences	Domain description
0	417	Very small $R \wedge$ very large $M \wedge$ large $S$
1	10 471	(Small $R \wedge$ small $M$ ) $\vee$ (very small $S$ )
2	484	Mid-range $R \wedge$ mid-range $M \wedge$ large $S$
3	2488	Mid-range $R \wedge$ mid-range $M \wedge$ mid-range $S$
4	209	Very large $R \wedge$ very large $M \wedge$ large $S$
5	374	Very large $R \wedge$ very large $M \wedge$ mid-range $S$

$N_{\text{min}} = 20$ ) and equal to the number of particles in the main halo. The separation distance is the distance between the centres of the two distributions that we sample from 101 linearly spaced values from 0 to 20 – so as to provide a reasonable range of possible hierarchies.

From Table 2 we see the dependencies of the size of the HALO-OPTICS hierarchy,  $|H|$ , upon the input space defined above. The exact number of input space combinations that correspond to a particular value of  $|H|$  is not as important as the domains within which these particular values of  $|H|$  occur. The case where  $|H| = 0$  reflects the scenario where HALO-OPTICS is unable to gather any grouping of  $N_{\text{min}}$  points within a radius of  $\epsilon$ . As such, this occurs at a small resolution, large mass fraction, and large separation distance – which effectively spreads a small number of points among two distinctly separate and equally massive distributions. The case of  $|H| = 1$  typically occurs due to either, or a combination, of a small resolution and a small mass fraction. The fact that it is the most commonly occurring case is simply an artefact of having performed a logarithmic sampling of both the resolution and mass fraction variables – which has artificially created a sampling bias towards the smaller values of these variables. A single cluster may also be returned for very small separation distances. As such, these variable ranges force HALO-OPTICS to ignore the satellite in the system and only find the points from the main halo to be significantly clustered. This is a consequence of the satellite having too few points associated with it and/or the two density profiles being indistinctly separated from each other.

The case where  $|H| = 2$  is simply a partitioning of the main and the satellite halo distributions into two root clusters and hence occurs at large separation distances – provided that the resolution and mass fraction variables are large enough to create significant samplings of both the main and satellite halo distributions. The case where  $|H| = 3$  generally occurs for the mid-range values of the separation distance – again, provided that the resolution and mass fraction variable values are suitable. A hierarchy consisting of three clusters marks the case where the main and satellite halo distributions are connected via the means of a particle bridge, whilst still having the two density peaks remain distinctly separate. In this scenario, HALO-OPTICS finds both the sampled distributions to be leaf clusters of an overarching root cluster. In a real astrophysical data set, such a root cluster would have a particular overdensity that is related to the mapping from  $\epsilon$  to the overdensity factor  $\Delta$  in the way that is outlined in Section 3.1.

The mock system we investigate here contains two NFW profiles that can only be hierarchically connected via a single overarching



**Figure 8.** The reachability plot of the stellar particles from the MW02 synthetic halo where the colours indicate different clusters as determined by the extraction process outlined in Section 3.2. The top panel is the full reachability plot and has only the root clusters coloured. The middle panel is the section of the reachability plot that is marked with the grey dashed lines in the top panel. The clusters in this panel are coloured at the first child level below the root level. The bottom panel is the section of the reachability plot that is marked with the grey dashed lines in the middle panel. The clusters in this panel are coloured at the second child level below the root level. The black regions within each panel correspond to unclustered points at that level of the hierarchy. The bold letter labels in this bottom panel correspond to the reachability profile of those clusters that have been similarly marked in Fig. 9.

root cluster, and as such our explanations thus far have covered all hierarchies that should be expected from the density profile of this system. However, since we perform a random and discrete sampling of these distributions in order to construct the system, the hierarchies that are feasibly possible here extend out to larger sizes than this – although they are strongly probabilistic in nature. All occurrences of hierarchies in which more than three clusters are found are due to the presence of randomly occurring clusters (ROCs) and are an artefact of the combination of the largest values of both the resolution and mass fraction variables. Together these effectively increase the probability that at least  $N_{\min}$  points will form a randomly positioned and distinct overdensity within the system. When such a ROC is found, it is forced to a deeper level of the hierarchy as a leaf cluster, alongside another leaf cluster that conforms to the remainder of the sampled distribution that occupies the region that is denser than the saddle point of the density field – this saddle point is produced due to the presence of the ROC.

It is apparent that the general domain of these larger hierarchies becomes increasingly restricted towards the largest values of the resolution and mass fraction variables as  $|H|$  increases. The general decrease in the occurrences for larger hierarchies is due to the probability of a ROC being found within this input space. As such, the cases where  $|H| = 4$  and where  $|H| = 5$  are effectively extensions of the  $|H| = 2$  and  $|H| = 3$  cases into this domain, respectively, i.e. the same separation distance ranges with the larger resolution/mass fraction values. Even larger values of  $|H|$  are also possible within this input space, although the probability of their occurrence is much lower. It should be noted that in this mock system, we could decrease the likelihood of the ROCs being found by simply increasing the HALO-OPTICS hyperparameter,  $\rho_{\text{threshold}}$ . However, the value of  $\rho_{\text{threshold}} = 2$  has a particular significance when pertaining to the detection and extraction of streams. Furthermore, the legitimacy of ROCs changes and becomes somewhat ambiguous when a true astrophysical data set is concerned. Nevertheless, this result may hint at the benefit that HALO-OPTICS stands to gain from some extra hierarchical cleaning processes.

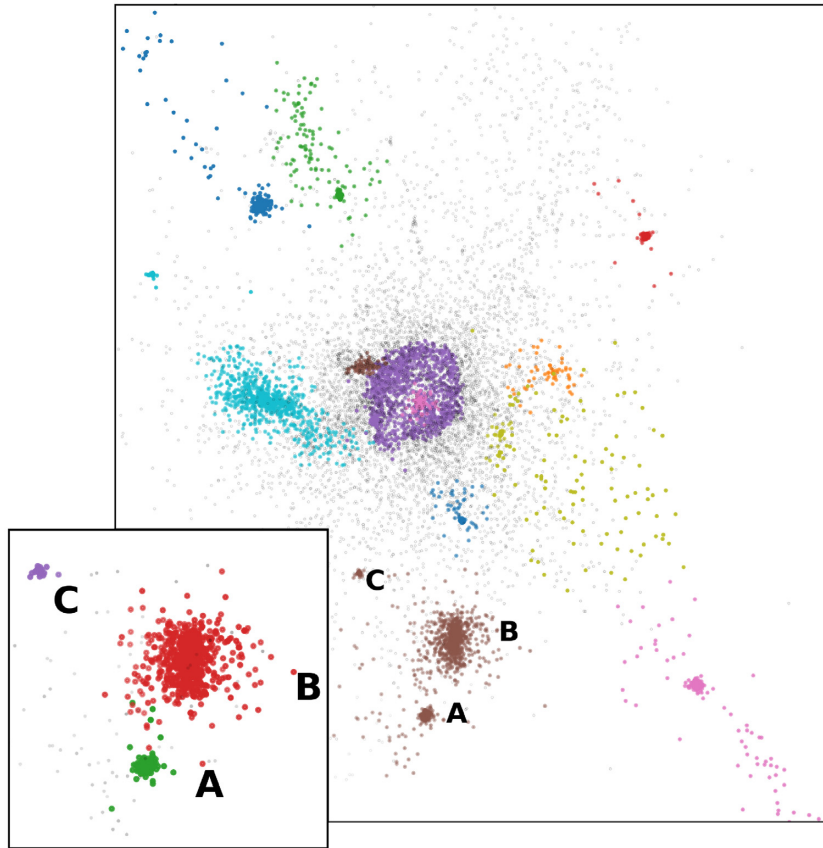
## 4 OUTPUT ANALYSIS

### 4.1 Visualizing the HALO-OPTICS output

The 3D positions and masses for the stellar and dark matter particles are taken from the MW02, MW03, MW04, and MW06 synthetic haloes simulated by Power & Robotham (2016). We use HALO-OPTICS on the stellar and dark matter particle types within each halo both separately and combined. HALO-OPTICS gives us a reachability plot (detailed in Section 2.2) and a list of significant clusters extracted from that (detailed in Section 3.2). The reachability plot for the stellar particles from the MW02 halo is shown in Fig. 8. The three panels display various ranges of the ordered index of particles, and the colours within each panel are cyclic between consecutive significant clusters of the root, first child, and second child levels from top to bottom, respectively. Fig. 9 contains a main panel and an inset one that depict the positions of the particles contained within the middle panel and bottom panels of Fig. 8, respectively. The colour scheme and bold letter labels within these two panels correspond to those used in the middle and bottom panels of Fig. 8, respectively.

It can be seen in Fig. 8 that even though the reachability plot is not smooth, our extraction process retrieves the more significant clusters while ignoring smaller undulations and noise. The reachability plot from each of the stellar and dark matter particle runs for each of the synthetic haloes always contain a very large valley relating to the denser region that surrounds the MW-type galaxy. Within this root level valley there is another very large valley – as well as many other smaller ones. This *other* very large valley is the inner halo (the right edge of which can be seen on the left of the middle panel in Fig. 8), which is the most massive leaf cluster present within each synthetic halo. The reachability distance of the inner halo is not only small, indicating a high density, but also quite smooth. The central region of Fig. 9 features the MW02 inner stellar halo’s least dense particles (i.e. those that possess the largest reachability distances). Towards the most dense parts of this inner halo, there are typically many small sharp peaks in reachability (not visible at the scale of





**Figure 9.** The main and inset panels are the positions – with the corresponding colour schemes – of all points in the middle and lower panels of Fig. 8, respectively. Here we showcase that HALO-OPTICS retrieves the relevant clusters and the appropriate hierarchy that they are contained within.

the top panel in Fig. 8), however these are artefacts of the OPTICS process and occur due to the reachability distance not being updated for points that have already been appended to the ordered list. Our extraction process successfully ignores these artefacts.

#### 4.2 A comparison with VELOCIRAPTOR

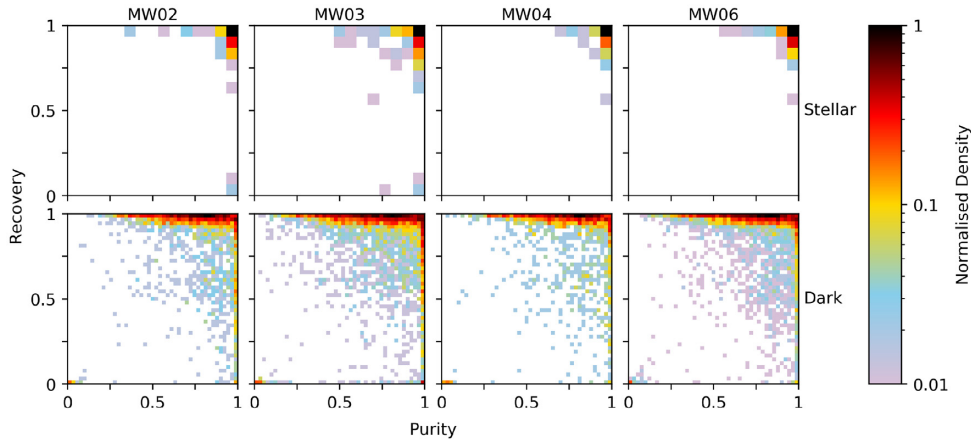
We now wish to inform the reader as to what the structures are that HALO-OPTICS has found. We do this via a comparison with the state-of-the-art galaxy/(sub)halo finder VELOCIRAPTOR. We apply VELOCIRAPTOR to both the stellar and dark matter particle types within each MW-type galaxy’s snapshot file. VELOCIRAPTOR first searches for field haloes from particle positions using a 3D FOF algorithm. Then, each field halo is searched for phase-space overdensities using an adaptive 6D FOF algorithm where the position and velocity linking lengths are based on the average spatial and kinematic dispersions of the parent cluster. These linking lengths are iteratively decreased in order to identify local maxima in phase-space density (cores) until no local maxima with enough particles are found. Particles of the root cluster are then iteratively assigned to their nearest core in phase space, according to the core’s phase-space dispersion tensor. A core’s phase-space dispersion tensor is updated as new particles are assigned. This process is similar to a Gaussian mixture model but where the number of distributions is fixed to the number of significant phase-space cores found.

To appropriately compare the two codes, we find the best-fitting match from the VELOCIRAPTOR catalogue of clusters for each HALO-OPTICS cluster, which we do by means of the maximum

Jaccard index, described in equation (7). Then for each HALO-OPTICS–VELOCIRAPTOR best-fitting pair we compute the recovery and purity fractions. Fig. 10 depicts the recovery versus purity fractions for all HALO-OPTICS clusters as compared to their best-fitting VELOCIRAPTOR cluster for both stellar and dark matter particle types within each MW-type galaxy’s snapshot file.

We see from these that for a large majority of HALO-OPTICS clusters there is a good match present within the corresponding VELOCIRAPTOR catalogue. Of the HALO-OPTICS clusters that have not been well matched by VELOCIRAPTOR, there is a small portion that have high purity and low recovery. We note that of these, many sit deep within the HALO-OPTICS hierarchy and have a best-fitting VELOCIRAPTOR cluster that sits comparatively towards the root clusters of the VELOCIRAPTOR hierarchy. The particles within these clusters are likely to be spatially clustered and not kinematically clustered, hence VELOCIRAPTOR does not find them to be significantly clustered at a deeper level of its hierarchy. In this scenario it is probable that, given the same particle information as VELOCIRAPTOR, HALO-OPTICS would too find these clusters to be insignificant.

It is particularly striking to see that HALO-OPTICS does quite well at retrieving a large majority of VELOCIRAPTOR clusters – an impressive result considering VELOCIRAPTOR’s use of particle kinematics in order to find many of these. These clusters must still have a significant spatial density contrast with respect to their background to be detected by HALO-OPTICS but some fine-tuning would be needed in order to retrieve these clusters with just a 3D FOF algorithm. This hints at the greater detection and extraction power of the comparatively adaptive HALO-OPTICS algorithm over that of



**Figure 10.** The recovery versus purity relations for all HALO-OPTICS clusters and their best-fitting (by means of the maximum Jaccard index) VELOCIRAPTOR clusters. Each column displays the relations for a different MW galaxy analogue and the rows correspond to the two particle types – stars and dark matter. The colour within each panel reflects the normalized density (such that the maximum is 1) of best-fitting cluster pairs within the cells of a 15 by 15 histogram for the stellar clusters and within a 45 by 45 histogram for the dark matter clusters. Within each panel the upper right, lower right, upper left, and lower left corners indicate the regions that a HALO-OPTICS cluster will be placed if it is well matched by contained within and comparatively smaller than, containing and is comparatively larger than, and mostly (or completely) unrecovered by its best-fitting VELOCIRAPTOR cluster. It is shown that a large portion of the clusters from HALO-OPTICS are well matched by the VELOCIRAPTOR catalogue with high recovery and purity. We also see here that the HALO-OPTICS dark matter clusters do have a more variable purity than that of their stellar counterparts when comparing to the VELOCIRAPTOR output, however these clusters do mostly have high recovery and high Jaccard indices.

a static FOF algorithm. VELOCIRAPTOR overcomes this problem by searching through a data-driven position–velocity phase space to further separate these clusters from their background – which does not require as much fine-tuning to achieve. Since HALO-OPTICS does not require this kind of fine-tuning, these results bode well for the performance of HALO-OPTICS in the event that it is applied using a more informative metric – inclusive of particle kinematics and metallicities for example.

Fig. 11 illustrates the 2D projections of a selected few clusters produced by HALO-OPTICS and their best-fitting VELOCIRAPTOR counterparts. The panels therein indicate the particles attributed to each cluster by only HALO-OPTICS (in blue), only VELOCIRAPTOR (in orange), and by both codes (in green). Various information about the cluster representations from the codes are annotated within each panel. We see that HALO-OPTICS provides a strong match to the predictions made from VELOCIRAPTOR with high recovery, purity, and Jaccard index between each of the representations.

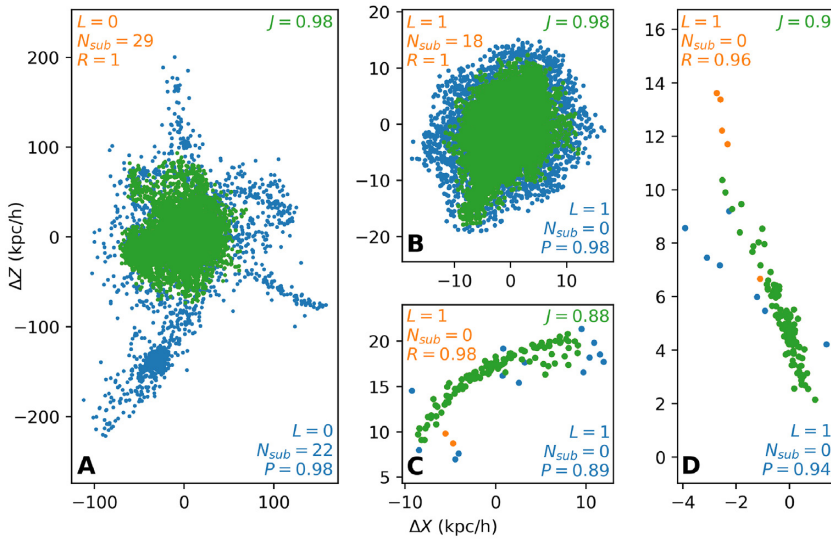
In panel (A), HALO-OPTICS has retrieved extended stellar components – that are likely of kinematic interest – associated with the galaxy’s surroundings that VELOCIRAPTOR has not attributed to the galaxy. The inner halo from HALO-OPTICS depicted in panel (B) also overextends that from VELOCIRAPTOR. Both of these overextensions, particularly the former, are largely resultant from the differences between the OPTICS and FOF algorithms, i.e. OPTICS is not affected by noisy interparticle spacing as it detects density fluctuations at the resolution of  $N_{\min}$  points. The overextension seen in panel (B) is also due to the differences between the featured structure’s spatial and phase-space densities – a disparity that can likely be mitigated with the inclusion of particle kinematics into the HALO-OPTICS distance metric.

Panels (C) and (D) show that HALO-OPTICS does remarkably well in retrieving and matching these streams without the knowledge of particle kinematics. Given this knowledge, HALO-OPTICS could provide better quality matches to these streams than it does in the application we present here and potentially find some associated particles that VELOCIRAPTOR does not.

We note that HALO-OPTICS does not find any of the substructures contained within the inner halo of the MW02 galaxy.<sup>7</sup> However, it should also be expected that HALO-OPTICS would do better in resolving these substructures with the knowledge of kinematics. Not indicated in Fig. 11 is that of the 22 substructures found within the HALO-OPTICS root cluster depicted in panel (A), 12 are contained exclusively within the best-fitting cluster found by VELOCIRAPTOR – which contains 29. By accounting for the known substructures in panels (B), (C), and (D) (and subsubstructures therein), we can deduce that there are precisely eight not visualized substructures from VELOCIRAPTOR within this region, and eight from HALO-OPTICS. These substructures are mostly the same between the codes, however there is disagreement between the codes, namely the grouping shown in the lower panel of Fig. 8 and the inset panel of Fig. 9. This grouping implies that there must be at least one other cluster found by HALO-OPTICS that is in dispute with those found by VELOCIRAPTOR. Such clusters are likely to be clustered spatially but not kinematically, and given the same phase-space information as VELOCIRAPTOR, we expect that HALO-OPTICS will find these clusters to be insignificant.

As mentioned in Section 2.2, a major drawback to OPTICS – and by extension HALO-OPTICS – is that it is computationally demanding. For example, to complete a clustering run over the MW02 galaxy’s stellar particles (209 834 particles), HALO-OPTICS takes  $\sim 10$  min to create the reachability plot and then  $\sim 7$  s to extract clusters from that. For HALO-OPTICS to complete a clustering run over the dark matter particles (2441 561 particles) within the MW02 galaxy, the runtime is  $\sim 5$  h to create the reachability plot and then  $\sim 1$  min to extract the clusters therein. In comparison, running VELOCIRAPTOR over the MW02 galaxy’s stellar particles only takes  $\sim 4.4$  s to search for substructure and  $\sim 25$  s to get the 6D FOF haloes. For the dark matter particles, VELOCIRAPTOR takes  $\sim 16$  s to search for substructure and  $\sim 37$  s to get the 6D FOF haloes.

<sup>7</sup>This is by definition since we choose the inner halo to be the largest leaf cluster in the hierarchy.



**Figure 11.** A series of 2D projections of a select few stellar clusters from the MW02 galaxy that have been found by both HALO-OPTICS and VELOCIRAPTOR. Particles colour: blue belong exclusively to the clusters as predicted by HALO-OPTICS; orange belong exclusively to the clusters as predicted by VELOCIRAPTOR; green belong to the intersection of the clusters from both codes. Panels (A)–(D) depict the HALO-OPTICS cluster and its best-fitting VELOCIRAPTOR candidate for the root stellar cluster that surrounds the MW02 galaxy barycentre (large pink valley in top panel of Fig. 8), the inner stellar halo of MW02 (partially shown in purple at the leftmost edge of the middle panel of Fig. 8), and two streams nearby the inner stellar halo (not explicitly shown in colour within Fig. 8 as they reside towards the left-hand edge of the inner stellar halo’s ordered list). The coordinate system of each panel is the same and is centred on the inner halo’s barycentre. Annotated in orange in the upper left-hand corner of each panel is the hierarchy level, number of substructures, and recovery of the cluster as predicted by VELOCIRAPTOR. Similarly, annotated in blue in the lower right-hand corner of each panel is the hierarchy level, number of substructures, and purity of the cluster as predicted by HALO-OPTICS. Annotated in green in the upper right-hand corner is the (maximum) Jaccard index of the two cluster representations.

These runtime discrepancies are partially due to our naive implementation of HALO-OPTICS being run with PYTHON-3 through a single core on an Intel Xeon E5-2698 v4 processor, whereas VELOCIRAPTOR is a ready compiled program written in C++11 with  $\geq -O2$  optimization using GCC that in this instance used a single core and single MPI on an Intel i7 vPro processor. However, the largest runtime setback for HALO-OPTICS comes from the fact that not only does its nearest neighbour radial search need to be much larger than the corresponding FOF nearest neighbour radial search, but HALO-OPTICS also needs to return the exact distances of each neighbour during this search, whereas VELOCIRAPTOR does not. It should also be noted that as our implementation currently exists, HALO-OPTICS has no parallelization capabilities and only uses optimized vectorized functions, whereas VELOCIRAPTOR has massively parallel capabilities. Using VELOCIRAPTOR in this way can dramatically reduce the overall runtimes. For example, by allowing VELOCIRAPTOR to use eight threads on the same dark matter particles as above, the substructure search time reduces to  $\sim 13$  s and the 6D FOF search reduces to  $\sim 26$  s.

### 4.3 Inside the high-resolution zone

Each synthetic halo has a virial mass ( $M_{200}$ ) of approximately  $2 \times 10^{12} M_{\odot} h^{-1}$  (details of each halo may be found in table 1 of Power & Robotham 2016). The proportion of  $M_{200}$  made up of stellar and dark matter particles also remains similar between haloes, however the total number of each type of particle within each galaxy snapshot does vary. The stellar and dark matter particles extend much further than  $R_{200}$  and as such the size and complexity of the reachability plot varies as well. To reduce this variability between haloes we now take only those clusters whose barycentres are contained within  $5R_{200} \approx 1 \text{ Mpc } h^{-1}$  from the barycentre of the inner halo of each galaxy. This radial cut is chosen as it represents the approximate boundary of

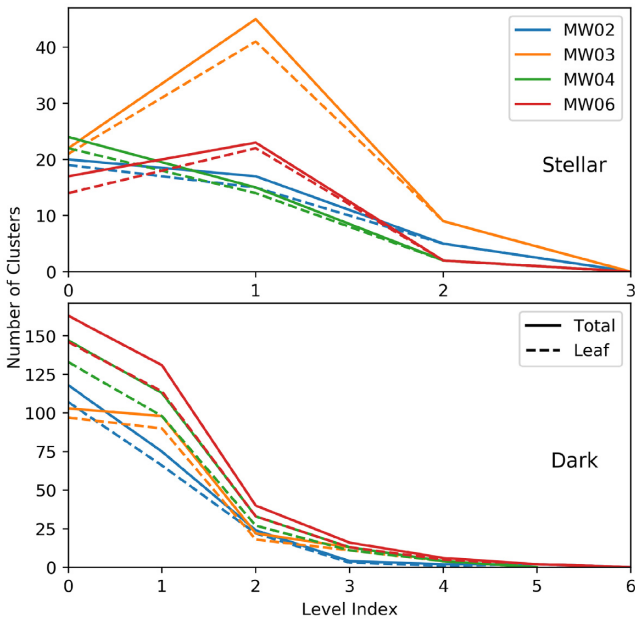
the high-resolution regions within each of the cosmological zoom simulations that contain each MW-type galaxy. Fig. 12 demonstrates the cluster hierarchy within this region of each of the galaxies for both the stellar and dark matter particles.

From Fig. 12 we see that within  $5R_{200}$  each galaxy’s hierarchy of clusters is similar. The number of clusters defined at the root level (level 0) is solely dependent on the OPTICS parameter  $\epsilon$ . By increasing  $\epsilon$ , the hierarchy will deepen overall and narrow at the zeroth level until eventually the zeroth level will only contain a single *cluster* – the entire data set. To some extent the shape of the hierarchy at each particular level will be influenced by the choice of  $\epsilon$ , although since we use a rigorous definition for  $\epsilon$  in our application of HALO-OPTICS the hierarchy shape between galaxies is meaningful.

Despite being born from the same simulation regime, the distinct hierarchy of stellar substructure within MW03 is seen in Fig. 12, where the peak is noticeably larger in magnitude than the other galaxies. The stellar component of MW03 appears to exhibit the most galaxy–galaxy variation as seen in fig. 8 of Power & Robotham (2016). Compared to the other galaxies the density of MW03 is large for small radii, drops for radii  $\sim 0.1R_{200}$ , and also features considerable spikes in density for radii approaching  $\sim R_{200}$ . The significant differences in stellar density within this region are certainly the reason for the large number of clusters at the first level of MW03’s stellar hierarchy as these clusters will have a lower background density and therefore be in contrast with it more so than for the other MW-type galaxies.

## 5 DISCUSSION

We have demonstrated that the HALO-OPTICS algorithm is a powerful tool to be used for the global identification of all meaningful clusters of a data set containing at least  $N_{\min}$  data points and the hierarchy within which they are embedded. When applied to



**Figure 12.** Stellar (top) and dark matter (bottom) cluster hierarchies for each galaxy. Levels 0–6 indicate the root to leaf layers of the hierarchy, respectively. The solid and dashed lines illustrate the total number of clusters and the number of leaf clusters at each level, respectively. The total number of leaf stellar clusters is 39, 71, 38, and 38 for the MW02, MW03, MW04, and MW06, respectively. Similarly, the total number of leaf dark matter clusters is 201, 220, 273, and 313 for the MW02, MW03, MW04, and MW06, respectively.

the physical clustering of particles, the ambiguity of a meaningful metric disappears and the output becomes particularly robust when compared to other algorithms that operate under the same metric. We applied HALO-OPTICS to the 3D positions of stellar and dark matter particles from four MW-type galaxies produced through a set of cosmological zoom simulations. We used HALO-OPTICS to detect and extract the significant clusters from these galaxies. We compared the output with VELOCIRAPTOR before then analysing the hierarchy of clusters that are situated within a radius of  $5R_{200}$  from their corresponding system’s galactic centre.

Through our comparison with VELOCIRAPTOR in Section 4.2, we have demonstrated that HALO-OPTICS retrieves the more significant clusters while electing to ignore those clusters whose density does not appreciably differ from their surroundings. Furthermore, this comparison indicates the power of adaptive hierarchical clustering algorithms such as HALO-OPTICS as it is able to uncover many clusters from only the 3D particle positions that VELOCIRAPTOR had identified by using both particle positions and kinematics.<sup>8</sup> For HALO-OPTICS to achieve this, these clusters must still have a significant spatial density contrast with their background, however for a 3D FOF algorithm to do the same, some level of fine-tuning would be needed.

The depth and shape of the hierarchy are influenced by the HALO-OPTICS hyperparameters. Those from the original OPTICS algorithm,

<sup>8</sup>This is not to say that VELOCIRAPTOR is not adaptive – it is – VELOCIRAPTOR iteratively uses a 6D FOF algorithm by locally adapting its phase-space metric to further separate clusters from their surroundings. This typically means that as the algorithm searches for clusters deeper within the hierarchy the phase space becomes more heavily weighted towards particle kinematics rather than particle positions.

$\epsilon$  and  $N_{\min}$ , are respectively responsible for the extents of lowest density and smallest size the clusters can be. Changing the HALO-OPTICS input parameter  $\Delta$  has similar affect as its less physical OPTICS counterpart,  $\epsilon$ . The additional extraction parameters exclusive to HALO-OPTICS –  $\rho_{\text{threshold}}$ ,  $f_{\text{reject}}$ , and  $S_{\text{outlier}}$  – are responsible for the number of divisions in between the root and leaf levels and which particles belong to each level and each cluster. However, the largest contributor to hierarchy is of course the physics of the interactions between the particles themselves. Being cold, the dark matter easily clumps together to form deep hierarchies by  $z = 0$ . Between baryonic feedback effects and the relative subdominance of stellar particles within the region defined by a radius of  $R_{200}$  as a whole (refer to table 1 of Power & Robotham 2016), the resultant stellar hierarchies are shallower than their CDM counterparts. However due to stellar particles being kinematically cold, we should expect the stellar hierarchy to deepen with the inclusion particle kinematics.

It is likely that the inclusion of extra localized information – i.e. velocities, chemical abundances, etc. – into the metric will have the largest impact on cluster yields in the inner regions of galaxies where large numbers of particles are very spatially dense and neighbouring local spatial densities are indistinct. This metric augmentation could conceivably be implemented as a non-linear combination of spatial, kinematic, and metallicity variables that are each weighted by a factor relative to their local variation within the data. Alternatively, HALO-OPTICS could perform its nearest neighbour searching over the spatial dimensions – preserving the maximum spatial scales defined by the overdensity factor  $\Delta$  – and then order points by a distance metric containing information about spatial, kinematic, and chemical variables – although this may not be necessary due to the adaptive nature of OPTICS.

Modifying the metric in this way will provide the means for determining clusters more distinctly from their background so long as the metric only includes good indicators of clustered data. The reachability plot will in general change shape for any given cluster, though not so significantly that we should not expect our cluster extraction method to still recover all relevant clusters. However, the optimal HALO-OPTICS hyperparameters may be different in a higher dimensional metric from those used in conjunction with a 3D spatial one.

The root levels of the hierarchy of clusters in a particular astrophysical data set will likely stay consistent across various metrics. Although, the hierarchy may deepen with the addition of extra clustering indicators since our cluster extraction process will be able to retrieve additional low spatial density and kinematically/metallicity coherent substructures at the leaf levels. Likewise, we may reasonably expect that changing the metric in these ways will not adversely affect the more massive substructures – nor will it resolve any new ones – and that the effect of an improved metric will predominantly modify the proportion of the less massive substructures compared to those that are larger.

## 6 CONCLUSIONS

We have shown HALO-OPTICS to be a robust cluster finder that is effective in determining a wide variety of cluster types, shapes, and sizes, even with a spatial distance metric as its only handle on localized information. Furthermore, we are satisfied that our extraction process is capable of determining these clusters without the need for supervised learning nor the restrictions of the more conventional extraction techniques. The ability for the HALO-OPTICS algorithm to retrieve the hierarchy of galaxies in this relatively fast and secure manner should pave the way for HALO-OPTICS to

be used complementary to a more traditional structure finder such as VELOCIRAPTOR, and as a simple and practical halo finder in astrophysics and its related fields. In a future work, we will extend HALO-OPTICS to use a multidimensional metric that is inclusive of extra localized information, such as particle kinematics and stellar metallicity. We also intend to build upon our extraction technique so that it incorporates more physical aspects of clusters such as particle boundedness. Among these changes, we leave the further optimization and potential parallelization of HALO-OPTICS for future work as well. These concepts, particularly the latter, present significant challenges due to the strongly sequential data access order that OPTICS makes use of.

## ACKNOWLEDGEMENTS

WHO gratefully acknowledges financial support through the Hunstead Student Support Scholarship from the Dick Hunstead Fund in the University of Sydney's School of Physics. This research benefitted from computation resources in the form of the Argus Virtual Research Desktop environment that was provided through the University of Sydney's Information and Communication Technologies and supported by the Sydney Informatics Hub.

## DATA AVAILABILITY

The data underlying this paper may be made available on reasonable request to the corresponding author.

## REFERENCES

- Angulo R. E., Lacey C. G., Baugh C. M., Frenk C. S., 2009, *MNRAS*, 399, 983
- Ankerst M., Breunig M. M., Kriegel H.-P., Sander J., 1999, in Delis A., Faloutsos C., Ghandeharizadeh S., eds, SIGMOD/PODS 1999: International Conference on Management of Data and Symposium on Principles of Database Systems. Association for Computing Machinery (ACM), New York, p. 49
- Behroozi P. S., Wechsler R. H., Wu H.-Y., 2012, *ApJ*, 762, 109
- Benson A. J., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2002, *MNRAS*, 333, 177
- Bentley J. L., 1975, *Commun. ACM*, 18, 509
- Blanton M. R. et al., 2017, *AJ*, 154, 28
- Bode P., Ostriker J. P., Turok N., 2001, *ApJ*, 556, 93
- Breunig M. M., Kriegel H.-P., Ng R. T., Sander J., 1999, in Żytkow J. M., Rauch J., eds, Principles of Data Mining and Knowledge Discovery. Springer-Verlag, Berlin, p. 262
- Brooks A. M., Zolotov A., 2014, *ApJ*, 786, 87
- Bullock J. S., Kravtsov A. V., Weinberg D. H., 2000, *ApJ*, 539, 517
- Bullock J. S., Stewart K. R., Kaplinghat M., Tollerud E. J., Wolf J., 2010, *ApJ*, 717, 1043
- Canovas H. et al., 2019, *A&A*, 626, A80
- Colin P., Avila-Reese V., Valenzuela O., 2000, *ApJ*, 542, 622
- Davis M., Efstathiou G., Frenk C. S., White S. D., 1985, *ApJ*, 292, 371
- Diemand J., Kuhlen M., Madau P., 2007, *ApJ*, 667, 859
- Dutton A. A., Maccio A. V., Frings J., Wang L., Stinson G. S., Penzo C., Kang X., 2016, *MNRAS*, 457, L74
- Elahi P. J., Thacker R. J., Widrow L. M., 2011, *MNRAS*, 418, 320
- Elahi P. J., Welker C., Power C., Lagos C. d. P., Robotham A. S. G., Cañas R., Poulton R., 2018, *MNRAS*, 475, 5338
- Elahi P. J., Canas R., Poulton R. J. J., Tobar R. J., Willis J. S., Lagos C. d. P., Power C., Robotham A. S. G., 2019, *Publ. Astron. Soc. Aust.*, 36, e021
- Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Simoudis E., Han J., Fayyad U., eds, KDD 1996: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, Palo Alto, CA, p. 226
- Fielder C. E., Mao Y.-Y., Newman J. A., Zentner A. R., Licquia T. C., 2019, *MNRAS*, 486, 4545
- Fuentes S. S., De Ridder J., Debosscher J., 2017, *A&A*, 599, A143
- Fukunaga K., Hostetler L., 1975, *IEEE Trans. Inf. Theory*, 21, 32
- Gaia Collaboration et al., 2018, *A&A*, 616, A1
- Gao L., White S. D. M., Jenkins A., Stoehr F., Springel V., 2004, *MNRAS*, 355, 819
- Garrison-Kimmel S., Boylan-Kolchin M., Bullock J. S., Lee K., 2014, *MNRAS*, 438, 2578
- Ghigna S., Moore B., Governato F., Lake G., Quinn T., Stadel J., 1998, *MNRAS*, 300, 146
- Hoffman Y., Metuki O., Yepes G., Gottlober S., Forero-Romero J. E., Libeskind N. I., Knebe A., 2012, *MNRAS*, 425, 2049
- Homma D. et al., 2019, *PASJ*, 71, 94
- Ishiyama T. et al., 2013, *ApJ*, 767, 146
- Kamionkowski M., Liddle A. R., 2000, *Phys. Rev. Lett.*, 84, 4525
- Kauffmann G., White S. D. M., Guiderdoni B., 1993, *MNRAS*, 264, 201
- Kim S. Y., Peter A. H. G., Hargis J. R., 2018, *Phys. Rev. Lett.*, 121, 211302
- Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, *ApJ*, 522, 82
- Knebe A. et al., 2011, *MNRAS*, 415, 2293
- Knollmann S. R., Knebe A., 2009, *ApJS*, 182, 608
- Koposov S. et al., 2008, *ApJ*, 686, 279
- Koposov S. E. et al., 2018, *MNRAS*, 479, 5343
- Lloyd S., 1982, *IEEE Trans. Inf. Theory*, 28, 129
- McConnachie A. W. et al., 2009, *Nature*, 461, 66
- McConnachie A. W. et al., 2018, *ApJ*, 868, 55
- Marquant J. F., Evins R., Bollinger L. A., Carmeliet J., 2017, *Appl. Energy*, 208, 935
- Martinez G. D., Minor Q. E., Bullock J., Kaplinghat M., Simon J. D., Geha M., 2011, *ApJ*, 738, 55
- Massaro F., Alvarez-Crespo N., Capetti A., Baldi R., Pillitteri I., Campana R., Paggi A., 2019, *ApJS*, 240, 20
- Mau S. et al., 2019, *ApJ*, 875, 154
- Monaghan J. J., 1992, *ARA&A*, 30, 543
- Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, *ApJ*, 524, L19
- Moore B., Diemand J., Madau P., Zemp M., Stadel J., 2006, *MNRAS*, 368, 563
- More S., Kravtsov A. V., Dalal N., Gottlöber S., 2011, *ApJS*, 195, 4
- Navarro J. F., Frenk C. S., White S. D. M., 1996, *ApJ*, 462, 563
- Newton O., Cautun M., Jenkins A., Frenk C. S., Helly J. C., 2018, *MNRAS*, 479, 2853
- Okabe A., 2016, in Richardson D. et al., eds, International Encyclopedia of Geography: People, the Earth, Environment and Technology. Wiley-Blackwell, New York, p. 1
- Patwary M. A., Palsetia D., Agrawal A., Liao W.-k., Manne F., Choudhary A., 2013, SC 2013: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. Association for Computing Machinery (ACM), New York, p. 1
- Pedregosa F. et al., 2011, *J. Machine Learning Res.*, 12, 2825
- Power C., Robotham A. S., 2016, *ApJ*, 825, 31
- Power C., Navarro J. F., Jenkins A., Frenk C. S., White S. D. M., Springel V., Stadel J., Quinn T., 2003, *MNRAS*, 338, 14
- Press W. H., Schechter P., 1974, *ApJ*, 187, 425
- Reed D., Governato F., Quinn T., Gardner J., Stadel J., Lake G., 2005, *MNRAS*, 359, 1537
- Ricotti M., Gnedin N. Y., 2005, *ApJ*, 629, 259
- Rodriguez-Puebla A., Behroozi P., Primack J., Klypin A., Lee C., Hellinger D., 2016, *MNRAS*, 462, 893
- Sander J., Qin X., Lu Z., Niu N., Kovarsky A., 2003, in Whang K.-Y., Jeon J., Shim K., Srivastava J., eds, Advances in Knowledge Discovery and Data Mining. Springer-Verlag, Berlin, p. 75
- Sawala T. et al., 2015, *MNRAS*, 456, 85
- Sawala T. et al., 2016, *MNRAS*, 457, 1931
- Sesar B. et al., 2014, *ApJ*, 793, 135
- Somerville R. S., 2002, *ApJ*, 572, L23
- Springel V., 2005, *MNRAS*, 364, 1105

- Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, *MNRAS*, 328, 726
- Springel V. et al., 2008, *MNRAS*, 391, 1685
- Starkenburg E. et al., 2017, *MNRAS*, 471, 2587
- Strigari L. E., Kaplinghat M., Bullock J. S., 2007, *Phys. Rev. D*, 75, 061303
- Tollerud E. J., Bullock J. S., Strigari L. E., Willman B., 2008, *ApJ*, 688, 277
- Tollerud E. J., Boylan-Kolchin M., Bullock J. S., 2014, *MNRAS*, 440, 3511
- Torrealba G. et al., 2019, *MNRAS*, 488, 2743
- Turk M. J., Smith B. D., Oishi J. S., Skory S., Skillman S. W., Abel T., Norman M. L., 2011, *ApJS*, 192, 9
- van de Weygaert R., 1994, *A&A*, 283, 361
- Voronoi G., 1908, *J. Reine Angewandte Math.*, 134, 198
- Walsh S. M., Willman B., Jerjen H., 2008, *AJ*, 137, 450
- Wang S. et al., 2014, *Plant Biotechnol. J.*, 12, 787
- Wetzel A. R., Hopkins P. F., Kim J.-h., Faucher-Giguere C.-A., Keres D., Quataert E., 2016, *ApJ*, 827, L23
- White S. D. M., Rees M. J., 1978, *MNRAS*, 183, 341
- Wolf J., Martinez G. D., Bullock J. S., Kaplinghat M., Geha M., Munoz R. R., Simon J. D., Avedo F. F., 2010, *MNRAS*, 406, 1220
- Xie L., Gao L., 2015, *MNRAS*, 454, 1697
- Zentner A. R., Bullock J. S., 2003, *ApJ*, 598, 49
- Zhang A. X., Noulas A., Scellato S., Mascolo C., 2013, *SocialCom 2013: Proceedings of the 2013 International Conference on Social Computing*. IEEE Computer Society, Washington, DC, p. 69
- Zhao G., Zhao Y.-H., Chu Y.-Q., Jing Y.-P., Deng L.-C., 2012, *Res. Astron. Astrophys.*, 12, 723
- Zheng Y., Li Q., Chen Y., Xie X., Ma W.-Y., 2008, in Youn H. Y., Cho W.-D., eds, *UbiComp 2008: Proceedings of the 10th International Conference on Ubiquitous Computing*. Association for Computing Machinery (ACM), New York, p. 312
- Zhu Q., Marinacci F., Maji M., Li Y., Springel V., Hernquist L., 2016, *MNRAS*, 458, 1559

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.

## Chapter 4

# A More Suitable Algorithm for Big Data

Following a clear demonstration by **HALO-OPTICS** of the power of using an adaptive-density clustering structure finder such as **OPTICS** in an astrophysical context, it has become obvious that the concept should be extended for application to any astrophysical data set. In order to be readily applicable to any astrophysical data set, however, the algorithm would need to facilitate structure finding with an arbitrary number of data points and be defined with an arbitrary number of features. The former of these two issues presents a unique problem for extending the algorithm, **OPTICS** and by extension **HALO-OPTICS**, both perform a radial nearest neighbour search about every point in the data set which as the data set grows can result in a very large run-time – even more so for data sets with a large number of features. The latter issue requires the extended algorithm to be capable of dealing with combinations of features with differing units and, preferably, in such a way that maximises the information available for producing relevant astrophysical clustering structure.

To overcome these downfalls I create **CLUSTAR-ND**. **CLUSTAR-ND** reduces the time-costly radial nearest neighbour search to a much faster  $k$  nearest neighbour search and then orders points in a similarly to **OPTICS** to produce a similar output of clusters. Since the ordering process of **CLUSTAR-ND** can be decoupled from the density estimation (unlike in **OPTICS** and **HALO-OPTICS**), I also have it compute the local density at each data point using the more robust Epanechnikov kernel [386] and balloon estimator [387]. So that it can operate on input data with any number of features, I give the user the option of using one of 3 metric adaptivity settings – corresponding to no transformation, a global PCA transformation, and an iterative PCA transformation. Using either of the PCA transformation settings allows for

CLUSTAR-ND to find structures from data whose features have differing units.

In the research presented in Sec. 4.1 I perform a comparison between CLUSTAR-ND and HALO-OPTICS finding excellent agreement between the two when operating on the same data and feature space. In this comparison, I also show that the run-time of CLUSTAR-ND is at least 3 orders of magnitude faster than that of HALO-OPTICS and has a modest time-complexity of  $\mathcal{O}(n \log n)$  instead of the inevitable  $\mathcal{O}(n^2)$  that arises in HALO-OPTICS from the large search radius needed to define galactic haloes. I then re-optimize the CLUSTAR-ND hyperparameters and find characteristic functions of their optimal parameters that allow for them to be automatically and optimally chosen given the input data. I find that the clustering power of the optimized CLUSTAR-ND algorithm increases with an increasing feature space and that it is able to retrieve a large portion of tidal debris from galactic haloes. The code for the CLUSTAR-ND algorithm can be found in App. B.2.

As with HALO-OPTICS, CLUSTAR-ND is designed to be a generalised astrophysical structure finder and, while these algorithms are aimed at revealing the hierarchical structure of galactic haloes, it is also the case that CLUSTAR-ND can be applied to other astrophysical data sets. Unlike HALO-OPTICS however, CLUSTAR-ND may be applied to much larger and more complex data sets with arbitrarily defined feature spaces. This extends the generalised nature of CLUSTAR-ND to be capable of identifying density-based clusters within chemo-dynamical data sets of objects or even (subject to parameter adjustments) time-domain clustering – so long as the definition of the predicted clusters remains appropriate.

## 4.1 Structure Finding with CluSTAR-ND

This section presents the published journal article:

2. *The Hierarchical Structure of Galactic Haloes: Generalised N-Dimensional Clustering with CluSTAR-ND*. **W. H. Oliver**, P. J. Elahi, & G. F. Lewis. *MNRAS* 514, 5767, 2022. [[arXiv:2201.10694](https://arxiv.org/abs/2201.10694)].

*Author Contributions:* I developed and trained the CLUSTAR-ND algorithm, produced the clustering outputs, drew comparisons between the outputs of CLUSTAR-ND and its predecessor HALO-OPTICS, and wrote the manuscript. Dr. Pascal J. Elahi made valuable contributions to the concept of the algorithm, its training, and the interpretation of the final results. The project was conducted under the supervision of Prof. Geraint F. Lewis, who also recommended using the GALAXIA code [388] to produce synthetic MW data. All authors reviewed and commented on the paper.





# The hierarchical structure of galactic haloes: generalized $N$ -dimensional clustering with CLUSTAR-ND

William H. Oliver,<sup>1</sup>★ Pascal J. Elahi<sup>2</sup> and Geraint F. Lewis<sup>1</sup>

<sup>1</sup>*Sydney Institute for Astronomy, School of Physics A28, The University of Sydney, NSW 2006, Australia*

<sup>2</sup>*Pawsey Supercomputing Research Centre, 1 Bryce Avenue, Kensington, WA 6151, Australia*

Accepted 2022 June 15. Received 2022 June 6; in original form 2022 January 25

## ABSTRACT

We present CLUSTAR-ND, a fast hierarchical galaxy/(sub)halo finder that produces **Clustering Structure via Transformative Aggregation and Rejection in N-Dimensions**. It is designed to improve upon HALO-OPTICS – an algorithm that automatically detects and extracts significant astrophysical clusters from the 3D spatial positions of simulation particles – by decreasing run-times, possessing the capability for metric adaptivity, and being readily applicable to data with any number of features. We directly compare these algorithms and find that not only does CLUSTAR-ND produce a similarly robust clustering structure, it does so in a run-time that is at least 3 orders of magnitude faster. In optimizing CLUSTAR-ND’s clustering performance, we have also carefully calibrated 4 of the 7 CLUSTAR-ND parameters which – unless specified by the user – will be automatically and optimally chosen based on the input data. We conclude that CLUSTAR-ND is a robust astrophysical clustering algorithm that can be leveraged to find stellar satellite groups on large synthetic or observational data sets.

**Key words:** methods: data analysis – methods: statistical – galaxies: star clusters: general – galaxies: structure.

## 1 INTRODUCTION

The process of identifying clusters – often referred to as clustering<sup>1</sup> – from a data set has been an ongoing data mining problem within the field of machine learning. The function of any algorithm tasked with undertaking this problem is to determine a set of statistically coherent groups within the intrinsic feature space of the data. Many such algorithms exist for this purpose, however, due to the innate subjectivity of the definition of a *cluster* it is not always obvious as to which of these may be useful for clustering of a given type.

The groupings found by a clustering algorithm can be categorized into one or more of the typical models, such as; centroid-based (e.g. K-MEANS; MacQueen et al. 1967; Lloyd 1982), distribution-based (e.g. EM; Dempster, Laird & Rubin 1977), density-based (e.g. DBSCAN; Ester et al. 1996), and others. The way in which a clustering algorithm partitions the data is also of importance and provides a similar categorization. These algorithms may return a flat or hierarchical clustering such that clusters are mutually exclusive or can be proper sub/supersets of one another – this is one of the differences between DBSCAN and HDBSCAN (Campello et al. 2015) for example. In addition to these, an algorithm may also return an overlapping set of clusters. A clustering may place all points within clusters (e.g. K-MEANS) or more commonly it may leave some points out of clusters classifying them as noise or outliers to the clusters that have been predicted. Moreover, these algorithms may return a hard or soft – also referred to as fuzzy – clustering. This distinction

clarifies whether the nature of a point being within a cluster is binary-based or probability-based (e.g. FUZZY C-MEANS; Dunn 1973). Since each of these separate class systems are codependent – in that an algorithm may be classified by multiple at a time – it suffices to say that choosing an appropriate clustering algorithm for the problem at hand is non-trivial.

In cosmology, a typical description of a galaxy and its halo is any spatial overdensity that is denser than the critical – or mean – density of the universe by some factor  $\Delta$ . Two common ways to identify such overdensities from both cosmological simulations and observed data sets are via the Spherical-Overdensity method (SO; Press & Schechter 1974) and the Friends-of-Friends algorithm (FOF; Davis et al. 1985). By themselves, these algorithms perform a density-based flat and hard clustering with noise which is well suited to the intended definition of galactic haloes – and as an additional constraint, the SO method can also be considered part of the distribution-based family due to ensuring that clusters adhere to a particular volumetric shape.

It is a primary prediction of the Lambda cold dark matter ( $\Lambda$ CDM) cosmological model that galaxies are formed hierarchically via continual accretion and merger events (White & Rees 1978; Kauffmann, White & Guiderdoni 1993; Ghigna et al. 1998). Depending upon the conditions of the satellite at the time of the merger as well as the ongoing conditions of the host halo (Bullock & Johnston 2005; Johnston et al. 2008), the particles within these infalling groups may become mixed in with, and indistinguishable from, the surrounding halo.<sup>2</sup> This prediction is mostly confirmed in nature barring a few exceptions such as the missing satellite problem (Klypin et al. 1999; Moore et al. 1999; Reed et al. 2005; Springel et al. 2008; Tollerud et al. 2008; Ishiyama et al. 2013) and the core-cusp problem

\* E-mail: [woli0618@uni.sydney.edu.au](mailto:woli0618@uni.sydney.edu.au)

<sup>1</sup>Note that we use this term throughout this paper to mean both the general process of finding clusters and the resultant set of classifications from this process – depending on the context. As such, our use of this term is much broader than its typical use within astro- and cosmo-related fields, e.g. referring specifically to large-scale structure.

<sup>2</sup>The phase-space volume of infalling groups is conserved in a fully collisionless Newtonian gravity simulation.

(Flores & Primack 1994; Moore 1994; Van Den Bosch et al. 2000). As such, an appropriate flavour of clustering algorithm for application to within galactic haloes is that of the density-based hierarchical type with noise. Many such algorithms exist for specifically this purpose, including but not limited to; SUBFIND (Springel et al. 2001), AHF (Knollmann & Knebe 2009), ROCKSTAR (Behroozi, Wechsler & Wu 2012), VELOCIRAPTOR (Elahi et al. 2019), and more. Comparisons of these and others tend to show the outputs of halo-finders to be similar (Knebe et al. 2011, 2013).

The above halo-finders mostly rely on the SO method and/or the FOF algorithm which both restrict the clustering to certain types of clusters. The varied shapes and density profiles of subhaloes often mean that these base algorithms need to be applied iteratively in order to find a range of subhalo types. In doing so, the run-times become larger although this does also allow some of these halo-finders to return hierarchical clusterings and even utilize adaptive metrics. Despite these measures of fine-tuning, SO-based halo-finders struggle to find curved structures such as streams and FOF-based halo-finders can still fall victim to point-point noise and fail to detect sparsely populated structures. The galaxy/(sub)halo finder HALO-OPTICS (Oliver et al. 2020) is based on OPTICS (Ankerst et al. 1999), a generalisation of DBSCAN, and is more robust to point-point noise. Coupled with an adaptive metric, HALO-OPTICS could be a powerful astrophysical clustering algorithm if not for the fact that the run-time of OPTICS can become unmanageable when operating in this way. Alternatives typically only use a set of nearest neighbours to perform locally adaptive metric computation. One such algorithm, the subhalo-finder ENLINK (Sharma & Johnston 2009), uses an entropy-based locally adaptive metric in combination with a group finder that is also robust to point-point noise in order to find a range of subhalo types. VELOCIRAPTOR and HSF (Maciejewski et al. 2009) also can attempt to generate locally adaptive metrics to identify clusters.

Having a fast clustering algorithm is not simply a matter of convenience. It is critical when clustering over extensive data sets with large feature spaces. It is also of particular importance when used to find a fuzzy clustering of an astrophysical data set. Unlike the output from the FUZZY C-MEANS algorithm, which is a fuzzy clustering of a hard data set, the fuzzy clustering found in this context is itself from a fuzzy data set and needs to have propagated the uncertainties of a point's features into the probability of that point's membership within a given cluster as well as the probability of that cluster's existence. Solving this problem typically requires taking many samplings of the data and comparing each of the clusterings therefrom (e.g. Fuentes, De Ridder & Debosscher 2017; Malhan et al. 2022). In order to be able to perform such a task whilst also maintaining a high-calibre clustering power, it is becoming necessary to use an algorithm that is: density-based; hierarchical; robust to point-point noise; equipped with adaptive metric capabilities; applicable to data sets with any number of features; and demonstrates exceptionally modest run-times.

We present CLUSTAR-ND, a derivative of HALO-OPTICS that possesses the above qualities, and apply it to synthetic survey data produced by GALAXIA (Sharma et al. 2011) in order to formalize the process of clustering  $N$ -dimensional data sets of galactic haloes. We first provide a relevant outline of the HALO-OPTICS algorithm (Section 2). We then describe the CLUSTAR-ND algorithm in Section 3. In doing so we give an overview of the concept and motivation behind the algorithm (Section 3.1), describe the root-haloes produced (Section 3.2), and the means by which the substructure therein is found (Section 3.3–3.6). In Section 4, we summarize the details of the synthetic galaxies from GALAXIA. Following this we make a

direct comparison between CLUSTAR-ND and the HALO-OPTICS algorithm in Section 5. In Section 6, we discuss the influence of the algorithm's parameters and optimize them by training them on the galaxies from GALAXIA. We then discuss the contrast in the information content available from the clusterings produced from different clustering scenarios in Section 7. Finally, we present our conclusions and directions for future work in Section 8.

## 2 AN OVERVIEW OF THE HALO-OPTICS ALGORITHM

In order to understand CLUSTAR-ND, it is first necessary to grasp the algorithm HALO-OPTICS (Oliver et al. 2020). HALO-OPTICS is a hierarchical galaxy/(sub)halo finder that can be used for the density-based determination of astrophysical clusters in a 3D spatial data set via a global distance metric, i.e. the Euclidean distance. It is an extension of the well-known hierarchical clustering algorithm, OPTICS (Ankerst et al. 1999) – which is itself an extension of DBSCAN (Ester et al. 1996). DBSCAN is non-hierarchical and will produce a flat clustering with noise of a given data set such that the points in each cluster are *densely connected* with a density greater than some threshold. All points with a density lower than this threshold are classified as noise and are not clustered. OPTICS not only connects all points in this manner, but also keeps track of how the points are connected together – i.e. in which order and with what measure of local density. This adjustment allows OPTICS to build a *reachability plot* which contains information of the clustering structure.

Both the OPTICS and DBSCAN algorithms require 2 parameters;  $\epsilon$ , a search radius, and  $N_{\min}$  (often denoted as *MinPts*), the minimum number of points a cluster can have. The algorithms also need a distance metric to be defined over the feature space of the input data. Conditioned upon this metric, OPTICS uses the concepts of *core distance* and *reachability distance* to produce its output – the reachability plot. The core distance of any given point is the distance between that point and its  $N_{\min}^{\text{th}}$  nearest neighbour. It then follows that a *core-point* is any point whose core distance is less than or equal to  $\epsilon$ . The reachability distance of any point,  $q$ , with respect to another,  $o$ , is the maximum of the core distance of  $o$  and the distance between  $q$  and  $o$ .

Initially, OPTICS computes the core distances of each point and sets the reachability distances of all points to infinity. The algorithm then iteratively orders each point in the data set. In each iteration of this ordering process, the next-to-be-ordered point is chosen as the point with the smallest reachability distance from the set of all unordered points – for this reason the first point is chosen at random. The chosen point is then removed from the list of unordered points and appended to the ordered list. If this point is a core-point, then all remaining unordered points within a radius of  $\epsilon$  of it are found. The reachability distance of each of these unordered neighbours is then set to be the minimum of their currently assigned reachability distance and their reachability distance with respect to the recently ordered core-point. Following this step, the ordering process continues on to the next iteration – repeating this cycle until all points are in the ordered list. It is in this manner that OPTICS creates an ordering of the data points, whilst concurrently seeking out regions of minimal reachability distance, i.e. highest density. After the ordering process, the reachability plot can then be constructed by plotting the final reachability distances as a function of the corresponding ordered indices, for all points. Within this plot, mutually clustered points appear as valleys since these points are both denser than their surrounds (smaller distances between points) and local to each other

(ordered consecutively). More details on the OPTICS algorithm can be found in Ankerst et al. (1999) and Oliver et al. (2020).

HALO-OPTICS takes the OPTICS algorithm and not only formalizes the way in which the OPTICS parameters are chosen but also provides a robust cluster extraction technique that operates on the reachability plot in order to extract a tree-structured hierarchy of suitable astrophysical clusters.<sup>3</sup> In HALO-OPTICS, the OPTICS parameter  $\epsilon$  is effectively converted to the more physical parameter  $\Delta$ , thereby classifying the root-level galaxy haloes using the factor by which they are denser than the critical (or mean) density of the Universe. HALO-OPTICS does this by using an approximate mapping between the FOF linking length  $l_x$ ,  $N_{\min}$ , and  $\epsilon$ . The technique of extracting clusters is based on the designs of Sander et al. (2003), Zhang et al. (2013), and McConnachie et al. (2018). The cluster extraction technique of HALO-OPTICS creates the hierarchy of clusters by finding the valleys within valleys that satisfy a range of conditions. These conditions assert that a cluster must; contain at least  $N_{\min}$  points; have a median density that is a factor of  $\rho_{\text{threshold}}$  denser than the cluster's surrounds; not be a single child cluster of its parent cluster; not share more than  $f_{\text{reject}}$  of its points with its parent; and it rejects local outliers that have local-outlier-factors (defined in equation 8 and Breunig et al. 1999) greater than or equal to  $S_{\text{outlier}}$  from each cluster. The HALO-OPTICS algorithm therefore transforms one OPTICS parameter and adds 3 unitless parameters for which near optimal values are  $\rho_{\text{threshold}} = 2$ ,  $f_{\text{reject}} = 0.9$ , and  $S_{\text{outlier}} = 2$ . More details on the HALO-OPTICS algorithm can be found in Oliver et al. (2020).

### 3 CLUSTAR-ND: A HIERARCHICAL GALAXY/(SUB)HALO FINDER

#### 3.1 Concept and motivation

Originally, our intention was to create an optimized version of the HALO-OPTICS algorithm (refer to Section 2 for details) that implemented a locally adaptive metric and could be applied to a data set with an arbitrary number of features with minimal additional input from the user. Ultimately, since HALO-OPTICS iteratively performs a radial search about each of the  $n$   $d$ -dimensional points in the data set, the cost of having it compute the distances via a locally adaptive metric is too great. For  $m$  points within some radius of the query point, this algorithm would effectively need to have computed the inverse and determinant of  $m-d \times d$  covariance matrices which makes the time complexity of the radial search and distance calculations  $\mathcal{O}(m \log(n))$  and  $\mathcal{O}(md^3)$ , respectively (each with different constant factors). In practice this increases the run-times dramatically such that even applying this to small galactic haloes becomes challenging.

We have designed CLUSTAR-ND with the intention of mimicking HALO-OPTICS along with the additional benefits of having; faster run-times, a variety of adaptive metric settings, and the capability of performing on a data set with any number of features. In addition to the input data, CLUSTAR-ND has 7 parameters that it uses to construct clusters:

(i)  $l_x$  ( $\in \mathbb{R}_{>0}$ ), the spatial linking length that is used to find root-level haloes as discussed in Section 3.2. May be given by the user

although is set to  $\infty$  by default – which makes CLUSTAR-ND treat the input data as a root-level halo.

(ii) *Adaptive* ( $\in \{0, 1, 2\}$ ), a flag that defines the behaviour of the metric used to calculate the Mahalanobis distances between points as discussed in Section 3.3. May be given by the user although is set to 1 by default – which provides consistently robust results as shown in Section 7.

(iii)  $k_{\text{den}}$  ( $\in \mathbb{N}_{\geq 7}$ ), the number of nearest neighbours that are used to calculate the measure of local density for each point as discussed in Section 3.4. May be given by the user although is set to 20 by default – which provides consistently robust results as shown in Sections 6 and 7.

(iv)  $k_{\text{link}}$  ( $\in \mathbb{N}_{\geq 7 \wedge \leq k_{\text{den}}}$ ), the number of nearest neighbours that are used to densely connect the points in the data as discussed in Section 3.5. May be given by the user although is automatically calculated to be optimal by default as in Section 6.1 – these optimal values are based on the input data and the value of  $k_{\text{den}}$  and ensure maximal cluster completeness and interneighbourhood connectivity.

(v)  $\rho_{\text{threshold}}$  ( $\in \mathbb{R}_{\geq 1}$ ), the factor by which the median density of a cluster must be denser than that cluster's surrounds as discussed in Section 3.6. May be given by the user although is automatically calculated to be optimal by default as in Section 6.2 – these optimal values are based on the input data and the value of  $k_{\text{den}}$  and ensure that the leaf level of the returned clusters are satellite-like overdensities.

(vi)  $f_{\text{reject}}$  ( $\in \mathbb{R}_{\geq 0 \wedge \leq 1}$ ), the maximum fraction of points that can be shared by parent-child clusters in the hierarchy as discussed in Section 3.6. May be given by the user although is set to 0.9 by default – which provides a simple to interpret hierarchy that agrees with analyses in Oliver et al. (2020) as discussed in Section 6.4.

(vii)  $S_{\text{outlier}}$  ( $\in \mathbb{R}_{\geq 1}$ ), the local-outlier-factor that is used to reject outlier points from a candidate cluster as discussed in Section 3.6. May be given by the user although is set to 2.5 by default – which provides a moderate level of outlier removal whilst maintaining good clustering results as is shown in Section 6.3.

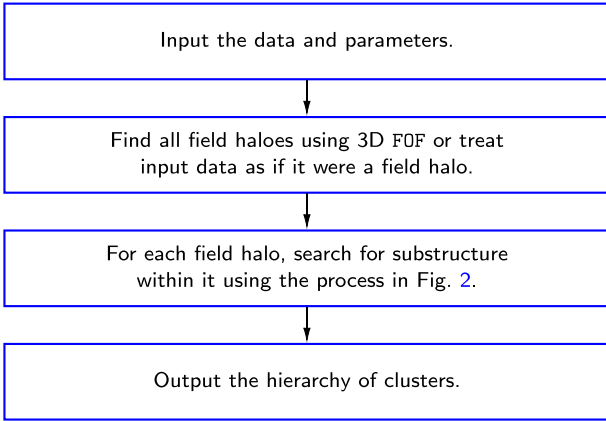
Since optimized values for four of these parameters are found, the user only *needs* to consider choosing values for  $l_x$ , *adaptive*, and  $k_{\text{den}}$ . The details of the algorithm CLUSTAR-ND and how to choose these three parameters is outlined in the subsections of Sections 3 and 6.

#### 3.2 Defining root-level clusters

One major distinction between the group-finding component of the CLUSTAR-ND algorithm and that of the HALO-OPTICS algorithm is that the implementation of CLUSTAR-ND does away with having to perform a radial search about each point. This is the largest factor in why CLUSTAR-ND is so much faster than HALO-OPTICS and is also why CLUSTAR-ND does not have an equivalent parameter to that of the OPTICS parameter  $\epsilon$ . In OPTICS this parameter not only aids in the ordering of points but also defines the maximum reachability distance that a clustered point can have – thereby prescribing an approximate minimum density that a cluster can have. The functionality of  $\epsilon$  that leads to the detection and extraction of clusters can be approximated without its presence (refer to Section 3.5 for details on this), however the effect of defining a cluster's minimum density requires another separate step.

In HALO-OPTICS, the choice of  $\epsilon$  is redirected to the choice of  $\Delta$  – the overdensity factor. This is the factor by which a field halo (root-level cluster) is denser than the critical density of the Universe,  $\rho_{\text{crit}}$ . This is accomplished via a mapping between the FOF linking length,  $l_x$ , and  $\epsilon$  and is used to approximate the FOF field haloes. In CLUSTAR-ND's implementation, we offer the functionality of

<sup>3</sup>The tree-structure of this hierarchy allows for clusters to be referred to by the typical terminology, e.g. root-level cluster (largest cluster in the tree), parent cluster (superset of the child), child cluster (subset of the parent), and leaf cluster (has no child clusters of its own).



**Figure 1.** An activity of the outer methods of the CLUSTAR-ND process. Before finding the substructure CLUSTAR-ND first decides on the root-level clusters – where the option is given to produce 3D FOF field haloes or to treat the input data as if it were a field halo.

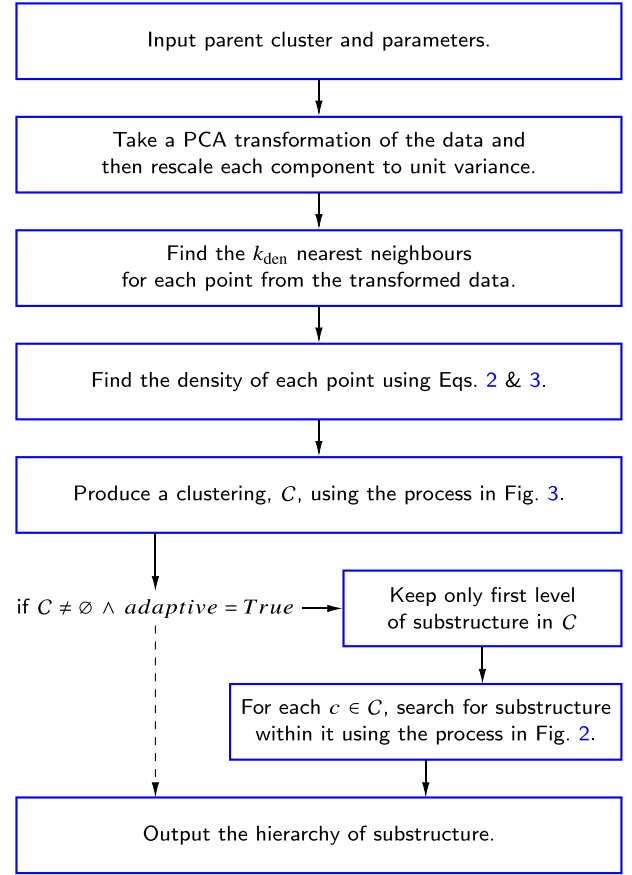
directly computing the 3DFOF field haloes from the positions<sup>4</sup> of the data set as a whole. This splits the input data into some number of field haloes and opens the opportunity for trivial parallelization over each of these haloes for the remaining steps.

These 3DFOF field haloes will be found by CLUSTAR-ND whenever  $l_x \neq \infty$ . A common choice for  $l_x$  when finding field haloes from cosmological simulations is  $l_x = 0.2L_{\text{box}}/N$  – where  $L_{\text{box}}$  is the side length of the simulation box and  $N$  is the number of particles within it – which corresponds to field halo overdensities of  $\gtrsim 100\bar{\rho}$  (Elahi et al. 2019). Alternatively when  $l_x = \infty$  (default), CLUSTAR-ND will neglect finding field haloes in this way which results in the root-level cluster becoming the entire data set. The choice of whether to find field haloes using the 3DFOF approach or not must be made under consideration of the input data and the intended output. This choice is outlined in the first few nodes of Fig. 1.

Other hierarchical astrophysical clustering algorithms such as VELOCIRAPTOR (Elahi et al. 2019) also perform this step in order to define the root-level clusters before then finding the substructure within them. However due to the possibility of not having to perform this step in CLUSTAR-ND, it is also possible for the user to apply some other algorithm to the data – or to not do so at all – in order to perform this step prior to using CLUSTAR-ND. By default  $l_x$  is set to  $\infty$  which forces CLUSTAR-ND to treat the input data as a single root-level halo from which the substructure therein is subsequently found.

Once the root-level cluster(s) have been found, we add them to a list of unsearched clusters in order to find the substructure within them. Each of the unsearched clusters are independent of each other, meaning that we can search for substructure in parallel. This remains true even when we set  $adaptive = 2$ , which ensures that CLUSTAR-ND iteratively searches within the top level of substructure until there are no more clusters found – as shown in Figs 1, 2, and 3. In this way, setting  $adaptive = 2$  creates a locally adaptive metric since the data are transformed before each instance of finding substructure. The following subsection provides details of how these transformations are performed.

<sup>4</sup>If this functionality is used, then the first 3 features of the input data must be the Cartesian spatial coordinates of the points.



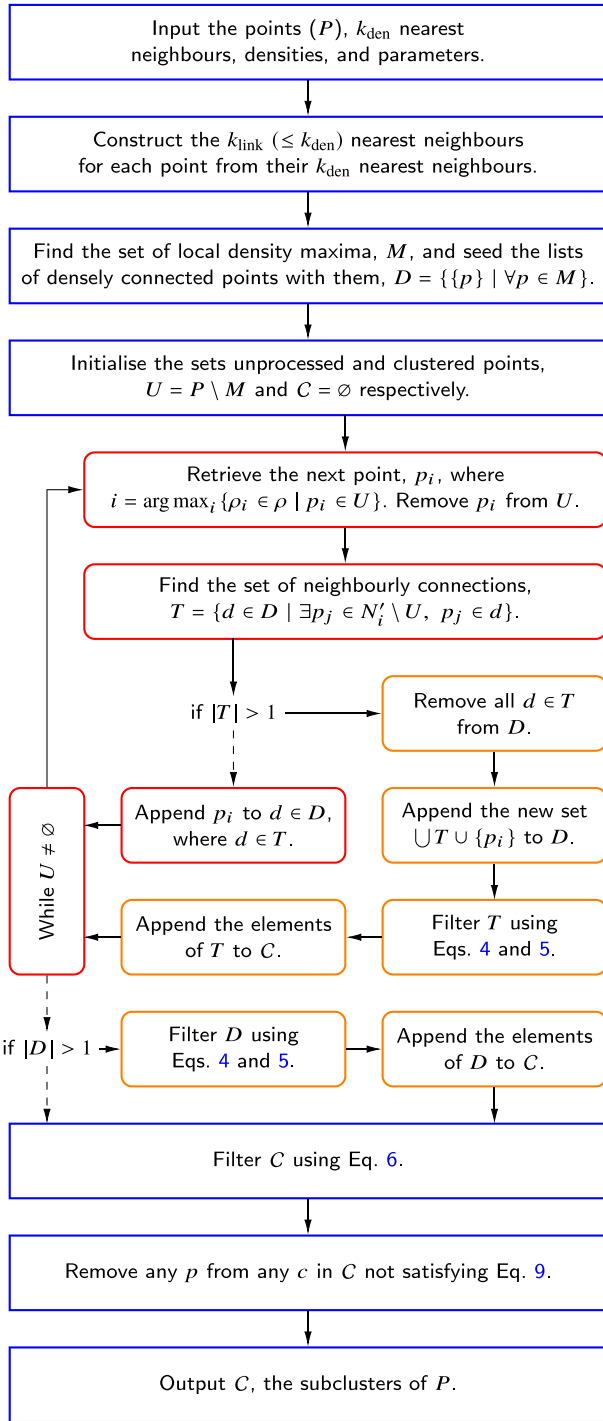
**Figure 2.** An activity chart of the methods concerned with the set up for finding substructure. A transformation of the data is taken, the nearest neighbour lists are found, and then from them the density of each point is calculated. This generalizes the approach to clustering the substructure. If such a clustering is found and  $adaptive$  parameter is set to 2, then the process in this activity chart is invoked again on first level of substructure. This occurs iteratively until there are no more significant clusters found, at which point the process produces a cascade of returns of the adaptive hierarchy of substructure. If  $adaptive$  is set to 0 or 1, the process returns the hierarchy of substructure that has been found using a parent cluster (globally) defined metric.

### 3.3 Data transformation

When searching for substructure within a previously unsearched cluster, we must implement a strategy that scales the various dimensions so that the clusters found are not overly dependent on a subset of the features. Such a strategy needs to be robust against the unit choice and coordinate orientation. So that we produce the desired effects, we choose to use a Principle Component Analysis (PCA) to first transform the data within the cluster. We then scale each PCA component to unit variance. If we calculate distances via the Euclidean distance metric on the transformed space, we are then effectively using a Mahalanobis distance metric (Mahalanobis 1936) such that

$$s^2(x_i, x_j) = (x_i - x_j)^T \Sigma^{-1} (x_i - x_j), \quad (1)$$

where  $\Sigma = E[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T]$  is the covariance matrix of the cluster before the transformation. This guarantees that substructure found within a cluster will be dense with regards to the cluster. When  $adaptive = 2$ , CLUSTAR-ND approximates a locally adaptive Mahalanobis metric since a new covariance matrix is redefined at



**Figure 3.** The activity chart for the aggregation and rejection process that constructs a hierarchy of subclusters similar to those from HALO-OPTICS. The nearest neighbour lists are reduced and from them the local density maxima are found. The lists of densely connected points are then seeded with these local maxima. In order of decreasing local density, the points are either appended to an existing list of densely connected points or used to merge multiple into a new one. In the event of the latter and subject to the conditions of equations (4) and (5), these lists are considered to be potential clusters. Following this inner loop, if not all points were densely connected, the remaining connected lists are subject to these same conditions in order to become potential clusters. The hierarchy of clusters is then cleaned using equation (6), and each remaining cluster is also cleaned subject to equation (9).

every level of the hierarchy.<sup>5</sup> Hence, the choice of whether to set *adaptive* to be 0, 1, or 2 must be made in consideration of the trade-off between run-time constraints and the additional clustering power that an adaptive metric gives over that of a global metric.

### 3.4 Density estimation

Following the transformation of the data, we now seek to find a measure of the local density surrounding each point. We do this by first finding the  $k_{\text{den}}$  nearest neighbours for each point within the transformed space. The density for each point is then approximated using the neighbour list and a multivariate kernel within a balloon estimator such that

$$\rho_i \propto \frac{1}{h_i^d} \sum_{j=1}^{k_{\text{den}}} K\left(\frac{s(x_i, x_j)}{h_i}\right). \quad (2)$$

Here,  $s(x_i, x_j)$  is defined in equation (1),  $d$  is the dimensionality of the feature space,  $K$  is a kernel function, and  $h_i$  is a smoothing length corresponding to point  $p_i$  – which we choose to be the distance from  $p_i$  to its  $k_{\text{den}}$ -most nearest neighbour. Choosing the smoothing length in this way is not atypical (e.g. Sain 2002), however, this also has the added bonus of being equal to the core distance from OPTICS. Together with an appropriate kernel function,  $K$ , this choice allows CLUSTAR-ND to compute a determinable and smoothed analogue of the rigid density estimator that is the reachability distance. For this we implement an Epanechnikov kernel (Epanechnikov 1969), which is defined as

$$K(u) \propto (1 - u^2), \quad (3)$$

and is 0 for  $u > 1$ . Typically there is a normalization constant that is defined such that the integral of the kernel over the space is 1. However, as we clarify in the following section, the process of determining significant substructure with CLUSTAR-ND only relies on density by processing points in order of descending density and by comparing the relative densities of points. For this reason, we do not need to compute the constant factors of equations (2) and (3).

### 3.5 Densely connecting points via neighbourly aggregation

Once the data has been transformed, the nearest neighbour lists found, and the local densities computed for any given unsearched cluster, the algorithm now seeks to find the substructure of that cluster by aggregating the points that are densely connected. The key to creating a similar output to that of HALO-OPTICS without performing any radial search is to densely connect points via their nearest neighbour lists. The OPTICS process allows for the algorithm to gain knowledge about the locality of points nearby to (and from the perspective of) the points that have already been appended to the ordered list. It uses this knowledge to constantly be seeking out regions of higher denser since the next-to-be-ordered point is always the point with the smallest reachability distance i.e. largest density.

<sup>5</sup>It should be noted that this is not equivalent to the locally adaptive Mahalanobis metric of ENLINK (Sharma & Johnston 2009), which is constructed via an entropy-based binary space partitioning algorithm and defines a unique covariance matrix for each point a priori to the clustering of the data. While this method is very effective for defining a locally adaptive metric based on the local distribution of points, it has a substantially larger run-time than the top-down method we implement in CLUSTAR-ND.

CLUSTAR-ND mimics this process by further restricting the nearest neighbour lists of each point to the nearest  $k_{\text{link}}$  points,<sup>6</sup> finding all points that are local density maxima with respect to their surrounds, and then using these points as seeds from which to densely connect other points to. Seeding the list of dense connections with each of the local maxima ensures that the aggregation process within CLUSTAR-ND does not need to seek out regions with a higher density than the points that have already been connected together. Instead it simply connects each point (that is not a local density maximum) in order of decreasing density to a set of already densely connected points – this process is similar to those of the core-search in VELOCIRAPTOR (Elahi et al. 2019) as well as the group-finders of SUBFIND (Springel et al. 2001) and ENLINK (Sharma & Johnston 2009).

The key condition that CLUSTAR-ND exploits here is that even though the OPTICS algorithm gains knowledge of all unordered points within a radius of  $\epsilon$ , the reachability distance of those points will typically be large if they are not directly connectable via the  $N_{\text{min}}$  nearest neighbours of either itself or the already ordered points. Hence at any given time during OPTICS ordering process, the next-to-be-ordered point is almost always a neighbour of a point (or vice versa) that has already been ordered. The only time that it is not as such, is if the next-to-be-ordered point cannot be directly connected through the  $N_{\text{min}}$ -sized neighbourhoods of those points that have already been ordered – or of such a neighbourhood of it's own. This does not mean that such a point can never be connected through  $N_{\text{min}}$ -sized neighbourhoods, however, doing so will require traversing points that are less dense than itself – densely connecting points concurrently. Such a scenario where the point with the smallest reachability distance cannot be densely connected through  $N_{\text{min}}$ -sized neighbourhoods is possible in OPTICS, although whether two points belonging to entirely separate sets of densely connected points are at all associated with somewhat uncertain in an astrophysical context. This distinction between CLUSTAR-ND and HALO-OPTICS is investigated further in Section 5.

The process of aggregating points is iteratively performed; by retrieving the unprocessed point with the highest density ( $p_i$ ), finding the set(s) of densely connected points ( $T$ ) that have members within the  $k_{\text{link}}$  nearest neighbours of  $p_i$ , then either appending  $p_i$  to an existing list of densely connected points or constructing clusters from  $T$  and merging the lists of  $T$  to create a new list of densely connected points (with  $p_i$  included). The condition for whether clusters are to be constructed and a new list of densely connected points created, is if  $|T| > 1$  i.e. if the point  $p_i$  can be densely connected to multiple sets of densely connected points. In this scenario, the previous sets of  $T$  are only retained as potential clusters if they satisfy equation (4) – a relative density condition dependent upon the trainable parameter  $\rho_{\text{threshold}}$ . More details on this condition and others that are responsible for cluster rejection are given in Section 3.6.

The circumstances under which  $|T| > 1$  for any given point depend strongly upon the value of  $k_{\text{link}}$  since this determines how many neighbours are considered for connecting the points. Decreasing  $k_{\text{link}}$  ensures that a greater number of points can be densely connected to the potential clusters before eventually finding a connecting point

<sup>6</sup>The reason for doing this is that the number of neighbours needed to densely connect points is not necessarily equal to number of neighbours that should be used to calculate the local density of a point. In fact, by restricting the nearest neighbour lists, CLUSTAR-ND is able to return a greater resolution on the clustering structure – as is described in Section 3.5.

whose neighbourhood satisfies  $|T| > 1$ . In effect, this gives a greater resolution to the clusters and, since  $k_{\text{link}}$  can be much less than  $k_{\text{den}}$ , is typically greater even than those from HALO-OPTICS. Of course, there is a reasonable lower limit to the value of  $k_{\text{link}}$  that can be found by ensuring that all points be densely connected under a uniform distribution of points. This lower limit is investigated further in Section 6.1. Finally, if all points within the previously unsearched parent cluster do not become densely connected through this aggregation process and if more than 1 of the densely connected subsets have at least  $k_{\text{link}}$  points, then we append those that do to the list of substructures as well.

### 3.6 Cluster and cluster member rejection

The cluster extraction algorithm of HALO-OPTICS is implemented in CLUSTAR-ND, however, it is not entirely its own separate process that simply takes place after the aggregation/ordering of points as it is in HALO-OPTICS. To detect and extract clusters from the reachability plot, HALO-OPTICS performs a set of routines – each of which are described in section 3.2 of Oliver et al. (2020). First, the algorithm determines all local maxima of the reachability plot from OPTICS and then using this, builds the hierarchy of clusters by taking contiguous subsets of the ordered list that are situated on either side of each local maxima – ensuring that each subset only contains points with reachability distances less than that at the corresponding local maxima. CLUSTAR-ND performs these steps by ensuring that points are aggregated in order of decreasing local density and then by considering lists of densely connected points that are connected to one another via the neighbourhood of a less dense point as potential clusters.

HALO-OPTICS then rejects all potential clusters that contain less than  $N_{\text{min}}$  points or that have median local densities less than  $\rho_{\text{threshold}}$  times the surrounding density of those potential clusters. Following this, the algorithm rejects all single leaf potential clusters. These criteria are fulfilled by CLUSTAR-ND in a single step during the aggregation process when multiple lists of densely connected points are considered to be potential clusters. As such, CLUSTAR-ND conducts the following:

$$\begin{aligned} &\text{reject any } d \text{ from } T \text{ if it has no children and either;} \\ &|d| < k_{\text{link}} \text{ or} \\ &\text{median}\{\rho_j \mid p_j \in d\} / \rho_i < \rho_{\text{threshold}}. \end{aligned} \quad (4)$$

Here,  $\rho_i$  is the local density of  $p_i$  – the point responsible for densely connecting the otherwise disconnected sets in  $T$ . Following the rejection of all  $d \in T$  that do not meet these criteria, the set  $T$  is then subject to the condition that

$$\text{if } |T| = 1, \text{ set } T = \emptyset. \quad (5)$$

The conditions in equations (4) and (5) are related to steps 3 and 4 from the cluster extraction process of HALO-OPTICS. Next, for all parent-child cluster pairs sharing at least  $f_{\text{reject}}$  of the parent's points; CLUSTAR-ND rejects the child if it has child clusters of its own, otherwise CLUSTAR-ND rejects the parent. To re-iterate, given a set of substructure  $S$  and any parent-child pair  $s_{\text{parent}}$  and  $s_{\text{child}}$ ,

$$\begin{aligned} &\text{if } |s_{\text{child}}| / |s_{\text{parent}}| \geq f_{\text{reject}} \text{ then,} \\ &\text{reject } s_{\text{child}} \text{ if; } \exists s \in S \text{ such that } s \subset s_{\text{child}}, \\ &\text{else; reject } s_{\text{parent}}. \end{aligned} \quad (6)$$

This is equivalent to step 5 of the extraction process in HALO-OPTICS. Steps 6 and 7 of the HALO-OPTICS extraction process are concerned with removing outliers based on how their local density

compares with the rest of their neighbours. While this approach works in the context of using a reachability distance as the measure of inverse density, it can be improved upon with a more robust measure of density such as that that CLUSTAR-ND computes. The measure used by HALO-OPTICS is the local-outlier-factor which relies upon the local-reachability-density (Breunig et al. 1999), both of which are defined in equations (3) and (4) of Oliver et al. (2020). In effect, the local-outlier-factor compares how reachable a point is from its neighbours to how reachable its neighbours from their own. We provide CLUSTAR-ND with a comparable measure with which to define outliers that incorporates the estimate of local density it uses from equation (2).

We define the kernel density estimate analogue of the local-reachability-density of a point  $p_i$  with respect to transformed space as

$$\text{lrd}(p_i) \propto \frac{1}{\sum_{p_j \in N_{k_{\text{den}}}} \min(\rho_i, \rho_j)^{-1/d}}, \quad (7)$$

where  $N_{k_{\text{den}}}$  is the  $k_{\text{den}}$  nearest neighbours of  $p_i$  within the previously unsearched cluster. CLUSTAR-ND then uses this to similarly compute an analogue of the local-outlier-factor

$$\text{lof}(p_i) = \frac{\sum_{p_j \in N_{k_{\text{den}}}} \frac{\text{lrd}(p_j)}{\text{lrd}(p_i)}}{k_{\text{den}}}. \quad (8)$$

In HALO-OPTICS, a point is considered an outlier from a cluster if the local-outlier-factor (calculated with respect to that cluster) is larger than  $S_{\text{outlier}}$ . In CLUSTAR-ND, we adjust this rule by first defining a cut-off density,

$$\rho_{\text{cut}} = \min\{\rho_i \mid \text{lof}(p_i) < S_{\text{outlier}}\}, \quad (9)$$

which we then use to reject all points from a cluster that have a density less than this. This ensures that only local-outliers at the outskirts of a cluster are rejected and not those that are embedded deeper within it – a more appropriate regime for outlier detection and rejection within a hierarchical set of clusters.

## 4 SYNTHETIC DATA

Within the remainder of this paper we compare (Section 5), train (Section 6), and scrutinize (Section 7) our algorithm against synthetic survey data of Milky Way (MW)-type galaxies produced by GALAXIA (Sharma et al. 2011). Given certain survey restrictions, such as one or more colour–magnitude bounds, a survey size, and geometry, GALAXIA is able to return a synthetic catalogue of stars in accordance with a given model of the MW. The code is generalized enough so that it can also accept star formation rates, age–metallicity relations, age–velocity–dispersion relations, and analytic density distribution functions from which it can also use to produce stellar catalogues.

Among the synthetic data that can be produced with GALAXIA are resamplings of the 11  $\Lambda$ CDM stellar haloes from Bullock & Johnston (2005) and the complementary 6 artificial stellar haloes from Johnston et al. (2008). The original  $\Lambda$ CDM haloes are simulated using a hybrid semi-analytical plus hydrodynamic  $N$ -body approach that replicates a density profile and satellite distribution that is similar to the MW. The simulation model assumes a  $\Lambda$ CDM cosmology with parameters  $\Omega_m = 0.3$ ,  $\Omega_\Lambda = 0.7$ ,  $\Omega_b h^2 = 0.024$ ,  $h = 0.7$ , and  $\sigma_8 = 0.9$ . The simulations generate, track, and evolve a number of individual satellites that are first modelled as  $N$ -body dark matter systems within a parent galaxy whose disc, bulge, and

halo are represented by time-dependent semi-analytical functions. Semi-analytical prescriptions are then used to assign star formation histories and leaky accreting boxes to each. A chemical enrichment model is also used to calculate the metallicities as a function of age for the stellar populations (Robertson et al. 2005; Font et al. 2006). Ultimately, the dark matter distributions of the satellites follow NFW profiles (Navarro, Frenk & White 1996) while the stellar distributions follow King profiles (King 1962) – the latter of which is constructed in order to reproduce an agreement with the structural properties of the Local Group’s dwarf galaxies.

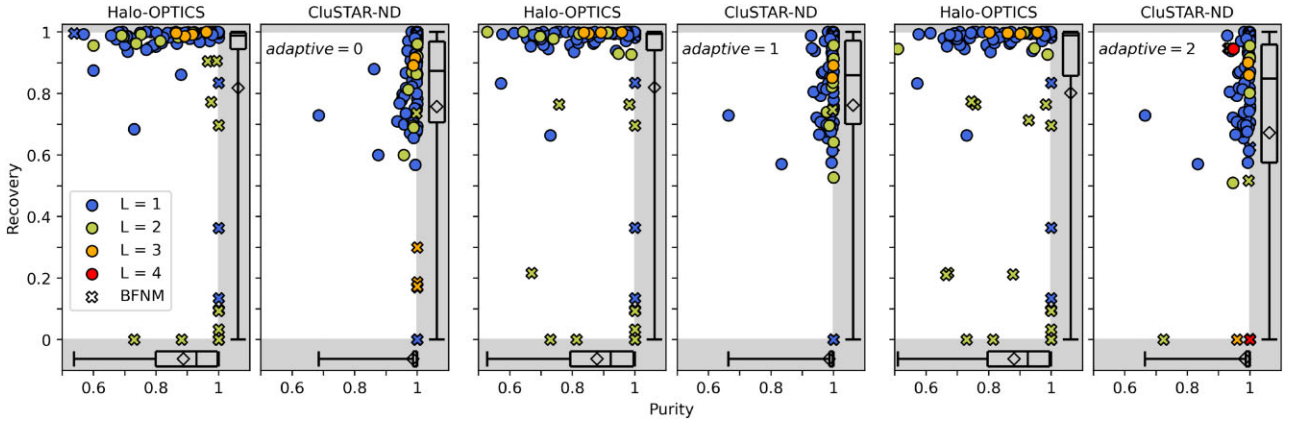
Each satellite created in this way has three main model parameters; the time since accretion ( $t_{\text{acc}}$ ), the luminosity ( $L_{\text{sat}}$ ), and the orbital circularity ( $\epsilon = J/J_{\text{circ}}$ ). The distribution of these parameters specify the accretion history of a halo and as such a further 6 artificial haloes were created in Johnston et al. (2008) for the purpose of studying the effects of different accretion histories on the properties of haloes. These 6 artificial haloes have accretion events that are predominantly; radial ( $\epsilon < 0.2$ ); circular ( $\epsilon > 0.7$ ); old ( $t_{\text{acc}} > 11$  Gyr); young ( $t_{\text{acc}} < 8$  Gyr); high luminosity ( $L_{\text{sat}} > 10^7 L_\odot$ ); and low luminosity ( $L_{\text{sat}} < 10^7 L_\odot$ ). Importantly, the output from GALAXIA also bestows each star with a label that corresponds to which satellite group it belonged to at the time that that satellite was created within the simulation – this allows us to test how well CLUSTAR-ND performs in Section 6. In addition to this, GALAXIA maintains a list of satellite properties that includes information on whether the satellite is self-bound or not. More details on these simulations can be found in section 3.4 of Sharma et al. (2011) and references therein.

We use GALAXIA to produce a random sample from each synthetic galaxy in both sets of the stellar haloes. From these we use various combinations of the spatial (denoted by  $\mathbf{x} = (x, y, z)$  in the later sections), kinematic (denoted by  $\mathbf{v} = (v_x, v_y, v_z)$  in the later sections), and chemical (denoted by  $\mathbf{m} = ([\text{Fe}/\text{H}], [\alpha/\text{Fe}])$  in the later sections) information in order to compare, optimize, and apply our algorithm. For each galaxy, the points are contained inside the 282 kpc virial radius of the corresponding host dark matter halo – these have an overdensity factor of  $\Delta \approx 337$  times the mean dark matter density of the universe. As such we treat these galaxies as field haloes and hence do not use CLUSTAR-ND to find field haloes before finding substructure. Since these field haloes and those that are found via 3D FOF are well-defined, the analysis in the following sections is an assessment of the substructure finding component of CLUSTAR-ND.

## 5 A COMPARISON WITH HALO-OPTICS

Our first assessment of CLUSTAR-ND’s performance is done by comparing it to HALO-OPTICS. We compare both the clustering and the run-times of these algorithms when applied to the synthetic haloes outlined in Section 4. Since HALO-OPTICS is designed for clustering on 3D spatial data, we only apply these algorithms to the 3D positions of the points in these haloes.

We apply CLUSTAR-ND in a way that treats each synthetic galaxy as a field halo ( $l_x = \infty$ ) – due to each galaxy only containing points within its virial radius. Similarly, for HALO-OPTICS we find the smallest value of  $\epsilon$  that will guarantee that every point in these reduced galaxies is a part of at least one core-point’s  $N_{\text{min}}$  nearest neighbours – thereby ensuring that every point will be given a reachability distance less than or equal to  $\epsilon$ . By applying these algorithms to ready-made galactic data sets, we can ensure that we are only comparing the substructure found and not the root-level clusters – which in both cases are equal to the entire data set due to the above.



**Figure 4.** A clustering comparison of the outputs of HALO-OPTICS and CLUSTAR-ND after having been applied to the 3D spatial dimensions of the synthetic galaxies from GALAXIA. Each panel belongs to one of three pairs and each pair corresponds to a different comparison wherein CLUSTAR-ND has been used with a different metric adaptivity setting – which is annotated on the relevant panels. Plotted within each panel are the recovery and purity fractions of each best-fitting cluster to either; the HALO-OPTICS predicted clusters (left-hand panel in each pair) or the CLUSTAR-ND predicted clusters (right-hand panel in each pair). Best-fitting clusters are those that produce the maximum Jaccard index in equation (10) (i.e.  $T^* = \arg \max_{T \in S_2} J(C, T)$ ), the recovery and purity are then found according to equation (11). To clarify, the label at the top of each panel indicates which clustering is used as  $S_1$  within these equations. Every cluster, except for the root-level clusters ( $L = 0$ ), appear in these panels as a marker coloured by its level within the hierarchy and shaped by whether or not its best fitting cluster is a mutually best fitting – note that BFNMs here stands for *best fitting not mutual*. The axes are extended beyond the possible recovery and purity values for easier readability, this extended area is coloured grey. Within this area on each panel and for both the recovery and purity distributions, there is a box and whisker plot that denotes the  $Q_0 - Q_4$  quartiles with mean values also indicated with diamonds markers.

In order to assess how adequately CLUSTAR-ND reproduces HALO-OPTICS, we set  $k_{\text{den}} = 20$  and similarly for HALO-OPTICS we set  $N_{\text{min}} = 20$ . For CLUSTAR-ND we set  $k_{\text{link}} = 7$ , as is described in Section 6.1. For both codes we use the near-optimal parameters found in Oliver et al. (2020) such that  $\rho_{\text{threshold}} = 2$ ,  $f_{\text{reject}} = 0.9$ . However since the outlier detection is different between the codes, we set  $S_{\text{outlier}} = \infty$  which ensures that there are no outliers removed from clusters. We review the optimal choice of these parameters in Sections 6.2–6.4.

### 5.1 Output similarity

To appropriately compare the output of the two codes, we find the best-fitting match from the HALO-OPTICS catalogue of clusters for each of the CLUSTAR-ND clusters (ignoring the root-level clusters which are fixed to be equal). As a measure of which pair is *best-fitting* we use the maximum Jaccard index (Jaccard 1912), which – given two clusterings of the same data set ( $S_1$  and  $S_2$ ) – allows us to not only find which clusters from  $S_1$  are best fitting to the clusters from  $S_2$  but to also compare how well the clusters in  $S_2$  are matched by those in  $S_1$ . Hence for two such clusterings the maximum Jaccard index is given by

$$J_{\text{max}}(C) = \max\{J(C, T) \mid \forall T \in S_2\}, \quad C \in S_1, \quad \text{where} \quad (10)$$

$$J(C, T) = \frac{|C \cap T|}{|C \cup T|}.$$

Once each cluster from  $S_1$  has been assigned a best-fitting cluster from  $S_2$ , we then compute the recovery and purity fractions of the pairs. We define these such that

$$R(C) = \frac{|C \cap T^*|}{|T^*|}, \quad \text{and} \quad (11)$$

$$P(C) = \frac{|C \cap T^*|}{|C|},$$

where  $T^* \in S_2$  is the best-fitting match to  $C \in S_1$  – i.e.  $T^* = \arg \max_{T \in S_2} J(C, T)$ . Neatly, this implies that

$$\frac{1}{J_{\text{max}}(C)} = \frac{1}{R(C)} + \frac{1}{P(C)} - 1, \quad \text{and hence} \quad (12)$$

$$J_{\text{max}}(C) \leq \min\{R(C), P(C)\}.$$

Fig. 4 depicts the recovery and purity fractions determined in this way for all of the best-fitting cluster pairs found within the synthetic haloes from GALAXIA for various values of the *adaptive* parameter. Here, we see that the clusterings from CLUSTAR-ND are best fitted by those from HALO-OPTICS in a way that most commonly leads to clusters from CLUSTAR-ND being encompassed by those from HALO-OPTICS – indicated in each comparison by high purity and varied recovery in the left-hand panel of each pair and the opposite in the right-hand panel of each pair. This matches the trends seen in both figs 6 and 10 of Oliver et al. (2020) where HALO-OPTICS was typically shown over-encompassing both mock cluster sets and predicted clusters found by VELOCIRAPTOR (Elahi et al. 2019). Regardless, this property of the best-fitting clusters from CLUSTAR-ND and HALO-OPTICS suggests that CLUSTAR-ND may benefit from an additional method to allocate points to the clusters that it already finds. Such a task could be achieved via an expectation-maximization-like technique that allocates previously noisy points to already existing clusters from CLUSTAR-ND depending on how well they would assimilate within them.

A small number of clusters are also shown in Fig. 4 to have high purity and very low recovery ( $< 0.2$ ). These clusters are those that have been found by one algorithm but not the other – hence they are not mutually best fitting – and are consequences of: (1) CLUSTAR-ND’s density estimate being more reliable than that of HALO-OPTICS; (2) CLUSTAR-ND being able to robustly detect clusters with a smaller number of points; and (3) the algorithmic differences between connecting points. We do also see a few high recovery and high purity clusters whose best-fitting cluster is not mutually best-fitting returned by both codes. These are clusters from one code that effectively sit between two levels of the hierarchy in



the other code – this can produce a good fit to a cluster that is not necessarily a mutually best-fitting cluster. These clusters arise for the same reasons as above.

Overall, it is clear that CLUSTAR-ND is able to reproduce the clusters and hierarchy of HALO-OPTICS with a similar clustering power. We see from Fig. 4 and the box and whisker plots therein that typically the clusters from CLUSTAR-ND (HALO-OPTICS) are mutually best fitting to those from HALO-OPTICS (CLUSTAR-ND) with both high recovery and high purity with medians of  $\sim 0.860$  ( $\sim 0.990$ ) and  $\sim 0.998$  ( $\sim 0.926$ ), respectively.

For completeness, we also discuss the clustering power of these outputs from HALO-OPTICS in comparison with the optimized outputs of CLUSTAR-ND in Section 7. In Section 7, we see that overall CLUSTAR-ND is expected to provide an equal clustering power to HALO-OPTICS, however the clustering power does tend to differ slightly on a per galaxy basis. This difference is due to the competing effects of CLUSTAR-ND having a more robust density estimation, while the ordering process of HALO-OPTICS can more easily gather like-satellite points before joining child clusters into their shared parent cluster. We will now define this measure of clustering power as it is used throughout both Sections 6 and 7.

### 5.1.1 Measuring clustering power

In order to measure the clustering power of CLUSTAR-ND we must construct a scalar function that reduces the complexity of comparing the quality of fit between an entire CLUSTAR-ND clustering output when applied to a synthetic galaxy and the ground truth labels of that synthetic galaxy. Since each galaxy has been constructed through the accretion of satellites (refer to Section 4 for details), the ground truth labels form a flat clustering without noise, i.e. each data point in the galaxy is a member of exactly one satellite. This is a fundamentally distinct clustering type from the hierarchical clustering with noise that CLUSTAR-ND produces – so we only find the goodness of fit between the leaf clusters produced by CLUSTAR-ND and the satellite labels within each galaxy.

In order to compare a leaf-clustering produced by CLUSTAR-ND ( $C_g$ ) to the ground truth clustering ( $T_g$ ) defined over the same galaxy ( $g$ ), we construct a mutual-information-based objective function similar to that built by Vinh, Epps & Bailey (2009). To do this we must first append an additional cluster to those in  $C_g$  to create  $C_g^*$  so that the sets  $C_g^*$  and  $T_g$  are defined over the same set of data points. The additional cluster is made of the remaining points in the data set that have not been clustered into the leaf clusters returned by CLUSTAR-ND. This means that  $C_g^*$  is an artificial construction of a flat clustering without noise using  $C_g$  which we can then easily match to the flat clustering without noise that is  $T_g$ .

The mutual information between  $C_g^*$  and  $T_g$ ,  $I(C_g^*; T_g)$ , is then the amount of information obtained about the true clusters by having observed the predicted clusters (Shannon 1948).

$$I(X; Y) = H(X) - H(Y|X), \text{ where}$$

$$H(X) \equiv - \sum_{x \in X} P(x) \log(P(x)), \text{ and}$$

$$H(Y|X) \equiv - \sum_{x \in X, y \in Y} P(x, y) \log\left(\frac{P(x, y)}{P(x)}\right). \quad (13)$$

This measure is non-negative and will have a different theoretical maximum depending on the synthetic galaxy in question. For our purposes, we normalize  $I(C_g^*; T_g)$  between 0 and 1 such that these values represent the absolute worst and best values that a clustering algorithm can be expected to have, respectively (see below). We

then also take the average of this normalized value over each of the synthetic galaxies so that we may compare sets of CLUSTAR-ND clusterings to sets of ground truth clusterings. Let  $\mathcal{S}$  represent some feature space combination on which the predicted clusters are dependent, then the resultant objective function we use is defined as

$$F(\mathcal{S}) = \frac{1}{N_{\text{galaxies}}} \sum_{g \in \text{galaxies}} F_g(\mathcal{S}), \text{ where}$$

$$F_g(\mathcal{S}) = \frac{I(C_g^*(\mathcal{S}); T_g) - E[I(R_g^*(\mathcal{S}); T_g)]}{H(T_g) - E[I(R_g^*(\mathcal{S}); T_g)]}. \quad (14)$$

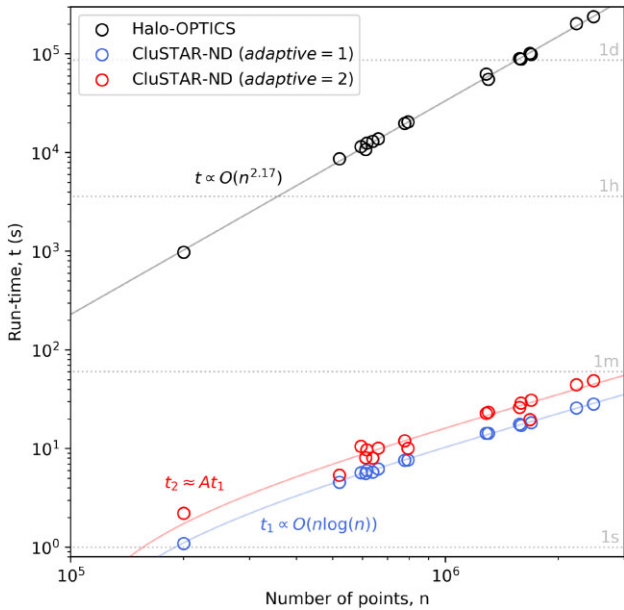
Here,  $H(T_g)$  is the entropy of  $T_g$  (this normalizes the maximum of  $F_g(\mathcal{S})$  to 1 since  $I(T_g; T_g) = H(T_g)$ ) and  $E[I(R_g^*(\mathcal{S}); T_g)]$  is the expected mutual information that arises when a clustering,  $R_g^*(\mathcal{S})$ , with equal cluster sizes as  $C_g^*(\mathcal{S})$  has been created via random assignment (this normalizes the minimum of  $F_g(\mathcal{S})$  to 0 since no clustering algorithm can ever do worse than random assignment). Unlike Vinh et al. (2009) (who relies on a hypergeometric model), we calculate  $E[I(R_g^*(\mathcal{S}); T_g)]$  empirically by taking the average of the mutual information that is found when points have been randomly assigned to a mock clustering,  $R_g^*(\mathcal{S})$ , with the number of clusters and number of points within each cluster being equal to those in  $C_g^*(\mathcal{S})$ . This average is calculated using 10 random realizations of  $R_g^*(\mathcal{S})$  which we find to be sufficient since the variance of  $I(R_g^*(\mathcal{S}); T_g)$  is small compared to its expectation.

The normalization adjustments we have made mean that  $F_g(\mathcal{S})$  can now be interpreted as the proportion of *relevant* information obtained about the true clusters of a galaxy by having observed the predicted leaf clusters produced from CLUSTAR-ND when applied to that galaxy. Similarly,  $F(\mathcal{S})$  is the average of this proportion across the galaxies. We use these mutual-information-based measures of clustering power throughout both Sections 6 and 7 as they provide the means to compare *entire* clusterings ( $F_g(\mathcal{S})$ ) and *sets* of entire clusterings ( $F(\mathcal{S})$ ).

While it is possible to construct an objective function with a similar purpose from the easy-to-interpret measures of recovery/purity/Jaccard index, it is non-trivial to reduce these to an appropriate scalar measure of clustering power as these are designed to be used to compare one cluster to another (rather than sets of clusters). Most critically, these measures do not account for true-negative predictions and so optimizing the parameters of CLUSTAR-ND (as is done throughout Section 6 with  $F(\mathcal{S})$ ) by using an objective function built from such measures will likely create an artificial incentive for CLUSTAR-ND to make fewer predictions on clusters. These measures also do not appropriately reward CLUSTAR-ND for having matched a true cluster with two predicted clusters – clearly this is not as nice as a perfect match, but it is better than a  $J_{\max}(C) = 0.5$  match for example. Nevertheless, maximizing  $F(\mathcal{S})$  does also maximize the recovery, purity, and Jaccard index that the predicted clusters will have – it just does so without promoting fewer predictions. In fact, a value of  $F_g(\mathcal{S}) = f$  is approximately equivalent to having  $\sim f \times |T_g|$  true clusters matched perfectly ( $J_{\max}(C) = 1$ ) by CLUSTAR-ND, of course, this relation is not one-to-one as  $F_g(\mathcal{S}) = f$  could also occur via a larger number of predicted clusters with smaller  $J_{\max}(C)$  values – although, by adjusting  $F_g(\mathcal{S})$  for random chance allocation we have limited this trade-off to meaningful predictions.

## 5.2 Run-time disparity

We now compare run-times of CLUSTAR-ND and HALO-OPTICS using synthetic galaxies from GALAXIA in Fig. 5. All runs were performed using a single core on an Intel i5 vPro processor and



**Figure 5.** The run-times of both CLUSTAR-ND and HALO-OPTICS when applied to the 3D spatial dimensions of the synthetic galaxies from GALAXIA. The lines of best fit correspond to an  $O(n \log(n))$  time complexity for CLUSTAR-ND and an  $O(n^{2.17})$  time complexity for HALO-OPTICS. We see that even when applied to the smallest size galaxy ( $\sim 2 \times 10^5$  points) in the suite CLUSTAR-ND is 3 orders of magnitude faster than HALO-OPTICS, and due to the difference in time complexities, this speed increase balloons outwards to nearly 5 orders of magnitude when applied to the largest of these galaxies ( $\sim 2.5 \times 10^6$  points). While some of this run-time disparity can be attributed to changing clustering structure (see Section 5.2), the modest run-time of CLUSTAR-ND makes it ideally suited for application to large data sets.

both codes are written using PYTHON3 but do make use of optimized numerical packages such as NUMPY (Harris et al. 2020), SCIPY (Virtanen et al. 2020), SCIKIT-LEARN (Pedregosa et al. 2011), and NUMBA (CLUSTAR-ND only; Lam, Pitrou & Seibert 2015). We see in Fig. 5 that for these galaxies the run-times of CLUSTAR-ND are at least 3 orders of magnitude faster than those of HALO-OPTICS and this disparity grows for larger data sets. CLUSTAR-ND follows an  $O(n \log(n))$  time complexity while HALO-OPTICS appears to follow a  $O(n^{2.17})$  time complexity. This is not the true time complexity of HALO-OPTICS and is partially an artefact of the changing clustering structuring within the galaxies – a feature that affects both algorithms albeit affecting CLUSTAR-ND to a much lesser degree.

Given two data sets with an equal number of points, the run-time of HALO-OPTICS will be larger for the data set that has a larger range of densities within the virial radius i.e. is more cusp-like rather than core-like. This is because there will be a larger fraction of points whose mutual distance is less than or equal to  $\epsilon$ . Ordinarily, the best-case run-time of HALO-OPTICS is  $O(n \log(n))$ , however as the larger synthetic galaxies we use in this paper are more cusp-like, we find that the time complexity of HALO-OPTICS approaches  $O(n^2)$  with increasing  $n$ . So while it is likely that the run-time of HALO-OPTICS applied to each galaxy individually is sub-quadratic, the overall trend of increasing cuspyness gives rise to a super-quadratic time complexity of  $O(n^{2.17})$ .

As CLUSTAR-ND does not perform a radial search it does not suffer from this same drawback and moreover, the clustering structure does not strongly affect the run-time of the  $k_{\text{den}}$  nearest neighbour search and density computation either. In reality the run-time of CLUSTAR-ND in Fig. 2 is dominated by the  $k_{\text{den}}$  nearest neighbour search –

taking up a constant  $\sim 89$  per cent of the run-time for this number of dimensions and this value of  $k_{\text{den}}$ . In this implementation, CLUSTAR-ND uses SCIPY’s cKDTree (Bentley 1975; Maneewongvatana & Mount 1999; Virtanen et al. 2020) which has an expected build and search run-time of  $O(n \log(n))$ . Such a run-time complexity is theoretically the fastest that can be achieved for this problem and hence CLUSTAR-ND’s overall run-time only stands to improve by some constant factor. Faster run-times can be achieved this way by running the  $k_{\text{den}}$  nearest neighbour search in parallel, by using a faster implementation of the kd-tree, or by using an approximate nearest neighbour algorithm rather than an exact one.

In the event that these run-time improvements are used and the dimensionality of the data set is sufficiently low, it is possible that the run-time of  $k_{\text{den}}$  nearest neighbour search becomes small enough that it is no longer the most time consuming component of the algorithm. The next most time consuming component of the CLUSTAR-ND algorithm is the aggregation and rejection process in Fig. 3 which takes up a constant  $\sim 10$  per cent of the total run-time for each of the runs in Fig. 5. This run-time is predominantly affected by the merging of potential clusters. In the best-case run-time scenario, there will be no mergers and the time complexity for this part of the algorithm would be  $O(n)$ . While technically it is possible for a data set to be structured in such a way that gives rise to no mergers during the aggregation process, for astrophysical data, it would be highly atypical. Most if not all points in the data set will be densely connected together during this process. As the mergers are discovered they are combined in a tree-like structure. By merging the densely connected lists in a compact binary-tree structure the aggregation process will be performed in an  $O(n \log(n))$  time. This is both the expected and worst case time complexity. The time complexity of the aggregation process is not only affected by the way in which the densely connected lists are merged but also by how many times this occurs. The values  $k_{\text{den}}$  and  $k_{\text{link}}$  both affect this number and will increase the run-time by some constant factor the smaller that they are.

The remaining  $\lesssim 1$  per cent of CLUSTAR-ND’s run-time can be attributed to the transformation of the data and general data manipulation – the former of which has an  $O(nd^2 + d^3)$  time complexity. Overall, these constituents make the run-time complexity of CLUSTAR-ND  $O(n \log(n))$ . In Fig. 5, we do not show the run-time of CLUSTAR-ND when *adaptive* = 0 since this is only slightly different ( $\lesssim 1$  per cent different as above) from when *adaptive* = 1. We see that when *adaptive* = 2 the run-time increases by some factor per data set. In practice, this factor that is less than the number of levels found within the clustering hierarchy. When applied to the galaxies we analyse in this paper, the depth of the hierarchy is typically capped at 3 or 4 levels. CLUSTAR-ND’s fast run-time makes it ideal for application to large data sets such as the Gaia DR3 catalogue.

## 6 PARAMETER INFLUENCE AND OPTIMIZATION

The CLUSTAR-ND algorithm has 7 parameters. Parameters  $l_x$ , *adaptive*, and  $k_{\text{den}}$  may be chosen by the user and are responsible for the spatial linking length that finds field haloes, the adaptivity of the distance metric, and the number of neighbours used to find the density of the points, respectively. A typical choice for  $l_x$  is 0.2 times the interparticle spacing and then whether *adaptive* is set to 0, 1, or 2 is decided in consideration of the run-time versus clustering power trade-off as well as whether the user wishes to apply a PCA transform to the input data.

The parameter  $k_{\text{den}}$  has a more loosely constrained range – if it is too small then the density fluctuations between neighbouring points

can be large and noisy; if it too large then significant differences in local density may be smoothed out enough that some structures may not be detected by CLUSTAR-ND. Typically, similar parameters to  $k_{\text{den}}$  in other clustering algorithms are chosen with values ranging from 20 to 500. It is commonly accepted that  $k_{\text{den}}$  should increase when using data with a larger feature space – however, this depends on the intended resolution of the density estimate and, in a clustering context, the expected number of points within the true clusters present in the data. A more thorough investigation of the effects of the choice of  $k_{\text{den}}$  is carried out in Sections 6.2–6.4.

Although the remaining 4 parameters in CLUSTAR-ND may be chosen by the user, these do have properties that allow for their optimal values to be determined. It should be mentioned that these parameters do have some co-dependence with regards to the state of the output, though this influence is small. Due to the fact that none of these parameters act within the same step of the algorithm, we are able to isolate their effects in order to find their optimal values. Furthermore and since the labels within the synthetic data provide a flat clustering without noise, we opt to assess the performance of CLUSTAR-ND – and hence the corresponding optimal values of some of the cluster extraction parameters ( $\rho_{\text{threshold}}$  and  $S_{\text{outlier}}$ ) – by determining how well the leaf clusters of CLUSTAR-ND are able to match a single label within the synthetic data sets.

While the affects of the cluster extraction parameters in CLUSTAR-ND are expected to be similar to those in HALO-OPTICS, the lists of points that they act upon are generally slightly different. The near-optimal values of HALO-OPTICS parameters were determined by maximizing the recovery and purity fractions of a set of hand-crafted 3D Gaussian distributions in both Cartesian and spherical coordinate systems. While this technique can be used to mimic returning the optimal set of astrophysical clusters, it of course is not equivalent to doing so. Furthermore, this investigation did not analyse the effects of higher dimensions and/or other subspaces on these parameters. We now investigate the effects of these parameters and deduce their preferred values.

### 6.1 Values for $k_{\text{link}}$

The parameter  $k_{\text{link}}$  is the number of neighbours used to aggregate the points together in order of decreasing density so that the substructure may be found. The consequence of decreasing  $k_{\text{link}}$  is that the clusters will be more complete up until their true boundary density. Irrespective of this, however, a practical lower limit exists. During the aggregation process, multiple potential clusters will be connected together into a new potential parent cluster whenever the current point being processed has multiple of its  $k_{\text{link}}$  nearest neighbours in different potential clusters. This has the consequence that at most  $k_{\text{link}} - 1$  potential clusters can be connected together during this step (since the point being processed is one of its own neighbours and does not yet belong to any potential cluster). The number of potential clusters that can be joined depends on the specific arrangement of points – although with a sufficiently large value of  $k_{\text{link}}$ , the number of these clusters that are observed to be joined together in these steps is most commonly 2, occasionally 3, and rarely 4. The is also a subtle dependency on  $k_{\text{den}}$  as well, however, the number of merging clusters only starts to increase for small values of this parameter i.e.  $<20$ . Given that typically  $k_{\text{den}} \geq 20$ , we suggest that for the choice of  $k_{\text{link}}$  not to affect this behaviour, its value needs to be at least as large as 5 or 6.

Furthermore, we must also consider that certain arrangements of points may give arise to the issue whereby the  $k_{\text{link}}$  nearest neighbours of each point is not an extensive enough neighbourhood for the process to be able to seamlessly aggregate most, if not all, points

together. Densely connecting all points via their neighbourhood in a data set is not always feasible. In particular, if all points in the data set are contained within dense clusters that are each sparsely distributed with respect to each other then connecting all points may require a value of  $k_{\text{link}}$  that approaches the size of the data set itself.

To manage this issue, we choose practical values for  $k_{\text{link}}$  by ensuring that CLUSTAR-ND is able to densely connect all points given that they are from a randomly sampled uniform distribution in  $d$ -dimensions when it is using a particular value for  $k_{\text{den}}$ . We vary both  $d$  consecutively from 1 to 9 and  $k_{\text{den}}$  within the set  $\{20, 30, 40, 60, 80, 120, 160, 240, 360, 480\}$ . For each  $d$ - $k_{\text{den}}$  combination we create 300 random samples of  $10^5$  points from a  $d$ -dimensional uniform distribution within the unit hypercube. For each of these we find the smallest value of  $k_{\text{link}}$  that will densely connect all  $10^5$  points together by the end of the aggregation process. The histograms of these smallest values of  $k_{\text{link}}$  can be seen in the top panel of Fig. 6 for various combinations of  $d$  and  $k_{\text{den}}$ . Within these histograms a value of  $k_{\text{link}} = x$  will contribute to the height of the bar in the interval  $(x - 1, x]$ .<sup>7</sup>

We then fit a skew-normal such that the sum of squared differences between the probability within each of the histogram’s intervals and the probability of the skew-Gaussian within those same intervals is minimized. The skew-normal distribution we fit has the standard form of

$$f(k_{\text{link}}) = \frac{2}{\omega} \phi\left(\frac{k_{\text{link}} - \xi}{\omega}\right) \Phi\left(\alpha \frac{k_{\text{link}} - \xi}{\omega}\right), \text{ where}$$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \text{ and}$$

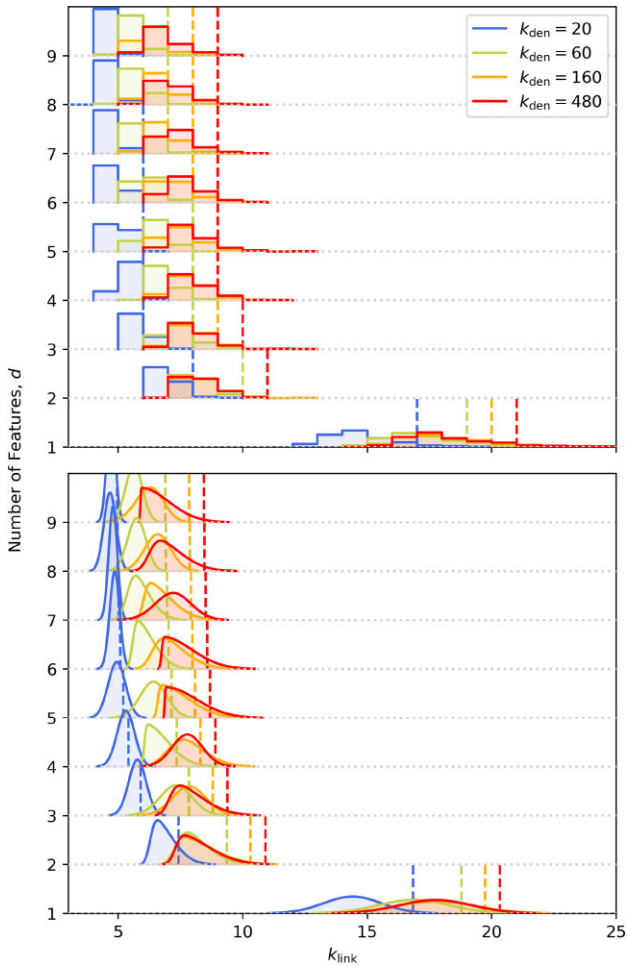
$$\Phi(x) = \int_{-\infty}^x \phi(y) dy = \frac{1}{2} \left[ 1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right) \right]. \quad (15)$$

Here,  $\xi$ ,  $\omega$ , and  $\alpha$  are the location, shape, and scale parameters, respectively. The continuous distributions shown in the bottom panel of Fig. 6 depicts the fitted skew-normal distributions for various combinations of  $d$  and  $k_{\text{den}}$ .

We now use the fitted skew-normal distributions to find the continuous analogue value for  $k_{\text{link}}$  that predicts that all  $10^5$  points within each uniform distribution will be densely connected together 95 per cent of the time. Once found, we then fit a function of the form  $k_{\text{link}, 95} = \alpha d^\beta + \gamma k_{\text{den}}^\delta + \epsilon$  by minimizing the squared differences between these distributional 95th percentile values and the function’s output – finding that  $\alpha \approx 12.0$ ,  $\beta \approx -2.2$ ,  $\gamma \approx -23.0$ ,  $\delta \approx -0.6$ , and  $\epsilon \approx 9.0$ . The predicted continuous analogue value of  $k_{\text{link}, 95}$  are shown as dashed vertical lines in the bottom panel of Fig. 6 for various combinations of  $d$  and  $k_{\text{den}}$ . We then convert these back to the original discrete regime by rounding up to the nearest integer. Likewise, the discrete values of  $k_{\text{link}, 95}$  are shown as dashed vertical lines in the top panel of Fig. 6 for various combinations of  $d$  and  $k_{\text{den}}$ .

The fitted values of  $k_{\text{link}, 95}$  in the continuous analogue do not always correspond well to the true 95th percentile of the fitted skew-normal distributions, however, we find that they do correspond well to the empirical 95th percentile of histograms produced over each combination of  $d$  and  $k_{\text{den}}$ . To be more certain that the automation of  $k_{\text{link}}$  will not artificially break densely connected regions we add 1 to these discrete values of the 95th percentile and – to keep the potential cluster merging behaviour unaffected as discussed above – always ensure that it is larger than 7. To be clear, unless the user specifies a

<sup>7</sup>Constructing the histogram in this way helps to provide a continuous analogue of the probability distribution since in such an analogue the value of  $k_{\text{link}} = x$  will have been drawn from this interval.



**Figure 6.** The empirical distributional relationships between the minimum  $k_{\text{link}}$  required to connect all  $10^5$  points of a  $d$ -dimensional uniform distribution contained to the unit hypercube given that  $k_{\text{den}}$  nearest neighbours are used to estimate the local density. The value of  $d$  is varied from 1 to 9 and the value of  $k_{\text{den}}$  is varied within the set  $\{20, 30, 40, 60, 80, 120, 160, 240, 360, 480\}$  – however, only specific values of the latter are shown. The histograms in the top panel are the binned minimum  $k_{\text{link}}$  values and are created from 300 re-samplings of the aforementioned data distribution for each  $d$  and  $k_{\text{den}}$  combination. The skew-Gaussian distributions in the bottom panel are the fitted continuous analogues of the histograms in the top panel. The dashed lines in the lower panel are the 95<sup>th</sup> percentile values of the fitted skew-Gaussian distributions and the dashed lines in the top panel are the discrete analogue of the latter.

value for  $k_{\text{link}}$  CLUSTAR-ND calculates it using automatically using

$$k_{\text{link}} = \max\{\text{ceil}(12.0d^{-2.2} - 23.0k_{\text{den}}^{-0.6} + 10.0), 7\}. \quad (16)$$

Constructing  $k_{\text{link}}$  this way standardizes the degree with which it is possible to connect the points of a data set whilst also ensuring maximal cluster completeness and as such these calculated values serve as a practical lower limit for  $k_{\text{link}}$ . The user may choose different values for  $k_{\text{link}}$ , although there are some considerations. The value of  $k_{\text{link}}$  must always be smaller than or equal to that of  $k_{\text{den}}$  – which applies to the user’s choice of  $k_{\text{den}}$  just as it does to that of  $k_{\text{link}}$ . In the event that  $k_{\text{link}}$  is chosen automatically by CLUSTAR-ND the above optimal values are effectively the lower limits of  $k_{\text{den}}$  given the number of dimensions  $d$ . The value of  $k_{\text{link}}$  could be chosen to be larger than the automated values for the purpose of decreasing the run-time, however, this comes at the detriment of cluster completeness. Other than these considerations, the optimal

values of  $k_{\text{link}}$  do not depend on the choice of  $k_{\text{den}}$  nor any other parameter from CLUSTAR-ND as they have been determined via the simple criterion that all points in a  $d$ -dimensional uniform distribution be densely connected.

## 6.2 Values for $\rho_{\text{threshold}}$

The parameter  $\rho_{\text{threshold}}$  is used to provide the functionality of determining which groupings of points are *significantly* dense with respect to their surrounds. More specifically, it is responsible for ensuring that the median density of any cluster is at least  $\rho_{\text{threshold}}$  times the density of that cluster’s surrounds – refer to equation (4).

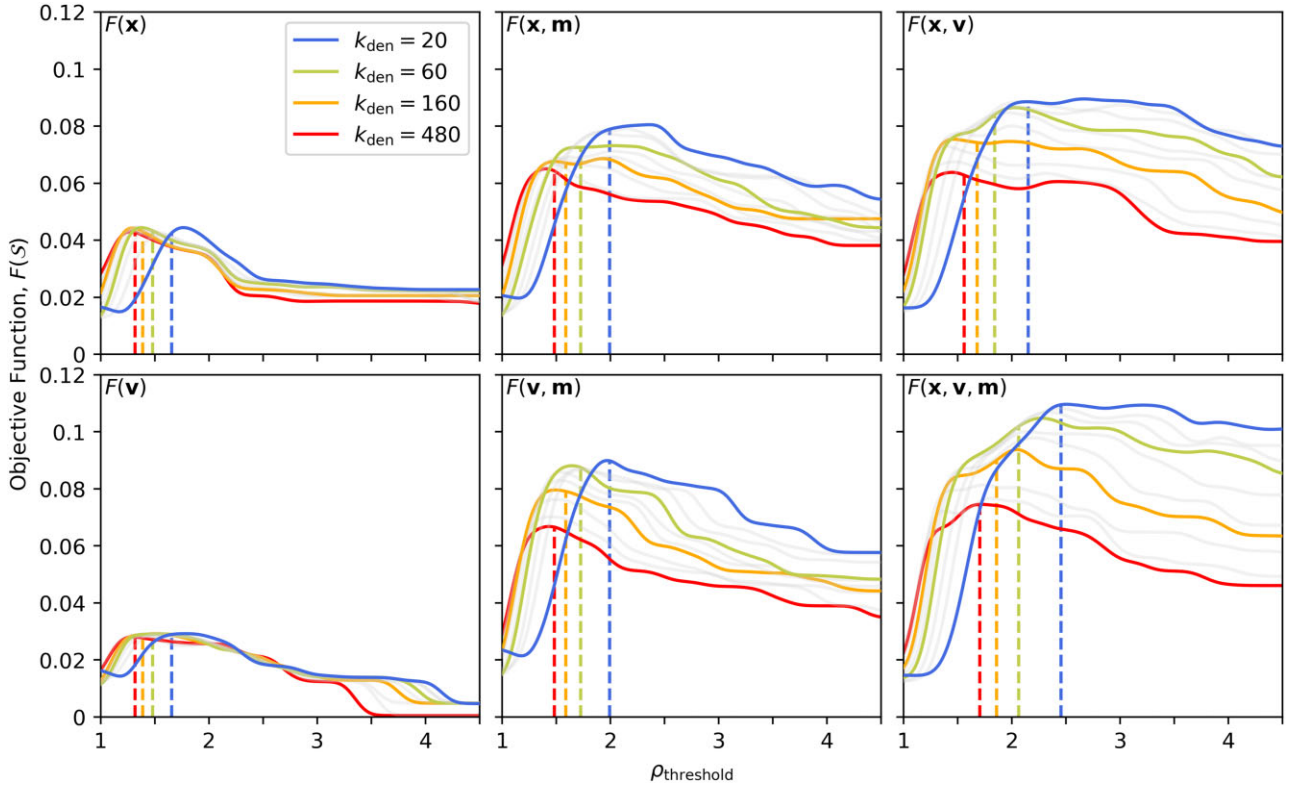
For its use in CLUSTAR-ND, we take a closer look at the  $\rho_{\text{threshold}}$  parameter by optimizing it on the synthetic galaxies described in Section 4. Since the labels within these data sets only provide a flat clustering of the synthetic galaxies without noise – instead of a hierarchical clustering with noise as is the case with CLUSTAR-ND’s output – we only use the leaf clusters found by CLUSTAR-ND in order to determine its performance with varying  $\rho_{\text{threshold}}$ . This approach then requires  $f_{\text{reject}}$  to be sufficiently large – to ensure that the largest leaf cluster shares less than  $f_{\text{reject}}$  of its points with the root cluster – so that the leaf clusters will be unaffected by the choice of  $f_{\text{reject}}$ . In fact, this poses a practical lower limit to  $f_{\text{reject}}$ , however, we revisit this with greater detail in Section 6.4. Here, we set  $f_{\text{reject}} = 1$ . Since the effect of  $S_{\text{outlier}}$  on the final clusters is only minimal, we simply set  $S_{\text{outlier}} = \infty$  to remove its influence – we take a closer look at  $S_{\text{outlier}}$  in Section 6.3. The  $k_{\text{link}}$  parameter is automatically chosen and accordingly to the scheme outlined in Section 6.1. For simplicity, we also optimize these parameters using  $\text{adaptive} = 1$ .

By setting the above parameters in this way, we are able to find the optimal values of  $\rho_{\text{threshold}}$  as they depend on  $k_{\text{den}}$  and the number of features,  $d$ . We vary  $\rho_{\text{threshold}}$  from 1 to 4.5 in intervals of 0.1 and  $k_{\text{den}}$  again from the set of  $\{20, 30, 40, 60, 80, 120, 160, 240, 360, 480\}$ . For each combination of  $\rho_{\text{threshold}}$  and  $k_{\text{den}}$ , we run CLUSTAR-ND over each of the various feature subspace combinations out of the spatial ( $\mathbf{x}$ ), kinematic ( $\mathbf{v}$ ), and chemical ( $\mathbf{m}$ ) subspaces.<sup>8</sup> Then for every run, we compare the leaf clusters found by CLUSTAR-ND to the ground truth labels from the relevant synthetic galaxy. We use the mutual-information-based objective function,  $F(S)$ , introduced in Section 5.1.1 and equation (14) to assess the *clustering power* of CLUSTAR-ND for each  $\rho_{\text{threshold}} - k_{\text{den}} - \text{subspace}$  combination.

Fig. 7 contains a panel for each feature subspace combination whereby  $F(S)$  is plotted against the values of  $\rho_{\text{threshold}}$  for each value of  $k_{\text{den}}$  (with specific values shown in colour). As is mentioned in Section 4, the spatial and kinematic features ( $\mathbf{x}$  and  $\mathbf{v}$ ) are each  $d = 3$

<sup>8</sup>Note that we do not analyse the clusterings over the  $\mathbf{m}$ -subspace alone as typically each data point shares the same values of both  $[\text{Fe}/\text{H}]$  and  $[\alpha/\text{Fe}]$  with a number of other data points. Producing clusterings over the  $\mathbf{m}$ -subspace thereby artificially predicts many thousands of spurious clusters that only exist in this context because each point they contain is effectively a duplicate – this dramatically increases the estimation of their local density and the density contrast between *nearby* neighbourhoods.

It is likely that this will also affect the clusterings in higher dimensional feature spaces that contain the  $\mathbf{m}$ -subspace to some extent. However, since the number of points per value of the  $([\text{Fe}/\text{H}], [\alpha/\text{Fe}])$  tuple is small (compared to the total number of points) in each synthetic galaxy, the global Mahalanobis metric used in CLUSTAR-ND here is effectively blind to the fact the data sets in higher dimensions are made up many hyperplanes separated along the  $[\text{Fe}/\text{H}]$  and  $[\alpha/\text{Fe}]$  directions. Specifically, the blindness that CLUSTAR-ND has to this bias in higher dimensional settings is a result of the PCA transformation where the resultant interpoint spacing within each local hyperplane is on the order of the interplane spacing between each hyperplane.



**Figure 7.** The clustering power of CLUSTAR-ND as it depends on the various feature subspaces,  $k_{\text{den}}$ , and  $\rho_{\text{threshold}}$ . The objective function used here is defined in equation (14); however, the plots show a Gaussian-smoothed version of this for simplicity. The dashed lines indicate the fitted values of  $\rho_{\text{threshold}}$  as a function of the number of features,  $d$ , and the value of  $k_{\text{den}}$ . These fitted values are calculated using equation (17).

subspaces and the chemical features ( $\mathbf{m}$ ) is a  $d = 2$  subspace. From these plots we can clearly see an interdependency between  $\rho_{\text{threshold}}$ ,  $k_{\text{den}}$ , and the number of features,  $d$ . As such, we wish to find a function that describes the optimal values for  $\rho_{\text{threshold}}$  such that this function is monotonically increasing with  $d$  and monotonically decreasing with  $k_{\text{den}}$ . We also need for  $\rho_{\text{threshold}}$  to always be at least 1 as there is no reason for choosing otherwise.

In light of these factors and  $\rho_{\text{threshold}}$ 's coupled dependency on  $k_{\text{den}}$  and  $d$  (unlike the interdependency between these variables and  $k_{\text{link}}$  investigated in Section 6.1), we choose to fit a function of the form  $\rho_{\text{threshold}}(k_{\text{den}}, d) = 1 + \alpha f(d)/g(k_{\text{den}})$  where  $f$  and  $g$  are monotonically increasing and strictly non-zero for positive inputs. We consider each power-law/logarithmic form combination for  $f$  and  $g$ , but ultimately find that setting  $f(d) = d^\beta$  and  $g(k_{\text{den}}) = \ln(k_{\text{den}})$  gives the best overall fit. In fitting, we maximize the function's output on all Gaussian-smoothed objective function curves (for all values of  $k_{\text{den}}$  not just those that are coloured) shown in Fig. 7 simultaneously with an equal weighting placed on each. Following this, we find that the optimal values of  $\rho_{\text{threshold}}$  are well-described by the function

$$\rho_{\text{threshold}}(k_{\text{den}}, d) = 1 + 0.81 \frac{d^{0.81}}{\ln(k_{\text{den}})}. \quad (17)$$

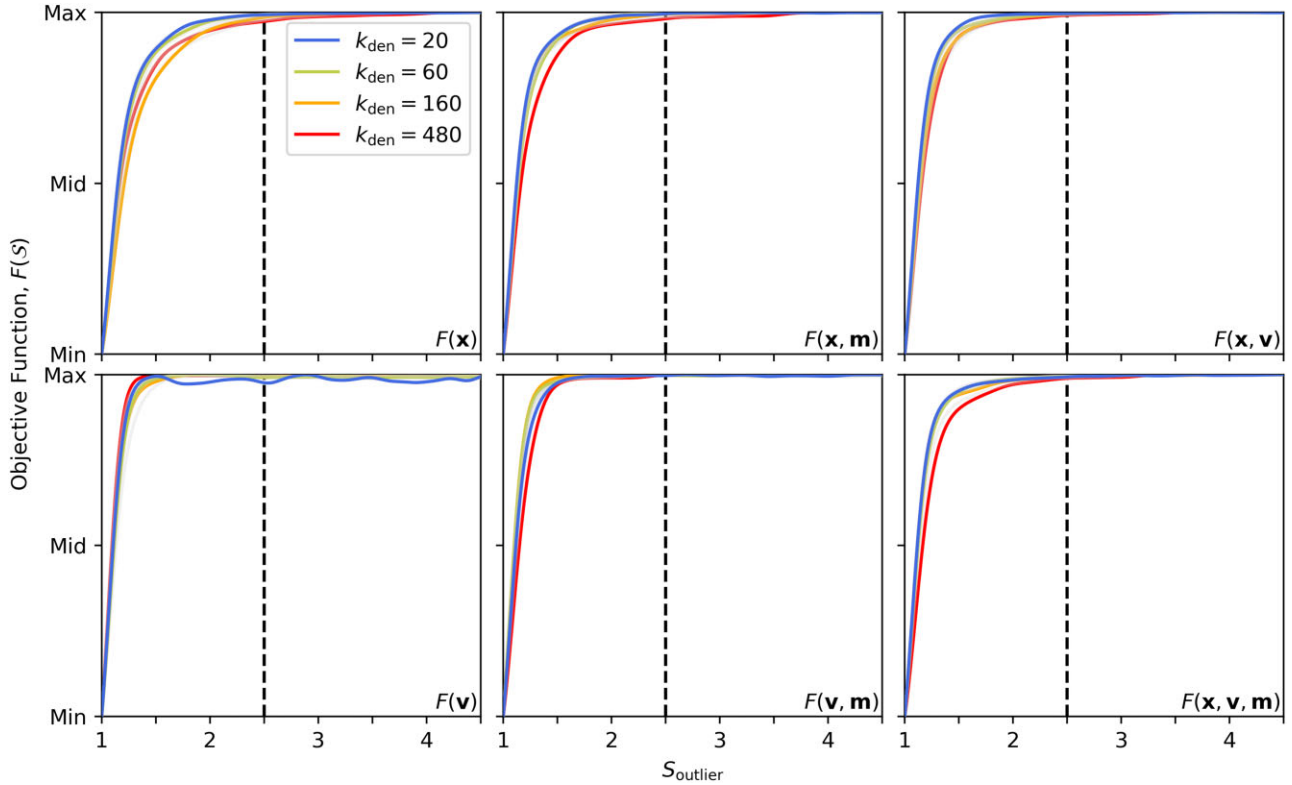
Some fitted values produced by equation (17) are shown in Fig. 7 as vertical dashed lines with colours that correspond to the appropriate value of  $k_{\text{den}}$ . While these fitted values are not always at the global maxima of the objective function curves, the function does not sacrifice much relevant information for the added convenience of having this parameter be chosen automatically from the user's choice of  $k_{\text{den}}$  and the number of features,  $d$ , in the data set that the user wishes to produce a clustering from.

In addition to being able to justify a set of optimal choices for  $\rho_{\text{threshold}}$ , we can also see from Fig. 7 the difference in information content as it depends on the specific feature subspace present within the input data. Here, we can easily see that including extra informative dimensions will increase clustering power. We can also see that smaller  $k_{\text{den}}$  values produce better clusterings in an astrophysical context since they increase the resolution with which CLUSTAR-ND can resolve structure. This is particularly prevalent in the larger feature spaces where fine-structure can be separated from background noise more easily – however choosing small values for  $k_{\text{den}}$  comes with diminishing returns as this also increases the effects of Poisson noise on the density estimation which can artificially introduce spurious structures. We take a closer look at the robustness of the CLUSTAR-ND output by comparing recovery and purity statistics of specific clusters across these same feature subspace combinations in Section 7.

### 6.3 Values for $S_{\text{outlier}}$

The parameter  $S_{\text{outlier}}$  is used to categorize the points of the input data as either local inliers or local outliers depending on whether their local-outlier-factor (or rather its density kernel analogue in equation (8)) is less than or greater than  $S_{\text{outlier}}$ . Following the aggregation process described in Section 3.5, a cut-off density is then defined for each cluster according to equation (9). This classification is then used to remove all points from each cluster whose local density is less than the cut-off density,  $\rho_{\text{cut}}$ , of that cluster – i.e. only local outliers whose density is less than that of all local inliers are removed.

To better understand the effect  $S_{\text{outlier}}$  has on the final clustering produced and gauge an optimal value for it from the syn-



**Figure 8.** The clustering power of CLUSTAR-ND as it depends on the various feature subspaces,  $k_{\text{den}}$ , and  $S_{\text{outlier}}$ . The objective function used here is defined in equation (14) and is re-normalized between each curve’s respective minimum and maximum over this range for  $S_{\text{outlier}}$  to better depict the global influence of this parameter. As with Fig. 7, these plots show a Gaussian-smoothed version of the objective function. We see that for small  $S_{\text{outlier}}$  values the clustering power decreases dramatically and for large values the clustering power asymptotically approaches a maximum (values of which are those shown in Fig. 7 at the relevant  $\rho_{\text{threshold}}$  values for each  $k_{\text{den}}$  and number of features,  $d$ ). The dashed lines indicate the value of  $S_{\text{outlier}} = 2.5$  which we set as the standard choice for CLUSTAR-ND. This value is both small enough to remove distinct outliers from clusters while being large enough to not have the resultant clustering breakdown. Ultimately, however, any value above about  $S_{\text{outlier}} = 1.5$ – $2$  will produce good clustering results – adjusting this further is dependent on how strict the user wishes to be on the removal of potential outliers from clusters.

thetic galaxies discussed in Section 4, we compute the objective function,  $F(S)$ , in equation (14) from the CLUSTAR-ND output for various combinations of  $S_{\text{outlier}}$ ,  $k_{\text{den}}$ , and feature subspace in largely the same way as in Section 6.2. We vary  $S_{\text{outlier}}$  from 1 to 4.5 in intervals of 0.1. We similarly vary  $k_{\text{den}}$  from the set of  $\{20, 30, 40, 60, 80, 120, 160, 240, 360, 480\}$  and the feature subspace combinations of the spatial ( $\mathbf{x}$ ), kinematic ( $\mathbf{v}$ ), and chemical ( $\mathbf{m}$ ) subspaces. Given these combinations of  $k_{\text{den}}$  and the number of features,  $d$ , we allow for  $k_{\text{link}}$  and  $\rho_{\text{threshold}}$  to be automatically chosen according to equations (16) and (17), respectively. In our investigation of  $S_{\text{outlier}}$  we again set  $f_{\text{reject}} = 1$  and  $\text{adaptive} = 1$ .

The relationship between the clustering power of CLUSTAR-ND and  $S_{\text{outlier}}$  is shown in Fig. 8 for each combination of  $k_{\text{den}}$  and  $d$ . More simplistic than that between the clustering power and  $\rho_{\text{threshold}}$ , the relationship here is very similar between these combinations and suggests that so long as  $S_{\text{outlier}}$  is *large enough* the CLUSTAR-ND output will be robust.<sup>9</sup> For values larger than about 1.5 or 2, the

<sup>9</sup>The  $F(\mathbf{v})$  panel of Fig. 8 appears to include a oscillation-like pattern for  $S_{\text{outlier}} \gtrsim 1.5$  and  $k_{\text{den}} = 20$ . This is an artefact of the objective function minimum being closer to its maximum for this combination. As such, the scale of the noise appears much larger and noticeable compared to the other  $k_{\text{den}}$  and  $d$  combinations. The exact reasoning for this noise is that for a given change in  $S_{\text{outlier}}$  the quality of some clusters will increase while others may decrease. These competing effects take place for every curve in Fig. 8 but are

choice of  $S_{\text{outlier}}$  has a minimal effect and will only serve to remove a select few or perhaps none of the points from each cluster. Of course changing  $S_{\text{outlier}}$  from 3 and 4 will still have an effect on which points are apart of the final clusters, however, this effect is minimal and cannot be robustly defined beyond the resultant clustering include more potential outliers per cluster than beforehand. In light of this ambiguity, we choose to use  $S_{\text{outlier}} = 2.5$  as the standard value within CLUSTAR-ND as this will provide strong clustering results with a moderate level of outlier detection.

#### 6.4 Values for $f_{\text{reject}}$

The parameter  $f_{\text{reject}}$  defines the maximum fraction of points (relative to the parent cluster) that can be shared between any of the parent-child cluster pairs that exist immediately following the aggregation process. A cluster is rejected from a parent-child pair if the child shares more than  $f_{\text{reject}}$  of the parent’s points – according to equation (6) – i.e. CLUSTAR-ND removes the smallest cluster of the pair that has child clusters of its own, unless it would be a root-level cluster in which case the child from the pair is removed. These rules mean that  $f_{\text{reject}}$  is implicitly responsible for determining the size and shape of the hierarchy between the terminating clusters of the hierarchy.

simply more noticeable on this curve due to the imposed scale between the minimum and maximum.

**Table 1.** A summary of the minima, medians, means, and maxima of the normalized mutual information values ( $F_g(\mathcal{S})$ ) for the various feature spaces ( $\mathcal{S}$ ) and *adaptive* settings shown within Fig. 9. The mean values are equivalent to the objective function defined  $F(\mathcal{S})$  in equation (14). For completeness, we also include these statistics for the clusterings produced with HALO-OPTICS using the  $(\mathbf{x})$  feature space in Section 5.

$\mathcal{S}$	<i>Adaptive</i>	Min	Median	Mean	Max
$(\mathbf{x})$	HALO-OPTICS	0	0.027	0.043	0.198
	0	0.001	0.028	0.043	0.197
	1	0.006	0.027	0.043	0.193
	2	0	0.023	0.039	0.193
$(\mathbf{v})$	0	0	0.009	0.029	0.261
	1	0	0.008	0.029	0.266
	2	0	0.006	0.028	0.265
$(\mathbf{x}, \mathbf{m})$	1	0.024	0.062	0.079	0.364
	2	0.017	0.066	0.075	0.341
$(\mathbf{v}, \mathbf{m})$	1	0	0.063	0.093	0.473
	2	0	0.070	0.102	0.469
$(\mathbf{x}, \mathbf{v})$	1	0	0.065	0.089	0.445
	2	0	0.067	0.092	0.443
$(\mathbf{x}, \mathbf{v}, \mathbf{m})$	1	0	0.077	0.110	0.485

Refer to Section 3.6 for more details on the rules surrounding the  $f_{\text{reject}}$  parameter.

Choosing extreme  $f_{\text{reject}}$  values of 0 and 1 would force CLUSTAR-ND to return ( $\leq^{10}$ ) $l$  and  $2(l - 1)$  clusters per root-level cluster, respectively – where  $l$  is the number of leaf clusters within that root-level cluster. In the same scenarios, the hierarchy returned by CLUSTAR-ND would have depths of 1 and somewhere in the range of  $[\log_2(l - 1), l - 1]$  per root-level cluster, respectively. To use the standard terminology, these values of  $f_{\text{reject}}$  create hierarchies that form perfect ( $\leq$ ) $l$ -ary trees and full binary trees, respectively – with the exact shape of the latter dependent upon  $l$ 's representation in base-2 and the clustering structure within the data set. Varying  $f_{\text{reject}}$  between these values transforms the shape of the hierarchy in a way is difficult to predict.

We should expect that a reasonable lower limit for  $f_{\text{reject}}$  is 0.5, since anything lower would suggest that equally sized groups can not be merged into a parent cluster. However, further constraining this within this paper would be a difficult task since the satellite labels within the synthetic data from GALAXIA (Section 4) provide a flat clustering without noise of each galaxy and this information alone does not indicate the optimal hierarchy shape. To properly determine the optimal  $f_{\text{reject}}$  from this data, we would need to assess the physics of the galaxies and their constituents – to determine how bound each satellite is to each other, and prepare an objective function that can be used to infer the most physical sense of what a *satellite within a satellite* should look like in an astrophysical context. Specifically, this would require analysis of the mass, orbits of the points, backwards integration, and analysis of the star formation histories within satellites and how this relates to each other satellite within a galaxy. Additionally, we cannot even be certain from the outset that we would see a single most-optimal value for  $f_{\text{reject}}$  that is able to consistently provide such a hierarchy because the rules that govern this may not even be suitable to do so.

<sup>10</sup>The  $\leq$  sign here indicates that the number of leaf clusters found in the  $f_{\text{reject}} = 0$  case may be less than or equal to the number of leaf clusters found in the  $f_{\text{reject}} = 1$ . The inequality will exist if there is one or more leaf clusters whose parent is the root-level cluster.

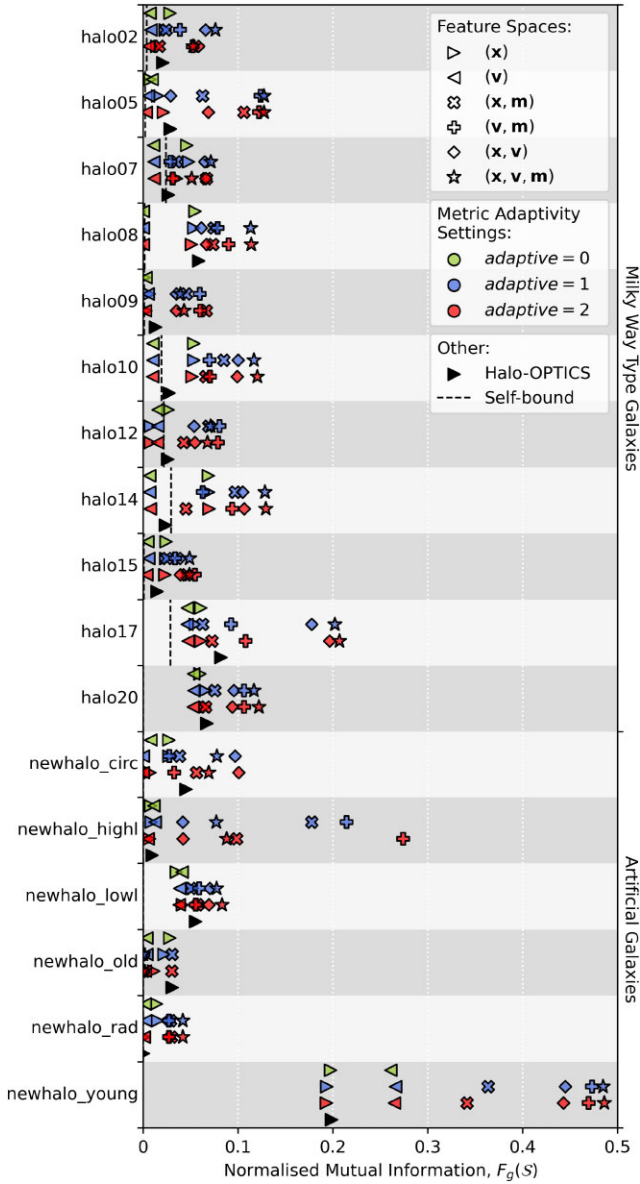
In light of these issues – and since we expect the hierarchy produced by CLUSTAR-ND to be similar to that of HALO-OPTICS – we choose to follow the verdicts with respect to the parameter  $f_{\text{reject}}$  in Oliver et al. (2020). Considering the results from Fig. 6 therein, a value of  $f_{\text{reject}} = 0.9$  is near-optimal and ensures good quality clustering results (high Jaccard index as well as high recovery and purity fractions) of the hierarchical combinations of hand-crafted leaf clusters. From a hierarchy interpret-ability stand point, running CLUSTAR-ND with this value of  $f_{\text{reject}}$  over the synthetic galaxies discussed in Section 4 creates consistently shaped hierarchies with levels  $L = 1, 2,$  and  $3 +$  containing  $\sim 70, \sim 20,$  and  $\sim 10$  per cent of all galactic substructure clusters ( $L > 0$ ). Furthermore, these percentages do not seem to change more than  $\sim 15$  per cent regardless of the value of  $k_{\text{den}}$ , the feature space of the input data, nor the choice synthetic galaxy – implying that we should expect the resultant hierarchy shape to remain simple to interpret and predictable when CLUSTAR-ND is applied to astrophysical data sets.

## 7 INFORMATION CONTENT OF DIFFERENT CLUSTERING SCENARIOS

In the preceding sections, we have presented the CLUSTAR-ND algorithm (Section 3), showing that it produces comparable clusters to HALO-OPTICS in a far more computationally efficient manner (Section 5), and optimized its parameters for generalized application to  $N$ -dimensional data sets (Section 6). In this section, we will further demonstrate the clustering power of CLUSTAR-ND and quantify the difference between the clusters that result when CLUSTAR-ND is applied to different galactic environments (Section 7.1), to different feature spaces (Section 7.2), and using metric adaptivity settings (Section 7.3).

The clustering power of the now-optimized CLUSTAR-ND is summarized in Table 1 and Fig. 9: Fig. 9 illustrates the breakdown of the objective function from equation (14) (i.e.  $F_g(\mathcal{S})$ ) over each synthetic galaxy from Section 4 – the measure of normalized mutual information is each galaxy's contribution to the objective function prior to having averaged over them; Table 1 displays the minima, medians, means, and maxima of the distribution of  $F_g(\mathcal{S})$  as it depends on the metric adaptivity setting and the feature space. With Fig. 9 we can also see how the clustering power differs depending on the galactic environment as well. While these values may seem low, both Table 1 and Fig. 9 show that CLUSTAR-ND has performed well given the circumstances of this data. As a reference point, Fig. 9 also shows the clustering power of a hypothetical clustering that exclusively and perfectly classifies ( $J_{\text{max}} = 1$ ) the self-bound satellites from within each MW-type galaxy. We see that in comparison, CLUSTAR-ND performs well and that the galaxies are composed of mostly disrupted tidal debris from previous mergers – which is mostly impossible to aggregate together without joining multiple unbound groups and effectively reducing our understanding of the catalogue of true satellites. We see that even with CLUSTAR-ND only using the positions ( $\mathcal{S} = (\mathbf{x})$ ) it will consistently return the self-bound satellites and some tidal debris to the effect of having an  $F_g(\mathcal{S})$  value that is at least a factor of  $\sim 1$ – $2$  times larger than for just classifying the self-bound satellites. This is unsurprising since HALO-OPTICS was shown to be able to do the same thing with only positional information (Oliver et al. 2020). As the feature space size increases CLUSTAR-ND's clustering power increases too such that when  $\mathcal{S} = (\mathbf{x}, \mathbf{v}, \mathbf{m})$ ,  $F_g(\mathcal{S})$  is typically at least a factor of  $\sim 3$ – $6$  times larger than that for self-bound satellites only.

All clusters analysed in this section have been produced using  $k_{\text{den}} = 20$  – as this consistently produced the best results in Section 6.



**Figure 9.** The normalized mutual information  $F_g(S)$  from equation (14) of various CLUSTAR-ND clustering scenarios involving different synthetic galaxies, feature spaces, and metric adaptivity settings. The  $F_g(S)$  values of both the HALO-OPTICS clusterings from Section 5 and a hypothetical set of clusterings that perfectly classify ( $J_{\max} = 1$ ) only the self-bound satellites within each MW type galaxy are also shown for reference. The vertical axes indicate the names of the synthetic galaxies as well as whether they are from the original MW-type galaxies of Bullock & Johnston (2005) or from the additional artificial galaxies of Johnston et al. (2008). It is easy to see that the galactic environ, the number and type of informative features, and the metric adaptivity all affect the power of the resultant clustering.

All other parameters are chosen automatically according to the analysis in Section 6 unless specified otherwise. We also highlight the impact of feature space and adaptivity setting using two satellites, shown in Figs 10 and 11.

### 7.1 The effects of galactic environment

Fig. 9 shows that there are distinct differences in CLUSTAR-ND’s clustering power that are dependent upon the *type* of galaxy that

the algorithm is applied to. This is not unexpected, the constituent satellites of any particular galaxy will be more/less difficult to classify for any clustering algorithm if a large number of these satellites are disrupted/intact. In essence, if a large number of a satellite’s points’  $k_{\text{link}}$  neighbourhoods contain points from multiple satellites then CLUSTAR-ND’s ability to provide a well-matched cluster to that satellite diminishes as any matching predicted cluster will also contain contamination points from other satellites. Such galactic environments will occur if the constituent satellites have experienced strong tidal forces throughout their history or if they are sparsely distributed at the time of their infall such that they encompass a large volume of the feature space. It is for these reasons that we see the clustering power of CLUSTAR-ND to be less when applied to galaxies created by old satellites, satellites on radial trajectories, and low luminosity satellites.

Contrarily, if most  $k_{\text{link}}$  neighbourhoods contain only points from a single satellite then the aggregation process can likely link together many points from a single satellite in sequence before the group becomes overly contaminated by members of other satellites. Of course, CLUSTAR-ND’s ability to create meaningful clusters from these neighbourhoods also depends on the density of the points and whether these neighbourhoods are isolated or connected to other like-neighbourhoods. Galactic environments that harbour such conditions will contain a larger proportion of intact satellites compared to those that do not. Accordingly, Fig. 9 shows that galaxies made of young and (for some feature spaces) high luminosity satellites are more easily categorized into their respective satellites.

As the MW-type galaxies are essentially a mix of disrupted and intact satellites we befittingly see a clustering power that lies somewhere in the middle of these extremes. Nevertheless, from these observations we can gain some understanding about what to expect if CLUSTAR-ND were to be applied to a data set of stars in the galactic bulge compared to if it were applied to the stars in the stellar halo.

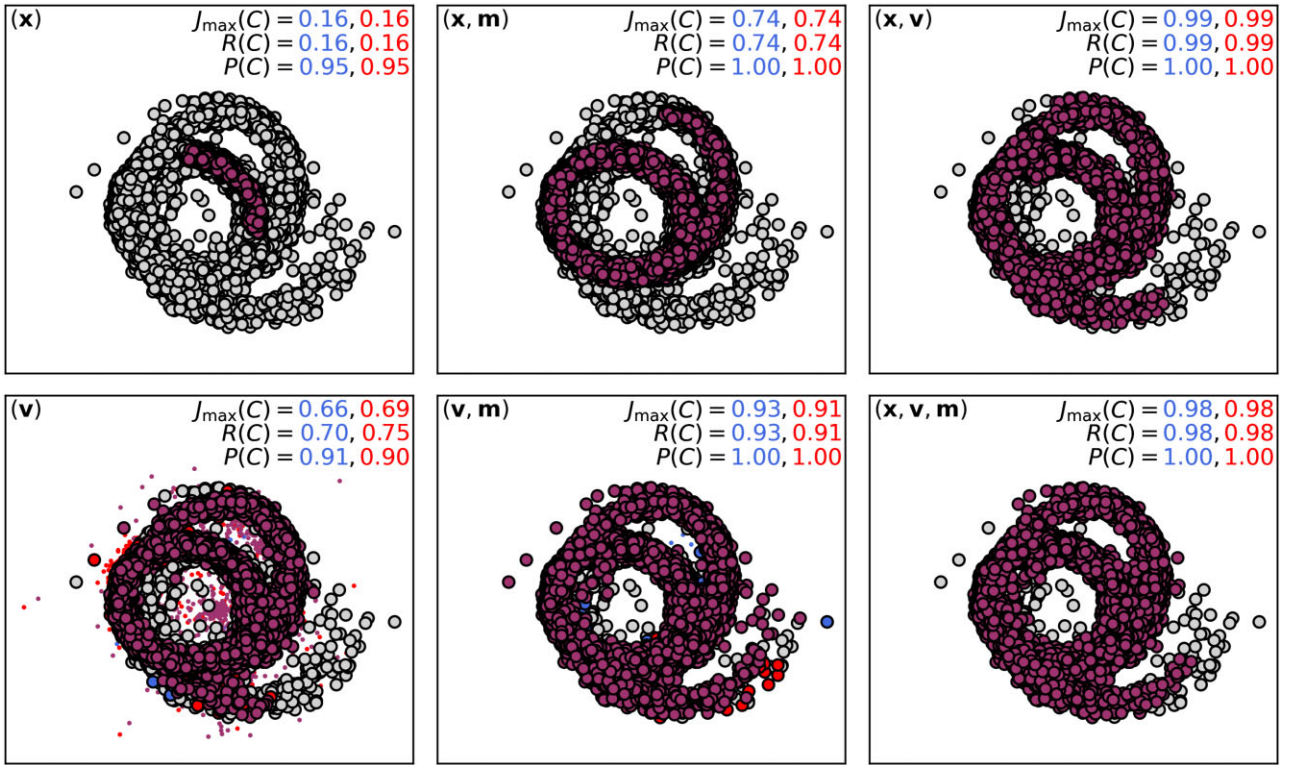
### 7.2 The effects of feature space

The effects of group-mixing in a data set are not only prevalent between different galactic environments, but are noticeable between different feature spaces as well. Generally speaking, Table 1 portrays that the CLUSTAR-ND clustering power is greater for larger feature spaces. One exception to this is that we find the  $(v, m)$  feature space more informative than the  $(x, v)$  feature space. Amongst equal-sized feature spaces, we see that spatial coordinates are generally more informative than kinematic coordinates although the opposite is true when chemical abundances are added in amongst these feature spaces.

From Fig. 9 however, we see that the above observations from Table 1 are not consistent across galaxies – i.e. the information content of any particular galaxy clustering is coupled to both the galactic environment and the underlying feature space of the data. Out of the synthetic galaxies we investigate in this paper, the clustering scenario that gives the most information content for each galaxy is most commonly derived from the largest feature space,  $(x, v, m)$ ; however, the title of *most informative feature space* is also sometimes held by the  $(x, m)$ ,  $(v, m)$ , or  $(x, v)$  feature spaces.

Adding in extra features will often increase clustering power by creating some separation between multiple satellites within the extra volume that is provided by the additional dimensions. However, this will not always be the case as it is possible that by adding in extra *uninformative* features the points belonging to multiple satellites will be brought closer together (relative to other points) – making it more difficult for CLUSTAR-ND to disentangle them from each other and





**Figure 10.** A satellite and the corresponding best-fitting clusters predicted by CLUSTAR-ND when using various feature spaces and metric adaptivity settings. The projection shown of the satellite in each panel is of the two spatial PCA components with the largest variances. The feature space used for the satellite’s prediction is shown in the top left corner of each panel. The metric adaptivity setting is indicated by colour; where blue and red corresponds to an *adaptive* value of 1 and 2, respectively – the purple indicates their intersection. Large points belong to the satellite, small points are incorrectly associated with the satellite. In the top right corner of each panel are the maximum Jaccard index (equation 10), recovery, and purity (equation 11) of the best-fitting clusters from the two *adaptive* values. These panels show the influence of the different feature spaces as well as that of the different metric adaptivity settings.

their surrounds. These effects are why the clustering of the satellite depicted in Fig. 10 is worse for the (x, v, m) feature space than it is for the (x, v) feature space, even though the former feature space is more voluminous. The same effect can be also be seen in Fig. 11, not only between these feature spaces but between the (x) and (x, m) feature spaces as well.

Irrespective of the above, Table 1 does confirm that this effect is outweighed by the opposite effect when including additional features to cluster over – i.e. that including additional features brings together  $k_{\text{link}}$  neighbourhoods made of points from the same satellite while separating those that contain points from other satellites. Specifically, we can observe this occur in Fig. 10 where the addition of the chemical features in the (x, m) feature space has dramatically improved the best-fitting predicted cluster to the satellite when compared to that found using the spatial features alone. The same can be said about the (v, m) and (v) feature spaces, and not only in Fig. 10 but also in Fig. 11. Moreover, we see that the amalgamation of the spatial and kinematic features allows CLUSTAR-ND to produce far better clusterings of the respective satellites in both Figs 10 and 11.

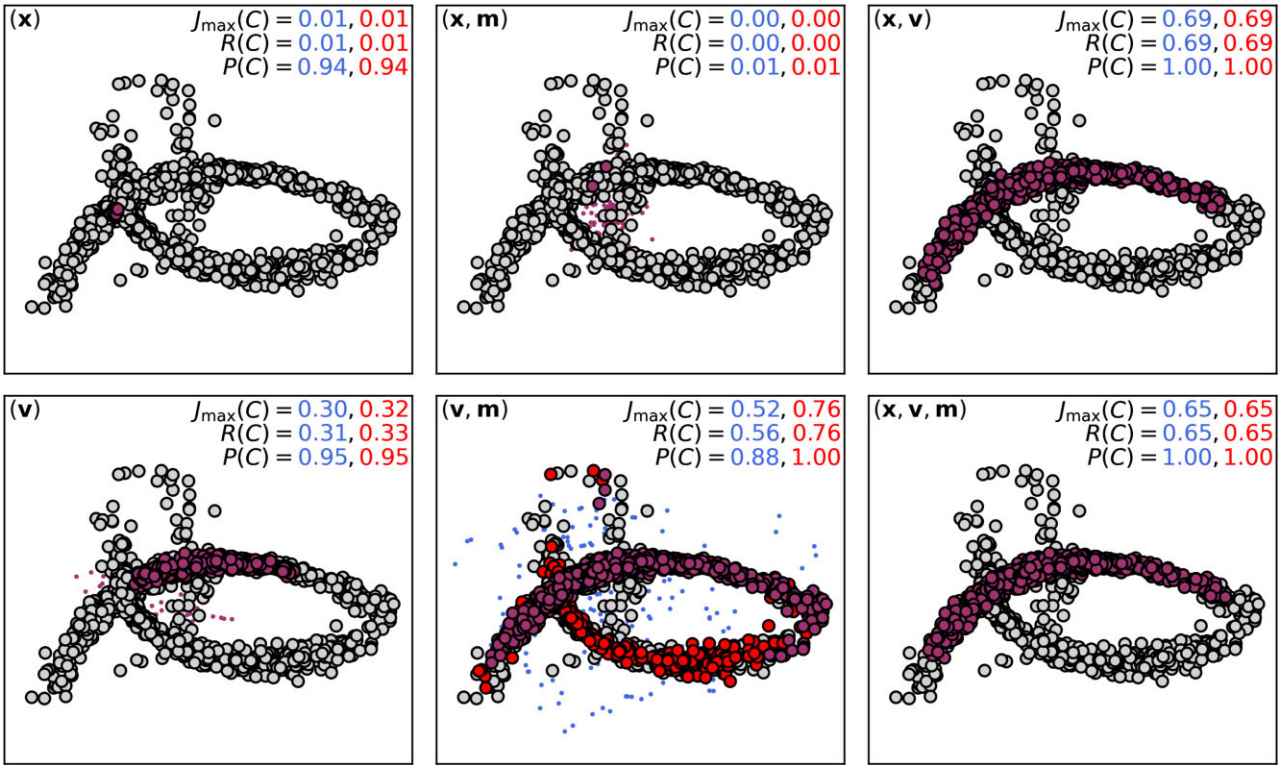
One method of finding the most informative feature space is by comparing the Shannon entropy (Shannon 1948) of the available feature spaces. The smaller the Shannon entropy of a particular feature space the less uniform the distribution of points within it and hence more clustered – this does not necessary guarantee that the feature space is appropriately clustered but it does hint to this possibility. Such a concept is already utilized within clustering algorithms such as ENLINK (Sharma & Johnston 2009) in order to create a meaningful locally adaptive metric and could be utilized

when using CLUSTAR-ND to narrow the list of possible feature spaces before applying the algorithm.

### 7.3 The effects of metric adaptivity settings

Another way to bring like-satellite points together while separating dislike-satellite points from each other is via the means of a adaptive metric. CLUSTAR-ND offers three metric adaptivity settings, however, only the *adaptive* values of 1 and 2 should be considered as adaptive metrics since there is no transformation performed by CLUSTAR-ND when *adaptive* = 0. It is for this reason that the latter should only be used if the feature space consists solely of similar coordinates such as the (x) or (v) feature spaces. The *adaptive* = 0 setting can also be used on compound feature spaces so long as some meaningful transformation of the data has been performed. The settings of *adaptive* = 1 and 2 are implementations of a globally adaptive metric and a locally adaptive metric, respectively.

In Table 1, we can see that running CLUSTAR-ND with the *adaptive* = 1 setting produces similar results to the *adaptive* = 0 in the feature spaces where this setting is applicable. This gives us confidence that the *adaptive* = 1 setting is appropriate for producing a globally adaptive metric for the larger compound feature spaces. The locally adaptive metric provides a similar clustering power overall to the globally adaptive metric. The medians and means of this measure appear to be higher for the locally adaptive metric when CLUSTAR-ND is applied to compound feature spaces, but it is difficult to say whether using it provides a better clustering in general. However, in Fig. 9 we see that blanket statements about the use of the locally



**Figure 11.** A satellite and the corresponding best-fitting clusters predicted by CLUSTAR-ND when using various feature spaces and metric adaptivity settings. This satellite is different to that shown in Fig. 10; however, the description of each panel’s information remains the same here. As with Fig. 10, these panels show the influence of the different feature spaces as well as that of the different metric adaptivity settings – however, the (v, m) panel in the lower middle is a pronounced case of the clustering difference that can occur between the globally adaptive metric setting (*adaptive* = 1) and the locally adaptive metric setting (*adaptive* = 2).

adaptive metric over the globally adaptive metric are not always true. With this it is clear that the optimal value of *adaptive* has a subtle dependency on the galactic environment as well as the feature space.

Fig. 11 depicts one such case where the use of the *adaptive* = 2 setting has caused CLUSTAR-ND to produce a distinctly better match to a satellite. We see here that by using the locally adaptive metric setting on the (v, m) feature space, the best-fitting match to the satellite has both an improved recovery and purity. However, it is possible that in other feature spaces, (x, v) and (x, v, m) for example, the lower arm (on the satellite projection) has been found as a sibling cluster(s) to the best-fitting cluster(s) shown in those panels. In this sense, it is likely that the locally adaptive metric has simply created a means to condense the lower density particle bridge between these regions.

Intuitively, the *adaptive* = 2 setting is only preferable to the *adaptive* = 1 if the most meaningful child cluster (in any parent-child cluster pair) is dense with respect to the shape of its parent cluster. It is not obvious as to whether this will be the case before using CLUSTAR-ND on a particular data set. Moreover, it is not obvious after the clustering has been found whether this technique has yielded the a more meaningful clustering overall. Without a recognizable way of appropriately using the *adaptive* = 2 setting,<sup>11</sup>

<sup>11</sup>Other than perhaps when using the (v, m) and (x, v) feature spaces since it is evident from Fig. 9 that the clustering power of CLUSTAR-ND is either similar or increased when using the *adaptive* = 2 setting compared to the *adaptive* = 1 on these feature spaces.

we can only expect to see a *different* clustering by using it over the *adaptive* = 1 setting.

Given that using CLUSTAR-ND with the locally adaptive metric is more computationally demanding and requires a longer run-time (by a factor of  $\sim 2$ ) with varied clustering results, it is left to the user’s discretion as to whether or not they wish to user it and by default the *adaptive* parameter will be set to 1 ensuring a globally adaptive metric. This aspect of the CLUSTAR-ND algorithm can be certainly be improved and we will endeavour to include a more appropriately determined locally adaptive metric in a future work. It is also possible to use CLUSTAR-ND on the *adaptive* = 0 setting in conjunction with other external manifold learning codes such as UMAP (McInnes, Healy & Melville 2018) – however, clustering over these results can be unpredictable and may also require a different  $\rho_{\text{threshold}}$  value than the optimal one we report in this paper.

## 8 CONCLUSIONS

We have presented the CLUSTAR-ND algorithm and have shown it to be well-suited to the hierarchical classification of stellar satellite groups within galaxies. Compared to its predecessor HALO-OPTICS, CLUSTAR-ND is not only capable of producing a similarly robust clustering output but it does so in an exceptionally efficient manner – outperforming the HALO-OPTICS run-times by 3 orders of magnitude at a minimum. Importantly, CLUSTAR-ND is readily applicable to any point-based astrophysical data set that is defined over an arbitrary number of features. In optimizing its clustering performance, we have removed the need for user-defined parameter

values such that the most optimal parameter settings will be chosen automatically based on the input data (unless the user specifies otherwise).

In this paper we also investigate the capacity of a computationally cheap design for producing a locally adaptive metric to improve CLUSTAR-ND's clustering power. The design iteratively re-applies CLUSTAR-ND's substructure searching component to each consecutive level of the hierarchy in a top-down fashion in order to adaptively modify the underlying distance metric and ensure that each level of clusters found are dense with respect to their parent cluster in the hierarchy. We find that while this approach does produce better results in a small number of clustering scenarios (predominantly on data defined with a combination of kinematic and chemical abundance features), it is mostly inconsistent with regards to the resultant clustering power and tends to simply provide a different clustering when compared to the globally adaptive metric scheme. Since the locally adaptive metric scheme requires more run-time than the globally adaptive metric scheme, we leave using it as a situational choice for the user to make.

In a future work, we will develop a more clustering-appropriate locally adaptive metric scheme for CLUSTAR-ND to make use of. We will also establish a method of producing fuzzy clusterings from fuzzy data such that the point-based uncertainties of the data may be propagated into cluster-based uncertainties. As of this paper, however, CLUSTAR-ND is ideally suited to large astrophysical data sets both synthetic and observational, and hence our further research will also include the application of CLUSTAR-ND to observational data sets of the MW.

## ACKNOWLEDGEMENTS

WHO gratefully acknowledges financial support through the Paulette Isabel Jones PhD Completion Scholarship at the University of Sydney. GFL received no funding to support this research.

## DATA AVAILABILITY

The data underlying this article may be made available on reasonable request to the corresponding author.

## REFERENCES

- Ankerst M., Breunig M. M., Kriegel H.-P., Sander J., 1999, in *ACM Sigmod Record*. ACM, New York, NY, p. 49
- Behroozi P. S., Wechsler R. H., Wu H.-Y., 2012, *ApJ*, 762, 109
- Bentley J. L., 1975, *Commun. ACM*, 18, 509
- Breunig M. M., Kriegel H.-P., Ng R. T., Sander J., 1999, *Principles of Data Mining and Knowledge Discovery*. Springer, Berlin Heidelberg, p. 262
- Bullock J. S., Johnston K. V., 2005, *ApJ*, 635, 931
- Campello R. J. G. B., Moulavi D., Zimek A., Sander J., 2015, *ACM Trans. Knowl. Discov. Data*, 10
- Davis M., Efstathiou G., Frenk C. S., White S. D., 1985, *ApJ*, 292, 371
- Dempster A. P., Laird N. M., Rubin D. B., 1977, *J. R. Stat. Soc.: Ser. B (Methodological)*, 39, 1
- Dunn J. C., 1973, *J. Cybern.*, 3, 32
- Elahi P. J., Canas R., Poulton R. J. J., Tobar R. J., Willis J. S., Lagos C. d. P., Power C., Robotham A. S. G., 2019, *PASP*, 36, e021
- Epanechnikov V. A., 1969, *Theory of Probability & Its Applications*, 14, 153
- Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Simoudis E., Han J., Fayyad U., eds, *KDD 1996: Proc. 2nd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Palo Alto, CA, p. 226
- Flores R. A., Primack J. R., 1994, *ApJ*, 427, L1

- Font A. S., Johnston K. V., Bullock J. S., Robertson B. E., 2006, *ApJ*, 638, 585
- Fuentes S. S., De Ridder J., Deboscher J., 2017, *A&A*, 599, A143
- Ghigna S., Moore B., Governato F., Lake G., Quinn T., Stadel J., 1998, *MNRAS*, 300, 146
- Harris C. R. et al., 2020, *Nature*, 585, 357
- Ishiyama T. et al., 2013, *ApJ*, 767, 146
- Jaccard P., 1912, *New phytologist*, 11, 37
- Johnston K. V., Bullock J. S., Sharma S., Font A., Robertson B. E., Leitner S. N., 2008, *ApJ*, 689, 936
- Kauffmann G., White S. D. M., Guiderdoni B., 1993, *MNRAS*, 264, 201
- King I., 1962, *AJ*, 67, 471
- Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, *ApJ*, 522, 82
- Knebe A. et al., 2011, *MNRAS*, 415, 2293
- Knebe A. et al., 2013, *MNRAS*, 428, 2039
- Knollmann S. R., Knebe A., 2009, *ApJSS*, 182, 608
- Lam S. K., Pitrou A., Seibert S., 2015, in *Proc. Second Workshop on the LLVM Compiler Infrastructure in HPC*. Association for Computing Machinery, New York, NY, p. 1
- Lloyd S., 1982, *IEEE Trans. Inform. Theory*, 28, 129
- Maciejewski M., Colombi S., Springel V., Alard C., Bouchet F. R., 2009, *MNRAS*, 396, 1329
- MacQueen J., 1967, in *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Univ. California Press, Berkeley, CA, p. 281
- Mahalanobis P. C., 1936, *Proc. National Institute of Science of India. National Institute of Science India*, Calcutta, p. 49
- Malhan K. et al., 2022, *ApJ*, 926, 107
- Maneevongvatana S., Mount D. M., 1999, in *Proceedings of the 4th Annual CGC Workshop on Computational Geometry*. Center for Geometric Computing, Santa Barbara, CA, p. 1
- McConnachie A. W. et al., 2018, *ApJ*, 868, 55
- McInnes L., Healy J., Melville J., 2018, preprint ([arXiv:1802.03426](https://arxiv.org/abs/1802.03426))
- Moore B., 1994, *Nature*, 370, 629
- Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, *ApJ*, 524, L19
- Navarro J. F., Frenk C. S., White S. D. M., 1996, *ApJ*, 462, 563
- Oliver W. H., Elahi P. J., Lewis G. F., Power C., 2020, *MNRAS*, 501, 4420
- Pedregosa F. et al., 2011, *J. Mach. Learn. Res.*, 12, 2825
- Press W. H., Schechter P., 1974, *ApJ*, 187, 425
- Reed D., Governato F., Quinn T., Gardner J., Stadel J., Lake G., 2005, *MNRAS*, 359, 1537
- Robertson B., Bullock J. S., Font A. S., Johnston K. V., Hernquist L., 2005, *ApJ*, 632, 872
- Sain S. R., 2002, *Comput. Stat. Data Anal.*, 39, 165
- Sander J., Qin X., Lu Z., Niu N., Kovarsky A., 2003, in Whang K.-Y., Jeon J., Shim K., Srivastava J., eds, *Advances in Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg, p. 75
- Shannon C. E., 1948, *Bell Syst. Tech. J.*, 27, 379
- Sharma S., Johnston K. V., 2009, *ApJ*, 703, 1061
- Sharma S., Bland-Hawthorn J., Johnston K. V., Binney J., 2011, *ApJ*, 730, 3
- Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, *MNRAS*, 328, 726
- Springel V. et al., 2008, *MNRAS*, 391, 1685
- Tollerud E. J., Bullock J. S., Strigari L. E., Willman B., 2008, *ApJ*, 688, 277
- Van Den Bosch F. C., Robertson B. E., Dalcanton J. J., De Blok W., 2000, *AJ*, 119, 1579
- Vinh N. X., Epps J., Bailey J., 2009, in *Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09*. Association for Computing Machinery, New York, NY, p. 1073
- Virtanen P. et al., 2020, *Nat. Methods*, 17, 261
- White S. D. M., Rees M. J., 1978, *MNRAS*, 183, 341
- Zhang A. X., Noulas A., Scellato S., Mascolo C., 2013, in *2013 International Conference on Social Computing*. IEEE, Washington, DC, p. 69

This paper has been typeset from a TeX/LaTeX file prepared by the author.

# Chapter 5

## An Entirely Data-Driven Method

Originally, I had planned to apply **CLUSTAR-ND** directly to observational data sets such as Gaia [360], SDSS [361], PAndAS [362], and GALAH [363]. However, upon attempting to provide a meaningful clustering of the foreground-background subtracted PAndAS data set I realised that I needed a finer and more intuitive control of the returned clusters in order to encapsulate all relevant structure. The problem is that observational data sets will somewhat unavoidably contain contamination noise within them. As such, the data is not a true sampling of the underlying density manifold that represents the real astrophysical structure that we are trying to uncover from the data. With a varying noise level in each data set, the fixed and pre-determined optimal values found for each of the **CLUSTAR-ND** will typically be in need of some adjustment. However, cluster extraction process of **CLUSTAR-ND** is complex – the effect of each parameter is coupled and as such it is difficult to adjust for varying noise and remain vigilant to the quality of the structure that is returned.

In light of this, I endeavour to strip back the cluster extraction process of **CLUSTAR-ND** and its 3 parameters and replace it with an extraction process governed by a single simple-to-interpret parameter that represents the minimum statistical significance that a cluster must have to be labelled as such. The new algorithm, **CLUSTARR-ND**, therefore has a parameter,  $S$ , that represents the distinction by which the extracted clusters must have from the noisy local density fluctuations that appear in the data. In doing this, **CLUSTARR-ND** must keep track of all overdense groups – including those that will later be classified as noise. Since the smoothing length used in the kernel density estimate is defined using a fixed number of nearest neighbours, the noisy density fluctuations do not adhere to those that would arise due to a Poisson point process. As such, I construct a measure called *prominence* that is defined as the logarithm of a signal-to-noise ratio for a given group. By fitting a distribution to the prominences of all noisy groups I can define which groups are

outliers from the noise and by how much.

To provide the user with additional information, I also have **CLUSTARR-ND** produce an ordered-density plot which – analogously to the reachability plot of **OPTICS** and **HALO-OPTICS** – is a visual indication of the clustering structure within the data. With these changes, I find that **CLUSTARR-ND** is a robust and *nearly* entirely data driven that is ideally suited for application to observational data sets. As the final installation in the series of astrophysical structure finding algorithms featured within this thesis, **CLUSTARR-ND** is the most readily applicable to any given astrophysical data set making it the truest version of a generalised astrophysical structure finder. **CLUSTARR-ND** will reliably find clusters so long as the input data set appropriately represents the underlying density manifold that describes those clusters. The code for the **CLUSTARR-ND** can be found in App. [B.3](#).

## 5.1 Structure Finding with CluSTARR-ND

This section presents the manuscript in preparation:

3. *The Hierarchical Structure of Galactic Haloes: Differentiating Clusters from Non-Poissonian Noise with CluSTARR-ND.* **W. H. Oliver**, P. J. Elahi, & G. F. Lewis. *in preparation*.

Given the time constraints surrounding the submission of this thesis, the paper presented here is a work-in-progress. As such, it is unfinished and may be subject to further changes throughout. At the time of writing however, Secs. 1, 2, and 3 are draft-complete while Secs. 4, 5, and 6 are not. In these remaining sections I will; discuss the limitations of algorithm; conduct a comparison between the outputs of **CLUSTARR-ND** and it’s predecessor **CLUSTAR-ND** as well as show that **CLUSTARR-ND** remains significantly more robust with varying noise levels; and present my conclusions about the algorithm, its effectiveness, and the future work that it can be used within. Although at present the paper does not depict the output of the algorithm on astrophysical data, Fig. [6.2](#) illustrates its effectiveness on the PAndAS data. Ultimately, I expect to have this work submitted for publication within  $\sim 2$  months of the submission of this thesis.

*Author Contributions:* I developed the **CLUSTARR-ND** algorithm and have written the manuscript. Dr. Pascal J. Elahi has made valuable contributions to the concept of the algorithm. The project has been conducted under the supervision of Prof. Geraint F. Lewis.

# The Hierarchical Structure of Galactic Haloes: Differentiating Clusters from Non-Poissonian Noise with CLUSTARR-ND

William H. Oliver<sup>1</sup>\*, Pascal J. Elahi<sup>2</sup>, and Geraint F. Lewis<sup>1</sup>

<sup>1</sup>*Sydney Institute for Astronomy, School of Physics A28, The University of Sydney, NSW, 2006, Australia*

<sup>2</sup>*Pawsey Supercomputing Research Centre, 1 Bryce Avenue, Kensington, WA, 6151, Australia*

Accepted XXX. Received YYY; in original form ZZZ

## ABSTRACT

This paper is a part of a series of papers that have sequentially built upon each other in order to produce fast and generalised clustering algorithms that are ideally suited to astrophysical clustering. In this paper, we build upon the hierarchical galaxy/(sub)halo finding algorithm CLUSTAR-ND to make CLUSTARR-ND (**Clustering Structure via Transformative Aggregation, Regression, and Rejection in N-Dimensions**). We do this by redefining its cluster extraction process to ensure that each resultant cluster is defined to be statistically distinct from random fluctuations in the estimated density field of the input data. This modification effectively exchanges the previously 3 cluster extraction parameters from CLUSTAR-ND with a single parameter,  $S$ , that is the lower threshold statistical significance of the extracted clusters. This change not only makes CLUSTARR-ND readily applicable to data of any size and shape but now also to any such data set with an arbitrary level of noise. We show the latter by demonstrating that CLUSTARR-ND produces a more robust clustering of the input data than CLUSTAR-ND for a range of background noise contamination levels. As such, CLUSTARR-ND is now suitable for application to synthetic or observational data sets and will not produce spurious clusters when applied to noisier-than-synthetic data sets.

**Key words:** galaxies: structure – galaxies: star clusters: general – methods: data analysis – methods: statistical

## 1 INTRODUCTION

The ability to correctly classify astrophysical structure from observational data sets plays a particularly significant role in the pursuit of knowledge about our Universe. Specifically, by revealing the spatial, kinematic, and/or chemical overdensities within these data sets we can begin to make predictions about the nature of structure formation and evolution. Historically, such structures have been discovered via a direct inspection of the data. Galaxies and clusters of galaxies have been found with photographic plates (Abell 1958), or more recently, galactic substructure has been uncovered through specific data projections (e.g. Arifyanto & Fuchs 2006; Duffau et al. 2006; Williams et al. 2011; Helmi, Amina et al. 2017; Belokurov et al. 2018). However, as we continue to gather data about our Universe it is increasingly important that we approach structure finding from the data mining perspective as inspection methods will fail to exhaust such large data sets of their clusters.

Data mining algorithms that find structure are referred to as clustering algorithms. The clustering algorithms built for astrophysical structure finding within observational data (particularly from those aimed at uncovering galactic substructure) will typically employ some physical model in order to classify groups. Commonly, this model includes some constraint on orbital motion due to the gravitational potential of the parent structure, e.g. STREAMFINDER (Malhan & Ibata 2018) and the xGC3 suite (Johnston et al. 1996; Mateu et al. 2011, 2017). These algorithms will also often produce projections or

transformations of the data so that they may target a specific type of structure, e.g. STARGO (Yuan et al. 2018), HSS (Pearson et al. 2022), and VIA MACHINAE (Shih et al. 2021). While these algorithms perform well when it comes to exposing the structure they are designed to target, these restrictions effectively enforce limitations as the algorithms can not uncover a range of structure types – nor the way in which these multiple structure types are related.

Clustering algorithms built for uncovering galaxies and their substructure from synthetic data – often called galaxy/(sub)halo finders – will naturally use some proportion of the information available within the simulation to find structure. As such, modern galaxy/(sub)halo finders fall into three categories: configuration space finders, phase space finders, and tracking finders. Configuration space finders – e.g. SUBFIND (Springel et al. 2001), AHF (Knollmann & Knebe 2009), and COMPASO (Hadzhiyska et al. 2021) – use the 3D spatial positions of particles to find physical overdensities (and then the velocities of particles are often also used to reduce the groups to self-bound haloes). Phase space finders – e.g. 6DFOF (Diemand et al. 2006), HSF (Maciejewski et al. 2009), ROCKSTAR (Behroozi et al. 2012), and VELOCIRAPTOR (Elahi et al. 2019) – use both the 3D spatial positions and 3D velocities of particles. Accordingly, tracking finders – e.g. SURV (Tormen et al. 2004; Giocoli et al. 2008) and HBT+ (Han et al. 2017) – use either configuration or phase space to construct haloes but are assisted by particle tracking in their determination of such groups at later times.

While simulation-specific astrophysical clustering algorithms have seen a lot of attention and development over the past few decades, the core of the vast majority of these algorithms is still based off of either the Spherical Overdensity (S0; Press & Schechter 1974)

\* E-mail: william.oliver@sydney.edu.au

or the Friends-Of-Friends (FOF; [Davis et al. 1985](#)) algorithms<sup>1</sup>. Through their consistent use, the structures found by these algorithms have effectively become the defining representations of galaxies and (sub)haloes (at least within synthetic data). This alone is not necessarily a negative point – a structure found by these methods that has also been subject to an additional unbinding process is a *true* self-bound group. Furthermore, a many comparison papers have found that most modern galaxy/(sub)halo finders (including those built off the SO and FOF algorithms) strongly agree on the structure they find ([Knebe et al. 2011, 2013](#); [Onions et al. 2012, 2013](#); [Elahi et al. 2013](#); [Avila et al. 2014](#); [Lee et al. 2014](#); [Behroozi et al. 2015](#)). However, the SO and FOF algorithms do not produce hierarchical clusterings and so typically they are applied iteratively in order to find structures with various densities/shapes/definitions. As such, modern galaxy/(sub)halo can struggle to return all relevant structure.

To overcome the downfalls of both observation and simulation specific astrophysical clustering algorithms whilst maintaining their advantages, a generalised method is sorely needed. Such a method needs to be readily applicable to; both observational and synthetic data sets; data sets with any number of points; and data sets with any number and type of features. Moreover, a generalised algorithm needs to be free of model constraints and provide an adaptive measure of hierarchical clustering structure so as to support to the discovery of all structure types defined with at any degree of overdensity. Generalised algorithms such as OPTICS ([Ankerst et al. 1999](#)) and HDBSCAN ([Campello et al. 2015](#); [McInnes et al. 2017](#)) have seen increased usage in recent years for astrophysical structure finding – e.g. [Costado et al. \(2016\)](#); [McConnachie et al. \(2018\)](#); [Canovas et al. \(2019\)](#); [Massaro et al. \(2019\)](#); [Ward et al. \(2020\)](#); [Higgs et al. \(2021\)](#); [Jensen et al. \(2021\)](#); [Soto et al. \(2022\)](#) for OPTICS and [Ruiz et al. \(2018\)](#); [Mahajan et al. \(2018\)](#); [Koppelman et al. \(2019\)](#); [Jayasinghe et al. \(2019\)](#); [Kounkel & Covey \(2019\)](#); [Webb et al. \(2020\)](#); [Kamdar et al. \(2021\)](#); [Walmsley et al. \(2022\)](#); [Lövdal et al. \(2022\)](#); [Casamiquela et al. \(2022\)](#) for HDBSCAN. However, neither OPTICS nor HDBSCAN are designed with astrophysics specifically in mind – in fact, only a few codes have been e.g. ENLINK ([Sharma & Johnston 2009](#)), FOPTICS<sup>2</sup> ([Fuentes et al. 2017](#)), and CLUSTAR-ND; [Oliver et al. 2022](#)).

We develop a novel and almost entirely data driven astrophysical clustering algorithm CLUSTARR-ND by improving upon the CLUSTAR-ND algorithm. First, we summarise the CLUSTAR-ND algorithm in Sec. 2. We then describe the CLUSTARR-ND algorithm throughout Sec. 3 by; outlining its similarities and differences to CLUSTAR-ND (Sec. 3.1); simplifying its aggregation process (Sec. 3.2); developing a statistical measure for clusteredness (Sec. 3.3); characterising the distribution of clusteredness among noisy density fluctuations (Sec. 3.4); establishing a method to detect statistically significant clusters (Sec. 3.5); and summarising its (short) list of operational parameters (Sec. 3.6). In Sec. 4 we discuss the effective limitations of this approach. Then in Sec. 5 we show the algorithm in practice and compare it to its predecessor, CLUSTAR-ND. Finally, we

<sup>1</sup> As the name suggests, the SO algorithm is used to construct galaxies and (sub)haloes from synthetic data by finding density peaks and then expanding spherical surfaces out from these until some pre-specified overdensity is achieved within the each of the volumes. The FOF algorithm is able to produce a similar clustering by grouping all particles that can be chained together through distances less than the linking length ( $l_x$ ) – which is typically chosen to be  $0.2l_{\text{mean}}$  (corresponding to  $\Delta \approx 200$  overdensities) where  $l_{\text{mean}}$  is the mean particle separation within the simulation box.

<sup>2</sup> Although this algorithm was designed to be used on the position-velocity phase-space of stars.

make our conclusions and present our ideas for future work in Sec. 6.

## 2 AN OVERVIEW OF THE CLUSTAR-ND ALGORITHM

The **Clustering Structure via Transformative Aggregation and Rejection in N-Dimensions** algorithm (CLUSTAR-ND; [Oliver et al. 2022](#)) takes input data of any size and number of features and produces a hierarchical clustering that represents the galaxies within that data and their substructures. It does this by either; first finding the Friends-Of-Friends (FOF; [Davis et al. 1985](#)) haloes from the spatial coordinates of the data; or, by treating the input data as such a halo. Once a list of galaxies/haloes is obtained, the substructure within each of them is found via an approach that;

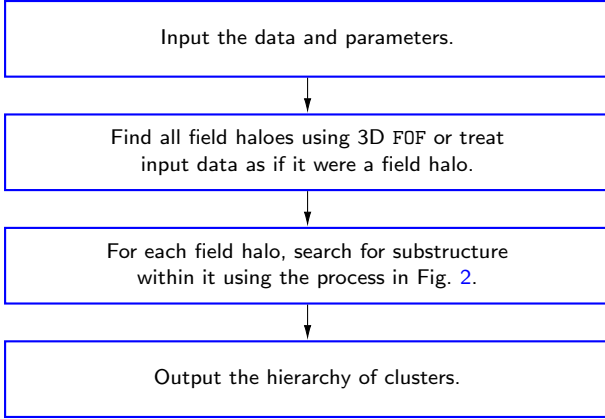
- (i) Optionally transforms the  $n$ -dimensional halo data via a PCA transformation;
- (ii) Estimates the  $n$ -dimensional density field from the transformed data using the set of each point's  $k_{\text{den}}$  nearest neighbours;
- (iii) Seeds groups with points that are situated at the local maxima within this field;
- (iv) Aggregates points in order of decreasing density to the group whose members belong to that point's neighbourhood;
- (v) Joins multiple groups whenever each of these groups have members that belong to a single next-to-be-aggregated point's neighbourhood;
- (vi) And concurrently keeps track of whether a group (just prior to being joined) satisfies the necessary conditions to be labelled a cluster.

Steps (i) and (ii) reduce the complexity of an arbitrarily scaled  $n$ -dimensional data set down to the simplicity of the neighbourhood linkage in order of decreasing density that follows – which allows CLUSTAR-ND to be generalisable to any  $n$ -dimensional data set. While steps (iii) – (vi) are effectively an algorithmic analogue to the process outlined in the HALO-OPTICS algorithm ([Oliver et al. 2021](#)) and, given the same input data (HALO-OPTICS has only been designed for 3-dimensional spatial data sets), will produce a clustering that is remarkably similar (refer to Sec. 5 and Figs. 4 & 9 of [Oliver et al. \(2022\)](#)).

HALO-OPTICS uses the OPTICS algorithm ([Ankerst et al. 1999](#)) and wraps it with a physically motivated parameter selection method and an astrophysically relevant cluster extraction method – since OPTICS alone only provides a 2-dimensional expression of the clustering structure within the input data<sup>3</sup>. OPTICS is itself a hierarchical extension of the DBSCAN algorithm ([Ester et al. 1996](#)) which finds a flat clustering of points that are clustered with densities above a specified threshold. Further still, DBSCAN can be thought of as an extension of the FOF algorithm, taking the point-point based linking scheme and swapping it for a more robust neighbourhood-neighbourhood linking scheme.

The clustering hierarchy of CLUSTAR-ND (and of HALO-OPTICS) is modelled to reflect the designs of the cluster extraction processes of [Sander et al. \(2003\)](#), [Zhang et al. \(2013\)](#), and [McConnachie et al.](#)

<sup>3</sup> This is called the reachability plot and is a plot of the reachability distance vs the ordered index. Within this plot, clusters appear as valleys since they are more dense than their surrounds (smaller distances between points) and are ordered consecutively (local to each other). For more details on the OPTICS algorithm, refer to the original paper ([Ankerst et al. 1999](#)) (and ([Oliver et al. 2021](#))).



**Figure 1.** An activity of the outer methods of both the CLUSTAR-ND and CLUSTARR-ND algorithms. Before finding the substructure each algorithm first decides on the root-level clusters - where the option is given to produce 3D FOF field haloes or to treat the input data as if it were a field halo. For CLUSTARR-ND, the cluster search is completed once the process in Fig. 2 is finished for each field halo.

(2018) that act on the OPTICS reachability plot. Within CLUSTAR-ND, these rules are enacted simultaneously to the aggregation process (steps (iv) & (v)) following which the final clusters satisfy a range of conditions, namely;

- (i) All clusters must have at least  $k_{\text{link}}$  points.
- (ii) All clusters must have median densities at least  $\rho_{\text{threshold}}$  times that at their boundaries.
- (iii) The hierarchy must not contain any lone leaf clusters.
- (iv) And any parent-child pair of clusters must not share more than  $f_{\text{reject}}$  of the parent's points.

Here,  $k_{\text{link}} (\leq k_{\text{den}})$  is the number of nearest neighbours that are used to connect points to already aggregated groups and  $\rho_{\text{threshold}}$  is the overdensity factor that specifies how much denser a cluster must be than its surrounds. These parameters have been optimised and will be automatically selected based on the user's choice of  $k_{\text{den}}$  as well as the dimensionality of the input data. The parameter  $f_{\text{reject}}$  is kept at 0.9 following the analysis performed in Sec. 3.3 in Oliver et al. (2021).

Following the construction of the hierarchy, CLUSTAR-ND also removes outliers from each cluster such that all remaining points have a density greater than  $\rho_{\text{cut}} := \min\{\rho_i \mid \text{lof}(\rho_i) < S_{\text{outlier}}\}$  – the  $\text{lof}(\rho_i)$  is defined in Eq. 8 of Oliver et al. (2022) and is a kernel density analogue of the local-outlier-factor formalised in Breunig et al. (1999). The parameter  $S_{\text{outlier}}$  is optimally set as 2.5 to allow for a moderate level of outlier removal without adversely affecting the clustering power of the algorithm. For more details on the CLUSTAR-ND algorithm refer to Oliver et al. (2022).

### 3 CLUSTARR-ND: BUILDING UPON CLUSTAR-ND

While the CLUSTAR-ND algorithm works well in most scenarios, the rigidity of its cluster extraction process leaves zero room for interpretation (i.e. a densely aggregated group is either found to be a cluster or isn't). Sometimes the user may wish to relax or tighten the constraints of what is a cluster – a functionality that is not easy to control with CLUSTAR-ND's extraction parameters. Furthermore, the optimal values for its extraction parameters  $\rho_{\text{threshold}}$ ,  $f_{\text{reject}}$ , and  $S_{\text{outlier}}$  may not remain optimal for data sets where a large amount

of noise is present. With many moving parts it can be difficult tuning this algorithm for the data at hand.

As such, we design CLUSTARR-ND to remedy these drawbacks and extract clusters with a single parameter,  $S$ , rather than the above 3 parameters from its predecessor CLUSTAR-ND. The parameter  $S$  is the lower threshold statistical significance that a cluster must have. Hence, it allows the user to adjust their criteria of *how clustered clusters need to be* while also implicitly adapting to any level of noise contamination within the input data.

#### 3.1 Algorithmic Similarities and Adjustments

The changes from CLUSTAR-ND to CLUSTARR-ND effectively only modify the functionality of step (vi) of CLUSTAR-ND, however, this step is actually interwoven amongst the processes of steps (iii) – (v) as well. So in the interest of maintaining algorithmic transparency we again present the CLUSTARR-ND equivalents of these steps in Secs. 3.2 – 3.4. For the complete and relevant details on finding root-level haloes as well as of the equivalent to steps (i) and (ii), we refer the reader to Secs. 3.1 – 3.3 of Oliver et al. (2022) as these details will remain largely unchanged – except for a small set of changes which we outline below. We now briefly outline these details and differences.

##### 3.1.1 Root-level Haloes

In finding root-level haloes – just as is done in CLUSTAR-ND – CLUSTARR-ND can optionally find 3D FOF haloes via the FOF algorithm (Davis et al. 1985). If provided with a spatial linking length,  $l_x$ , CLUSTARR-ND will apply the FOF algorithm with this length and the first 3 features of the input data as its input and find the corresponding haloes accordingly<sup>4</sup>. In simulation data, a typical choice for the linking length is  $l_x = 0.2L_{\text{box}}/N$  – where  $L_{\text{box}}$  is the side length of the simulation box and  $N$  is the total number of particles within it. Such a linking length corresponds well to field halo overdensities  $\geq 100\bar{\rho}$  (Elahi et al. 2019). If the user does not provide a value for  $l_x$ , then it is effectively set to  $l_x = \infty$  and instead the input data is treated as if it were a field halo for the purposes of the remainder of the algorithm. This step is shown in Fig. 1.

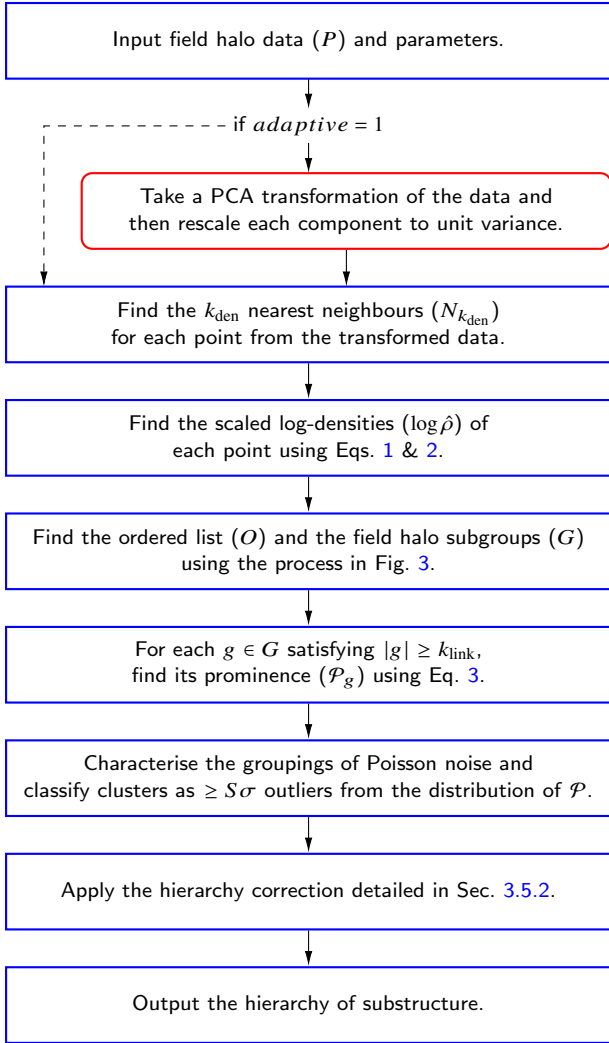
##### 3.1.2 Data Transformation

Once a list of field haloes have been obtained (or if  $l_x = \infty$ , then this list will simply contain the input data) then if  $adaptive = 1$  a data transformation is applied to each before a search for substructure is conducted within them. This can be used to remove any unwanted clustering over-dependencies upon a subset of the features which can occur if some features are effectively weighted more heavily with respect to the others. As in CLUSTAR-ND, we use a Principle Component Analysis (PCA) transform and re-scale the output so that each component has a unit variance. By calculating Euclidean distances on the transformed data we are effectively calculating Mahalanobis distances (Mahalanobis 1936) on the input data – which guarantees that the substructure found will be dense with respect to the field halo shape.

In CLUSTAR-ND the setting that controls the data transformation is governed by the value of the  $adaptive$  parameter – where;  $adaptive = 0$  incurs no transformation;  $adaptive = 1$  incurs the

<sup>4</sup> In this case, the first 3 features must be the Cartesian spatial coordinates of the points.





**Figure 2.** An activity chart of the methods concerned with the set up for finding substructure. A transformation of the data is (optionally) taken, the nearest neighbour lists are found, and then from them the scaled log-density of each point is calculated. This generalises the approach to clustering the substructure. Following this, the ordered list and the field halo’s subgroups are found via the process outlined in Fig. 3. The set of prominences is found for each subgroup using Eq. 3 to which a distribution is then fit. The hierarchy of substructure is then determined by those subgroups whose prominences are  $\geq S\sigma$  outliers to this fitted distribution.

PCA transformation described above; and  $adaptive = 2$  incurs the use of a recursive PCA transformation to every subsequent level of substructure. The latter setting could be used to effectively update the governing distance metric for each level of substructure, however this was not shown to provide a consistent boost in clustering power due to the hierarchy being too coarse grained. As such (and due to the separation of the aggregation and cluster extraction processes), we remove this functionality from CLUSTARR-ND and leave its *adaptive* parameter with two settings – 0 and 1 – whose functionalities are as described above. This step is shown in Fig. 2.

### 3.1.3 Density Estimation

The technique of density estimation will remain mostly unchanged in CLUSTARR-ND. First a set of  $k_{den}$  nearest neighbours,  $N_{k_{den}}$ , will be found for each point within each field halo. Using these sets

in conjunction with the use of a multivariate Epanechnikov kernel (Epanechnikov 1969) inside a balloon estimator (Sain 2002) the density is estimated such that;

$$\rho_i \propto \frac{1}{h_i^d} \sum_{j \in N_{k_{den}}} K\left(\frac{s(x_i, x_j)}{h_i}\right) \text{ where}$$

$$h_i = \max\{s(x_i, x_j) \mid j \in N_{k_{den}}\},$$

$$K(u) \propto (1 - u^2), \quad (1)$$

$d$  is the dimensionality of the feature space, and  $s(x_i, x_j)$  is the Euclidean (Mahalanobis) distance between points  $i$  and  $j$  in the transformed (input) data space. Note that  $K(u) := 0$  for  $u > 1$ .

In CLUSTAR-ND this estimate of the density was used directly, however in CLUSTARR-ND we take its logarithm and re-scale it between 0 and 1 such that

$$\log \hat{\rho}_i := \frac{\log(\rho_i/\rho_{\min})}{\log(\rho_{\max}/\rho_{\min})}. \quad (2)$$

For simplicity, we now refer to the set of these scaled log-density values for all points as  $\log \hat{\rho}$ . The transform not only renders all noisy fluctuations on the same scale regardless of their real density but it also allows us to see how large the affect of these fluctuations are compared to the range of densities within the data. This step is shown in Fig. 2.

### 3.1.4 Connectivity during the Aggregation Process

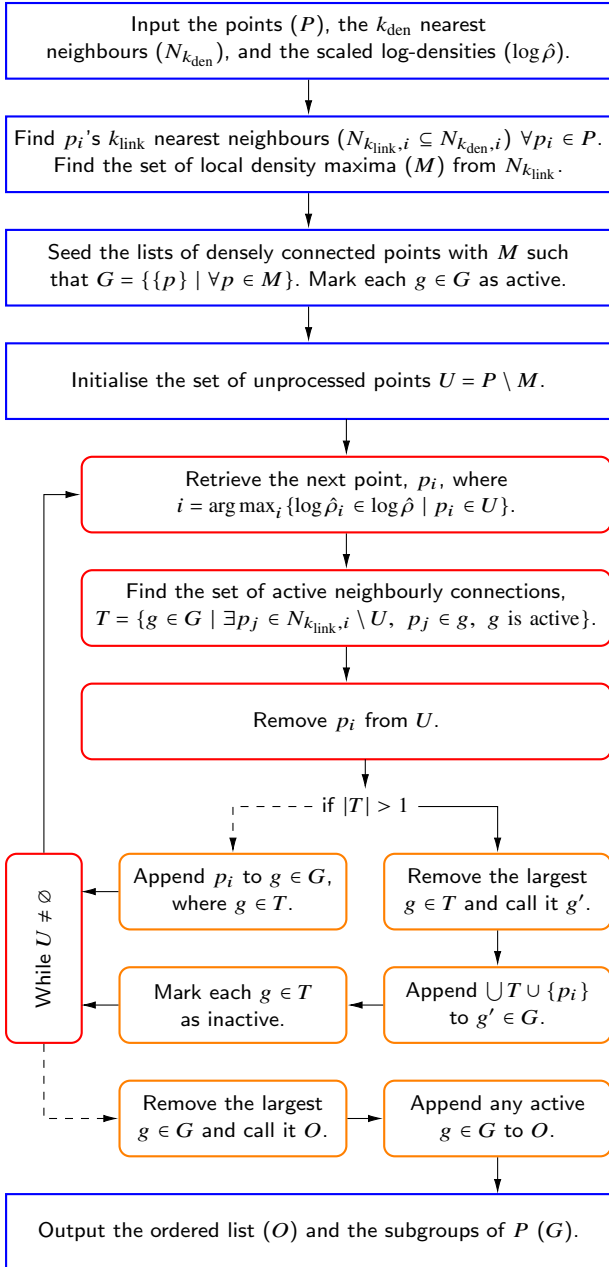
During the aggregation processes of both CLUSTAR-ND (refer to Sec. 3.5 of Oliver et al. (2022)) and CLUSTARR-ND (detailed in Sec. 3.2) the points of the input data are connected via a neighbourhood linkage scheme – i.e. two points will eventually be merged into a single group so long as an unbroken path can be taken through a series of mutually-shared and inter-connectable nearest neighbours. The parameter that is responsible how connectable the set of points will be is  $k_{link}$ .

If  $k_{link}$  is large enough then any two points of the data will be connectable via neighbourhood linkage. However, if it is too large then the resolution of clusters can decrease and the quality of the clustering can be reduced. In Oliver et al. (2022) the optimal value of  $k_{link}$  was determined by ensuring that a  $10^5$ -point data set that had been sampled from a uniform distribution on the unit hypercube would be entirely inter-connectable 95% of the time. This optimal value has a dependency upon both the dimensionality of the data set ( $d$ ) and the value of  $k_{den}$  – since this is implicitly responsible for the number of seeds that are used to begin the aggregation process (refer to Sec. 6.1 of Oliver et al. (2022) for details on this).

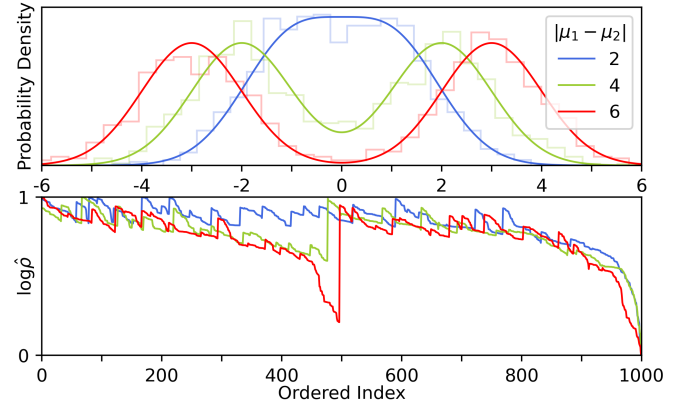
It was found that an optimal value for  $k_{link}$  could be chosen using  $k_{link} = \max\{\text{ceil}(12.0d^{-2.2} - 23.0k_{den}^{-0.6} + 10.0), 7\}$ , which ensures the inter-connectability conditions above were satisfied whilst also maximising the resolution of the resultant clusters. The role of the  $k_{link}$  parameter remains the same in CLUSTARR-ND as is does in CLUSTAR-ND and hence its optimal value can be chosen automatically dependent on  $d$  and  $k_{den}$  as above (unless the user chooses it otherwise).

## 3.2 Simplifying the Aggregation Process

In CLUSTAR-ND, the aggregation process is performed alongside much of the assessment about whether a group qualifies as a cluster or not. For CLUSTARR-ND to be adaptable to any level of noise contamination, this assessment must come after the aggregation process so that it may gauge which of the groups are more noise-like and



**Figure 3.** The activity chart for the CLUSTARR-ND simplified aggregation process that constructs a hierarchy of subgroups whilst also ordering points in a manner that is similar to OPTICS. The nearest neighbour lists are then reduced and from them the local density maxima ( $M$ ) are found. The lists of densely connected points ( $G$ ) are then seeded with these local maxima and are marked as being *active*. In order of decreasing local density, the points are then either appended to an existing list of densely connected points or are used to merge  $\geq 1$  subgroups into a main connecting group ( $g'$ ). Following this inner loop, the ordered list ( $O$ ) is retrieved from  $G$  and if not all points were appended to this, then the remaining connected lists are appended as well. The function that composes the set of scaled log-densities with the ordered list ( $f(i) = \log \hat{\rho}_i, \forall i \in O$ ) is then the CLUSTARR-ND equivalent of the reachability plot produced from OPTICS (examples of which can be seen in Fig. 4). The hierarchy of  $P$ 's subgroups is then given by the remaining densely connected groups in  $G$  and the merger tree that they form (indicated for the same example in Fig. 5).



**Figure 4.** An example of how the ordered-density plots from CLUSTARR-ND are an indicator of clustering structure. The top panel depicts three 1000-point data sets each consisting of two 1-dimensional standard normal distributions whose means are separated by 2, 4, and 6 units respectively where the probability density distributions are plotted over histograms of the sampled data. We have applied CLUSTARR-ND to each of these data sets using  $k_{\text{den}} = 20$ ,  $\text{adaptive} = 0$ , and  $k_{\text{link}} = \text{auto}$  (which results is  $k_{\text{link}} = 18$  for this scenario). The bottom panel shows the corresponding ordered-density plot for each of these data sets. We can see that as the separation between the two distributions grows, the easier it becomes to visually distinguish the two clusters as separate contiguous groups in the ordered-density plot.

which are more cluster-like (refer to Secs. 3.3 & 3.4 for these details), given the state of all the aggregated groups. To do this, we simplify the aggregation as is shown in Fig. 3 so that it is only responsible for building the ordered list of points ( $O$ ) and finding the hierarchy of subgroups ( $G$ ) through a merger tree.

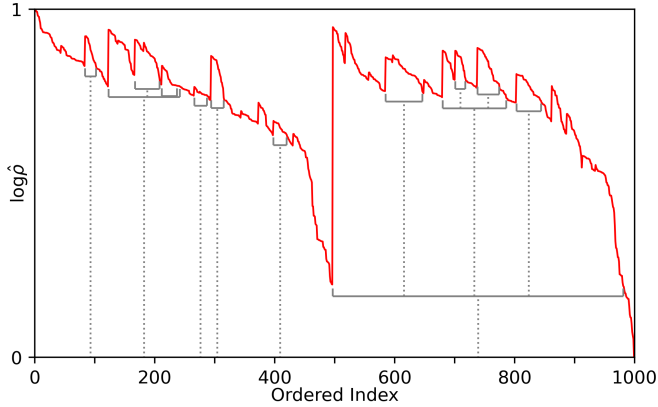
### 3.2.1 The Ordered-Density Plot

The ordered list can be used to construct the CLUSTARR-ND analogue to the OPTICS reachability plot (which we'll call the ordered-density plot), which is an easy way to visualise the clustering structure as it reduces the complexity of the  $n$ -dimensional feature space down to the simplicity of a 2-dimensional dendrogram. To construct the ordered-density plot, we need to compose the set of scaled log-densities with the ordered list which gives a function such that  $f(i) = \log \hat{\rho}_i, \forall i \in O$ . Fig. 4 depicts the ordered-density plot that follows the aggregation process for an input data set consisting of two 1-dimensional standard normal distributions at various separations.

The ordered-density plot consists of groups of points that have been contiguously ordered in terms of decreasing density and joined on the basis of having shared a common nearest neighbour. As such, overdensities will appear as peaks in the plot since they are both denser than their surrounds and ordered consecutively due to being locally connected to each other. In Fig. 4 we can see that the ordered-density plot reveals two main peaks that become increasingly prominent as the separation between the distributions grows. Within these peaks are a series of smaller peaks that arise due to the Poisson noise that arises due to the random sampled of the distributions. However, these become much less prominent than the two main peaks which indicates that a set of meaningful clusters should be retrievable using the ordered-density plot.

### 3.2.2 The Merger Tree

Created alongside the ordered-density plot is a hierarchy of subgroups whose relation to each other is determined by the aggregation



**Figure 5.** An illustration of the CLUSTARR-ND merger tree and how it relates to the ordered-density plot. The merger tree and ordered-density plot used here are taken from the 1-dimensional data set with the largest separation seen in Fig. 4. Marked in grey are all subgroups with at least  $k_{\text{link}} = 18$  points. It is these groups that are subsequently used to find the distribution of subgroup prominences – as indicated in Fig. 2 and as detailed in Sec. 3.3.

process and the order with which they merge. By definition, this merger tree is unbalanced, as for every merger that occurs the largest list of densely connected points is not marked as a subgroup. This tree has a similar structure to the hierarchy that forms during the SUBFIND (Springel et al. 2001) and ENLINK (Sharma & Johnston 2009) algorithms. Fig. 5 displays the merger tree for all subgroups with  $\geq k_{\text{link}}$  ( $= 18$ ) points.

There are clustering scenarios where this hierarchy design is useful. For instance, when clustering over a data set of a galaxy it may be beneficial to only retrieve clusters from these subgroups as the larger group in every merger will often represent the galactic halo (or some contiguously connected dense region of it). However, by only considering clusters that can be retrieved directly from this hierarchy CLUSTARR-ND can not retrieve any of the larger subgroups that complement every node (except the root node) at every level of the tree. We could simply adjust the aggregation process to also record the largest group in every merger although this slows the aggregation and makes the noisy groups more difficult to characterise (refer to Sec. 3.4.3 for details on this).

In reality it is obvious from Fig. 5 that without considering the major group in every merger, CLUSTARR-ND could not provide an appropriate clustering over data sets where the characterisation of the larger merger is necessary. As such, we have CLUSTARR-ND perform a hierarchy correction as its final step. Refer to Sec. 3.5.2 for details on this.

### 3.3 A Statistical Measure to Identify Overdensities

Following the aggregation process, CLUSTARR-ND has produced an ordering of the points within each root-level cluster that allows us to create the ordered-density plot (Fig. 4 for examples). The process has also produced a hierarchical merger tree consisting of a series of subgroups – some number of which may be considered as clusters. Before a subgroup can be considered a cluster or not, we need to assign each of them a measure of *how clustered* they are.

#### 3.3.1 Problems with using a Boolean Overdensity Condition to Classify Clusters

In CLUSTARR-ND the role of deciding whether a group was sufficiently clustered or not is given to the  $\rho_{\text{threshold}}$  parameter – whereby if the median density of the group was at least  $\rho_{\text{threshold}}$  times the density at its boundary then the group would be considered a potential cluster (before then being subjected to other cleaning processes). This is a simplistic definition, which can be beneficial and is easily optimisable (refer to Sec. 6.2 of Oliver et al. (2022) for details), but this does not take into account the full density profile of the group nor does it allow for an intuitively adjustable measure of *how clustered* a group is.

By describing clusters with an expression that does not consider the density profile, some true clusters may be missed – simply because their density profiles are not well-modelled by this rule. Likewise, it is possible that some more spurious clusters may be included in the final list of clusters if they happen to satisfy this condition.

#### 3.3.2 The Prominence of a Group

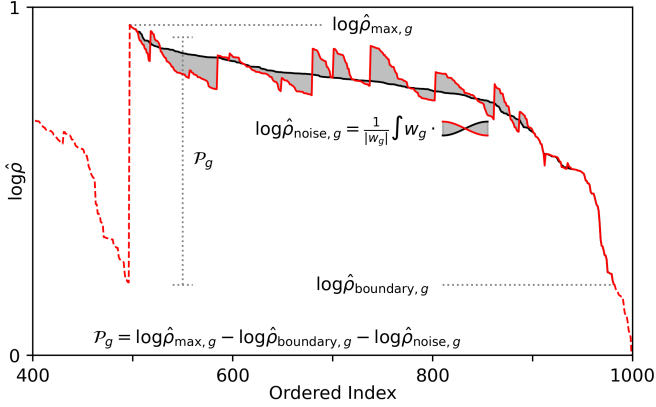
Fortunately, the ordered-density plot contains all the relevant information that is needed to determine *how clustered* a group is. Even by looking at the plot we can see whether a group is more clustered than another. We construct a statistical measure – which we call the *prominence* – that captures both the visual and physical elements of what it means to be *more clustered*.

Using the same notation as in Figs. 2 & 3, we find the prominence for each group  $g \in G$  with  $|g| \geq k_{\text{link}}$  by first taking the maximum scaled log-density of points within the group ( $\log \hat{\rho}_{\text{max},g}$ ) and subtracting the scaled log-density at the group boundary ( $\log \hat{\rho}_{\text{boundary},g}$  – which is the scaled log-density of the point that merged the group with a larger one). As it appears on the ordered-density plot, this is the maximum height difference with which a group stands out from the ordered-density continuum of its parent group. Physically, this is the scaled log-density ratio of the maximum density of the group to the density at the saddle point of its surrounds – i.e.  $\log \hat{\rho}_{\text{max},g} - \log \hat{\rho}_{\text{boundary},g} \propto \log(\rho_{\text{max},g}/\rho_{\text{boundary},g})$ . However this alone does account for the density profile of the group and hence does not adjust for the magnitude of the density fluctuations within the group either.

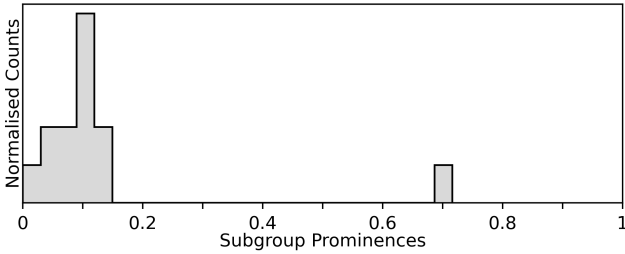
#### 3.3.3 Accounting for Noise

We adjust for the adverse effects of a fluctuating density profile by comparing the estimated density profile to an ideal one where no noise exists. As such, when calculating the prominence of a group we also subtract a weighted average difference of the scaled log-densities between these two profiles ( $\log \hat{\rho}_{\text{noise},g}$ ). An ideal density profile should be strictly monotonically decreasing on the ordered-density plot. We can predict what this might look like for each group by taking the estimated scaled log-densities of the group and then sorting them into an order of decreasing density. The weightings are calculated from the index-distance between the sorted and unsorted densities i.e.  $w_g = |\text{argsort}(\log \hat{\rho}_g)_i - i|$ ,  $\forall p_i \in g$ .

By accounting for noise in this way we encapsulate the amount by which a group would need to change in order to have an ideal density profile. CLUSTARR-ND can then accordingly penalise the  $\log \hat{\rho}_{\text{max},g} - \log \hat{\rho}_{\text{boundary},g}$  value in the case that a group has proportionally large density fluctuations within it. Hence, the prominence of a group is



**Figure 6.** An example of how the calculation of the prominence ( $\mathcal{P}$ ) relates to the ordered-density plot for the largest subgroup featured in Fig. 5. The 1-dimensional Gaussian distribution that this subgroup represents only has small density fluctuations within it and as such is not penalised heavily by the  $\log \hat{\rho}_{\text{noise},g}$  term. Refer to Sec. 3.3 and Eq. 3 for more details on this calculation.



**Figure 7.** The distribution of prominences from the subgroups shown in Fig. 5. Here the histogram has been binned according to the Freedman–Diaconis rule (Freedman & Diaconis 1981). We can see that the one *objectively clustered* subgroup (whose prominence calculation is shown visually in Fig. 6) is a clear outlier from the rest of the prominence distribution.

given by

$$\begin{aligned} \mathcal{P}_g &= \log \hat{\rho}_{\text{max},g} - \log \hat{\rho}_{\text{boundary},g} - \log \hat{\rho}_{\text{noise},g} \\ &= \max\{\log \hat{\rho}_i \mid \forall p_i \in g\} \\ &\quad - \max\{\log \hat{\rho}_j \mid p_j \in N_{k_{\text{link}},i} \cap g, \forall p_i \in g\} \\ &\quad - \text{sum}\{w_g \log \hat{\rho}_g - (\log \hat{\rho}_g)_{\text{desc}}\} / |w_g|, \end{aligned} \quad (3)$$

where  $(\log \hat{\rho}_g)_{\text{desc}}$  is the set  $\log \hat{\rho}_g = \{\log \hat{\rho}_i \mid \forall p_i \in g\}$  that has been sorted into descending order and  $|w_g|$  is the 1-norm of the weightings  $w_g$ . Fig. 6 shows a visualisation of the calculation of  $\mathcal{P}_g$  for the largest subgroup shown in Fig. 5.

### 3.4 Adaptively Characterising Noise

Once the prominences ( $\mathcal{P}_g$ ) for each subgroup ( $g$ ) have been found, CLUSTARR–ND then uses the distribution that these measures to distinguish between *real* clusters and noise. Since there are typically many more noisy subgroups than clustered ones (refer to Sec. 4 for cases where this is not true), we wish to fit an appropriate distribution to  $\mathcal{P}$  that is descriptive of the noisy subgroups in the input data set and/or field halo root cluster. Shown in Fig. 7 is an example of the binned distribution of  $\mathcal{P}$  from the subgroups shown in Fig. 5 where we can see that the clustered subgroup is distinct from the noisy ones.

#### 3.4.1 Systematic Model Fitting

In fitting a probability model ( $M$ ) to the distribution of subgroup prominences  $\mathcal{P}$ , we have CLUSTARR–ND maximise the likelihood function  $\mathcal{L}(M; \mathcal{P}) = \prod_{i=1}^N f(\mathcal{P}_i; M)$  (although in reality it minimises the negative of the logarithm of the likelihood function for improved numeric stability). However, since the  $\mathcal{P}$  distribution will (typically) include prominences from both noisy and clustered subgroups we need to provide  $M$  with the capacity to be descriptive of both – otherwise the distribution that CLUSTARR–ND uses to describe the noisy subgroups would become heavily skewed with the presence of clusters in the data.

For this fitting technique to still yield a description of the noisy subgroups, we construct  $M$  using a linear combination of two probability distributions – one for noisy subgroups and one for clustered subgroups such that  $M = aM_{\text{noisy}} + (1-a)M_{\text{clusters}}$  (for  $a \in [0, 1]$ ). This way, we can easily identify the clustered subgroups from the noisy ones by treating as outliers to the distribution of noisy prominences.

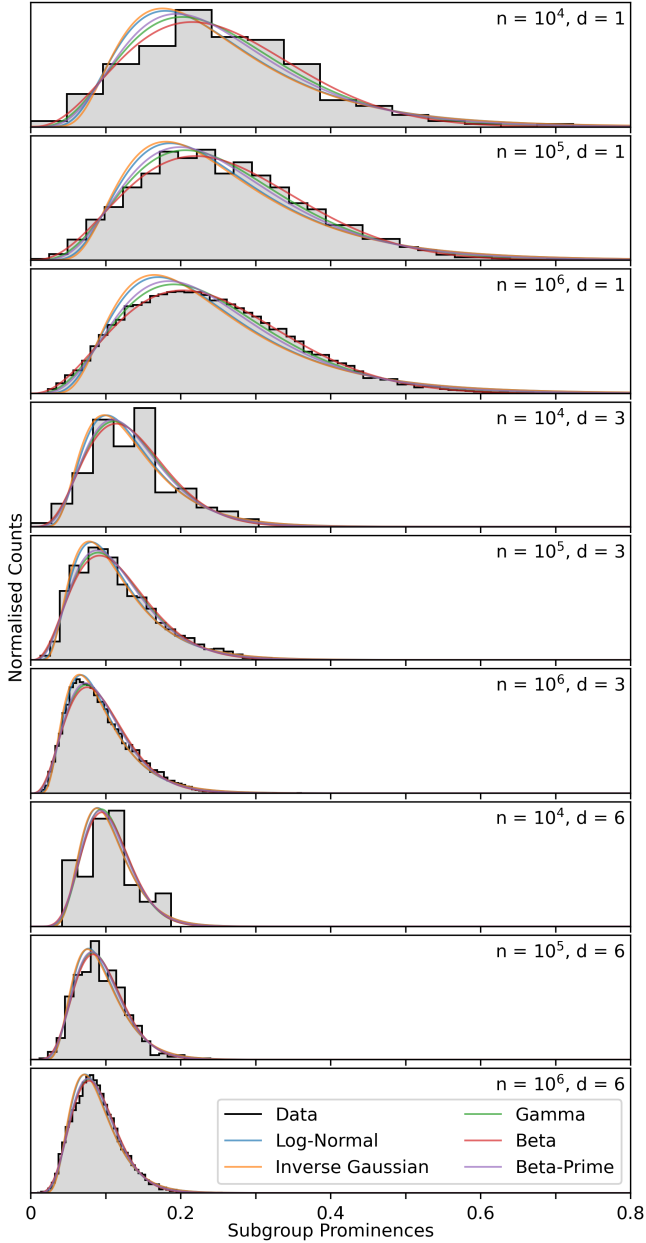
#### 3.4.2 Choosing an Appropriate Model for Subgroup Prominences

So that the model remains unassuming of what clusters may be found, we simply use a continuous uniform distribution on the unit interval – since the prominence distribution is defined there. Setting  $M_{\text{clusters}} = U(0, 1)$  implies that clusters can exist with prominences less than that of noisy subgroups. Such *clusters* would be indistinguishable from noise however this statement is true since we could artificially construct a data set consisting of only noise and insert a cluster whose overdensity gives it an arbitrarily small prominence.

For the noisy subgroups however, we know that it is lower-bounded at 0, uni-modal and positively skewed based on the nature of Poisson noise<sup>5</sup>. Since the upper bound of the prominence distribution is artificial due to the scaling of  $\log \hat{\rho}$  performed in Eq. 2, we should expect that any uni-modal distribution with positive skew supported on the open interval  $(0, a)$  for  $a \in [1, \infty)$  might be suitable. We assess to suitability of the following 5 such distributions for describing the noisy subgroups.

- (i) *Log-Normal*:  
 $\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$  with  $x, \sigma \in (0, \infty)$  and  $\mu \in (-\infty, \infty)$ .  
 $f_X(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$
- (ii) *Inverse Gaussian*:  
 $X \sim IG(\mu, \lambda)$  with  $x, \mu, \lambda \in (0, \infty)$ .  
 $f_X(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right)$
- (iii) *Gamma*:  
 $X \sim \Gamma(k, \theta)$  with  $x, k, \theta \in (0, \infty)$ .  
 $f_X(x; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp\left(-\frac{x}{\theta}\right)$
- (iv) *Beta*:  
 $X \sim \beta(a, b)$  with  $x \in (0, 1)$  and  $a, b \in (0, \infty)$ .  
 $f_X(x; a, b) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}$

<sup>5</sup> If we instead designed CLUSTARR–ND’s aggregation process to return all merging subgroups – regardless of the number of points within them – then the prominence distribution would resemble an exponential distribution. In this case however, the peak of the prominence distribution is suppressed compared to a true exponential distribution. This is due to CLUSTARR–ND having seeded the initial subgroups with the points situated at local density maxima – which are defined using their  $k_{\text{link}}$  nearest neighbours leading to the under-representation of subgroups with less than  $k_{\text{link}}$  points. As such, using an exponential distribution to describe the prominences of noisy subgroups is not practical for the purpose of finding clusters.

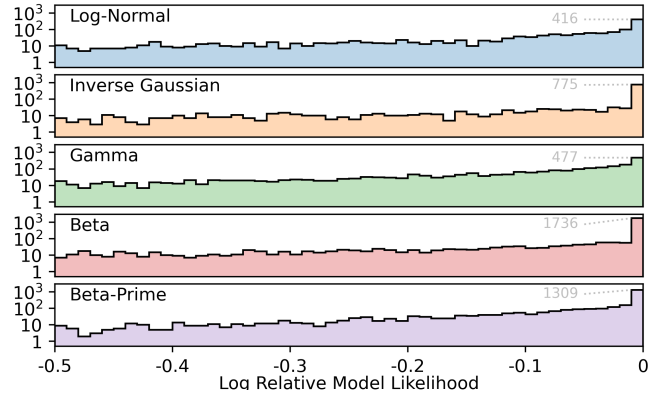


**Figure 8.** Example distributions of subgroup prominences that CLUSTARR-ND produces (when  $k_{\text{den}} = 20$  and  $\text{adaptive} = 0$ ) from various  $n$ -point  $d$ -dimensional data sets that have each been sampled from a uniform distribution defined over the volume of the unit hypercube. The different distribution models (a linear combination of the distributions (i) – (v) and a uniform distribution) are shown fitted to this data in the way described in Sec. 3.4.1. Visually, each noise model could be considered a reasonable choice and as such it is not immediately obvious which is best.

(v) *Beta-Prime:*

$$X \sim \beta'(a, b) \text{ with } x, a, b \in (0, \infty). \\ f_X(x; a, b) = \frac{1}{B(a, b)} x^{a-1} (1+x)^{-a-b}$$

Fig. 8 shows how the compound model  $M$  fits to various prominence distributions when  $M_{\text{noise}}$  is defined with the probability density functions (i) – (v). Here the prominences have been produced by CLUSTARR-ND using  $n$ -point  $d$ -dimensional uniform distributions defined over the unit hypercube as the input data. Each model has the three parameters



**Figure 9.** The distributions of the relative likelihoods of the models outlined in Sec. 3.4.2. The relative likelihoods have been found for various  $n$ -point  $d$ -dimensional data sets of uniform distributions. Each distribution has a peak at  $\sim 0$  (indicating the frequency with which the model is the most suitable choice, or at least very nearly the most suitable choice) and a long negatively skewed tail. We see here that the model defined by a linear combination of these distributions is most likely to be the most suitable of these models for the purpose of describing the prominences of noisy subgroups.

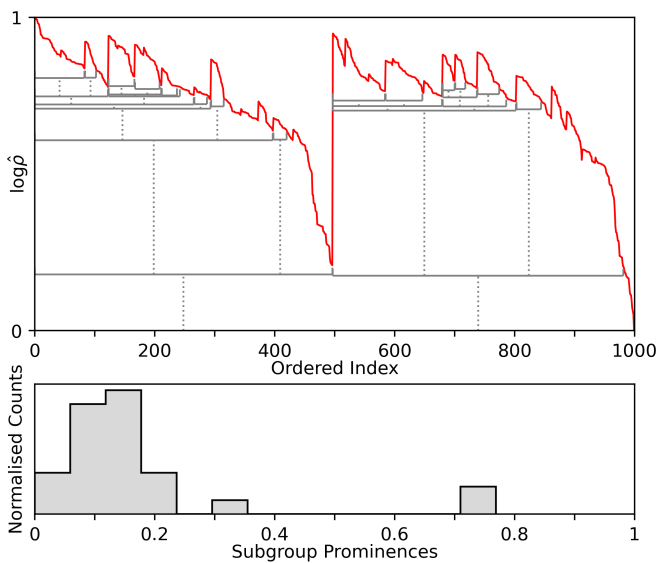
using the different probability density functions (i) – (v) for  $M_{\text{noise}}$ . Here  $M$  has been fit to various sets of subgroup prominences produced from various data sets sampled from uniform distributions. In doing so, we first create 100  $n$ -point data sets that are each randomly sampled from a  $d$ -dimensional uniform distribution defined over the volume of the unit hypercube. We do this for each  $n$ - $d$  combination with  $n \in \{10^4, 10^5, 10^6\}$  and  $d \in \{1, 2, 3, 4, 5, 6\}$ . Then, we apply CLUSTARR-ND (with  $k_{\text{den}} = 20$ ,  $\text{adaptive} = 0$ , and  $k_{\text{link}} = \text{auto}$ ) to each of these data sets to get the subgroup prominences. We then use these to calculate the relative likelihood of each model to the best-fitting model on a per data set basis<sup>6</sup>.

Fig. 9 indicates that most suitable model is that of the Beta distribution – (iv) – which is closely followed by the Log-Normal distribution – (i). From simple observational tests it would also appear that the Log-Normal distribution is more easily skewed by the presence of clusters in the data than the Beta distribution is. So in the interest of good fitting results, we implement the Beta distribution within CLUSTARR-ND as the model that systematically characterises the prominence distribution of noisy subgroups and allows CLUSTARR-ND to separate cluster from noise (refer to Sec. 3.5 for more on this).

### 3.4.3 The Reason for Using an Unbalanced Merger Tree

Having now shown that the CLUSTARR-ND regression technique (Sec. 3.4.1 and prominence distribution model Sec. 3.4.2) is a reasonable approach to classifying noisy subgroups, it is now sensible to explain to the reader why the merger tree constructed from the aggregation

<sup>6</sup> We did also test the suitability of the generalised gamma (3 parameters) and generalised beta-prime (4 parameters) distributions by comparing both the second-order correction estimate of the Akaike information criterion (AIC & AICc; Akaike 1974; Hurvich & Tsai 1989) and the Bayesian information criterion (BIC; Schwarz 1978). We did not include these tests in our final results as these models were consistently more difficult to fit to the data with the method described in Sec. 3.4.1 whilst still not providing a better fitting to the data than the Log-Normal and Beta distributions. Nevertheless, the relative likelihood that is derived from both the AICc and the BIC reduces to a simple ratio of likelihoods when considering models with the same number of parameters.



**Figure 10.** A hypothetical example of a *complete* merger tree and the corresponding prominence distribution that would occur when clustering over the same input data as is used in Figs. 5 & 7. We can see here that if CLUSTARR-ND were made to keep track of the largest group in every merger, the prominence distribution would become comparatively heavier in its tails and make the detection of clusters more difficult given that they would then present as less significant outliers in this distribution. Even though the clusters would still be detectable as outliers in the prominence distribution shown here, for other highly structured data sets this will not always be the case.

process does not keep track of the largest group within each merger (described in Sec. 3.2.2). The most pressing reason for this (apart from decreasing the computational costs) is that by maintaining a more balanced merger tree – whereby every connected group that is a part of the merging process is recorded as a subgroup – the prominence distribution becomes difficult to fit a characterising model to.

Primarily, this is because such a merger tree would then be a record of many cascading subgroups that may share a large portion of their points. As such, it is difficult to construct a measure of prominence that appropriately penalises this attribute for some of these groups while indicating exactly which of these is the most reasonable choice for being labelled a cluster – and certainly the measure we use in Eq. 3 does not help to distinguish the *correct* group from a series of very similar cascading groups. The top panel of Fig. 10 depicts what such a merger tree would look like in the context of the example given in Fig. 5.

Also in Fig. 10, the bottom panel depicts the corresponding binned prominence distribution of the hypothetical merger tree that is shown in the top panel. We can see here that the notion of detecting clusters as outliers in this distribution becomes less robust since there are now a series of subgroups whose prominences effectively extend the tail of this distribution. This is an unwanted effect since these are mostly neither *clusters* nor noise – instead they (mostly) represent a contiguously ordered conglomerate of the majority of points whose boundary density is larger than a particular value (defined at the saddle point of each merger).

Not all such cascading subgroups are necessarily irrelevant when constructing a hierarchy of clusters though. Fig. 5 shows that 1 of the existing Gaussian distributions can not be recovered in the same sense as the other, whereas this subgroup is possible to retrieve from the altered merger tree shown in Fig. 10. CLUSTARR-ND extracts these

clusters via the  $f_{\text{reject}}$  parameter – which whilst doing a reasonable job, is a difficult parameter to optimise given its complex (and somewhat nonphysical) functionality. In CLUSTARR-ND, we implement a new strategy to extract these relevant clusters which we detail in Sec. 3.5.2.

### 3.5 Extracting Clusters from the Data

CLUSTARR-ND differentiates clusters in the data by distinguishing them from noise. Having now created a measure of a group’s prominence (Eq. 3) and a way of characterising the distribution of this measure for all the noisy subgroups (Sec. 3.4), we can begin to extract clusters from the data.

#### 3.5.1 The Statistical Significance of Clusters

To label groups as clusters we now simply identify all subgroups that have a prominence-based statistical significance ( $S$ ) that is greater than  $S\sigma_{\text{noise}}$ . To be clear, any subgroup ( $g$ ) will be labelled a cluster if its prominence ( $\mathcal{P}_g$ ) satisfies

$$S_g = F_{N(0,1)}^{-1} [F_{\beta(a,b)}(\mathcal{P}_g)] \geq S, \quad (4)$$

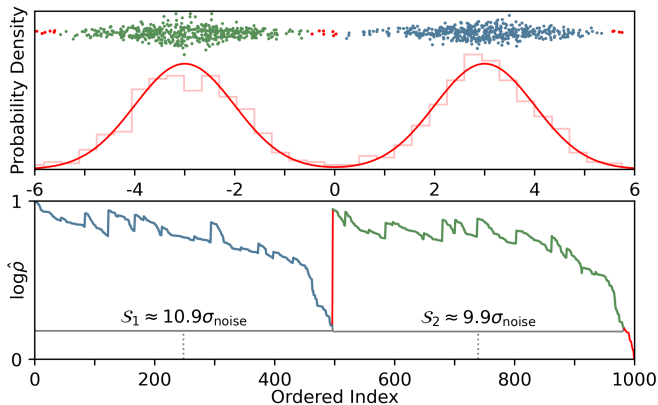
where  $F_{N(0,1)}$  and  $F_{\beta(a,b)}$  are the cumulative distribution functions of the standard normal and the beta distribution respectively, and the parameters  $a$  and  $b$  are derived by fitting to the prominence distribution in the way described in Sec. 3.4.1.

In this sense, we transform the measure of prominence to one of significance such that the distribution of subgroup significances is akin to a standard normal with a long positive tail that contains the significances of increasingly more clustered subgroups. The  $S$  parameter is then a measure of how clustered – compared to the noise present within the data – the user wishes the resultant clusters to be. As such, there is no correct or optimised value for  $S$ . However in Sec. XX, we do show how the clustering power of CLUSTARR-ND varies with  $S$  for a series of synthetic galaxies which we then use to suggest a range of reasonable values that suffice in most use cases.

#### 3.5.2 Correcting the Hierarchy

By simply classifying clusters as the subgroups of the CLUSTARR-ND merger tree that satisfy the condition in Eq. 4, CLUSTARR-ND creates a hierarchy that is similar to SUBFIND and ENLINK (as is mentioned in Sec. 3.2.2) and can be used if desired or if the user will apply some other disentangling method to reduce the hierarchy. In order to construct a hierarchy that is instead similar to its predecessor CLUSTARR-ND, we must make a correction following the initial identification of these clusters. In CLUSTARR-ND clusters are first created in pairs and then the hierarchy is cleaned. We implement a similar design within CLUSTARR-ND although we do so without the need for any additional parameters. This hierarchy style is optional within CLUSTARR-ND and will only be created if the  $h\_style$  parameter is set as *True*.

We first find the clustering hierarchy using the process described in Sec. 3.5.1. Then for each of the extracted clusters that have a parent cluster, CLUSTARR-ND takes the contiguously ordered points that precede them in the ordered list and that are contained within their parent clusters – we call these complementary subgroups. Using the set of complementary subgroups, CLUSTARR-ND calculates each of their prominences and keeps only the newly constructed groups that satisfy Eq. 4 – i.e. now the only remaining complementary subgroups



**Figure 11.** An example clustering the toy example data set from Fig. 4. The top panel depicts the 1-dimensional scatter (with random vertical spread for visual clarity), the histogram, and the combined Gaussian distributions from which the data was sampled from. The bottom panel shows the ordered-density plot, the final hierarchy tree, the cluster significances, and plotted over the ordered-density plot is a colour key that matches the points attributed to the cluster (shown in the top panel). By applying CLUSTARR-ND to this data with  $k_{\text{den}} = 20$ ,  $adaptive = 0$ ,  $k_{\text{link}} = auto$ ,  $h\_style = 1$ , and  $S = 5$  (although in this case it could be chosen from  $S \in (\sim 1.3, \sim 9.8)$  to achieve the same clustering result) it has differentiated the two Gaussian distributions from the noisy density fluctuations within them with a statistical significance of  $10.9\sigma_{\text{noise}}$  and  $9.9\sigma_{\text{noise}}$  respectively.

are significantly clustered compared to the noisy fluctuations of the input data<sup>7</sup>.

Each of these could be potential clusters however to avoid a series of cascading clusters, we again remove all such complementary subgroups that have a child complementary subgroups – this now leaves only the smallest complementary subgroups within each branch of the hierarchy. These are added to the hierarchy of clusters and as an additional cleaning step, CLUSTARR-ND also removes the parent clusters of these if their prominence is smaller than that of their corresponding newly added complementary cluster. This last step ensures that a parent-child pair of clusters is not too similar. Specifically, it ensures that a parent cluster (with a complementary cluster as at least 1 of its child clusters) is not less distinguishable from the noise within it than its newly born complementary child cluster is.

This process yields a hierarchy of clusters that is similarly styled to CLUSTARR-ND, however now each cluster has a more meaningful and statistically interpretable significance. Using the same running toy example from Figs. 4–7 & 10, Fig. 11 now depicts the final hierarchy of extracted clusters from the data set. We can see that CLUSTARR-ND has picked out the over-dense segments of the two Gaussian distributions nicely and – contingent upon the use of locally adaptive metrics and differing density estimators/kernels – this is essentially the highest recovery that a density-based algorithm could hope to achieve without some further process that assigns points to each of these clusters. Similar depictions for synthetic galaxy data can be seen in Figs. XX.

### 3.6 Summary of Parameters

We have now completely described the inner workings of the CLUSTARR-ND algorithm. It takes any (refer to Sec. 4.1 for a set of

<sup>7</sup> At this stage, if the initial set of clusters were found using  $S = -\infty$  – i.e. letting all subgroups be clusters – then the resultant merger tree would effectively be constructed in the same fashion as that that is shown in Fig. 10.

conditions that the input data should satisfy)  $n$ -point  $d$ -dimensional data set and produces a clustering that is contingent on a small number of intuitive, or otherwise pre-optimised, parameters. In doing this, CLUSTARR-ND is able to; first find galaxies using the FOF algorithm; apply a transformation to the intra-galaxy data; calculate an estimate of the local density of all points with each galaxy; aggregate the set of a galaxy’s points in a manner that reveals the clustering structure; determine the prominence of each densely connected subgroup; use these to extract cluster significant clusters; and produce a meaningful hierarchy. Tab. 1 outlines the parameters responsible for these functionalities, the valid values they can take as well as their default values, and provides a comment on the behaviour they control. Unless stated otherwise, the parameters will be set using their default setting throughout the remainder of the paper.

## 4 LIMITATIONS OF THE APPROACH

### 4.1 Conditions Imposed on the Input Data

While the CLUSTARR-ND algorithm is generally applicable to most data sets, there are two main restrictions that must be taken into account. The first of these is simple and obvious – the number of points in the input data must be larger than  $k_{\text{den}}$ . However for the new cluster extraction method implemented within CLUSTARR-ND to work, there must be a sufficiently large number of noisy subgroups found within aggregation process. This is a critical must-have for CLUSTARR-ND to be able to extract clusters from the data since without enough noisy subgroups, clusters can not be deemed outliers to the noisy prominence distribution. The inability to classify clusters as outliers in the prominence distribution can occur in two scenarios; if the input data does not contain enough points; or if the number of truly clustered subgroups significantly outweighs the number of noisy subgroups. We now provide examples of when these scenarios can occur in Secs 4.1.1 & 4.1.2 respectively.

#### 4.1.1 Effective Lower Limits on Data Set Size

To probe the data set size lower limit of extracting clusters with CLUSTARR-ND, we find the number of bins in the prominence histogram that is used to fit the beta distribution that CLUSTARR-ND uses to describe the noisy subgroups. If this distribution can not be reliably fit to this histogram, then clusters can not be dependably extracted using the method described in Sec. 3.5.1. Since there are effectively 4 free parameters in the fitting model (2 in the beta distribution, 1 scaling parameter, and an additional free parameter due to assuming the conditions needed for least squares fitting), the lower limit should be the number of points in the data set that is expected to give at least 5 bins in the prominence distribution.

We record the number of prominence bins ( $|H|$ ) for various numbers of points ( $n$  from  $10^2$  to  $10^4$  in steps of 100) and dimensionalities ( $d \in \{1, 2, 3, 4, 5, 6\}$ ) of a randomly sampled uniform distribution data set. For each combination we also vary the value of  $k_{\text{den}} \in \{20, 40, 80\}$  (while using  $k_{\text{link}} = auto$ ). Since  $|H|$  is a probabilistic measure, we do this for 50 re-samplings in each combination and take the average to represent expected number of bins given the input data and the value of  $k_{\text{den}}$ .

Note that the effective lower limit may vary with other distributions, however the lower limit we depict here should serve as a guide

**Table 1.** A short summary of the CLUSTARR-ND parameters that are relevant to its clustering behaviour. CLUSTARR-ND has a total of 6 parameters, as opposed to the 7 parameters of its predecessor CLUSTAR-ND, and for most clustering purposes it is only the choice of  $l_x$ , *adaptive*,  $k_{\text{den}}$ , and  $S$  that need to be considered. However, deciding their value is mostly a simple task.  $l_x$  should only be set to a value other than  $\infty$  when the user wishes for the root-level clusters to be defined by 3D FOF groups – a common definition for galaxies in cosmological simulation. *adaptive* should only be set to 0 when the features of the data have the same units or when a meaningful transform has been applied to the data prior to its input to CLUSTARR-ND.  $k_{\text{den}}$  should be chosen according to the resolution of structure that the user wishes to find – although a small value such as  $k_{\text{den}}$  was shown to produce the best clustering results in Oliver et al. (2022).  $S$  should be chosen to represent the strictness of confidence in the resulting clusters – the clustering power over synthetic galaxies is shown as a function of  $S$  in Fig. XX.

Parameter	Default Value	Valid Values	Functionality
$l_x$	$\infty$	$\mathbb{R}_{>0}$	The 3D FOF linking length used to find field haloes. Refer to Sec. 3.1.1 for more details.
<i>adaptive</i>	1	{0, 1}	A flag to apply a PCA transformation to field halo data. Refer to Sec. 3.1.2 for more details.
$k_{\text{den}}$	20	$\mathbb{N}_{\geq k_{\text{link}}}$	The number of nearest neighbours used to estimate local density. Refer to Sec. 3.1.3 for more details.
$k_{\text{link}}$	<i>auto</i>	$\mathbb{N}_{\geq 2 \wedge \leq k_{\text{den}}}$	The number of nearest neighbours used to densely connect points. Refer to Sec. 3.1.4 for more details.
$S$	5	$\mathbb{R}$	The minimum statistical significance of extracted clusters. Refer to Sec. 3.5.1 for more details.
<i>h_style</i>	1	{0, 1}	A flag denoting the style of the extracted hierarchy. Refer to Sec. 3.5.2 for more details.

#### 4.1.2 Effective Upper Limits on Data Set Clusteredness

### 4.2 Restrictions on Recoverable Clusters

## 5 A COMPARISON TO CLUSTAR-ND

### 5.1 Synthetic Data

### 5.2 Method

### 5.3 Results

## 6 CONCLUSIONS

### ACKNOWLEDGEMENTS

WHO gratefully acknowledges financial support through the Paulette Isabel Jones PhD Completion Scholarship at the University of Sydney.

### DATA AVAILABILITY

The data underlying this article may be made available on reasonable request to the corresponding author.

### REFERENCES

Abell G. O., 1958, *The Astrophysical Journal Supplement Series*, 3, 211  
 Akaike H., 1974, *IEEE transactions on automatic control*, 19, 716  
 Ankerst M., Breunig M. M., Kriegel H.-P., Sander J., 1999, in *ACM Sigmod record*. ACM, pp 49–60, doi:10.1145/304182.304187  
 Arifyanto M. I., Fuchs B., 2006, *A&A*, 449, 533  
 Avila S., et al., 2014, *Monthly Notices of the Royal Astronomical Society*, 441, 3488  
 Behroozi P. S., Wechsler R. H., Wu H.-Y., 2012, *The Astrophysical Journal*, 762, 109  
 Behroozi P., et al., 2015, *Monthly Notices of the Royal Astronomical Society*, 454, 3020  
 Belokurov V., Erkal D., Evans N. W., Koposov S. E., Deason A. J., 2018, *Monthly Notices of the Royal Astronomical Society*, 478, 611  
 Breunig M. M., Kriegel H.-P., Ng R. T., Sander J., 1999, *Principles of Data Mining and Knowledge Discovery*. Springer Berlin Heidelberg, pp 262–270, doi:10.1145/342009.335388  
 Campello R. J. G. B., Moulavi D., Zimek A., Sander J., 2015, *ACM Trans. Knowl. Discov. Data*, 10  
 Canovas H., et al., 2019, *A&A*, 626

Casamiquela L., et al., 2022, arXiv preprint arXiv:2206.03777  
 Costado M. T., Alfaro E. J., González M., Sampedro L., 2016, *Monthly Notices of the Royal Astronomical Society*, 465, 3879  
 Davis M., Efstathiou G., Frenk C. S., White S. D., 1985, *The Astrophysical Journal*, 292, 371  
 Diemand J., Kuhlen M., Madau P., 2006, *The Astrophysical Journal*, 649, 1  
 Duffau S., Zinn R., Vivas A. K., Carraro G., Méndez R. A., Winnick R., Gallart C., 2006, *ApJ*, 636, L97  
 Elahi P. J., et al., 2013, *Monthly Notices of the Royal Astronomical Society*, 433, 1537  
 Elahi P. J., Canas R., Poulton R. J. J., Tobar R. J., Willis J. S., Lagos C. d. P., Power C., Robotham A. S. G., 2019, *Publications of the Astronomical Society of Australia*, 36, e021  
 Epanechnikov V. A., 1969, *Theory of Probability & Its Applications*, 14, 153  
 Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in *Kdd*. pp 226–231  
 Freedman D., Diaconis P., 1981, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57, 453  
 Fuentes S. S., De Ridder J., Deboscher J., 2017, *Astronomy & Astrophysics*, 599, A143  
 Giocoli C., Tormen G., van den Bosch F. C., 2008, *MNRAS*, 386, 2135  
 Hadzhyiska B., Eisenstein D., Bose S., Garrison L. H., Maksimova N., 2021, *Monthly Notices of the Royal Astronomical Society*, 509, 501  
 Han J., Cole S., Frenk C. S., Benitez-Llambay A., Helly J., 2017, *Monthly Notices of the Royal Astronomical Society*, 474, 604  
 Helmi, Amina Veljanoski, Jovan Breddels, Maarten A. Tian, Hao Sales, Laura V. 2017, *A&A*, 598, A58  
 Higgs C., McConnachie A., Annau N., Irwin M., Battaglia G., Côté P., Lewis G., Venn K., 2021, *Monthly Notices of the Royal Astronomical Society*, 503, 176  
 Hurvich C. M., Tsai C.-L., 1989, *Biometrika*, 76, 297  
 Jayasinghe T., et al., 2019, *Monthly Notices of the Royal Astronomical Society*, 488, 1141  
 Jensen J., et al., 2021, *Monthly Notices of the Royal Astronomical Society*, 507, 1923  
 Johnston K. V., Hernquist L., Bolte M., 1996, *ApJ*, 465, 278  
 Kamdar H., Conroy C., Ting Y.-S., 2021, arXiv preprint arXiv:2106.02050  
 Knebe A., et al., 2011, *Monthly Notices of the Royal Astronomical Society*, 415, 2293  
 Knebe A., et al., 2013, *Monthly Notices of the Royal Astronomical Society*, 428, 2039  
 Knollmann S. R., Knebe A., 2009, *ApJS*, 182, 608  
 Koppelman H. H., Helmi A., Massari D., Price-Whelan A. M., Starkenburg T. K., 2019, *Astronomy & Astrophysics*, 631, L9  
 Kounkel M., Covey K., 2019, *The Astronomical Journal*, 158, 122  
 Lee J., et al., 2014, *Monthly Notices of the Royal Astronomical Society*, 445, 4197



- Lövdal S. S., Ruiz-Lara T., Koppelman H. H., Matsuno T., Dodd E., Helmi A., 2022, arXiv preprint arXiv:2201.02404
- Maciejewski M., Colombi S., Springel V., Alard C., Bouchet F. R., 2009, *MNRAS*, **396**, 1329
- Mahajan S., Singh A., Shobhana D., 2018, *Monthly Notices of the Royal Astronomical Society*, **478**, 4336
- Mahalanobis P. C., 1936.
- Malhan K., Ibata R. A., 2018, *Monthly Notices of the Royal Astronomical Society*, **477**, 4063
- Massaro F., Alvarez-Crespo N., Capetti A., Baldi R., Pillitteri I., Campana R., Paggi A., 2019, *The Astrophysical Journal Supplement Series*, **240**, 20
- Mateu C., Bruzual G., Aguilar L., Brown A. G. A., Valenzuela O., Carigi L., Velázquez H., Hernández F., 2011, *Monthly Notices of the Royal Astronomical Society*, **415**, 214
- Mateu C., Read J. I., Kawata D., 2017, *Monthly Notices of the Royal Astronomical Society*, **474**, 4112
- McConnachie A. W., et al., 2018, *The Astrophysical Journal*, **868**, 55
- McInnes L., Healy J., Astels S., 2017, *Journal of Open Source Software*, **2**, 205
- Oliver W. H., Elahi P. J., Lewis G. F., Power C., 2021, *Monthly Notices of the Royal Astronomical Society*, **501**, 4420
- Oliver W. H., Elahi P. J., Lewis G. F., 2022, The Hierarchical Structure of Galactic Haloes: Generalised N-Dimensional Clustering with CluSTAR-ND, doi:10.48550/ARXIV.2201.10694, <https://arxiv.org/abs/2201.10694>
- Onions J., et al., 2012, *Monthly Notices of the Royal Astronomical Society*, **423**, 1200
- Onions J., et al., 2013, *Monthly Notices of the Royal Astronomical Society*, **429**, 2739
- Pearson S., Clark S. E., Demirjian A. J., Johnston K. V., Ness M. K., Starkenburg T. K., Williams B. F., Ibata R. A., 2022, *The Astrophysical Journal*, **926**, 166
- Press W. H., Schechter P., 1974, *The Astrophysical Journal*, **187**, 425
- Ruiz A., Corral A., Mountrichas G., Georgantopoulos I., 2018, *Astronomy & Astrophysics*, **618**, A52
- Sain S. R., 2002, *Computational Statistics & Data Analysis*, **39**, 165
- Sander J., Qin X., Lu Z., Niu N., Kovarsky A., 2003, in Whang K.-Y., Jeon J., Shim K., Srivastava J., eds, *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, pp 75–87
- Schwarz G., 1978, *The annals of statistics*, pp 461–464
- Sharma S., Johnston K. V., 2009, *The Astrophysical Journal*, **703**, 1061
- Shih D., Buckley M. R., Necib L., Tamasas J., 2021, *Monthly Notices of the Royal Astronomical Society*, **509**, 5992
- Soto M., et al., 2022, *Monthly Notices of the Royal Astronomical Society*, **513**, 2747
- Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, *Monthly Notices of the Royal Astronomical Society*, **328**, 726
- Tormen G., Moscardini L., Yoshida N., 2004, *Monthly Notices of the Royal Astronomical Society*, **350**, 1397
- Walmsley M., et al., 2022, *Monthly Notices of the Royal Astronomical Society*, **509**, 3966
- Ward J. L., Kruijssen J. D., Rix H.-W., 2020, *Monthly Notices of the Royal Astronomical Society*, **495**, 663
- Webb S., et al., 2020, *Monthly Notices of the Royal Astronomical Society*, **498**, 3077
- Williams M. E. K., et al., 2011, *The Astrophysical Journal*, **728**, 102
- Yuan Z., Chang J., Banerjee P., Han J., Kang X., Smith M. C., 2018, *ApJ*, **863**, 26
- Zhang A. X., Noulas A., Scellato S., Mascolo C., 2013, in 2013 International Conference on Social Computing. IEEE, pp 69–74

This paper has been typeset from a  $\text{\LaTeX}$  file prepared by the author.

# Chapter 6

## Conclusions

In this thesis, I have developed a series of novel generalised astrophysical clustering algorithms that are distinct from the algorithm families of current simulation and observation specific structure finders. With a particular focus on revealing the hierarchical structure of galactic haloes, I have shown that these algorithms produce robust classifications of astrophysical structure and are ideally suitable for application to both simulated and observational data sets of galactic haloes.

In Chapter 2, I discuss the many prevailing aspects of common use clustering algorithms (such as similarity measurement, cluster models, computational methods, and statistical evaluation) as well as of those that are specifically designed and used for the analysis and discovery of astrophysical structures such as galaxies, haloes, subhaloes, and tidal debris. With this I determine that, while individually the simulation and observation specific structure finders are powerful and useful methods of uncovering specific structure types within their respective contexts, there exists a division between the two algorithm types. With this in mind, I assert that a generalised structure finding algorithm that does not reduce the available clustering information and is robust to noise is sorely needed.

In Chapter 3, I develop the configuration space based astrophysical clustering structure finder **HALO-OPTICS** by using the general-purpose density-based clustering algorithm **OPTICS**. After having designed and optimised the algorithm, I then compare its output to the state-of-the-art galaxy/(sub)halo finder **VELOCIRAPTOR**. Even though **VELOCIRAPTOR** is a phase-space finder and therefore also uses velocities to find structures where **HALO-OPTICS**, I find excellent agreement between the two codes with **HALO-OPTICS** even being to uncover kinematically coherent tidal streams. Hence displaying the power of using an adaptive density clustering algorithm over static density clustering algorithm such as **FOF** (the base algorithm of **VELOCIRAPTOR**).

In Chapter 4, I build upon **HALO-OPTICS** to create the generalised astrophysical

structure finder **CLUSTAR-ND**. I find that by reducing the **HALO-OPTICS** radial search to a simple  $k$  nearest neighbours search, **CLUSTAR-ND** can achieve a  $\geq 3$  orders of magnitude run-time reduction over **HALO-OPTICS** while boasting greater sensitivity to clustering structure. I also find that a global PCA transform is typically sufficient for producing near optimal clustering results over multidimensional data with various units among its features, and that an iterative PCA constructs a locally adaptive metric that produces mixed results. In applying **CLUSTAR-ND** in various clustering scenarios to semi-analytic simulated galaxies, I find that it is able to classify a large portion of the tidal debris associated with disrupted satellite mergers and that this portion increases with an increase to the feature space complexity provided.

In Chapter 5 I present **CLUSTARR-ND** – a work in progress. **CLUSTARR-ND** replaces the complex 3 parameter cluster extraction process used in **HALO-OPTICS** and **CLUSTAR-ND** for a much simpler, single-parameter, statistical-significance-motivated procedure. As a result, **CLUSTARR-ND** is able to extract clusters from data sets with varied levels of noise without the need for adjusting the controlling parameters. **CLUSTARR-ND** also produces an ordered-density plot – analogous to the reachability plot of **OPTICS** – that allows for the user to visualise the clustering structure in its entirety without the loss of structural information.

As the culmination of the sequential algorithmic development conducted throughout this thesis, the **CLUSTARR-ND** algorithm is an almost entirely data-driven and adaptive clustering algorithm – effectively only requiring the user to define the number of nearest neighbours used to estimate the local density about each data point and the threshold level of statistical significance that each extracted cluster must have. It is blind to cluster size and shape, is more sensitive to fine structure variations than current methods, and is applicable to any size and dimensionality data set with any level of noise contamination. As such, the **CLUSTARR-ND** algorithm is ideally suited for generalised astrophysical structure finding in the contexts of both synthetic and observational data sets.

## 6.1 Future Outlook

With a series of high performance astrophysical clustering algorithms in-hand, I now set out to apply them to various large-scale observational data sets. For this I will re-write these algorithms in a low-level coding language such as C++ and utilise parallel computations to improve their performance. In their application, some additional approaches can be incorporated to improve the recovery of some cluster types. These approaches and the data sets I intend on applying these techniques

and methods to are now outlined in the following subsections.

### 6.1.1 Additional Machine Learning Approaches

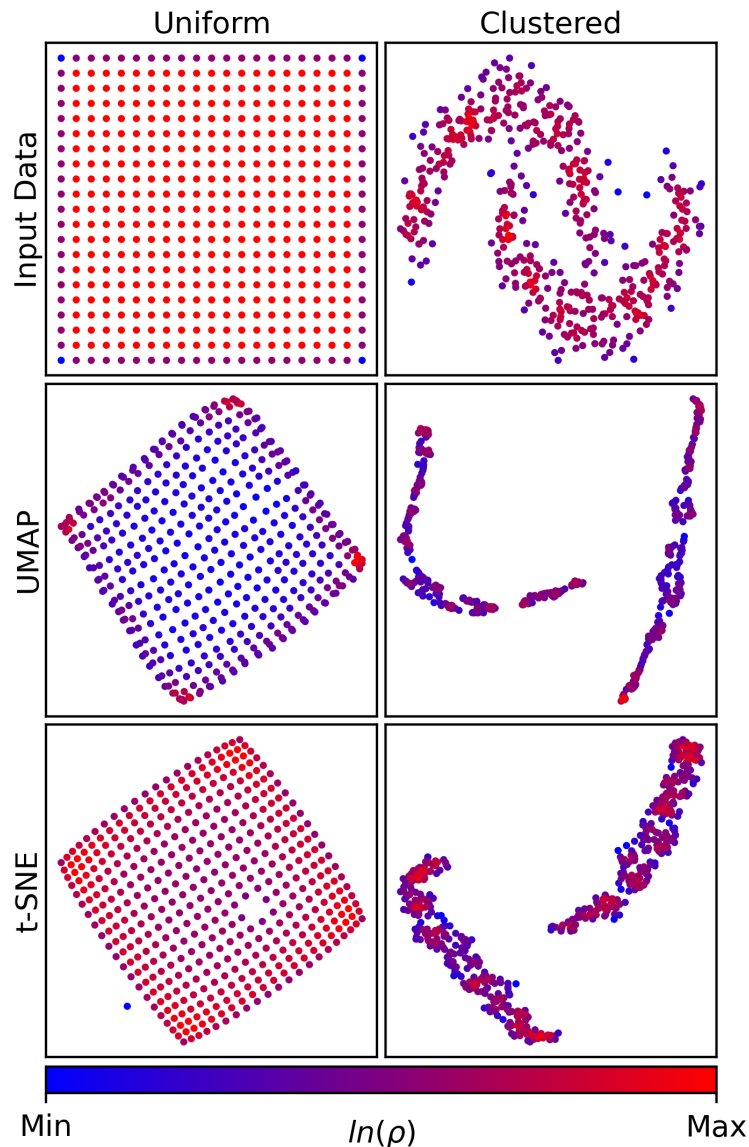
To further improve the efficacy of these algorithms, during this project I will initially develop two additional techniques for providing cluster existence/membership probabilities as well as for enhancing the returned clustering results and providing greater recovery of tidal streams/diffuse features.

#### Producing Fuzzy Clusters from Fuzzy Data

Very few astrophysical clustering algorithms can extract a set of fuzzy clusters from an input of fuzzy data i.e. clusters of a probabilistic nature from data of the same nature. The **STREAMFINDER** algorithm [354] incorporates data uncertainties in this way and just recently a statistical procedure was coupled with the use of the **ENLINK** algorithm to do the same [374]. However, these approaches come with their own pitfalls. The clustering results of the **STREAMFINDER** algorithm are heavily dependent upon the mass distribution model of the MW (and hence it has only been used in the case of substructure finding within the MW) and is also only capable of detecting dynamically cold (stream-like) substructures. Furthermore, the **STREAMFINDER** algorithm only considers the values and uncertainties of a fixed set of measured quantities in the data and not how these change or project into different feature spaces. The statistical approach used by Malhan et al. [374] (described in Sec. 4.1 – 4.3 therein) does not suffer from the latter, however, it is not feasibly scalable to data sets with a large number of data points as it relies on the clustering being non-hierarchical, which is not the case in many astrophysical clustering scenarios.

I will create a fuzzy clustering method that builds upon the concepts of these methods whilst overcoming their pitfalls and simultaneously being completely generalisable and equally prolific when applied to any data set. The **CLUSTARR-ND** algorithm currently provides a hard clustering – i.e. the data points are either contained within a cluster or they are not. This novel method will involve the comparison of  $N$  such hard clusterings each produced from a random sample of the data set. From this comparison, the resultant cluster existence and membership probabilities will be found by considering how the inherent conditional statistical significances that **CLUSTARR-ND** provides propagate in a Bayesian inference driven manner. Such a method will prove exceedingly useful in its application across a wide range of data sets as current methods of predicting cluster existence/membership probabilities typically fall short of encapsulating *all* probable clusters. This is due to current

techniques often relying on a post-process analysis of a single clustering realisation, rather than the intra-process analysis of many clustering realisations.



**Figure 6.1:** Estimated manifold projections from UMAP and  $\tau$ -SNE of both uniform and clustered data sets in 2 dimensions. Data points are coloured by density. These algorithms can increase clustering power at the cost of additional noise and the loss of global structure.

### Improving Cluster Separability with Riemannian Manifold Estimation

Most density-based clustering algorithms use a globally constant metric to extract clusters, such that for any two points in the feature space domain, the distance between them remains invariant under translation (and rotation for the case of Euclidean metrics). While this works well if the in-situ clusters are easily separable

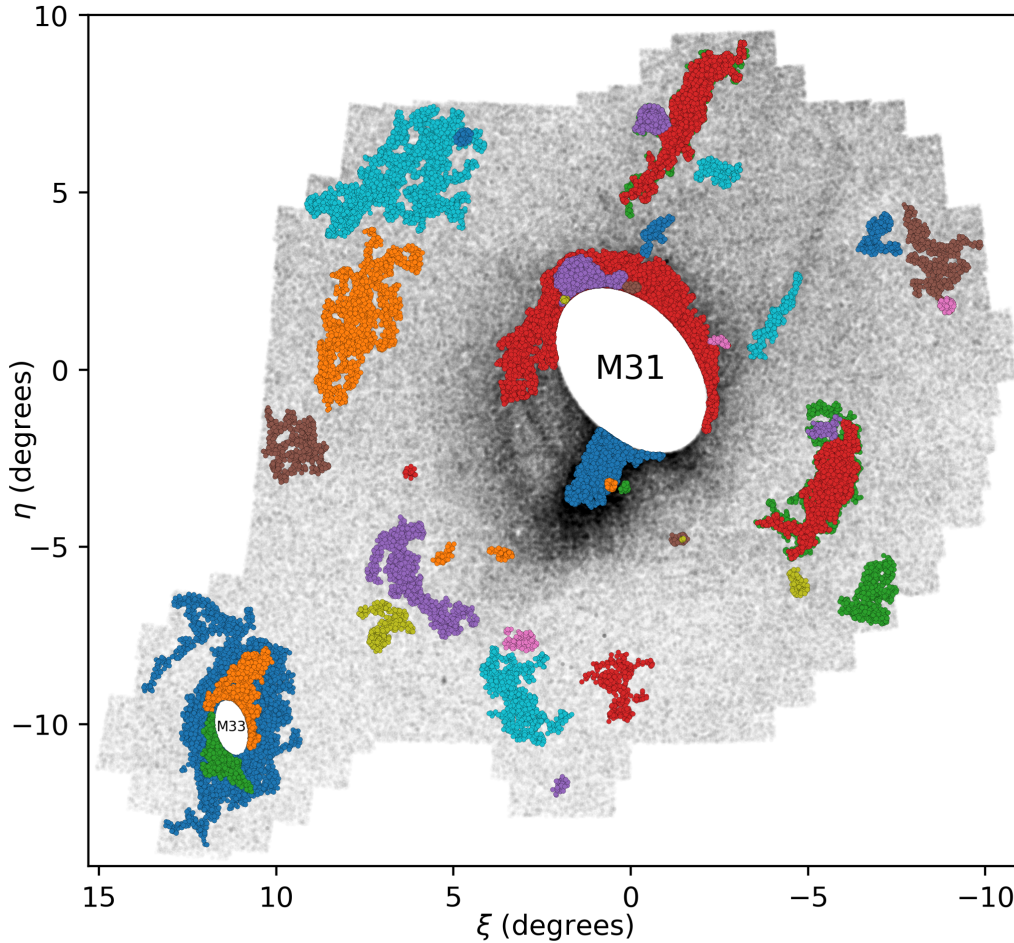
as overdensities under this global metric, the approach begins to break down when attempting to find highly anisotropic clusters. In large data sets where not all physical features are known about an object, this can result in some clusters not being detected or not being separated from other nearby overdensities (a good example of this can be seen in Fig. 6.2). I will use transformation-invariant quantities and concepts from fuzzy topology to produce a robust locally-adaptive-metric algorithm that preserves the relevant (sub)structure.

Of the few unsupervised codes to attempt this, the functional assumption is that all points within the data set are uniformly distributed over some intrinsic Riemannian manifold and that the manifold is Euclidean only locally – e.g. ENBID [256], UMAP [257], and t-SNE [258]. A comparison of the UMAP and t-SNE outputs is shown in Fig. 6.1 for uniform and clustered data sets. These algorithms can have mixed results and are extremely computationally expensive.

Following suit, I will formulate our locally-adaptive-metric algorithm to quickly transform the empirical cumulative distribution along each dimension of the input data set into a pseudo-random uniform-like cumulative distribution – thereby constructing a suitable map between the underlying Riemannian manifold and the data. The effect of embedding the data within such a manifold (in an astrophysical context) is that the distances between data points within the same structure are contracted, increasing their density, in turn resolving the structure more clearly in contrast to the surrounding structures. The preliminary algorithm that achieves this has already been created and while operational improvements can still yet be made, this will prove simple to implement and will greatly improve clustering results from the CLUSTARR-ND algorithm.

### 6.1.2 Applications to Observational Data Sets

With a suite of purpose built and unsupervised state-of-the-art machine learning techniques in-hand, I will then perform a series of robust clustering analyses of the MW and the surrounding Local Group revealing the in-situ clustering structure of the local Universe. Since each of these algorithms have/will be built explicitly for generalised applicability, I need not perform extensive pre-process data reduction steps or remove any of the available information to ensure that our methods produce high quality results. Typically, the only reduction step that will need to be performed will be the removal of background/foreground contamination so that the input data is representative of the environment that I wish to probe for clustering structure. The following sub-subsections contain the details of various data sets to which I intend on applying our methods and the insights that finding the clustering structure



**Figure 6.2:** A clustering of the foreground-subtracted PAndAS data set produced by CLUSTARR-ND. The regions containing M31 and M33 have been masked out to reduce the numerical artefacts that occur due to the competing effects of high spatial density and the limiting resolution of the survey. The input data includes the two spatial dimensions and the photometric metallicity,  $[\text{Fe}/\text{H}]$ . Each cluster is shown in a different colour. Clusters shown here have a statistical significance of at least  $3\sigma_{\text{noise}}$ .

is expected to reveal.

### The Pan-Andromeda Archaeological Survey

The Pan-Andromeda Archaeological Survey (PAndAS; [362]) is a large-scale panoramic astronomical survey conducted over the surrounds of our nearest galactic neighbour, the Andromeda galaxy (M31), and extends to include the Triangulum galaxy (M33). The positions on the sky are reported for all stars in this survey as well as a measure of the metallicity (photometric  $[\text{Fe}/\text{H}]$ ). Analysis of the clustering structure within the M31 – M33 system was performed by McConnachie et al. [138] using the foreground-subtracted data of stellar positions. However, this was done using the

general purpose clustering algorithm, **OPTICS**.

The **HALO-OPTICS** clustering algorithm is an improvement on **OPTICS** (for astrophysical data sets) and likewise **CLUSTAR-ND** and **CLUSTARR-ND** are also both significant improvements to it, with the latter being superior to each of the others – particularly in the context of astrophysical clustering with observational data. By applying **CLUSTARR-ND** and the algorithms described in Sec. 6.1.1 to the PAndAS data (which I have in-hand), I will robustly improve upon the clustering structure found by McConnachie et al. [138] by producing a clustering with; a higher sensitivity to small-scale structure variations; a more robust and deterministic estimate of local density; and a more statistically relevant framework. Furthermore, where **OPTICS** was only applied to the on-the-sky positions of stars, I will find a clustering that incorporates the metallicity of these stars too – effectively increasing the feature space (and hence available clustering information) by a factor of  $\sim 50\%$ .

As an example of how readily achievable this is, Fig. 6.2 is a preliminary clustering of the ‘foreground-subtracted’ (refer to [138] for details on this) PAndAS data set with M31 and M33 both masked to reduce the compound effects of increased spatial density and the limiting resolution of the survey. The input data features the positions on the sky and the photometric metallicity,  $[\text{Fe}/\text{H}]$ . Some extra care could be taken to reduce the noise and better prepare this data set for clustering, however, I see already that the clustering produced from this depicts many known substructures including dwarf galaxies and stellar substructures in the M31-M33 system. It also hints at a possible newfound stream-like substructure to the North-East of M33 that has not been completely separated from the M33 halo – the first detection of this substructure.

It is this lack of separability that the locally-adaptive-metric algorithm discussed in Sec. 6.1.1 will improve. The algorithm will also improve the recovery of the obvious streams situated close to M31 that have only been partially recovered as a result of them not being separable from the galactic halo by a saddle-point in the density field. Specifically, an appropriate locally-adaptive-metric will improve these clustering results by artificially creating a preference for connecting data points together along the length of the streams. The algorithm will do this by re-shaping the surface of equal distance surrounding each data point within the streams. Currently, these streams could be better resolved simply by adjusting the global metric, however, this would come at the cost of resolving other substructures that are not well defined by the adjusted metric. The ill-defined and jagged edges of these substructures will also be improved by this, however, these will also benefit from the fuzzy clustering method discussed in Sec. 6.1.1 – assuming this is due to noise effects.



### The Gaia Data Release 3

The third regular data release from the Gaia mission (Gaia DR3; [360]) has been recently released and boasts an enormous sampling of stars from the MW galaxy. Among the features included are those of the full astrometric solution (i.e. positions on-the-sky, parallaxes, and proper motions) for  $\sim 1.46$  billion stars. Many attempts at finding clusters in Gaia data have been made on earlier data releases, most notably with the **STREAMFINDER** algorithm on the Gaia DR2 and EDR3 catalogues [139]. While this application has unveiled much about the tidal stream structure of the MW, the algorithm is highly dependent upon the model of the MW mass distribution as well as upon the numerous user-tuned input parameters that are needed for its effective operation. Other clusterings have also been found from these data, however, the algorithms responsible for constructing these are not capable of targeting all substructure types simultaneously. Specifically, current clustering algorithms are not able to completely encapsulate both self-bound and unbound substructures as well as those with highly anisotropic and arbitrary shapes concurrently. Hence, the clusters returned from these algorithms are not entirely descriptive of the intimate link between a satellite, its tidally disrupted debris, and its host galaxy.

By applying **CLUSTARR-ND** and the techniques described in Sec. 6.1.1, I will achieve the latter as these methods are/will be less restrictive and unbiased towards cluster shapes and sizes. With such an extensive data set and a higher sensitivity to the finer clustering structure than has been possible with previous attempts, I will not only provide independent predictions of known substructures but will also uncover as yet unknown substructures, greatly expanding the catalogue of coherent substructures within the MW – a huge step towards a greater understanding of the evolution of our own galaxy.

### The Galactic Archaeology with HERMES Survey

The GALactic Archaeology with HERMES survey (GALAH; [363]) is a spectroscopic stellar survey which has its third data release available publicly. The release extends the Gaia catalogue by providing the line-of-sight velocity, additional stellar parameters, and up to 30 elemental abundances for each star. A previous GALAH catalogue has been used for chemical tagging [389], which requires an initial clustering analysis (using the astrometric solution from Gaia) followed by a comparison to the known spectra of each cluster’s members.

For the reasons discussed throughout this thesis and in Sec. 6.1.1, our techniques will provide improved insight into the chemical evolution of the MW’s merger history found with Gaia DR3 when chemically tagged with the spectra from GALAH.

Furthermore, I will apply `CLUSTARR-ND` and the algorithms from Sec. 6.1.1 directly to the catalogue of stellar spectra within GALAH DR3. A clustering over GALAH's high dimensional set of elemental abundances has not yet been successfully attempted, however, with a much higher sensitivity to clustering structure I will uncover local variations in chemical evolution that have not yet been studied.

### **The Sloan Digital Sky Survey**

The Sloan Digital Sky Survey is a large scale survey whose fifth generation (SDSS-V; [361]) is expected to release its first data release later this year. SDSS-V is to be the first all-sky time-domain spectroscopic survey boasting the near infra-red and/or optical spectra of more than 4 million stars situated throughout the MW and Local Group. The stellar parameters and elemental abundances derived from these data will enable a unique global map of the MW's fossil records that survive in its stars and interstellar material. The SDSS-V survey will make use of the BOSS and APOGEE spectrographs to provide a high resolution chemo-dynamical map of the region surrounding the Sun. With such fine detail and a large feature space, the SDSS-V data will be ideally suited to extract clusters from.

By applying our clustering methods to this data I will extract new substructures defined by their high dimensional chemo-dynamical coherence – which is a feature space that is currently challenging to obtain substructures from. The identification of these anisotropic substructures will prove invaluable for studying the chemo-dynamical evolution of the MW and the Local Group.

# Appendix A

## Contributing Publications

Chapters 3, 4, and 5 are assembled from research papers where I am lead author, and as such, make up the core of my thesis. Here in this appendix, I present the publications that I have contributed to as a co-author but whose research is not predominantly my own. These publications are provided as a complete record of my research activities during my PhD tenure.




### A.1 The Asymmetric Dwarf Galaxy Distribution around Andromeda

This section presents the published journal article:

- A1. *On the Origin of the Asymmetric Dwarf Galaxy Distribution around Andromeda.* Z. Wan, **W. H. Oliver**, G. F. Lewis, J. I. Read, & M. L. M. Collins. *MNRAS* **492**, 456, 2020. [[arXiv:1912.02393](https://arxiv.org/abs/1912.02393)].

I contributed to this paper by assisting in the design of the analysis and by interpreting the results that were produced therefrom. I also contributed by co-writing the manuscript with the lead author Dr. Zhen Wan.

# On the origin of the asymmetric dwarf galaxy distribution around andromeda

Zhen Wan <sup>1</sup>★, William H. Oliver,<sup>1</sup> Geraint F. Lewis <sup>1</sup>, Justin I. Read<sup>2</sup>  
and Michelle L. M. Collins <sup>2</sup>

<sup>1</sup>*Sydney Institute for Astronomy, School of Physics A28, The University of Sydney, NSW 2006, Australia*

<sup>2</sup>*Department of Physics, University of Surrey, Guildford, Surrey GU2 7XH, UK*

Accepted 2019 December 4. Received 2019 November 15; in original form 2019 August 21

## ABSTRACT

The dwarf galaxy distribution surrounding M31 is significantly anisotropic in nature. Of the 30 dwarf galaxies in this distribution, 15 form a disc-like structure and 23 are contained within the hemisphere facing the Milky Way. Using a realistic local potential, we analyse the conditions required to produce and maintain these asymmetries. We find that some dwarf galaxies are required to have highly eccentric orbits in order to preserve the presence of the hemispherical asymmetry with an appropriately large radial dispersion. Under the assumption that the dwarf galaxies originate from a single association or accretion event, we find that the initial size and specific energy of that association must both be relatively large in order to produce the observed hemispherical asymmetry. However if the association was large in physical size, the very high-energy required would enable several dwarf galaxies to escape from the M31 and be captured by the Milky Way. Furthermore, we find that associations that result in this structure have total specific energies concentrated around  $E = V_{\text{esc}}^2 - V_{\text{init}}^2 \sim 200^2 - 300^2 \text{ km}^2 \text{ s}^{-2}$ , implying that the initial velocity and initial position needed to produce the structure are strongly correlated. The overlap of initial conditions required to produce the radial dispersion, angular dispersion, and the planar structure is small and suggests that either they did not originate from a single accretion event, or that these asymmetric structures are short-lived.

**Key words:** galaxies: evolution – galaxies: kinematics and dynamics.

## 1 INTRODUCTION

Early evidence of structures in the distributions of dwarf galaxies dates back several decades when Lynden-Bell (1976) discovered that several globular clusters and dwarf galaxies surrounding the Milky Way (MW) lay in streams of high-velocity clouds that were thought to form a planar structure (Lynden-Bell & Lynden-Bell 1995). The suspicion that a great disc of MW dwarf galaxies existed was independently confirmed through follow-up studies (Kroupa, Theis & Boily 2005; Metz, Kroupa & Jerjen 2007; Metz, Kroupa & Libeskind 2008; Pawlowski, Pflamm-Altenburg & Kroupa 2012). This then raised the question of the origin of such structures, since the likelihood that they would assemble from a previously isotropic distribution is extremely small. However, studies have suggested that the structure may not rotate coherently (Cautun et al. 2015a; Phillips et al. 2015) and also that the statistical relevance of disc configurations is heavily influenced by detection bias (Cautun et al. 2015b; Buck, Dutton & Macciò 2016; Maji et al. 2017). More recent papers have created further tension with these findings claiming that

the MW satellites as a whole do not lie in a thin plane, although there is strong evidence that their distribution is anisotropic (Gaia Collaboration et al. 2018; Simon 2018).

Such an anisotropic distribution of dwarf galaxies has been seen in M31, with the Pan-Andromeda Archaeological Survey (McConnachie et al. 2009), claiming that 15 of the 27 observed dwarf galaxies constitute a great disc, all with same sense of rotation about M31 (Conn et al. 2013; Ibata et al. 2013). The size of this disc is at least 400 kpc in diameter with perpendicular scatter of less than  $\sim 14$  kpc. Adding to the complexity, Conn et al. (2012, 2013) have found that the dwarf galaxies surrounding M31 possess a significant hemispherical anisotropy, with 21 of the 27 dwarf galaxies are contained within the same hemisphere. However, the radial distribution of those dwarf galaxies is less special than the directional distribution, with the distances of the dwarf galaxies to the M31 ranging from 40 to 400 kpc.

The origin of these anisotropic structures has been the subject of much debate. The nature of the cosmic web imposes some coherence on the accretion of neighbouring galactic structures (Zentner et al. 2005; Libeskind et al. 2011, 2015), and consequently, galaxies often fall into larger structures as part of a group (D’Onghia & Lake 2008; Read et al. 2008; Li & Helmi 2009). In principal, these effects

\* E-mail: zwan3791@uni.sydney.edu.au

could be responsible for manifesting dwarf galaxy disc structures that surround larger host galaxies within the  $\Lambda$  cold dark matter ( $\Lambda$ CDM) model of cosmology (e.g. Lovell et al. 2011; Goerd, Burkert & Ceverino 2013; Wang, Frenk & Cooper 2013; Bahl & Baumgardt 2014; Buck, Macciò & Dutton 2015; Gillet et al. 2015). However, some studies have suggested that the discrepancy between the anisotropy in observation and simulation is significant and is not easily explained with  $\Lambda$ CDM cosmologies (Kroupa et al. 2005; Pawlowski et al. 2012; Pawlowski & Kroupa 2013; Ibata et al. 2014; Forero-Romero & Arias 2018; Pawlowski et al. 2019). Other studies advocate that 10 (Cautun et al. 2015b), or even 20 per cent (Shao et al. 2016) of  $\Lambda$ CDM haloes have even more prominent planes than those present in the Local Group. It has been proposed that pairs of large galaxies similar to those in the Local Group can impose a shape alignment of satellite galaxies (Wang et al. 2019). Furthermore, Libeskind et al. (2016) and Gong et al. (2019) have suggested that it is statistically likely for satellite distributions surrounding pairs of galaxies – such as the M31–MW system – to be lopsided, though these satellites are primarily on their first infall. Inferring the whereabouts of dwarf galaxies in the past can be challenging, particularly if a dwarf’s position is to be tracked over more than a full orbit of its host, due to many dwarfs falling in as part of associations (Lux, Read & Lake 2010).

Other attempts to explain disc-like structures of dwarf galaxies from a smaller scale perspective have also been made. Pasetto & Chiosi (2009) have tested the feasibility of disc structures in the Local Group as a result of tidal effects, finding that these could account for the planar structure excluding those tightly bound dwarf galaxies. Bowden, Evans & Belokurov (2013) show that in a triaxial Navarro–Frenk–White (NFW; Navarro, Frenk & White 1996) potential, it is possible for a thin-disc structure to persist over cosmological time-scales if and only if it lies in the planes perpendicular to the long or short axis of a triaxial halo, else it will double in thickness within  $\sim 5$  Gyr. Later Bowden, Evans & Belokurov (2014) calculated the life-times of inward falling associations in various potentials and found that asymmetric structures could survive longer than the current age of the universe in the outer regions of nearly spherical potentials.

Given this groundwork, our investigation aims to numerically investigate how a more realistic dynamic potential configuration contributes to the formation of these asymmetric structures. We construct the M31–MW system potential by considering the disc, bulge, and halo components of these galaxies in Section 2. Then in Section 3.1 we place all observed dwarf galaxies surrounding M31 into this potential at their current positions and integrate backwards with various tangential velocities so as to obtain the orbital properties of each dwarf galaxy. In Section 3.2, we also integrate the orbits of numerous dwarf galaxy associations forwards in time to identify the set of initial conditions required to assemble the currently observed structures surrounding M31 from a single association of dwarf galaxies. Finally, we discuss our findings in Section 4.

## 2 METHOD

### 2.1 Potential

To appropriately consider the dynamic behaviour of the dwarf galaxies surrounding M31, a superposition of both the M31 and the MW gravitational potentials are modelled. For the MW, we use the MW POTENTIAL 2014 in *galpy* (Bovy 2015), which is composed of equations (1)–(3).

We use a spherically symmetric power-law potential with an exponential cut-off for the bulge. This is derived from the mass–density model,

$$\rho(r) = \rho_0 r^{-\alpha} \exp((-r/r_c)^2), \quad (1)$$

for which we use a power-law index of  $\alpha = 1.8$ , and a cut-off radius of  $r_c = 1.9$  kpc.

The disc is modelled using the axisymmetric Miyamoto–Nagai potential,

$$\Phi(R, z) = -\frac{\Phi_0}{\sqrt{R^2 + (a + \sqrt{z^2 + b^2})^2}}, \quad (2)$$

where  $R = \sqrt{x^2 + y^2}$  in galactocentric coordinates. Here, we use potential parameters  $a = 3$  kpc and  $b = 0.28$  kpc.

To model the effect of the dark-matter halo, we use the NFW potential with the density profile,

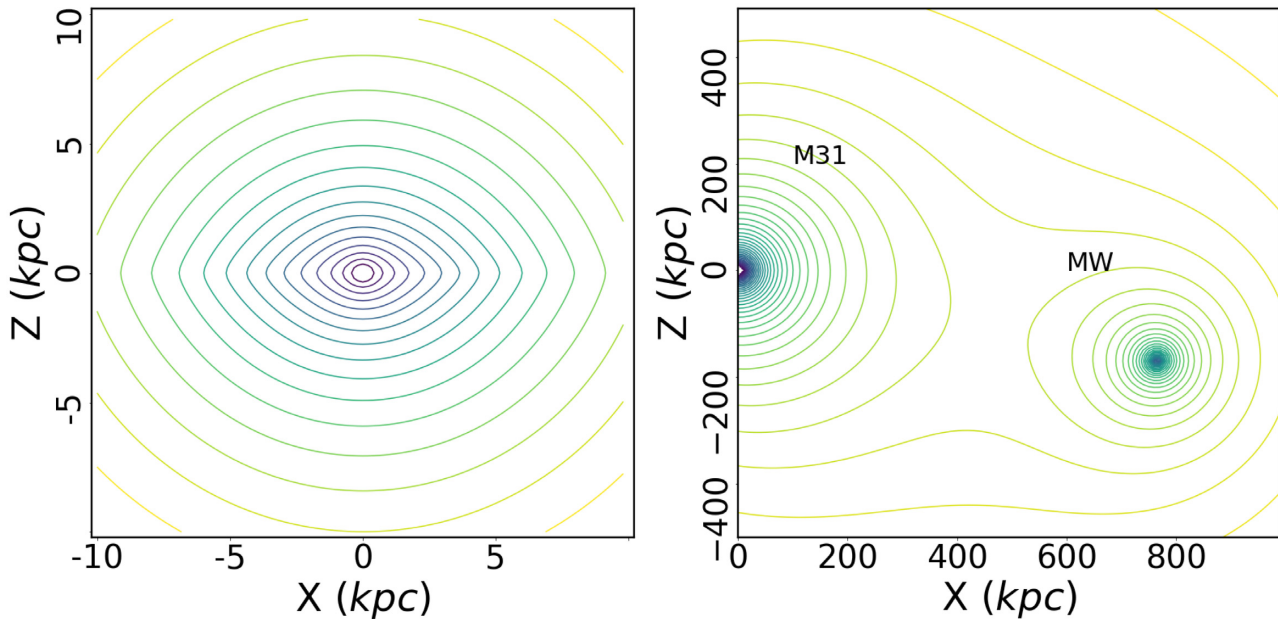
$$\rho(r) = \frac{\rho_0}{(r/h)(1 + r/h)^2}, \quad (3)$$

and characteristic radius  $h = 16$  kpc.

This potential is scaled so that the circular velocity at  $r = 8$  kpc away from the galactic centre in the disc ( $z = 0$  kpc) is set to  $220 \text{ km s}^{-1}$ . In addition, the potential parameters are also tuned to match multiple data sets that include observations of the velocity dispersion, vertical force, terminal-velocity, mid-plane density profile slope, and total mass (Clemens 1985; Dehnen & Binney 1998; Holmberg & Flynn 2000; McClure–Griffiths & Dickey 2007; Binney & Tremaine 2008; Xue et al. 2008; Bovy et al. 2012; Bovy & Rix 2013; Zhang et al. 2013). The corresponding virial mass of this potential is  $0.8 \times 10^{12} M_\odot$  that agrees well with the virial mass from dynamical analyses (e.g. Xue et al. 2008; Deason et al. 2012; Kafle et al. 2012, 2014), and at the low end of a massive Milky Way (e.g. Li & White 2008; Watkins, Evans & An 2010; Sohn et al. 2018; Posti & Helmi 2019).

For M31, we use the same potential form as for the MW; however, we set  $a = 5.09$  kpc,  $b = 0.28$  kpc in equation (1), and  $h = 20$  kpc in equation (3) (Seigar, Barth & Bullock 2008), and the total mass of M31 is 1.5 times of the MW total mass. In the integration, we also examine the effects of the oblate and prolate NFW profiles using the TRIAXIAL NFW POTENTIAL with the equi-density radius defined as  $r = \sqrt{x^2 + y^2 + (0.5z)^2}$  and  $r = \sqrt{x^2 + y^2 + (1.5z)^2}$ , respectively, for the M31 potential. This increases the asymmetry of the potential (e.g. Dubinski 1994; Debattista et al. 2008) and might lead to the anisotropic distribution of the dwarf galaxies (Hayashi & Chiba 2014). To focus on the shape, we set all the dark halo profiles to have same mass. The mass of the M31–MW system ( $\sim 2 \times 10^{12} M_\odot$ ) we have constructed is consistent with recent timing argument constraints (Penarrubia et al. 2015). The MW halo is assumed to be spherically symmetric, since the shape of the MW potential should be less significant due to the distance between the MW and M31. Furthermore, within  $\sim 50$  kpc, the MW halo appears to be rather round (e.g. Ibata et al. 1998; Read 2014; Wegg, Gerhard & Bieth 2019).

The left-hand panel of Fig. 1 depicts the galactic potential contour in  $X$ – $Z$  plane within 10 kpc of which M31 is at the centre. Then by adopting the M31 configuration as position angle  $\theta = 39.8^\circ$  and inclination  $i = 77.5^\circ$  (de Vaucouleurs 1958; McConnachie & Irwin 2006), we place another galactic potential at the position of the MW as seen from M31 with the corresponding configuration to include the effect of the MW. The right-hand panel of Fig. 1 portrays an overview of the M31–MW system potential we use



**Figure 1.** *Left:* Contour lines of equi-potential in  $X$ - $Z$  plane. This potential includes a bulge, a disc, and a halo, where the bulge and halo are spherical symmetric and the disc is axisymmetric. Here, the  $X$ - $Y$  plane lies in the disc plane. *Right:* The potential of the M31–MW system in the  $X$ - $Z$  plane, which is centred on the galactic centre of M31. Here, the  $X$ - $Y$  plane coincides with the M31 disc plane. The MW lies to the right of the figure and due to its presence, a slight deformation of the M31 potential is visible at  $\sim 400$  kpc (see the Fig. A1 for the demonstration of the prolate/oblate profiles).

in our calculations. Here, it is clear that the contours deform at  $\sim 400$  kpc from each galaxy’s centre.

## 2.2 Integration

To calculate the orbits of the dwarf galaxies within the potential we build, we use the `SOLVE_IVP` function from the `scipy` package with an adaptive step-size (Jones et al. 2001) to solve the differential equation of motion numerically. Here, we set the relative tolerance to be  $10^{-11}$ . Prior to integration, we calculate the current MW velocity based on the M31 line-of-sight velocity from McConnachie (2012) and the solar reflex motion as (11.1, 232.24, 7.25)  $\text{km s}^{-1}$  (Schönrich, Binney & Dehnen 2010; Bovy 2015). We then use this relative motion to calculate the position of the MW for all times during the past 10 Gyr. This then realistically ensures that the effect of the potential from the MW will change dynamically throughout any integration we perform. Given the position and velocity of a dwarf galaxy, we are then able to integrate its orbit both forwards and backwards within the M31–MW potential. In this paper, we ignore the interaction between dwarf galaxies so that the orbits of the dwarf galaxies only depend on the system potential. The results then reflect how the M31–MW potential contributes to the observed anisotropic distribution of the dwarf galaxies surrounding M31.

## 3 RESULTS

### 3.1 Backward integration of orbits

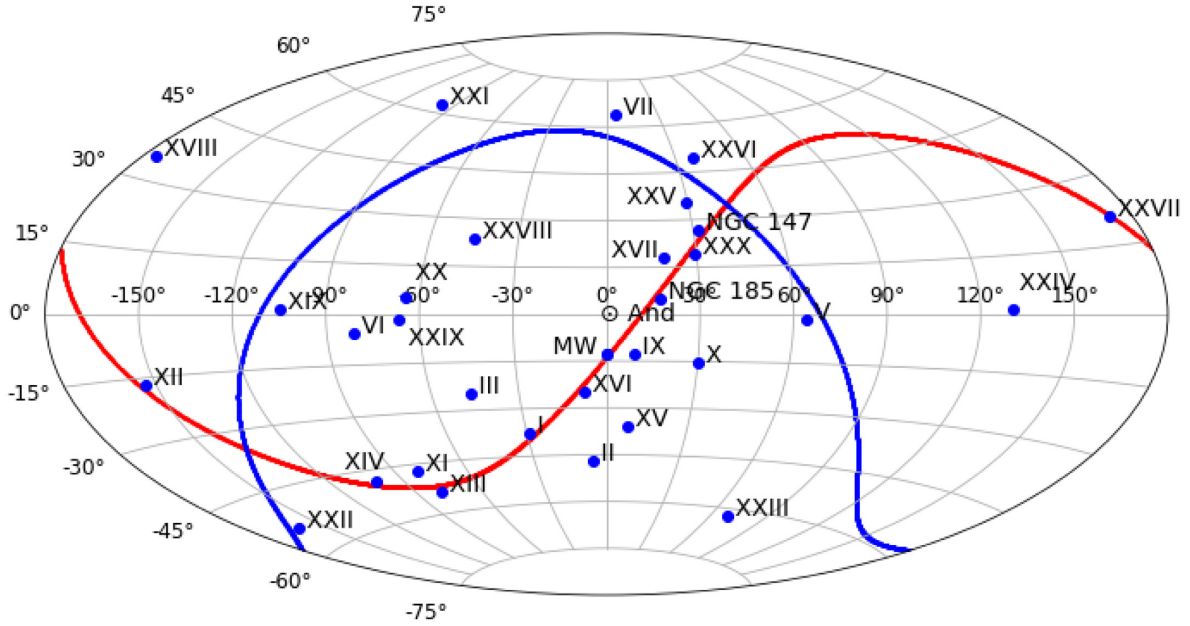
The position and line-of-sight velocities of the dwarf galaxies surrounding M31 are adopted from McConnachie (2012), and the distances to each dwarf galaxy are taken from Conn et al. (2012). We consider these parameters as a current snapshot. Fig. 2 displays the Aitoff projection of these dwarf galaxies in the M31-centred coordinate system. This figure also indicates the angular asymmetry

(Conn et al. 2013) as well as the disc structure (Conn et al. 2013; Ibata et al. 2013).

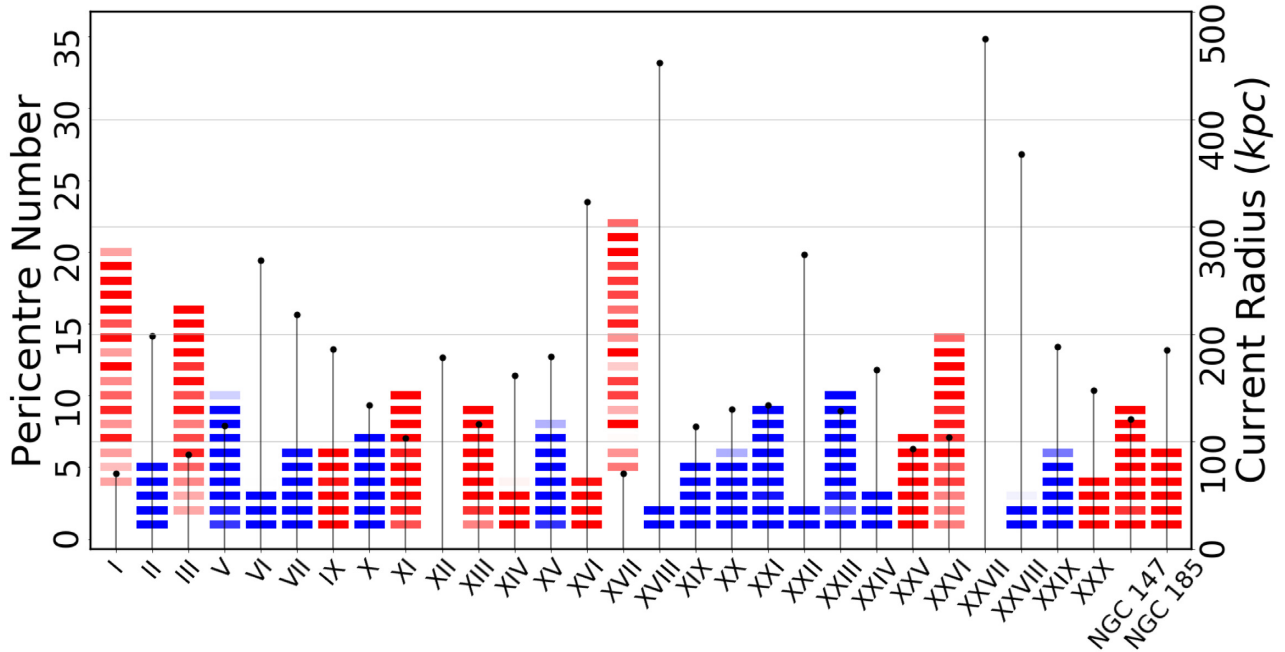
This data set however does not provide the tangential components of the velocities of these dwarf galaxies. Without knowledge of these proper motions, we simulate a range of current conditions for each dwarf galaxy. We sample the magnitude of the tangential velocities in  $30 \text{ km s}^{-1}$  steps from the interval  $30$ – $240 \text{ km s}^{-1}$ . For each magnitude, the angular direction is also sampled at a resolution of  $0.02$  rad over a full  $2\pi$  range. This step size ensures a high resolution as well as an affordable calculation time. In total, 2520 current tangential velocities are sampled for each dwarf galaxy. For each current condition, the dwarf galaxy’s orbit is integrated into the past for 10 Gyr. Note that during this integration, the MW will move away from M31.

As a measure of orbital frequency, we count the number of times that each dwarf galaxy passes through a pericentre along each orbit integration. Since the current three-dimensional position and line-of-sight velocity of each dwarf galaxy are fixed, an individual dwarf galaxy’s pericentre number is only dependent on the current tangential velocity we assign. The pericentre number tends to be larger for a dwarf galaxy that is currently closer to the centre of M31. It is these same dwarf galaxies whose pericentre number for a particular orbit is more strongly affected by the choice of that orbit’s current tangential velocity. Dwarf galaxies closer to the centre of M31 will therefore show a larger range of possible pericentre numbers when compared with dwarf galaxies currently far from the centre of M31. Furthermore, any particular dwarf galaxy’s pericentre number will decrease with increasing tangential velocity magnitude. That is until the total velocity reaches the upper limit for bounded orbits of that dwarf galaxy (above which the dwarf galaxy will never pass through a pericentre).

In Fig. 3, we show the possible range of each dwarf galaxy’s pericentre number by considering their orbits for all sampled current tangential velocities. The transparency of each bar represents the



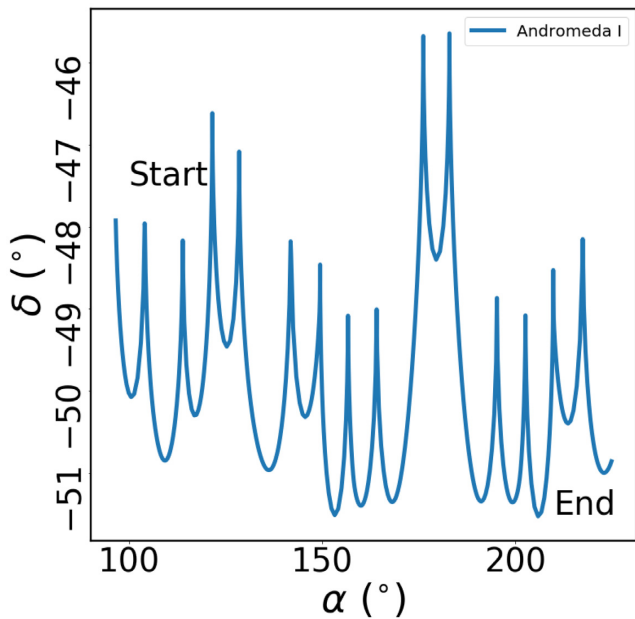
**Figure 2.** An Aitoff projection of the positions of the dwarf galaxies in the M31-centred coordinate system. The positions are taken from McConnachie (2012) and Conn et al. (2012). The red line indicates the great plane of 15 dwarf galaxies (Ibata et al. 2013), and the blue line separates the hemispheres of largest asymmetry where one side includes the 21 dwarf galaxies in Conn et al. (2013). We also include a Cartesian projection of the dwarf galaxies in Fig. A2.



**Figure 3.** The number of times each dwarf galaxy passes through a pericentre during 10 Gyr of backwards integration with different tangential velocities. Here the red bars indicate those 15 dwarf galaxies that belong to the great plane structure (Andromeda XII is considered to be a part of this structure as well). The point markers represent the current radius from M31 of each dwarf galaxy (right axis). It is possible that some dwarf galaxies that are close to the centre of M31 with low line-of-sight velocity could be in low-energy orbit. These dwarf galaxies are closely bound to M31 and will conclude more than 20 passes of M31 during 10 Gyr. Other dwarf galaxies that are far from M31 may only exist on high-energy orbits and consequentially only be able to pass M31 up to a few times.

frequency of that particular number of pericentre passages over all sampled conditions for that dwarf galaxy. Those that are coloured in red (as well as Andromeda XII) are the 15 dwarf galaxies that lie in the great circle (Ibata et al. 2013). Dwarf galaxies whose current position places them far from the centre of M31 are only able to

finish their first several cycles through a pericentre regardless of the tangential velocity we assume. Whereas other dwarf galaxies (e.g. Andromeda I, Andromeda IX) that are close to the M31 centre, are likely to pass through a pericentre far more frequently than those far away from the centre, given that they are in a low-energy orbit. These



**Figure 4.** The direction of angular momentum of one possible low-energy orbit of Andromeda I over 10 Gyr backwards integration. The orientation of the orbit plane changes substantially throughout the integration due to the precession of the angular momentum. This implies that a dwarf galaxy in this orbit is not able to stay within one plane during a 10 Gyr time-scale.

closer dwarf galaxies will hence explore much more of the phase space and mix their dynamic information. Furthermore, due to the asymmetric nature of the potential, the angular momentum of those dwarf galaxies will not be conserved. For the spherical potential, the asymmetric nature of the potential comes from the disc, hence the effect is more significant when a dwarf galaxy orbits close to the M31 centre. Naturally, this asymmetry is the reason that the prolate and oblate potentials shift the angular momentum even faster. For example, Fig. 4 shows variation of the angular momentum direction of a single Andromeda I orbit during 10 Gyr backwards integration throughout the spherical NFW profile. This particular dwarf galaxy will deviate from its plane of origin quickly due to the precession of its orbit. This effect dictates that it is unlikely that dwarf galaxies with high angular velocity and those with low angular velocity can remain contained within one stable asymmetric structure.

In consideration of this, we can determine that the 15 dwarf galaxies within the great plane are unlikely to be coherent or that this structure is much younger than 10 Gyr if some of its members have high angular velocity. A similar phenomenon happens to the asymmetric structure of the 23 dwarf galaxies that lie in the same hemisphere. Even though the dwarf galaxies far from M31 will stay in this hemisphere for a long time, dwarf galaxies close to M31 will leave it within a much shorter time-scale. One possible solution to satisfy the longevity of this structure over larger time-scales occurs if the dwarf galaxies close to M31 are in highly eccentric orbits with large tangential velocities. This would increase the orbit energy and decrease the angular velocities of those galaxies. As Fig. 3 shows, the dwarf galaxies could always have low angular velocity as long as we assume that they have large enough tangential velocity.

The different positions of each dwarf galaxy within the potential require varying escape velocities. In our potential, both Andromeda XII and Andromeda XXVII do not exhibit any pericentre passings since they have very low escape velocities at their current position. Chapman et al. (2007) suggested that due to its large line-of-sight

velocity relative to that of M31, Andromeda XII must be on its first infall into the system. Andromeda XXVII, is unbound to the M31 in our model, which implies that it will never be a part of any long-term structure associated with the system. However, there is evidence that it is actually closer to us (and M31) (e.g. Richardson et al. 2011; Conn et al. 2012; and Preston et al. 2019) than the distance proposed in Conn et al. (2012). This would allow its lowest possible energy to be lower so that it becomes bound to the M31. In this work though, these two dwarf galaxies will always be in long period orbits regardless of the tangential velocity we assign.

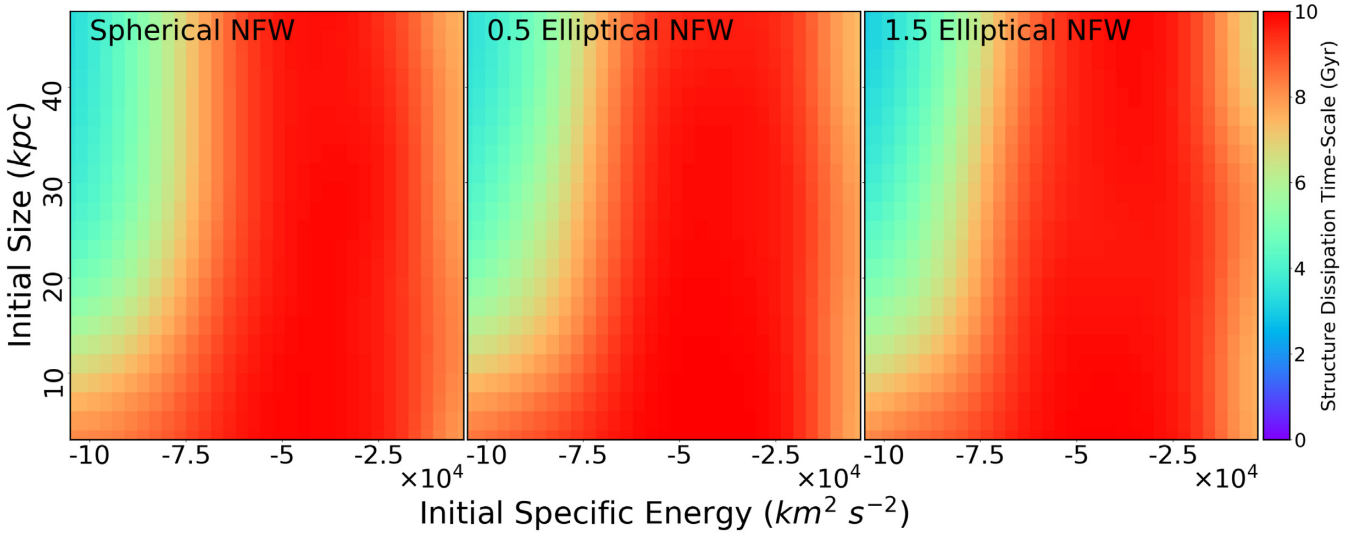
### 3.2 Forward integration of orbits

We attempt to construct the asymmetric distribution of dwarf galaxies around M31 by considering various associations of dwarf galaxies with different initial conditions and allow them to evolve for 10 Gyr. For each association, we place a dwarf galaxy at all six vertices of a regular octahedron and another dwarf galaxy at the centre of this group as a primary reference. We use the distance between the vertex dwarf galaxies and the central dwarf galaxy to represent the size of the association. The initial association sizes we take can be as large as 50 kpc, with which we find the effect of a large association size is already clear. Then, under the condition that the central reference dwarf galaxy of each association is bound to the M31–MW system, we place these associations randomly within our potential and initialize them with a random velocity. Note that the seven dwarf galaxies in each association (1 at the centre and 6 at the vertices) all possess the same initial velocity. We integrate each of the dwarf’s orbits by setting the M31 halo potential to spherical, oblate, and prolate separately to investigate the effect of a non-spherical NFW profile as described in Section 2.

During each 10 Gyr integration, the associations will deform from their initial octahedral configuration to a more distorted shape. This change occurs due to the difference in potential energy across the association, since each dwarf galaxy within an association is initialized with the same velocity. The potential difference within each association is dependent upon the size of the association and the potential gradient surrounding it. The dispersion of the association is an effect of time-evolution that arises from the difference of this potential gradient. Based on the gravitational potential model (equations 1–3), this difference is largest around the centre of M31. In the case where the association starts close to the centre of M31 with a low velocity, the dwarf galaxies in the association will be moving on low-energy orbits with a large gravitational potential and will complete a higher number of revolutions. Throughout a 10 Gyr integration of a dwarf galaxy association of this kind, the accumulative effect of the dispersion will be largely due to the considerable potential gradient in this region. Contrarily, an association that starts from a large radius and is moving on a high-energy orbit will exhibit a smaller dispersive effect since the potential gradient is shallower along this orbit.

From these analyses, we take the initial size and the total energy (the addition of the potential energy and the dynamic energy such that the total specific energy  $\geq 0 \text{ km}^2/\text{s}^2$  represents an unbounded orbit) as characteristics of the associations. We then consider the time-scale it takes for the association to dissipate to the extent that half of the galaxies in the association are at least  $90^\circ$  away from the central reference in angular distance. The relationship between the size, energy, and dissipation time is presented in Fig. 5. Associations that are moving on low-energy orbits with large sizes will be easily disrupted within shorter time-scales. The oblate and prolate NFW





**Figure 5.** The time that associations with different initial sizes and specific energies need to dissipate to the extent that half of the dwarf galaxies within it are at least  $90^\circ$  away from the association centre in angular distance. This plot is a Gaussian smoothed bin-map of our simulation results where each pixel spans  $3200 \text{ km}^2 \text{ s}^{-2}$  in specific energy and 2 kpc in size, with a Gaussian kernel  $\sigma = 3$  pixels. A high-energy orbit is necessary for the association to be long-lived.

profiles have similar effects. For each case, it is clear from this figure that for an association to remain compact after a 10 Gyr integration, it initially needs to be on a high-energy orbit with a preference towards having a small association size. The top panel of Fig. 6 similarly shows that to have most of the dwarf galaxies in one hemisphere after 10 Gyr, as are the observed M31 dwarf galaxies, the association is required to be compact.

However, if the initial size is too small or the starting radius too large, the association will be unlikely to end up with a large difference in radius between its members. The middle panel of Fig. 6 shows the radial dispersion of the associations within the three potentials at the 10 Gyr snapshot given varying initial sizes and energies. For an association to obtain a large radial dispersion, it needs to initially exhibit both a high-energy and a large size that will not typically complete many revolutions of M31 within a 10 Gyr time-scale. We note that on the other hand, this energy cannot be too large otherwise most of the dwarf galaxies in the association will easily escape to  $\geq 1$  Mpc away from the M31 within 10 Gyr.

We select associations that have final radial dispersion larger than 90 kpc and a final median angular separation smaller than  $60^\circ$ . The initial position and velocity distribution of these associations are shown in Fig. 7. We use orange crosses to indicate the escape velocity of each dwarf galaxy (indicated as  $V_{\text{esc}}$ ) and green crosses to indicate  $\sqrt{V_{\text{esc}}^2 - V_{\text{init}}^2}$  of each association, a quantity which is the square root of the (negative) total specific energy of the association (note that for bounded orbits, this quantity is always positive). For all potentials, the selected association parameters typically occupy the high-energy region where the total specific energy  $E = V_{\text{esc}}^2 - V_{\text{init}}^2$  is roughly concentrated around  $200^2 - 300^2 \text{ km}^2 \text{ s}^{-2}$ .<sup>1</sup>

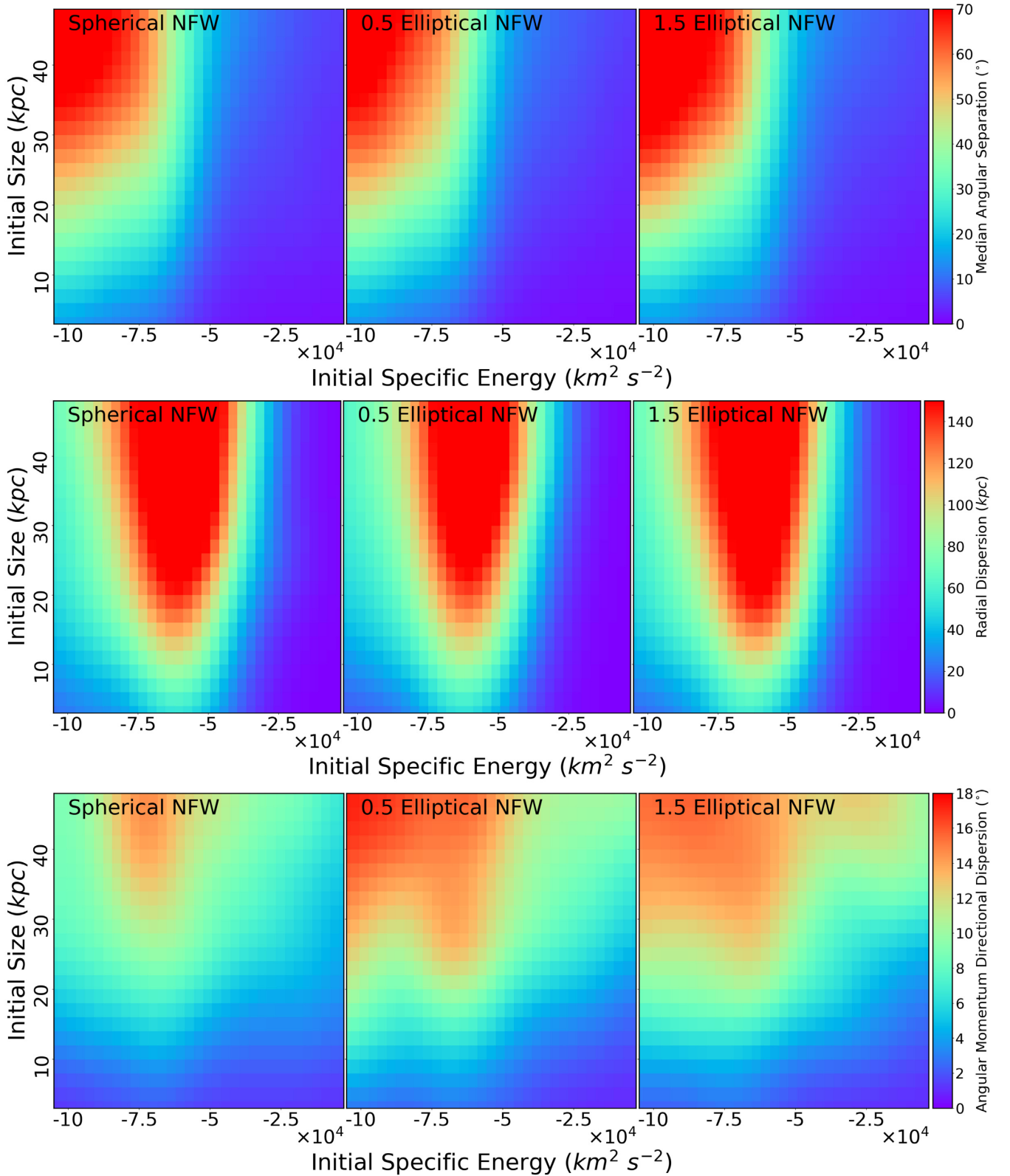
The left-hand panel of Fig. 8 depicts the orbits of the members of one such association that has been placed 110 kpc away from the M31 centre with an initial velocity of  $262.75 \text{ km s}^{-1}$  and an initial size parameter of 41.9 kpc. Some of the dwarf galaxies within this association are in a position of higher potential, and after 10 Gyr, 5 out of 7 dwarf galaxies in this association have acquired

enough energy to escape to large radii and remain in the North hemisphere. The other two dwarf galaxies remain closely bound, within 100 kpc from the M31 centre. All of the dwarf galaxies in this association are in eccentric orbits. In fact, to construct the asymmetric distribution observed around M31 with a model that ignores dwarf–dwarf interaction such as this, a highly eccentric orbit is preferred so that some of the dwarf galaxies are close to the M31 centre while others are far away. Those associations with eccentric orbits either start close to the M31 centre with large initial velocities or far from the M31 centre with small velocities. Under both conditions, the associations will pass close to the M31 centre where the potential gradient is steep enough to radially separate the dwarf galaxies.

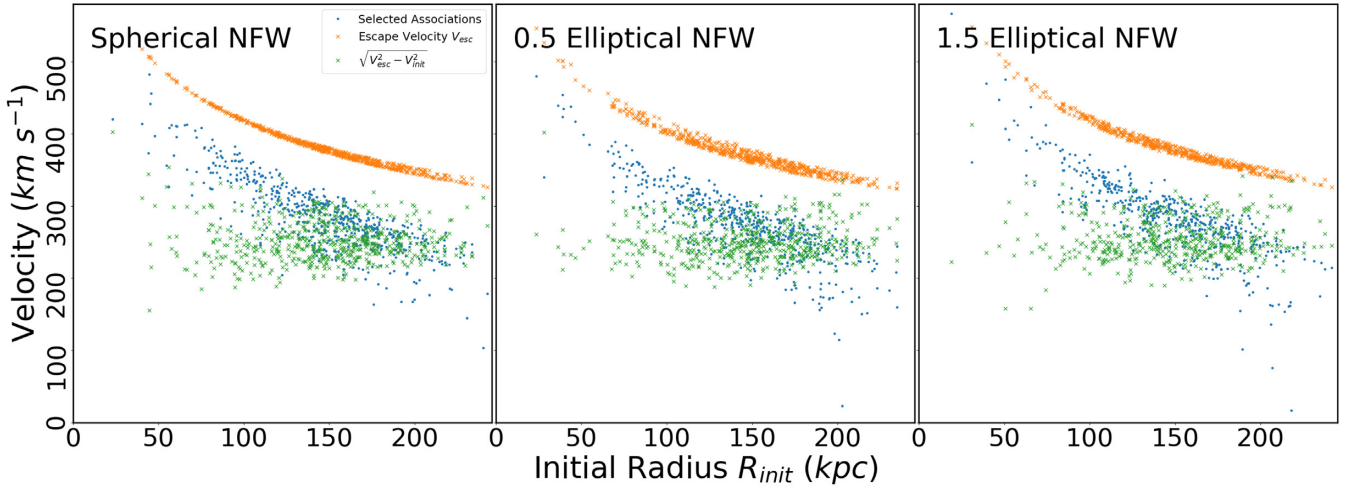
Since there is no dwarf–dwarf interaction in our model, the precession of the angular momentum of each dwarf galaxy is due to the asymmetric nature of the potentials. We show the final angular momentum directional dispersion of the associations in the bottom panel of Fig. 6. Both the prolate and oblate NFW potentials lead to a significantly larger dispersion than when compared to the spherical potential. The low-energy associations have larger angular velocities and revolve many times around the M31 during integration, so the accumulative effect is significant. For initial positions further from the M31 centre, and larger initial specific energies, the associations will have smaller angular velocities. Typically, the potential field is weaker along the orbits of these dwarf galaxies, and hence the angular momentum directional dispersion is smaller than that of associations moving on low-energy orbits. Some of these high-energy associations with large initial sizes, however, can acquire enough energy to escape the M31 potential and be captured by the MW. The right-hand panel of the Fig. 8 shows one of these kinds of associations, where some dwarf galaxies in this association are captured by the MW and change the direction of their velocity. This significantly changes the direction of their angular momentum. For an association to have a smaller final dispersion, a high initial energy and a small size are preferred because

(i) We give the same initial velocity to each dwarf galaxy in an association. So a larger initial size will result in a larger initial angular momentum dispersion.

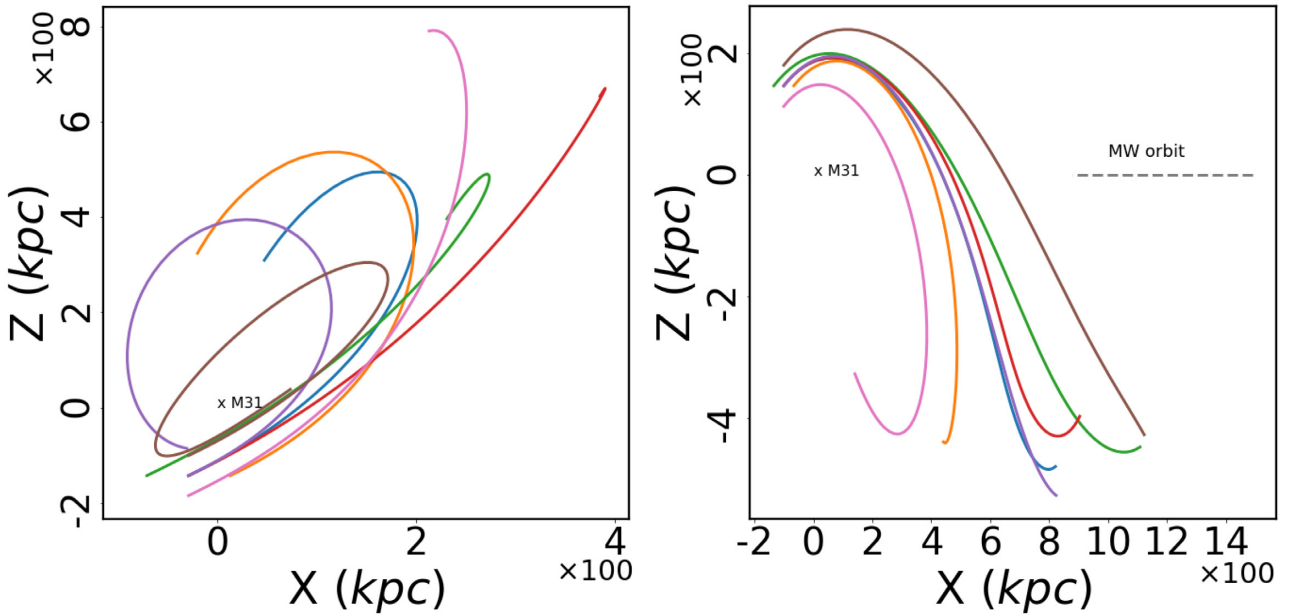
<sup>1</sup>Note that the quantity  $\frac{1}{2}(V_{\text{esc}}^2 - V_{\text{init}}^2)$  indicates the *negative* total specific energy.



**Figure 6.** *Top:* The median angular distance of each dwarf galaxy to the centre of its association after 10 Gyr integration, sampled over associations varying in initial sizes and specific energies. *Middle:* The radial dispersion of associations after the 10 Gyr integration, sampled over associations varying in initial sizes and specific energies. A substantial radial dispersion occurs within an association after 10 Gyr if its initial size and energy are large. *Bottom:* The final angular momentum directional dispersion of all associations. For associations with initial sizes and energies that allow for the observed asymmetrical distributions (visualized in Figs 5 and 6), the dispersion of the angular momentum direction is generally small. Note that we give the same velocity for each dwarf galaxy in an association, so this angular momentum dispersion is entirely dependent on the initial size of an association and the asymmetrical nature of the potential along its orbital path.



**Figure 7.** The initial radius ( $R_{\text{init}}$ ) and initial velocity ( $V_{\text{init}}$ ) of the selected associations that have final radial dispersion larger than 90 kpc and median angular separation smaller than  $60^\circ$  (blue dots). The radial dispersion of association scales with the asymmetry of the potential, and an association’s angular separation is dependent on the steepness of the potential gradient. The orange crosses indicate the escape velocity ( $V_{\text{esc}}$ ) of each association. The green crosses are the quantity  $\sqrt{V_{\text{esc}}^2 - V_{\text{init}}^2}$ , which indicates the specific energy needed for the association to escape the potential well. This quantity also roughly indicates how far the association could reach from the M31 centre.



**Figure 8.** *Left:* Traced orbits of the members of an example association with an initial specific energy of  $-28081 \text{ km}^2 \text{ s}^{-2}$  and an initial size of 41.9 kpc. The starting points of each dwarf galaxy are marked with crosses. This association is initially close to the M31 centre with a high velocity of  $262.75 \text{ km s}^{-1}$ . Some members of the association acquire enough energy to escape to a large radius with highly eccentric orbits, while others remain closely bound to M31. *Right:* Another example association with an initial specific energy of  $-19381 \text{ km}^2 \text{ s}^{-2}$  and an initial size of 33.8 kpc. Some of the dwarf galaxies escape from M31 and fall towards the MW (the grey dashed line indicates the orbit of the MW) due to their high initial energy, which increases the angular momentum directional dispersion of these kinds of associations.

(ii) For a spherical NFW profile, high-energy orbits will be far away from the centre where the asymmetry of the disc component of the potential is small. We find that associations that result in an angular momentum distribution with a preferred direction are in high-energy orbits. However, if their energy is too high, some dwarf galaxies will be captured by the MW and significantly change the direction of their angular momentum.

Similarly, we integrated the association orbits with the Hernquist potential (Hernquist 1990) as the form of the underlying dark halo, where the profile follows:

$$\rho(r) = \frac{\rho_0}{(r/a)(1+r/a)^3}. \quad (4)$$

This profile is scaled so that the total mass of the halo is  $1 \times 10^{12} M_\odot$  and that the circular velocity at  $(x, y, z) = (8, 0, 0) \text{ kpc}$

of the potential is  $220 \text{ km s}^{-1}$ . The angular momentum directional dispersions are comparable to the spherical NFW profile. However, the resultant dissipation time-scale is much shorter, the median angular separation is more pronounced at higher specific energies, and the dwarf associations can conform to similar radial dispersions to the NFW profiles, though typically at larger energies too. This is expected since the Hernquist profile has a steeper gradient than compared to that of the spherical NFW potential. The initial size and specific energy parameter space that would allow for long-lasting coherent structures is extremely small if not non-existent and as such the Hernquist profile will destroy these structures within shorter-time scales.

#### 4 DISCUSSION

The observed distribution of the 30 dwarf galaxies surrounding M31 is noticeably distinct, where 23 dwarf galaxies reside in one hemisphere and 15 dwarf galaxies are contained within a planar structure surrounding M31. The radial distribution of these dwarf galaxies, which ranges from 40 to 400 kpc, is less special. The asymmetric nature of this distribution raises the question of its origin. In previous sections, we explored the possible orbits of the dwarf galaxies around M31 in an attempt to examine the longevity of these structures as well as the possibility that these dwarf galaxies come from a single association.

In our simulations, we include the gravitational potential of the MW that deforms the otherwise axisymmetric potential of M31 as in Fig. 1, though its effect within 300 kpc is small. The integration results (e.g. Fig. 8) indicate that the MW potential will have some significant influence on the dwarf galaxy orbits from those high-energy associations with a large initial size.

The data we use for our analysis includes the sky position, distance, and line-of-sight velocity of each dwarf galaxy (as well as M31), giving us a three-dimensional map and one component of the velocity of their distribution. Using this distribution as a current snapshot of the dwarf galaxies, we performed numerous integrations into the past for each dwarf galaxy's orbit by sampling over various tangential velocities. Through this method we find that there are only a limited number of possible bound orbits for those dwarf galaxies that are either far away from the M31 centre, or that have a high line-of-sight velocity. Lower energy orbits will be closer to the M31 centre with a higher angular velocity. From Fig. 3, we see that some dwarf galaxies could possibly revolve the M31 centre in excess of 20 times within 10 Gyr corresponding to less than 500 Myr per revolution. In comparison, a few other dwarf galaxies could only complete up to two revolutions. Under the condition that these dwarf galaxies have come from a single large association where some members have been drawn into low-energy orbits and others into high-energy orbits, we find that the observed asymmetric structure will be short-lived (the lifetime may be as short as  $\sim 500$  Myr during which some of the dwarf galaxies can complete one revolution of the M31). For this scenario to resemble the current snapshot of dwarf galaxies, the tangential velocity of those dwarf galaxies on lower energy orbits must be large. This is because there is little room to adjust the tangential velocities of those dwarf galaxies whose orbital energies have a high lower limit without making them unbound to the system. An increase in the magnitude of the tangential velocities of these low orbital energy dwarf galaxies is necessary to result in their orbits becoming highly eccentric with long periods. If the 23 dwarf galaxies that are in the same hemisphere are co-rotating around the M31, then their angular velocity (or the pericentre number in Fig. 3) should be roughly same as each other. By assuming all

of the dwarf galaxies in the same hemisphere have the same specific energy as the Andromeda XXVIII (which has the largest specific energy without considering proper motion), we may calculate a rough estimation of the upper limit of the proper motion of the dwarf galaxies. As such, most of these dwarf galaxies will have a tangential velocity of  $\sim 150\text{--}350 \text{ km s}^{-1}$  that corresponds to a magnitude of proper motion  $\sim 45\text{--}110 \mu\text{as yr}^{-1}$ , which agrees with a recent proper motion estimation based on the planer structure (Hodkinson & Scholtz 2019). Under these conditions Andromeda XVII, which is currently the closest dwarf to M31, would need to have a tangential velocity of  $\sim 382 \text{ km s}^{-1}$ . Note that these estimations only take into account the dwarfs' motion relative to the M31 and as such, the actual proper motion we observe would differ from this and would include the effects of both the magnitude and direction of the proper motion of M31 (e.g. van der Marel et al. 2019). These values are the contribution of the proper motion from the dwarfs motion with respect to the M31. Given the magnitude of these proper motions, this may be detectable within the next generation of telescopes.

Having established that the observed asymmetries are likely short lived, we explored whether they could have formed from the recent infall of a single association. To do this, we sampled associations with different initial conditions, placed them in the M31–MW potential, and integrated their orbits forwards for 10 Gyr. Here, we use the initial size and specific energy to characterize each association. We find that, for an association to result in the observed small angular dispersion after integration, it requires a high initial energy, and that the observed large radial dispersion requires a high initial energy and a large initial size. We also find that the orbital energies of the associations that result in the same hemisphere structure are concentrated where the total specific energy  $E = V_{\text{esc}}^2 - V_{\text{init}}^2 \sim 200^2\text{--}300^2 \text{ km}^2 \text{ s}^{-2}$ . This energy is high enough that the dwarf galaxies could reach further than 500 kpc from the centre, but still be bound to M31. Compared to lower energies, this high-energy keeps the angular momentum directional dispersion small so that the disc structure can also survive. Because we give the same velocity to all dwarf galaxies in an association, the initial angular momentum differences are due to the size of the association as well as its position within our potential model. Then the precession of the dwarf galaxy orbits originates from the asymmetry of the potential as well as the comparative initial size of the association relative to its initial distance from the M31 centre – which magnifies the differential apsidal precession within the group. Far from the centre of M31 (with the spherical NFW potential), the potential is approximately spherically symmetric and the initial size of the association contributes less to the differential apsidal precession. As such, the precession of those high-energy orbits is relatively small. However, both the prolate and oblate NFW potentials will significantly destroy the planar structure by changing the direction of the angular momentum of the dwarf galaxies with their asymmetric shape.

In Fig. 8, we show one example association orbit that shows angular asymmetry. Here, five of the seven dwarf galaxies of that association end up in the Northern hemisphere and some dwarf galaxies remain within 200 kpc from the M31 centre, while others are far away. The high-energy initial condition is necessary from two aspects. First, some of the dwarf galaxies in this association need to end up far from the M31 centre by the end of the integration; and secondly, the angular velocity of the association cannot be so large that the dwarf galaxies become well mixed. From both the forward and backward integrations, we find that for a long-living asymmetric structure to exist, the association likely needs to have recently completed or currently be in its first revolution of M31.

In reality, it seems unlikely that a structure with a size of  $\sim 40$  kpc has formed 110 kpc away from the M31 centre with a velocity of  $\sim 260$  km s $^{-1}$ . One possible case is that such an association has formed earlier and subsequently fallen inwards to the M31 centre from large radius. In this scenario, the structure of the association would become disturbed by the M31 potential and may be capable of developing into the initial condition of the association modelled in Fig. 8. The infalling orbit would need to be eccentric to a large enough degree that the association could approach close to the M31 centre. This way the stronger tidal forces would cause some dwarf galaxies to become more bound to M31. Additionally, the size of this association could not be too small to ensure that enough dwarf galaxies are still able to escape to large radii and resemble the currently observed large radial dispersion. In conclusion, the intersection of initial condition regions required by the observed radial dispersion, angular asymmetry and planar structure is small. The asymmetric structures – especially the planar structure, which will be easily destroyed by both the prolate and oblate potentials – are less likely to have come from a single association than not, or they are short-lived.

We note that these results and subsequent discussion are based on the assumption that this asymmetric structure has originated from a single association that has been able to last for up to 10 Gyr. The initial conditions may be utterly distinct from these assumptions if this is a young structure, or if this asymmetric structure is merely a coincidence. Another premise of the simulation results presented here is that the dwarf galaxies are non-interacting throughout the course of their orbits. Our results demonstrate how the M31–MW potential alone could contribute to the observed asymmetric distribution under the assumptions previously discussed. However, the interaction between each dwarf galaxy may establish more self-bound associations than those that appear in our analysis. This interaction could also provide a means of transferring energy and angular momentum between the dwarf galaxies of an association so that a large radial dispersion can evolve. The interaction may also provide a mechanism for a longer-lasting distributional asymmetry of the dwarfs. An association with sufficient self-gravity could be held together during its first infall so that it may exhibit a condition similar to the initial conditions that we have shown could result in the observed distribution of dwarf galaxies. A possible candidate for the progenitor of such an association is NGC3190 (Bellazzini et al. 2013), as it is large and currently far from M31. It is possible that the observed distribution of dwarf galaxies could have resulted from an association centred on a younger NGC3190 that had passed close enough to M31 to be tidally disrupted. We leave this concept and the inclusion of dwarf-dwarf interaction in our model for future work. In addition, the M33 could have a significant contribution to the potential as well, similar to the effect of Large Magellanic Cloud on the MW. A more detailed model of this system with M33 could also be investigated in the future and would complement the research we present here. We leave other dynamical effects such as the time dependence of the potential components and the incorporation of an appropriate cosmological setting for future work as well. Thus, the research we present here, whilst based on simple assumptions, is a map for future works from which we may compare the effects of these various factors.

## ACKNOWLEDGEMENTS

ZW gratefully acknowledges financial support through the Dean's International Postgraduate Research Scholarship from the Physics School of the University of Sydney. GFL acknowledges support

from the Institute of Advanced Studies Santander Fellowship at the University of Surrey and thanks them for their hospitality where the initial idea for this project was conceived.

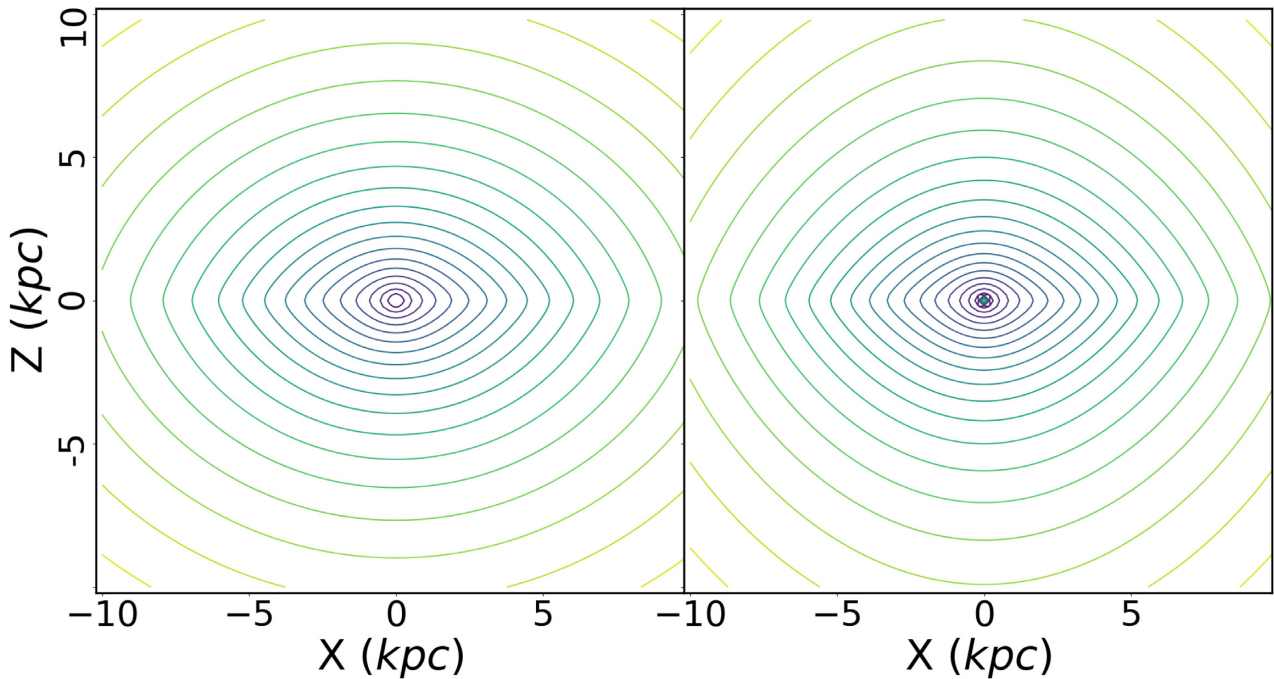
## REFERENCES

- Bahl H., Baumgardt H., 2014, *MNRAS*, 438, 2916  
 Bellazzini M., Oosterloo T., Fraternali F., Beccari G., 2013, *A&A*, 559, L11  
 Binney J., Tremaine S., 2008, *Galactic Dynamics*, 2nd edn. Princeton Univ. Press, Princeton, NJ  
 Bovy J., 2015, *ApJS*, 216, 29  
 Bovy J., Rix H.-W., 2013, *ApJ*, 779, 115  
 Bovy J. et al., 2012, *ApJ*, 759, 131  
 Bowden A., Evans N. W., Belokurov V., 2013, *MNRAS*, 435, 928  
 Bowden A., Evans N. W., Belokurov V., 2014, *ApJ*, 793, L42  
 Buck T., Dutton A. A., Macciò A. V., 2016, *MNRAS*, 460, 4348  
 Buck T., Macciò A. V., Dutton A. A., 2015, *ApJ*, 809, 49  
 Cautun M., Bose S., Frenk C. S., Guo Q., Han J., Hellwing W. A., Sawala T., Wang W., 2015b, *MNRAS*, 452, 3838  
 Cautun M., Wang W., Frenk C. S., Sawala T., 2015a, *MNRAS*, 449, 2576  
 Chapman S. C. et al., 2007, *ApJ*, 662, L79  
 Clemens D. P., 1985, *ApJ*, 295, 422  
 Conn A. R. et al., 2012, *ApJ*, 758, 11  
 Conn A. R. et al., 2013, *ApJ*, 766, 120  
 Deason A. J. et al., 2012, *MNRAS*, 425, 2840  
 Debattista V. P., Moore B., Quinn T., Kazantzidis S., Maas R., Mayer L., Read J., Stadel J., 2008, *ApJ*, 681, 1076  
 Dehnen W., Binney J., 1998, *MNRAS*, 294, 429  
 de Vaucouleurs G., 1958, *ApJ*, 128, 465  
 Dubinski J., 1994, *ApJ*, 431, 617  
 D'Onghia E., Lake G., 2008, *ApJ*, 686, L61  
 Forero-Romero J. E., Arias V., 2018, *MNRAS*, 478, 5533  
 Gaia Collaboration et al., 2018, *A&A*, 616, A12  
 Gillet N., Ocvirk P., Aubert D., Knebe A., Libeskind N., Yepes G., Gottlöber S., Hoffman Y., 2015, *ApJ*, 800, 34G  
 Goerdt T., Burkert A., Ceverino D., 2013, preprint (arXiv:1307.2102)  
 Gong C. C. et al., 2019, *MNRAS*, 488, 3100G  
 Hayashi K., Chiba M., 2014, *ApJ*, 789, 62  
 Hernquist L., 1990, *ApJ*, 356, 359  
 Hodkinson B., Scholtz J., 2019, *MNRAS*, 488, 3231  
 Holmberg J., Flynn C., 2000, *MNRAS*, 313, 209  
 Ibata R. A., Ibata N. G., Lewis G. F., Martin N. F., Conn A., Elahi P., Arias V., Fernando N., 2014, *ApJ*, 784, L6  
 Ibata R. A., Lewis G. F., Totten E., Irwin M. J., 1998, in Kroupa P., Palous J., Spurzem R., eds, *Dynamical Studies of Star Clusters and Galaxies*. p. 178, ESA Publ. Division, c/o ESTEC, Noordwijk, The Netherlands  
 Ibata R. A. et al., 2013, *Nature*, 493, 62  
 Jones E., Oliphant T., Peterson P., 2001, *SciPy: Open source scientific tools for Python*. <http://www.scipy.org/>  
 Kafle P. R., Sharma S., Lewis G. F., Bland-Hawthorn J., 2012, *ApJ*, 761, 98  
 Kafle P. R., Sharma S., Lewis G. F., Bland-Hawthorn J., 2014, *ApJ*, 794, 59  
 Kroupa P., Theis C., Boily C. M., 2005, *A&A*, 431, 517  
 Libeskind N. I., Guo Q., Tempel E., Ibata R., 2016, *ApJ*, 830, 121  
 Libeskind N. I., Hoffman Y., Tully R. B., Courtois H. M., Pomarède D., Gottlöber S., Steinmetz M., 2015, *MNRAS*, 452, 1052  
 Libeskind N. I., Knebe A., Hoffman Y., Gottlöber S., Yepes G., Steinmetz M., 2011, *MNRAS*, 411, 1525  
 Li Y.-S., Helmi A., 2009, in Andersen J., Nordström B., Bland-Hawthorn J., eds, *Proc. IAU Symp. 254, The Galaxy Disk in Cosmological Context*. p. 263, Cambridge University Press, Cambridge England, preprint (arXiv:0807.2780)  
 Li Y.-S., White S. D. M., 2008, *MNRAS*, 384, 1459  
 Lovell M. R., Eke V. R., Frenk C. S., Jenkins A., 2011, *MNRAS*, 413, 3013  
 Lux H., Read J. I., Lake G., 2010, in Debattista V. P., Popescu C. C., eds, *AIP Conf. Proc. Vol. 1240, Hunting for the Dark: the Hidden Side of Galaxy Formation*. Am. Inst. Phys., New York, p. 415

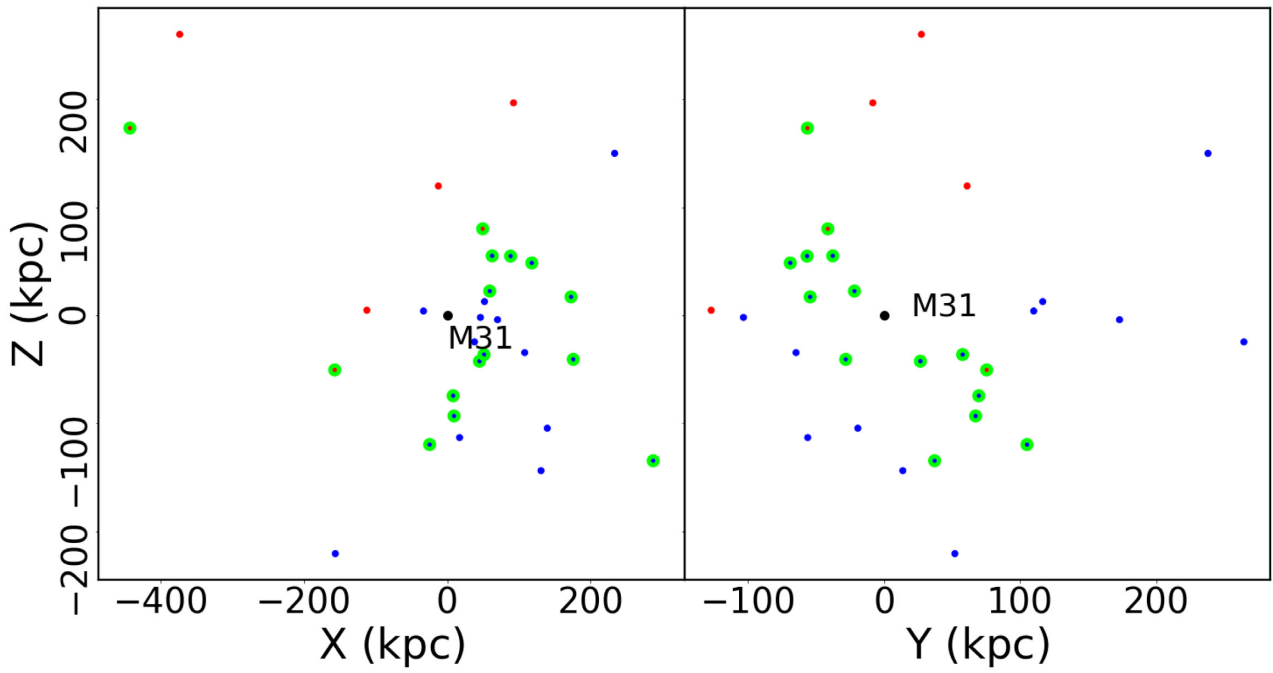
- Lynden-Bell D., 1976, *MNRAS*, 174, 695  
 Lynden-Bell D., Lynden-Bell R. M., 1995, *MNRAS*, 275, 429  
 Maji M., Zhu Q., Marinacci F., Li Y., 2017, *ApJ*, 843, 62  
 McClure-Griffiths N. M., Dickey J. M., 2007, *ApJ*, 671, 427  
 McConnachie A. W., 2012, *AJ*, 144, 4  
 McConnachie A. W., Irwin M. J., 2006, *MNRAS*, 365, 902  
 McConnachie A. W. et al., 2009, *Nature*, 461, 66  
 Metz M., Kroupa P., Jerjen H., 2007, *MNRAS*, 374, 1125  
 Metz M., Kroupa P., Libeskind N. I., 2008, *ApJ*, 680, 287  
 Navarro J. F., Frenk C. S., White S. D. M., 1996, *ApJ*, 462, 563  
 Pasetto S., Chiosi C., 2009, *A&A*, 499, 385  
 Pawlowski M. S., Bullock J. S., Kelley T., Famaey B., 2019, *ApJ*, 875, 105  
 Pawlowski M. S., Kroupa P., 2013, *MNRAS*, 435, 2116  
 Pawlowski M. S., Pflamm-Altenburg J., Kroupa P., 2012, *MNRAS*, 423, 1109  
 Penarrubia J., Gómez F. A., Besla G., Erkal D., Ma Y.-Z., 2015, *MNRAS*, 456, L54  
 Phillips J. I., Cooper M. C., Bullock J. S., Boylan-Kolchin M., 2015, *MNRAS*, 453, 3839  
 Posti L., Helmi A., 2019, *A&A*, 621, A56  
 Preston Janet. et al., 2019, *MNRAS*, 490, 2905  
 Read J. I., 2014, *J. Phys. G Nucl. Phys.*, 41, 063101  
 Read J. I., Lake G., Agertz O., Debattista V. P., 2008, *MNRAS*, 389, 1041  
 Richardson J. C. et al., 2011, *ApJ*, 732, 76  
 Schönrich R., Binney J., Dehnen W., 2010, *MNRAS*, 403, 1829  
 Seigar M. S., Barth A. J., Bullock J. S., 2008, *MNRAS*, 389, 1911  
 Shao S., Cautun M., Frenk C. S., Gao L., Crain R. A., Schaller M., Schaye J., Theuns T., 2016, *MNRAS*, 460, 3772  
 Simon J. D., 2018, *ApJ*, 863, 89  
 Sohn S. T., Watkins L. L., Fardal M. A., van der Marel R. P., Deason A. J., Besla G., Bellini A., 2018, *ApJ*, 862, 52  
 van der Marel R. P., Fardal M. A., Sohn S. T., Patel E., Besla G., del Pino A., Sahlmann J., Watkins L. L., 2019, *ApJ*, 872, 24  
 Wang J., Frenk C. S., Cooper A. P., 2013, *MNRAS*, 429, 1502  
 Wang P., Guo Q., Libeskind N. I., Tempel E., Wei C., Kang X., 2019, *MNRAS*, 484, 4325  
 Watkins L. L., Evans N. W., An J. H., 2010, *MNRAS*, 406, 264  
 Wegg C., Gerhard O., Bieth M., 2019, *MNRAS*, 485, 3296  
 Xue X. X. et al., 2008, *ApJ*, 684, 1143  
 Zentner A. R., Kravtsov A. V., Gnedin O. Y., Klypin A. A., 2005, *ApJ*, 629, 219  
 Zhang L., Rix H.-W., van de Ven G., Bovy J., Liu C., Zhao G., 2013, *ApJ*, 772, 108

### APPENDIX: MORE FIGURES

See Figs A1 and A2.



**Figure A1.** Equipotential contours for the oblate NFW profile (*left*) and the prolate NFW profile (*right*).



**Figure A2.** The Cartesian projection of the dwarf galaxies in the M31. The points centred on the lime circles are the 15 dwarf galaxies contained within the thin disc, and the blue dots are the 23 dwarf galaxies that are within the same hemisphere.

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.

## A.2 The Dynamics of NGC3201









This section presents the published journal article:

- [A2.](#) *The Dynamics of the Globular Cluster NGC 3201 out to the Jacobi Radius.* Z. Wan, **W. H. Oliver**, H. Baumgardt, G. F. Lewis, M. Gieles, V. Hénault-Brunet, T. de Boer, E. Balbinot, G. Da Costa, & D. Mackey. *MNRAS* **502**, 4513, 2021. [[arXiv:2102.01472](#)].

For this paper I contributed to the observing of the globular clusters through the AAT, the analysis design, the interpretation of the results, and the writing of the paper. A follow-up paper, which I am also a coauthor of, is also currently in preparation.



# The dynamics of the globular cluster NGC 3201 out to the Jacobi radius

Zhen Wan <sup>1</sup>★, William H. Oliver,<sup>1</sup> Holger Baumgardt <sup>2</sup>, Geraint F. Lewis <sup>1</sup>, Mark Gieles <sup>3,4</sup>,  
Vincent Hénault-Brunet <sup>5</sup>, Thomas de Boer <sup>6,7</sup>, Eduardo Balbinot <sup>8</sup>, Gary Da Costa <sup>9</sup> and  
Dougal Mackey<sup>9</sup>

<sup>1</sup>Sydney Institute for Astronomy, School of Physics A28, The University of Sydney, Sydney, NSW, 2006, Australia

<sup>2</sup>School of Mathematics and Physics, The University of Queensland, St Lucia, QLD 4072, Australia

<sup>3</sup>ICREA, Pg. Lluís Companys 23, E-08010 Barcelona, Spain

<sup>4</sup>Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí Franquès 1, E-08028 Barcelona, Spain

<sup>5</sup>Department of Astronomy and Physics, Saint Mary's University, 923 Robie Street, Halifax, NS B3H 3C 3, Canada

<sup>6</sup>Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA

<sup>7</sup>Department of Physics, University of Surrey, Guildford GU2 7XH, UK

<sup>8</sup>Kapteyn Astronomical Institute, University of Groningen, Postbus 800, NL-9700AV Groningen, the Netherlands

<sup>9</sup>Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia

Accepted 2021 January 30. Received 2021 January 28; in original form 2020 November 23

## ABSTRACT

As part of a chemodynamical survey of five nearby globular clusters with 2dF/AAOmega on the Anglo-Australian Telescope (AAT), we have obtained kinematic information for the globular cluster NGC 3201. Our new observations confirm the presence of a significant velocity gradient across the cluster which can almost entirely be explained by the high proper motion of the cluster ( $\sim 9 \text{ mas yr}^{-1}$ ). After subtracting the contribution of this perspective rotation, we found a remaining rotation signal with an amplitude of  $\sim 1 \text{ km s}^{-1}$  around a different axis to what we expect from the tidal tails and the potential escapers, suggesting that this rotation is internal and can be a remnant of its formation process. At the outer part, we found a rotational signal that is likely a result from potential escapers. The proper motion dispersion at large radii reported by Bianchini et al. ( $3.5 \pm 0.9 \text{ km s}^{-1}$ ) has previously been attributed to dark matter. Here, we show that the LOS dispersion between 0.5 and 1 Jacobi radius is lower ( $2.01 \pm 0.18 \text{ km s}^{-1}$ ), yet above the predictions from an  $N$ -body model of NGC 3201 that we ran for this study ( $1.48 \pm 0.14 \text{ km s}^{-1}$ ). Based on the simulation, we find that potential escapers cannot fully explain the observed velocity dispersion. We also estimate the effect on the velocity dispersion of different amounts of stellar-mass black holes and unbound stars from the tidal tails with varying escape rates and find that these effects cannot explain the difference between the LOS dispersion and the  $N$ -body model. Given the recent discovery of tidal tail stars at large distances from the cluster, a dark matter halo is an unlikely explanation. We show that the effect of binary stars, which is not included in the  $N$ -body model, is important and can explain part of the difference in dispersion. We speculate that the remaining difference must be the result of effects not included in the  $N$ -body model, such as initial cluster rotation, velocity anisotropy, and Galactic substructure.

**Key words:** stars: kinematics and dynamics – globular clusters: individual: NGC 3201 – dark matter.

## 1 INTRODUCTION

The formation and evolution of globular clusters (GCs) remain an open question in astrophysics. Clues from dynamical signatures have proven to be useful to unravelling this, with structures in GC phase-space, such as tidal arms and velocity gradients along the length of these arms (e.g. Chun et al. 2010; Jordi & Grebel 2010; Sollima et al. 2011; Hansen et al. 2020) representing evidence of the interaction between GCs and their host galaxies. Similarly, internal dynamical features including rotation (e.g. Bellazzini et al. 2012; Bianchini et al. 2018; Ferraro et al. 2018; Kamann et al. 2018; Gaia Collaboration et al. 2018c; Sollima, Baumgardt & Hilker 2019; Vasiliev 2019b; Lanzoni et al. 2018a, b), stellar envelopes (Marino et al. 2014; Kuzma, Da Costa & Mackey 2018; de Boer et al. 2019), and the

velocity dispersion profile (e.g. Scarpa, Marconi & Gilmozzi 2003; Scarpa et al. 2007; Küpper et al. 2010; Baumgardt & Hilker 2018; Baumgardt et al. 2019) are thought to trace both the formation and evolution of GCs.

However, in answering questions surrounding whether GCs are born *in situ* or *ex situ*, key information such as a GC's formation environment, or the time taken for a GC to be accreted into its galactic host, still remain unclear. In particular, whether or not GCs are born within dark matter mini-haloes is still under debate, although given their extreme age, theoretical models have suggested that GC formation occurs within a dark matter mini-halo of a mass of  $\sim 10^8 M_{\odot}$  (Peebles 1984; Trenti, Padoan & Jimenez 2015). The presence of stellar envelopes surrounding some GCs – where stars are confined to the GC over a long time-period – is in agreement with this theory (Carballo-Bello et al. 2012; Peñarrubia et al. 2017; Kuzma et al. 2018), whereas the presence of tidal features (e.g. Odenkirchen et al. 2001) is not (Moore 1996), and the absence

\* E-mail: zwan3791@uni.sydney.edu.au

of tidal tails in some GCs can in some cases be explained by the preferential loss of low-mass stars due to mass segregation (Balbinot & Gieles 2018). Furthermore, whilst the comparison between the dynamics and stellar luminosity in the inner parts of GCs suggests that it is not necessary to include non-baryonic dark matter (e.g. Conroy, Loeb & Spergel 2011; Kimmig et al. 2015; Watkins et al. 2015; Baumgardt 2017; Gieles et al. 2018), this is not evidence of the absence of dark matter in the outer regions of GCs; collisional relaxation can push the dark matter to the periphery where tidal interaction with the Milky Way (MW) is effective in stripping the entire dark matter content (Mashchenko & Sills 2005a,b; Baumgardt & Mieske 2008).

The dynamics at the periphery of GCs is where different models can be distinguished since the presence of dark matter inevitably inflates the velocity dispersion here. The existence of ‘potential escapers’ within this region creates some difficulty when testing these models. These stars are located within the Jacobi radius, but have energies above the critical energy for escape (Fukushige & Heggie 2000; Baumgardt 2001; Claydon, Gieles & Zocchi 2017; Daniel, Heggie & Varri 2017). As a result, the outer density profiles of GCs can be very similar to the dark matter prediction (Küpper et al. 2010), though finding the retrograde rotation of potential escapers (Tiongco, Vesperini & Varri 2016) would strongly support the scenario where GCs do not possess dark matter, at least not at present.

Direct imaging of stars in many GCs out to large radii is available (e.g. Simioni et al. 2018), although spectroscopic observations are still lacking for many stars beyond half the Jacobi radius (Claydon et al. 2017). This results from target selection based solely on colour–magnitude diagrams (CMDs) where MW stars significantly outnumber the cluster members in the low-density outskirts of GCs, resulting in a low efficiency when allocating spectroscopic fibres. However, this changed with the arrival of *Gaia* Data Release 2 (Gaia Collaboration et al. 2018a, b), which includes the precise proper motion measurements of distant halo stars, allowing the isolation of GC members using both their photometry and their astrometry. Using this catalogue, we have selected a sample of GC members and performed a spectroscopic survey of five nearby GCs – NGC 3201, NGC 1904, NGC 1851, NGC 1261, and NGC 4590 – with 2dF/AAOmega on the 3.9-m Anglo-Australian Telescope. We direct our efforts towards stars situated beyond half the Jacobi radius with the aim of understanding the dynamics of these GCs with the resulting moderate-resolution spectra of their members.

In this paper, we present a brief summary of our survey, and with the longest exposure time, we present the first scientific results on NGC 3201. This cluster is an interesting GC with its retrograde orbit, which is assigned as accreted in the Gaia-Enceladus/Sequoia event by Massari, Koppelman & Helmi (2019). We discuss the details of the survey in Section 2, including the target selection, observations and data reduction. To interpret the observations of NGC 3201, we compare our data to an  $N$ -body simulation in Section 2.4. We discuss our first results on NGC 3201 as well as the effects from binary stars and black holes (BHs) in Section 3 and present our conclusions in Section 4. We note that results on the remainder of the survey GCs will be published in later contributions.

## 2 OBSERVATIONS AND DATA REDUCTION

### 2.1 Target selection

The selection of targets for spectroscopic follow-up is based on the samples of GC members using data from *Gaia* DR2 produced by de Boer et al. (2019). The GC samples are extracted through the

application of a ‘matched-filter’ algorithm to the CMDs, using an isochrone from the Padova library (Marigo et al. 2017), as queried from <http://stev.oapd.inaf.it/cmd>. The GC metallicity ( $[\text{Fe}/\text{H}] = -1.59$ ) and distance (4.9 kpc) are taken from Harris (2010), with the age (11.5 Gyr) taken from Marín-Franch et al. (2009), Vandenberg et al. (2013). For a more secure membership selection, we consider only stars in a region around the isochrone with  $|(G_{\text{BP}} - G_{\text{RP}}) - (G_{\text{BP}} - G_{\text{RP}})_0| < 2 \times \delta(G_{\text{BP}} - G_{\text{RP}})$  at each  $G$  magnitude, with a minimum colour error of 0.03. Here, the  $G_{\text{BP}}$  and  $G_{\text{RP}}$  represent the magnitude in the *Gaia* GP and RP bands, and the  $\delta(G_{\text{BP}} - G_{\text{RP}})$  is the colour error.

The sample is further cleaned using *Gaia* DR2 proper motions to compute the membership probability of each star. The proper motions are fit using a Gaussian mixture model consisting of a cluster distribution and an MW foreground distribution. Initial guesses for the cluster Gaussian centres are taken from Helmi et al. (2018), before distributions are fit using the `emcee` python MCMC package (Foreman-Mackey et al. 2013). The parameters of the symmetric 2D Gaussian for NGC 3201 are  $[\mu_{\text{ra}}, \mu_{\text{dec}}, \sigma] = [8.37 \pm 0.12, -1.96 \pm 0.12, 0.47 \pm 0.15]$  mas yr $^{-1}$ . The final member samples are then selected by adopting a cut of 0.5 for the proper motion membership probability. To assess the importance of the various selection cuts, we note that the initial sample of 623583 stars is reduced to 79480 following the colour cuts and reduced further to 9913 given the proper motion selection. Therefore, applying the *Gaia* DR2 proper motion cuts is instrumental in obtaining a robust sample of high-probability members that can reasonably be followed up with spectroscopic facilities.

The resulting samples of members cover the entire spatial extent of the GCs, we intend to study, with proper motion errors of 0.6 mas yr $^{-1}$  at *Gaia*  $G = 19$  mag. For this survey, we focus on the Ca II triplet (CaT) at 8498.02, 8542.09, and 8662.14 Å (Edlén & Risberg 1956), hence all targets were selected from the red giant branch in each GC.

For our sample of NGC 3201 members, we find that there are  $\approx 10\,000$  member stars available within 2dF’s 2° field of view with 1944 members beyond  $0.25r_{\text{Jacobi}}$ . This ensures that enough fibres can be allocated outside of the densely crowded central regions of the GCs. The radii probed by 2dF are well outside the range of currently available data (within 15 arcmin from the GC centre) and contain sufficient numbers of cluster members to measure a possible bulk cluster rotation that is retrograde with respect to the orbit of the GC (due to potential escapers).

### 2.2 Observations with AAT and 2dF/AAOmega

The AAT is a 3.9-m optical telescope located at Siding Spring Observatory near Coonabarabran, New South Wales, Australia. For this study, we used 2dF/AAOmega, which is a fibre positioner with a field of view of 2° coupled to a dual-arm spectrograph. We use the 580V grating for the blue arm and the 1700D grating for the red arm, corresponding to resolutions of  $\sim 1300$  and  $\sim 10000$ , and wavelength ranges of 3800–5800 and 8400–8820 Å, respectively. The red arm setting enables the coverage of the target CaT lines, and, given the pixel resolution, results in velocity uncertainties of  $\sim 1$  km s $^{-1}$  with a signal-to-noise ratio (S/N) of 10.

We use CONFIGURE (Miszalski et al. 2006) to produce the configuration file for 2dF for each target list, and acquire sufficient biases and calibrations for the data reduction during observing nights. The total integration time per field pointing was 2 h, split into four individual exposures of 30 min to mitigate the effects of cosmic rays.

When possible, to mitigate the effects of binaries, we split the observations into multiple epochs with a separation of about a month

**Table 1.** The observational details of our GC survey, demonstrating multiepoch observations for three of our GC targets. Stars were observed at multiple times across different epochs so as to mitigate the effect of binaries. During the third epoch, due to scheduling constraints, we split the targets of NGC 3201 into two exposures.

Target name	Epochs	Exposure time	$N_{\text{target}}$	$N_{\text{goodstar}}$	MJD
NGC 3201	3	7200 s/ 7200 s/ 9000 s + 3600 s	248/ 252/ 321 + 147	207/ 213/ 320 + 146	58452.63/ 58479.63/ 58875.59
NGC 1904	2	7800 s/ 7200 s	188/ 77	121/ 45	58452.53/ 58875.49
NGC 1851	2	7200 s/ 1800 s	126/74	95/58	58479.54/ 58876.53
NGC 1261	1	7800 s	138	78	58452.42
NGC 4590	1	8400 s	92	76	58876.62

or more. More specifically, observations were performed in three blocks: 2018 November 29–30, 2018 December 27–30, and 2020 January 27–28. Overall, we obtained  $\sim 4$  nights-worth of useful observing time, during which the seeing ranged from 1.7– $\sim 3$  arcsec. We obtained a single epoch for each of NGC 1261, and NGC 4590, and multiple epochs for NGC 3201, NGC 1904, and NGC 1851. The observational epochs and exposure information are summarized in Table 1, and the survey footprint across NGC 3201 is presented in Fig. 1.

### 2.3 Data reduction

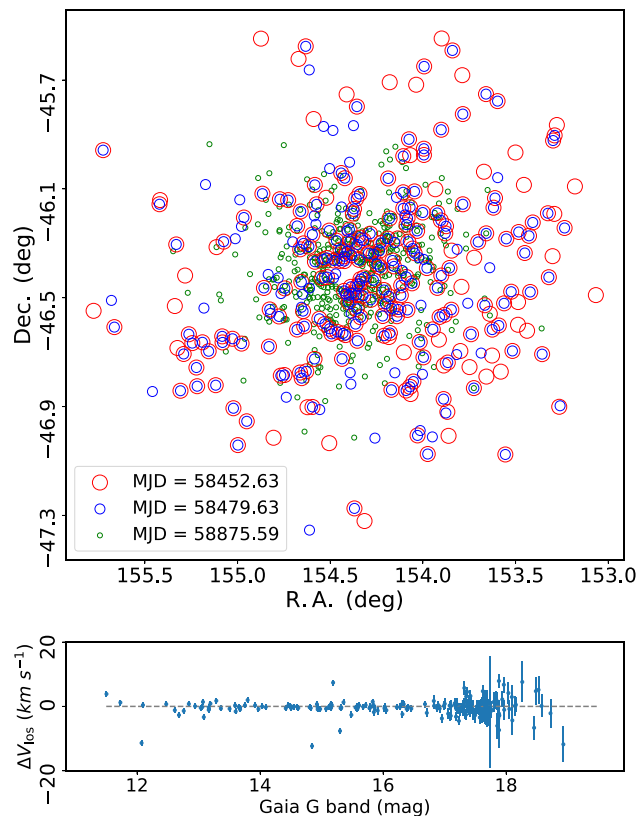
The raw data are primarily reduced with the 2DFDR<sup>1</sup> pipeline (AAO Software Team 2015) default setting `aaomega1700D` provided by the AAT, which automatically subtracts the bias, calibrates the pixel-to-pixel sensitivity using the fibre flats, and calculates wavelengths with the arc lines. In addition, the 2DFDR pipeline removes the sky spectrum, which is significant in the 1700D region.

The chemodynamical information, including radial velocity, are extracted using the CaT absorption lines. For this, we model each spectrum as consisting of the CaT lines and a continuum. The continuum is fit by means of a sixth-order polynomial with major spectral lines being masked out. We then normalize the flux of each spectrum to the best-fitting continuum. Then, we represent each line with a pseudo-Voigt profile (the summation of a Gaussian and a Lorentzian profile) as following:

$$\begin{aligned}
 \mathcal{F}(\lambda) &= A_0 \mathcal{G}(\lambda, \lambda_0, \sigma_g) + A_1 \mathcal{L}(\lambda, \lambda_0, \sigma_l), \\
 \mathcal{G}(\lambda, \lambda_0, \sigma_g) &= \frac{1}{\sqrt{2\pi}\sigma_g} e^{-\frac{(\lambda - \lambda_0)^2}{2\sigma_g^2}}, \\
 \mathcal{L}(\lambda, \lambda_0, \sigma_l) &= \frac{\sigma_l}{\pi((\lambda - \lambda_0)^2 + \sigma_l^2)}, \\
 \lambda_0 &= \lambda_{\text{LAB}} \times (1 + z).
 \end{aligned} \tag{1}$$

where  $z$  is the redshift, which is related to the velocity in the low-velocity regime through  $z = v/c$ . The  $A_0$  and  $A_1$  parameters are the strength of the Gaussian and Lorentzian profiles, respectively;  $\lambda$  is the wavelength, and  $\lambda_0$  is the spectral-line centre;  $\sigma_g$  and  $\sigma_l$  indicate the linewidth from the Gaussian and Lorentzian profiles, respectively. The spectral template is constructed with three pseudo-Voigt profiles, whose line centres are correlated by the redshift.

We fit each spectrum with the CaT profile above. The data uncertainties come from the *variance* from 2DFDR, which are taken into account by convolving with the spectrum profile parameters' probability distribution. The best-fitting line profile parameters and their uncertainties (defined as the mean and the  $1\sigma$  quantiles) were derived by MCMC sampling of the posterior using EMCEE (Foreman-Mackey et al. 2013). The systematic uncertainties are derived from



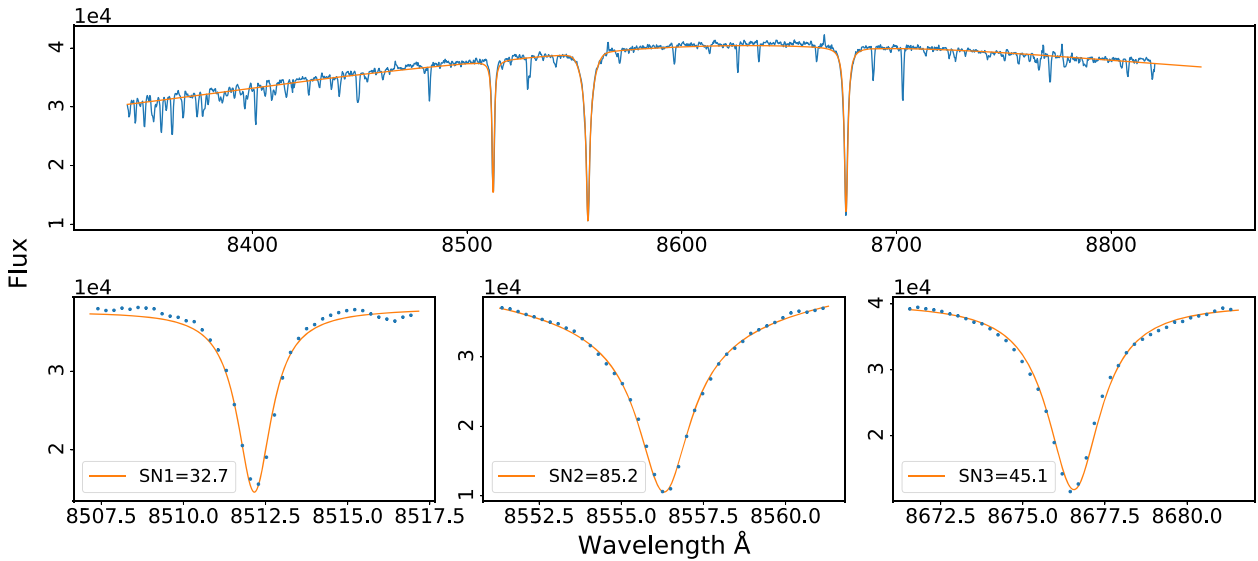
**Figure 1.** The top panel shows the footprint of our survey of NGC 3201, where different epochs are marked in different coloured and sized circles; some stars have multiepoch observations so that we can analyse the impact of binaries. The bottom panel presents the velocity difference between the first and second epochs, where some binaries deviate significantly from the zero line.

the comparison between multiple epoch observations. The  $S/N$  are defined as the ratio of the absorption-line strength to the residual surrounding the absorption lines ( $\pm 5 \text{ \AA}$ ). Here we define a *good\_star* when

$$\begin{aligned}
 S/N &> 3 \text{ and} \\
 \sigma_{\text{vlos}} &< 3 \text{ km s}^{-1}.
 \end{aligned} \tag{2}$$

As for stars with multiple observations, those with a velocity difference larger than  $5 \text{ km s}^{-1}$  are clearly binaries and are excluded from the sample, otherwise we only adopt the velocity information from the spectrum that has the highest  $S/N$ . The equivalent width (EW) of each of the lines is calculated by integrating  $\pm 20 \text{ \AA}$  over the line centre and the uncertainties of the EW are derived by repeating the integration 100 times with random noise. Fig. 2 presents one

<sup>1</sup><https://www.aao.gov.au/science/software/2dfdr>



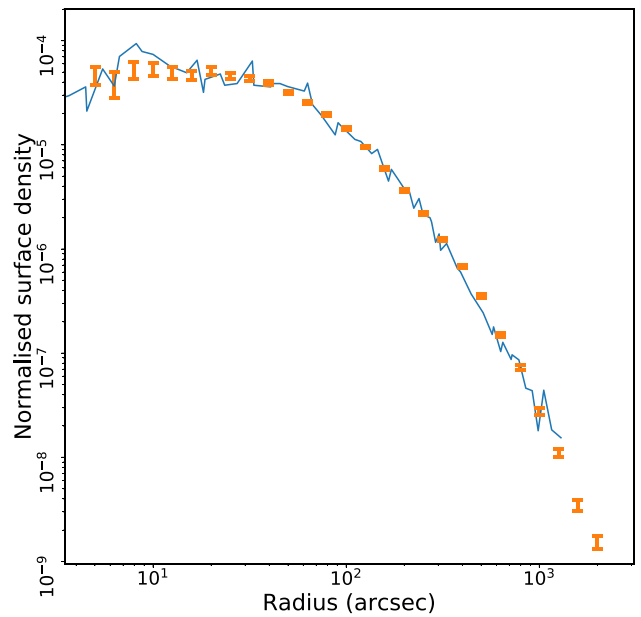
**Figure 2.** An example of a spectrum with a high S/N. The top panel shows the spectra in blue as well as the best-fitting CaT profile in orange. The bottom three panels are zoomed-in views of the three absorption lines.

example of a high-S/N star and a zoomed-in view of the three lines as well as the best-fitting model. Table 1 lists the number of targets and good stars for each GC along with other observational information.

#### 2.4 *N*-body simulations

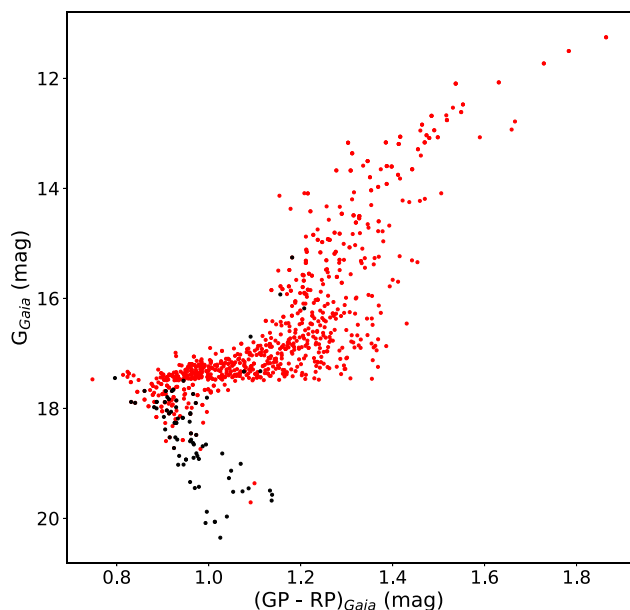
In order to interpret the observation and estimate the influence of the external tidal field of the MW on the outer dynamical profile of NGC 3201, we performed a series of direct *N*-body simulations, which were made with the direct *N*-body code NBODY7 (Nitadori & Aarseth 2012) on the OzSTAR GPU cluster of Swinburne University and the GPU cluster of the University of Queensland. We have implemented the MW potential of Irrgang et al. (2013) as an additional option for an external tidal field in NBODY7, in order to model the influence of the MW on NGC 3201.

For our simulations, we first integrated the orbit of NGC 3201 backwards in time for 4 Gyr in the MW potential of Irrgang et al. (2013) using a fourth-order Runge–Kutta integrator, with the initial phase-space parameters from Baumgardt et al. (2019). We then set up an *N*-body model of NGC 3201 that is non-rotating in an inertial reference frame and integrated the orbit of NGC 3201 forward in time to the present-day position using NBODY7. The initial *N*-body model was created based on the grid of *N*-body models described in Baumgardt & Hilker (2018), where the initial number of stars in the models of Baumgardt & Hilker (2018) was 100 000 and do not contain primordial binaries. These models started from King (1962) density profiles with varying concentration parameters  $c$ . The models of Baumgardt & Hilker (2018) followed a range of initial mass functions, starting with those from Kroupa (2001) and extending towards those that are more strongly depleted in low-mass stars. The initial cluster models of Baumgardt & Hilker (2018) were unsegregated, however, mass segregation developed dynamically over time, so the simulations presented here started from already mass segregated models. This mass segregation increased further during the 4 Gyr duration of the simulations and developed into a cluster that is segregated in the same way as seen for NGC 3201. Since NGC 3201 loses about 5 per cent of its stars due to interaction with the tidal field of the MW during the 4 Gyr of the simulation and



**Figure 3.** The number density profile of bright stars from the simulation (blue, solid line) compared to the observed surface brightness profile of NGC 3201 from Trager, King & Djorgovski (1995) (orange error bars). The simulation agrees excellently with the observed surface density profile.

also shrinks its core size by about 20 per cent due to the two-body relaxation driven evolution towards core collapse, we varied both the initial cluster mass and density profile slightly until we found the best match to the present-day observations of NGC 3201. Fig. 3 compares the surface density of bright stars in the simulation and the estimation based on the surface brightness from Trager et al. (1995). Also, comparing to a 2 Gyr simulation, we obtain similar results in terms of the final cluster size and the final velocity dispersion profile; hence, we expect that an even longer simulation time will not change our final results. After the simulation finished, we extracted the particle data from the simulation, projected the cluster on to the



**Figure 4.** The *Gaia* CMD of the NGC 3201 targets. The red dots are the *good\_star* targets (see the definition in the text). The targets of the third epoch are brighter than 17.5 mag, and most of the *good\_star* targets are brighter than 18 mag.

sky and analysed the observations in the same way in which we analysed the observational data.

### 3 RESULTS AND DISCUSSION

As a result of the above observations and analysis for NGC 3201, we successfully extracted 886 good stellar spectra for 694 stars, and we excluded 11 clear binaries. Among the final sample, we have multiple epoch observations for 170 stars, and 94 stars (51 of them have multiple epoch observations) located beyond one half of the Jacobi radius, enabling us to characterize the dynamics out to a large distance from the cluster centre. Fig. 4 presents the CMD of the cluster, where the *good\_stars* are colour-coded in red. At  $V = 18$  mag, the number of good stars decreases significantly. The mean LOS velocity of NGC 3201 is  $496.47 \pm 0.11$  km s $^{-1}$  based on the observations. A summary of these results is presented in Fig. 5, which shows the LOS velocity distribution for the target stars, including their individual values as a function of cluster-centric radius. With this, we address several key scientific questions about NGC 3201.

#### 3.1 Is NGC 3201 rotating?

One of the most prominent dynamical signatures would be rotation, which is a record of the cumulative effects from the birth of the GC, two-body relaxation and interaction with the tidal field of the host galaxy. A good understanding of the rotation is also important for determining the velocity dispersion (e.g. Cote et al. 1995; Bellazzini et al. 2012; Ferraro et al. 2018; Sollima et al. 2019). With precise LOS velocity measurements, internal rotation has been found in different clusters (e.g. Kamann et al. (e.g. Bianchini et al. 2018; Kamann et al. 2018; Sollima et al. 2019, and references therein) and is typically measured using the LOS velocity difference on both sides of a central axis. The left panel of Fig. 6 shows the tangent plane projection of stars in NGC 3201 coloured by the measured LOS velocities (with

the mean LOS velocity subtracted), where the inner and outer dashed circles indicate the King tidal radius (36.1 pc or 1520 arcsec, Harris 1996) and an estimation of the Jacobi radius (83.46 pc, or 3513 arcsec, Balbinot & Gieles 2018), respectively. Similarly, the right-hand panel shows the best-fitting simulation, colour-coded with the LOS velocity.

The velocity variation of member stars in NGC 3201 can be reproduced by the best-fitting simulation. As shown in Fig. 6, there is an obvious rotation signal with amplitude of  $\sim 5$  km s $^{-1}$ , with stars in the east (positive  $\Delta RA$ ) are moving away from the observer (relative to the cluster centre), while stars in the west (negative  $\Delta RA$ ) are moving towards us (relative to the cluster centre). This signal is mostly due to perspective rotation, which is important for objects with a large angular diameter that have a high systemic proper motion. For the distance, systemic proper motion, and diameter of NGC 3201 ( $\sim 5$  kpc,  $\sim 10$  mas yr $^{-1}$ ,  $\sim 100$  arcmin, respectively) the magnitude of this effect is a velocity difference of  $\sim 7$  km s $^{-1}$  (van de Ven et al. 2006). Perspective rotation is therefore important and responsible for most of the signal seen in Fig. 6.

In addition to the perspective rotation effects, several other scenarios can also lead to rotation. For example, as the GC evolves within the MW potential the LOS velocity varies along the GC orbit, which will be most obvious for the unbound stars in the tidal tails. Furthermore, potential escapers at large radii can contribute to the rotation signal. Finally, internal rotation from the formation of the GC could naturally lead to an observable rotation.

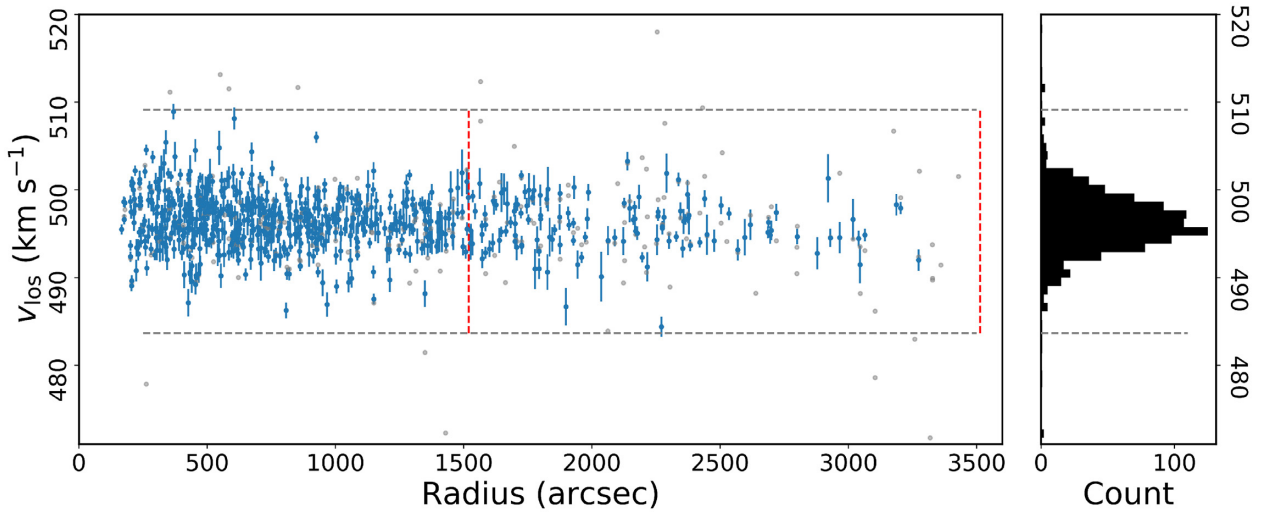
To calculate the perspective rotation effect on the observed LOS velocities of NGC 3201, we assume that NGC 3201 is centred on (RA, Dec.) = (154:40, -46:41) at a heliocentric distance of 4.9 kpc (Harris 2010), with a systemic LOS velocity of 496.47 km s $^{-1}$  and proper motion of  $(\mu_\alpha, \mu_\delta) = (8.37, -1.96)$  mas yr $^{-1}$ . We calculate the systemic velocity of the simulation by taking the mean of the stars within  $0.5$  around the GC centre. The perspective rotation effect can then be calculated for each star from the systemic velocity using equation (6) of van de Ven et al. (2006). To adjust for this effect in the observed and simulated stars, it is then subtracted from each star's corresponding observed LOS velocity. Fig. 7 shows the LOS velocity of the simulation and the observation after having adjusted for the perspective rotation effect.

A simple relation that includes a rotation component

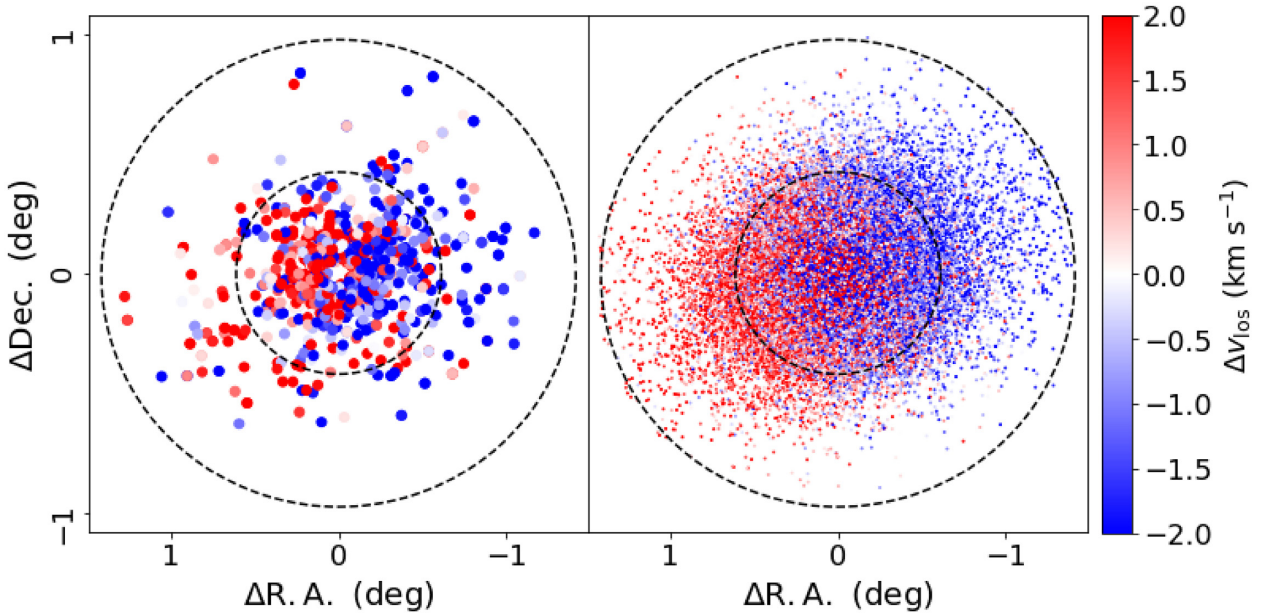
$$v_{\text{los},0} = A_{\text{rot}} \sin(\phi - \phi_0) + v_{\text{sys}},$$

$$p(v_{\text{los}}) = \frac{1}{\sqrt{2\pi}\sigma_{v_{\text{los}}}} \exp\left[-\frac{(v_{\text{los}} - v_{\text{los},0})^2}{2\sigma_{v_{\text{los}}}^2}\right], \quad (3)$$

is fitted to the residual velocity within radial bins for both observation and simulation including the velocity errors in quadrature. Here,  $A_{\text{rot}}$  is the amplitude of the velocity difference;  $\phi$  is the directional angle from the rotation axis increasing from the north to the east and  $\phi_0$  is the reference positional angle (PA);  $v_{\text{sys}}$  is the systemic velocity along the line of sight and  $\sigma_{v_{\text{los}}}$  is the intrinsic LOS velocity dispersion. The posterior parameters space is sampled with an MCMC approach, and the best-fitting parameters and  $1\sigma$  uncertainties are summarized in Table 2. In the inner part of the cluster (for stars within  $\sim 900$  arcsec around the GC centre), we found a signal of rotation with amplitude of around 1 km s $^{-1}$  (see Table 2 for detailed profile). The amplitude of the rotation becomes weaker at larger radius. For stars beyond 900-arcsec radius, the amplitude decreases to  $A_{\text{rot}} \approx 0.35^{+0.32}_{-0.24}$  km s $^{-1}$ . The PA of the rotation axis in the inner part is  $\sim 133^\circ$ – $170^\circ$ , which is significantly different from the simulation (PA =  $30.3^{+62.6}_{-68.1}$  with a very weak amplitude  $A_{\text{rot}} = 0.13^{+0.11}_{-0.09}$  km s $^{-1}$  for stars with radius between 300 and 900 arcsec, and PA =  $22.5^{+27.6}_{-30.7}$  with an amplitude



**Figure 5.** The LOS velocity of target stars, where the blue stars are those targets with S/N larger than 3, and grey dots are low S/N stars. Given the high radial velocity, the members of NGC 3201 are well separated from the contaminating MW halo stars. The right panel is the velocity distribution of all targets in the observation, which peaks at  $496.4 \text{ km s}^{-1}$ . The two horizontal dashed lines indicate the  $5\sigma$  range of the  $v_{\text{los}}$ , and the two vertical red dashed lines indicate the King tidal radius (Harris 1996) and the Jacobi radius (Balbinot & Gieles 2018) correspondingly.

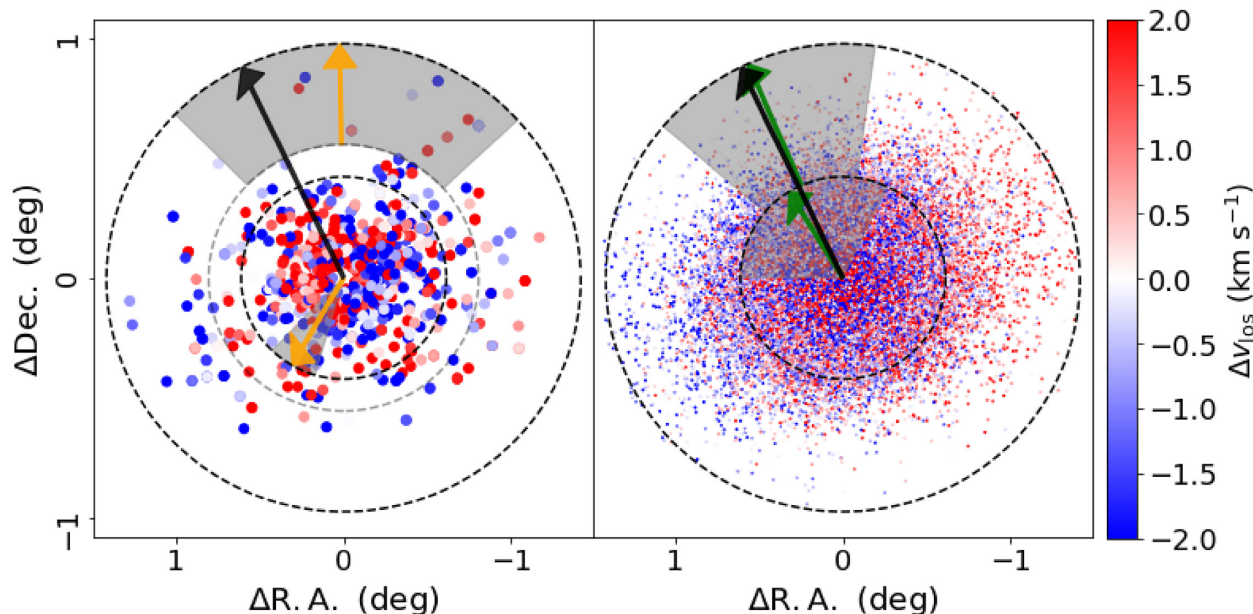


**Figure 6.** The projection on the sky of the observations (left-hand panel) and the best-fitting simulation (right panel) colour-coded by their LOS velocities. The inner and outer dashed circles in both panels indicate the King tidal radius (Harris 1996) and Jacobi radius (Balbinot & Gieles 2018), respectively. The simulation exhibits a similar apparent rotational velocity pattern to the observed data, which is due to the effect of perspective view effect.

$A_{\text{rot}} = 0.64^{+0.33}_{-0.32} \text{ km s}^{-1}$  for stars between the King tidal radius and the Jacobi radius). At the outer part of the GC, where the radius  $r > 2000 \text{ arcsec}$ , we found an opposite rotational signal compared to the rotation at the inner part, with  $A_{\text{rot}} = 0.80^{+0.49}_{-0.47} \text{ km s}^{-1}$  and  $\text{PA} = 0.6^{\circ+45.8^{\circ}}_{-49.2^{\circ}}$ , which agrees with the rotational direction of the simulation. This counterrotation at the outermost region relative to the inner one is in good agreement with the prediction of a tidally perturbed, rotating stellar cluster from Tiongco, Vesperini & Varri (2018). The visualization of the results is presented in Fig. 7.

The rotation could be characterized by angular momentum relative to the GC centre. Though we do not have the full phase-space information for the observation data, we can explore the dynamical

features of the simulation. Hence, we calculated the present-day angular momentum of the stars in the simulation with respect to the GC centre. Here, we use the coordinate system that has the  $z$ -axis perpendicular the orbital plane, and the  $x$ -axis aligned with the systemic velocity of the GC, but the frame is inertial, i.e. non-rotating. Fig. 8 presents the angular momentum of stars – with  $r < 2000 \text{ arcsec}$  and  $2000 < r < 3600 \text{ arcsec}$  around the GC centre. For all stars at the inner part, the mean and dispersion of each components of the specific angular momentum are  $(L_x, L_y, L_z) = (0.00 \pm 0.02, 0.00 \pm 0.02, 0.00 \pm 0.02) \text{ kpc km s}^{-1}$ ; for stars at the outer part, we find  $(L_x, L_y, L_z) = (0.01 \pm 0.06, -0.01 \pm 0.06, 0.04 \pm 0.06) \text{ kpc km s}^{-1}$ . The results at the inner part indicate that the GC

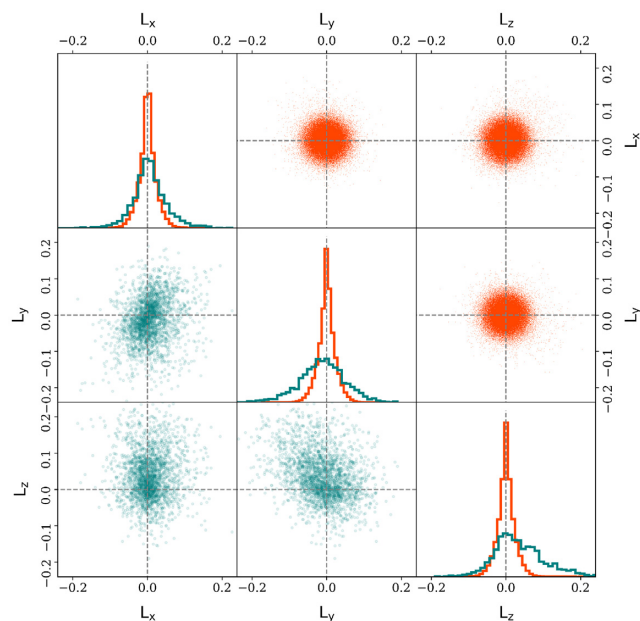


**Figure 7.** The comparison of LOS velocity with the perspective rotation effect subtracted between the observation (left-hand panel) and simulation (right-hand panel). The orange arrows (left) show the best-fitting PA for the stars with radius  $433.2 < r < 956.7$  arcsec and  $r > 2000$  arcsec (see the values in the Table 2) in the observations, and the grey region shows the  $1\sigma$  range. Correspondingly, the green arrows (right) show the best-fitting PA for the stars in simulation, where the results from inner part (within King tidal radius) and outer part (beyond King tidal radius) are shown separately, and the grey regions again show the  $1\sigma$  range. The black arrows in both panels indicate the expected rotational axis for the potential escapers (PA =  $26.3^\circ$ ). Within the inner part, the disagreement between the simulation and observation suggests that this velocity variation comes from the internal rotation. At the outer part, there is some signal of unbounded stars that present the rotational direction aligned with the potential escapers.

**Table 2.** The estimated velocity dispersion profile. The first and second columns show the range in radius; the third column give the mean radius in each bin; the fourth column shows the estimated dispersion and  $1\sigma$  uncertainties; and the last column gives the number of stars within each bin. The last two rows present the fitting results of the rotational signal from the inner and outer parts of the GC, respectively.

$R_{\text{low}}$ (arcsec)	$R_{\text{high}}$ (arcsec)	$\langle R \rangle$ (arcsec)	$A_{\text{rot}}$ ( $\text{km s}^{-1}$ )	$\phi_0$ $^\circ$	$\sigma_{\text{los}}$ ( $\text{km s}^{-1}$ )	$N$
166.6	433.2	312.5	$0.39^{+0.37}_{-0.27}$	$89.9^{+64.8}_{-73.4}$	$3.21^{+0.22}_{-0.20}$	136
433.2	669.3	544.8	$0.88^{+0.41}_{-0.42}$	$169.8^{+26.9}_{-24.4}$	$2.90^{+0.22}_{-0.19}$	136
669.3	956.7	804.0	$1.09^{+0.37}_{-0.37}$	$132.4^{+18.5}_{-16.6}$	$2.54^{+0.19}_{-0.17}$	135
956.7	1434.6	1181.0	$0.35^{+0.32}_{-0.24}$	$49.3^{+50.6}_{-80.8}$	$2.48^{+0.19}_{-0.16}$	136
1435.6	3273.8	2019.0	$0.51^{+0.35}_{-0.32}$	$312.1^{+35.6}_{-33.4}$	$2.01^{+0.18}_{-0.16}$	136
433.2	956.7	665.3	$0.96^{+0.26}_{-0.26}$	$149.8^{+15.5}_{-14.3}$	$2.70^{+0.14}_{-0.13}$	272
2000	3600	2468	$0.80^{+0.49}_{-0.47}$	$0.6^{+45.8}_{-49.3}$	$2.40^{+0.33}_{-0.29}$	54

has no internal rotation after 4 Gyr evolution, and the clear bias from zero at the outer part is due to the potential escapers. We can compare these values to what is expected from potential escapers. For circular orbits, in a reference frame that co-rotates with the orbit, prograde stars are preferentially lost, resulting in a net retrograde solid-body rotation of potential escapers (Claydon et al. 2017; Daniel et al. 2017). Tiongco et al. (2016) showed the average angular frequency of the potential escapers is  $\langle \Omega_{\text{PE}} \rangle = -0.5\Omega_{\text{orb}}$ , where  $\Omega_{\text{orb}}$  is the angular frequency of the orbit. In a non-rotating frame  $\langle \Omega_{\text{PE}} \rangle = +0.5\Omega_{\text{orb}}$ . We can approximate the eccentric orbit of NGC 3201 by a circular orbit at Galactocentric radius  $R_p(1+e) \simeq 13$  kpc (Baumgardt & Makino 2003; Cai et al. 2016), where  $R_p \simeq 9$  kpc is the pericentre distance and  $e \simeq 0.5$  the eccentricity (Gaia Collaboration et al. 2018c). Assuming



**Figure 8.** The components of the angular momentum of stars of the simulation with  $r < 2000$  arcsec (orange-red) and  $2000 < r < 3600$  arcsec (green) from the GC centre. The inner part shows no significant angular momentum signal, suggesting that the GC does not have internal rotation. At the outer part, the angular momentum bias comes from the potential escapers.

a flat rotation curve of  $220 \text{ km s}^{-1}$ ,  $\Omega_{\text{orb}} \simeq 0.017 \text{ Myr}^{-1}$ , and thus  $\Omega_{\text{PE}} \simeq 8.7 \times 10^{-3} \text{ Myr}^{-1}$ . The average orbital velocity at  $0.75r_{\text{Jacobi}} \simeq 63$  pc is then  $0.55 \text{ km s}^{-1}$ , and in the non-rotating frame the average angular momentum is  $\langle L_z \rangle \simeq 0.035 \text{ kpc km s}^{-1}$ , i.e. as we find in the

$N$ -body model, suggesting that the rotation we see in the outskirts of the  $N$ -body model is due to potential escapers.

The rotation axis of the potential escapers is aligned with the angular momentum vector of the Galactic orbit, which is well constrained by the systemic proper motion, distance and line-of-sight velocity of NGC 3201. In Fig. 7 we show in both panels with a black arrow the projection of the angular momentum vector of the Galactic orbit. As Fig. 7 shows, the direction of the rotational axes in the inner part of the observation indicates that the rotational signal within  $r < 900$  arcsec (the orange arrow) is different from what is expected from potential escapers (the black arrow). Hence this signal at the inner part of NGC 3201 is likely to be the internal rotation of the GC. At the outer part ( $r > 2000$  arcsec), we found that the direction of the rotational signal is aligned with the potential escapers, which suggests that those stars at the outskirts are likely energetically unbound, yet associated to the cluster (Henon 1970). As far as we are aware, this is the first detection of this predicted signal of potential escapers in a star cluster. In the  $N$ -body simulation, the rotation in both inner and outer regions aligns with the potential escaper prediction. This is because the model started without rotation, so all the rotational signal is imposed by the tides.

### 3.2 Velocity dispersion

For a pressure-supported system in dynamical equilibrium, the dispersion is directly related to the average internal kinetic energy. This is then related to gravitational potential energy as based upon the virial theorem. Hence, the mass profile of the stars, as well as any dark content, can in principle be estimated by measuring the dispersion profile. A typical way of interpreting the result is to compare the measured profile to a model (e.g. Bianchini, Ibata & Famaey 2019; Hénault-Brunet et al. 2019; Vasiliev 2019a). In this section, we present our estimate of the dispersion profile with a higher precision and discuss the effects of the MW potential, binaries and the escape rate (extratidal stars, considering the effect of stellar-mass BHs).

As noted in the previous section, NGC 3201 presents a pattern where the LOS velocities are larger on the east side than on the west side. This systematic variation of velocity has to be taken into consideration when estimating the dispersion. Here, the dispersion was included in the rotation model (equation 3), and the best-fitting intrinsic dispersions are listed in Table 2. As a consistency check and to demonstrate the dispersion profile in the inner part of the GC, we also included previously published  $v_{\text{los}}$  data (Baumgardt & Hilker 2018; Giesers et al. 2019), and the proper motion dispersion profiles from Bianchini et al. (2019) and Vasiliev (2019a). Meanwhile, we fit the same relation to the  $N$ -body simulation described in Section 2.4, and include the dispersion profiles from the LIMEPY models (Gieles & Zocchi 2015) and SPES models (Claydon et al. 2019) from de Boer et al. (2019) as comparisons.

Fig. 10 shows the velocity dispersion profile from the observations and simulation, where the top panel shows the dispersion in LOS, and the bottom panel shows the dispersion in the tangent plane. The two dashed lines in both panels indicate the King tidal radius and Jacobi radius, respectively. The velocity dispersion within  $\sim 500$  arcsec from the GC centre can be well reproduced by the simulation, and our observations agree with the data from the literature. However, the simulation is significantly lower than the observations at larger radius. At radii beyond 2000 arcsec, we find that the LOS velocity dispersion is  $\sim 2.01 \pm 0.18 \text{ km s}^{-1}$ , and tends to flatten outwards, whereas the dispersion of the simulation decrease faster with radius and is about  $1.48 \pm 0.14 \text{ km s}^{-1}$ , i.e.  $\sim 2\sigma$  lower. The dispersion

profiles of the LIMEPY and SPES models are presented as red and blue regions respectively. Compared to King (1966) models, the LIMEPY models have an additional degree of freedom that describes the ‘sharpness’ of the energy truncation, and these models are therefore more flexible in describing the outer density profiles. The SPES model includes a prescription for potential escapers.

The model parameters are taken from de Boer et al. (2019). The cluster mass is 27 per cent smaller than the mass from Baumgardt et al. (2019), so that the models have the best fit to the inner part data from Baumgardt & Hilker (2018), Giesers et al. (2019). We note that the  $N$ -body models from Baumgardt et al. (2019) are multimass, where the massive stars move a bit slower due to equipartition, which allows for a larger mass than the LIMEPY and SPES models. The dispersion profiles from these two models agree with the  $N$ -body simulation. However, both models underestimate the dispersion of the GC at the outer part. As for the dispersion in the tangent plane, the dispersion profile from Vasiliev (2019a) agrees with the tangential dispersion profile from Bianchini et al. (2019). The tangential dispersion is lower than the radial dispersion profile, which is expected if the cluster has radially biased velocity anisotropy. Our results agree well with the observations in the inner part of the GC, especially, the dispersion profile from Vasiliev (2019a) and the tangential dispersion profile from Bianchini et al. (2019). In the outer part of the GC, the observed proper motion dispersion in the tangent plane is significantly larger than the simulation. Similarly, Bianchini et al. (2019) compare the proper motion dispersion profiles to the model for the dispersion of potential escapers from Claydon et al. (2017), finding that the observed dispersion out to the Jacobi radius is approximately half the model prediction of Claydon et al. (2017). This is perhaps not too surprising, because the potential escaper model of Claydon et al. (2017) was derived for circular orbits, and NGC 3201 is near pericentre, where the dispersion of potential escapers is about twice as high as near apocentre for an eccentricity of 0.5 (see fig. 10 in Claydon et al. 2019). However, the potential escapers are present in the  $N$ -body simulation and their effects are included in the dispersion profile of the simulation. Hence, the potential escapers are unlikely to cause the observed large dispersion.

Several scenarios could potentially explain the discrepancy between observations and simulations. The large dispersion might relate to the interaction with the galactic potential. The heating when the GC crosses the galactic disc might also increase the dispersion (e.g.  $\omega$  Cen, Da Costa 2012). However, compared to  $\omega$  Cen, NGC 3201 has a much larger peri-galacticon radius (Baumgardt et al. 2019), where the disc heating is insignificant. In addition, the  $N$ -body model, we adopted includes the influence from the MW (as well as the disc). We conclude that the excess dispersion is unlikely to be a result of the interaction between the GC and the potential field of the MW.

In the following sections, we will discuss two additional mechanisms that might lead to the flattened dispersion profile in the outer parts of NGC 3201, and our estimation of their effects on the observations.

### 3.3 The effect of binaries

The observed velocity of a binary can be significantly different from the systemic velocity of the GC due to its internal orbital velocity. Some binaries with a large velocity deviation from the GC velocity or with a short period would be easily identified (see Figs 1 and 5), while some long-period binaries are difficult to detect. The presence of



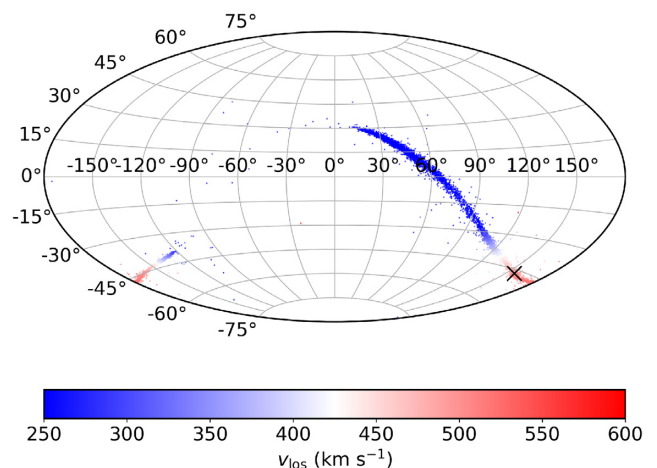
undetected binaries would inevitably influence the measured velocity and hence the estimated dispersion.

The velocity of a star in a binary system varies periodically, and with our multipoch observation, short-period binaries with a significant velocity difference at different epochs are easily identified as binaries. Fig. 1 shows the velocity difference between epochs, where some binaries deviate significantly from zero. Long-period binaries, however, are difficult to identify. We also note that a fraction of our targets were only observed for a single epoch, and binaries (even short-period) within that sub-sample cannot be identified. Hence, a proper estimation of the effect of binaries is necessary for the dynamical analysis.

To estimate the effect of binaries, we first randomly sampled binaries with VELBIN (Cottaar, Meyer & Parker 2012; Cottaar & Hénault-Brunet 2014) from distributions of period, mass ratio and eccentricities appropriate for solar-type binaries (Raghavan et al. 2010). Given that our target stars are either near the main-sequence turnoff or RGB stars, we assume a mass of  $0.8 M_{\odot}$  for the primary star in the binary systems. Since softer binaries would be disrupted in a cluster environment, we retained only hard binaries. Giesers et al. (2019) found that all the binaries for which they secured orbital solutions in NGC 3201 have energies a factor of  $\sim 5$  or more above the hard-soft boundary. We therefore kept only hard binaries with an orbital velocity larger than three times the current central velocity dispersion of the cluster. This translates into a higher minimum period for the binaries, and accounts for the possibility that binaries with an energy just above the present-day hard-soft boundary have been destroyed in the past when the cluster was more massive and more compact.

Based on the binary sample, we constructed mock radial velocity data sets with time baselines, radial velocity uncertainties, and numbers of epochs that mimic our observed cluster member stars in the two outer bins shown in Fig. 10. The cluster velocity dispersion is initialized to be  $1.5 \text{ km s}^{-1}$ , comparable to the velocity dispersion of the best-fitting simulation in the outermost radial bins. We adopted different binary fractions and just like in the real observations we excluded from the final data sets the stars that would have been identified as binaries, as well as stars that would have not been retained as likely cluster members based on a significantly discrepant single-epoch measurement of  $v_{\text{los}}$ . We calculated the resultant dispersion and repeated the experiment for a large number of random samples and mock data sets. We kept track of the rate of radial velocity variables that would have been detected. For realistic binary fractions and orbital parameter distributions, this should be consistent with the observed rate of variables observed in our sample of cluster members ( $\sim 2.5$  per cent in the radial region of the two outermost bins in the top panel of Fig. 10).

Fig. 11 shows the probability distribution of the final dispersion for different binary fractions. Different binary fractions result in different measured velocity dispersion differences. With the initial cluster dispersion of  $1.5 \text{ km s}^{-1}$ , the probability of producing a dispersion larger than  $\sim 2 \text{ km s}^{-1}$  is 1.6 per cent given a 5 per cent binary fraction, 23 per cent given a 10 per cent binary fraction, and a velocity dispersion of  $2 \text{ km s}^{-1}$  or more is easily obtained for a binary fraction higher than 20 per cent. The probability of producing a dispersion larger than  $2.5 \text{ km s}^{-1}$  (as observed at a projected radius of  $\sim 1000$  arcsec; Fig. 10) is 0 per cent given a 5 per cent binary fraction, 0.04 per cent given a 10 per cent binary fraction, and 12.4 per cent given a 20 per cent binary fraction. However, Giesers et al. (2019) show the core binary fraction in NGC 3201 is  $6.75 \text{ per cent} \pm 0.72 \text{ per cent}$ , which decreases outwards (Milone et al. 2016) with radius. In addition, with deep field observations



**Figure 9.** An all-sky Aitoff projection of the output of the simulation of NGC 3201, colour-coded with respect to the LOS velocity. The cross indicates the location of the GC. The plot shows that the change of the LOS velocity is not a local effect, but extends continually along the tidal arms around the MW. The LOS velocity of the GC is close to the maximum, and decreases significantly along the stream. At distances larger than  $5^{\circ}$  from the centre of NGC 3201, the LOS velocity is smaller than  $480 \text{ km s}^{-1}$ .

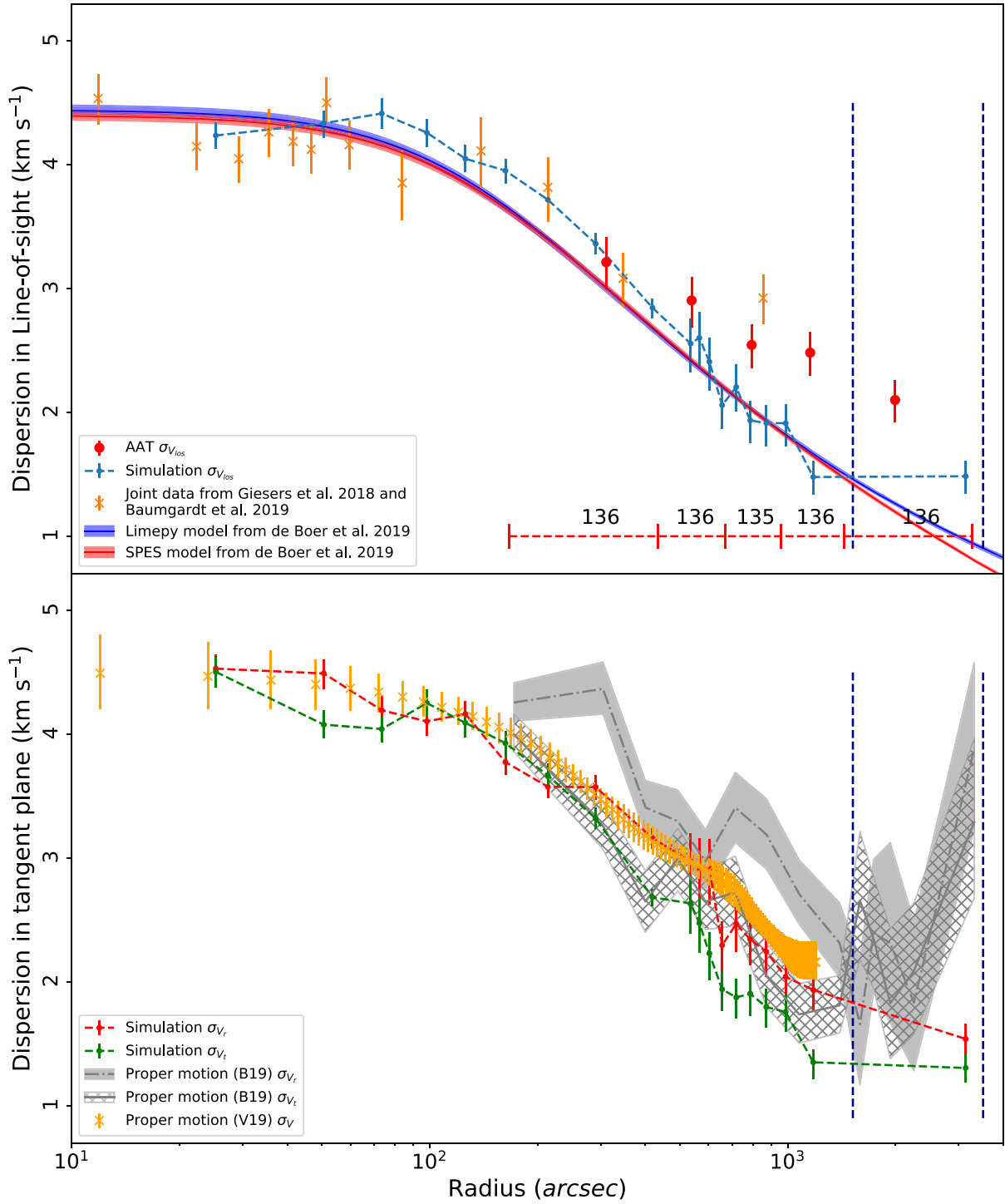
out to 8 arcmin (Simioni et al. 2018), the tight main-sequence track argues against a high binary fraction. The observed rate of radial velocity variables (2.5 per cent) in our sample of cluster members in the two outer radial bins of Fig. 10 also argues against a binary fraction significantly larger than 10 per cent. Adopting a binary fraction of 20 per cent or higher in our mock radial velocity experiments yields a typical rate of detected radial velocity variables in excess of 4 per cent. A binary fraction large enough to significantly inflate the velocity dispersion would also overpopulate the wings of the velocity distribution compared to the observed sample. For example, with a 20 per cent binary fraction, we would expect in excess of 15 stars outside the  $5\sigma$  range shown in Fig. 5 at radii beyond 1000 arcsec even before considering non-members, which is already more than we observe.

Hence, although we cannot exclude that undetected binaries contribute to inflating the observed velocity dispersion, we conclude that the underestimation of the dispersion in the outer parts of the GC is unlikely to be purely due to the effect of undetected binaries given that the binary fraction is likely to be smaller than 10 per cent.

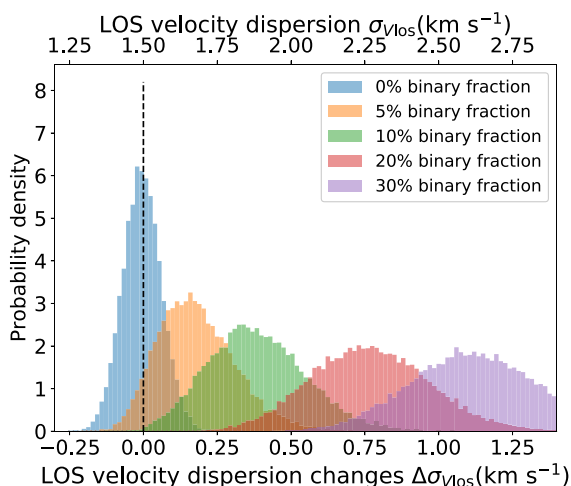
### 3.4 The effect of different escape rates

As we can see from Fig. 9, interacting with the MW produces the tidal tails from the escaped stars. The unbound stars in the tidal tails might increase the measured LOS velocity dispersion depending on the viewing angle. The escape rate, which describes the efficiency with which stars escape from the GC, will determine the number of stars inside the tidal tails, and thus might change the dispersion.

Stellar-mass BHs are believed to be able to shape the core profiles of GCs and increase the escape rate of stars. The GCs in the MW possess a clear separation in the distribution of core radii into ‘core collapsed’ and ‘non core collapsed’ clusters, defined by small and large core radii, respectively. With strong gravitational interaction, the BHs effectively deposit energy into the GC bulk population, leading to a ‘puffier’ core (e.g. Merritt et al. 2004; Mackey et al. 2007, 2008; Peuten et al. 2016). In the outer parts, BHs can increase the escape rate of the cluster by close interaction with other stars



**Figure 10.** The top panel depicts the dispersion in the LOS from the N-body model (blue), from the AAT 2dF/AAOmega observations (red), and from the previously published data (orange). The LIMEPY and SPES models with  $1\sigma$  uncertainties from de Boer et al. (2019), where the mass of NGC 3201 is 27 per cent smaller than the mass from Baumgardt et al. (2019), which leads to a best fit to the previously published data (Baumgardt & Hilker 2018; Giesers et al. 2019). The red dash lines at the bottom indicate the radial range of each bin and the number above them are the number of stars in each bin. The error bars indicate the  $1\sigma$  uncertainty. The bottom panel depicts the dispersion in the tangent plane from the N-body model, and from Vasiliev (2019b, hereafter V19) and Bianchini et al. (2019, hereafter B19) [including the radial (grey) and tangential (grey hatched) components, where the regions indicate the  $1\sigma$  ranges, respectively]. In both panels, the left and right dark vertical dashed lines mark the King tidal radius and Jacobi radius, respectively. Within  $r = 500$  arcsec, the simulation agrees well with observation, however, the observed dispersion is significantly larger than that in the simulation at large radii.



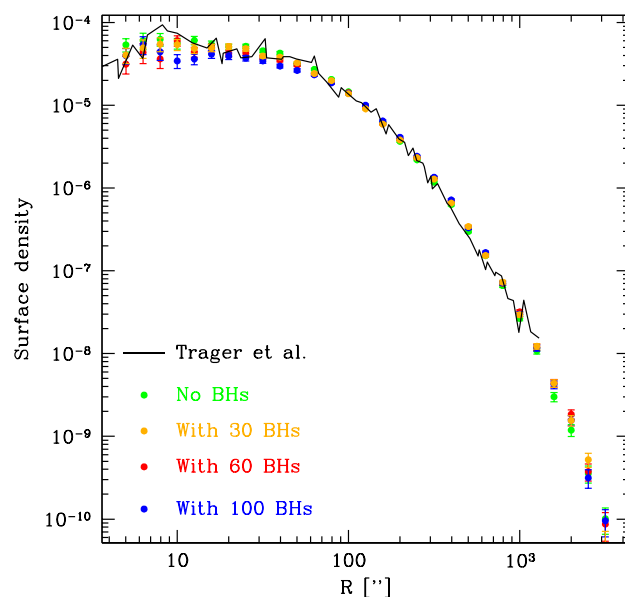
**Figure 11.** The effect of binaries on the velocity dispersion profile. With the initial dispersion set to be  $1.5 \text{ km s}^{-1}$ , this figure shows the distribution of the final dispersion. Given a reasonable binary fraction of less than 15 per cent, the presence of binaries cannot significantly change the estimated dispersion.

(Giersz et al. 2019; Wang 2020), resulting in a larger number of stars in the tidal tails. NGC 3201 is known to host stellar mass BHs from radial velocity measurements (Giesers et al. 2018), and the luminosity profile of the core region (Askar, Arca Sedda & Giersz 2018; Kremer et al. 2019). However, the effect on the LOS dispersion from the BHs at large radii is unknown.

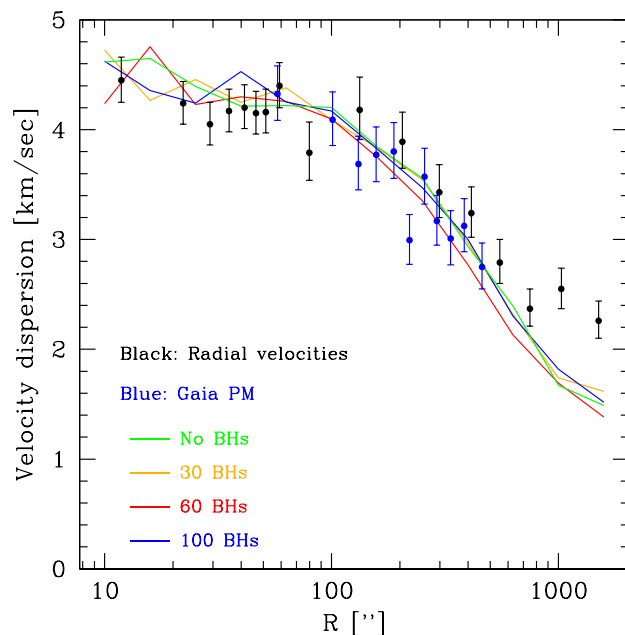
To explore the effects from BHs, we included extra BHs in the  $N$ -body simulation, but kept all the other parameters the same as the best-fitting model. Following 4 Gyr of evolution of the simulations with varying BH numbers, the 5 per cent Lagrangian radius (the radius which contains 5 per cent of the bound mass of the cluster) of the GC differs significantly, whereby the core of the GC with 100 BHs is 3 per cent larger than that of the GC without BHs. Correspondingly, the surface density profile in the inner part of the GC with BHs is slightly lower than the GC without BHs. However, Fig. 12 shows that the GCs with 30 and 60 BHs still fit the data reasonably well within the core region.

Compared to the cluster without BHs, the escape rate is about 1.3 per cent higher for the cluster with 30 BHs, and is about 2.6 per cent higher for the cluster with 60 BHs. However, the effects on the dispersion profile are insignificant. As Fig. 13 shows, the clusters with 30 and 60 BHs have slightly lower dispersion, whereas the cluster with 100 BHs has a higher dispersion, suggesting that the effects from BHs are less significant than the systematic uncertainties on the dispersion profile. In the more extreme simulation with 150 BHs, we found that the cluster core is strongly heated, where the 5 per cent Lagrangian radius is about 0.22 pc after 4 Gyr of evolution. However, the escape rate is not significantly different to the other simulations. Hence, we conclude that BHs are not able to produce the observed dispersion.

We also estimated the effect of a large escape rate directly with a mock tidal tail model. We adopt a Lagrange Stripping technique (Fardal, Huang & Weinberg 2015; Küpper et al. 2015) to produce an oversampled tidal tail model. The progenitor was assumed to have a mass of  $1.5 \times 10^5 M_{\odot}$  and used the same initial conditions as the  $N$ -body model. The tails were evolved for 200 Myr, releasing particles every 0.005 Myr in the MWPotential2014 (Bovy 2015), which is nearly identical to the Irrgang potential in the inner MW and matches the  $N$ -body simulations quite well. The simulation was



**Figure 12.** The comparison of the surface density from the simulations with different numbers of BHs. The inner part of the surface density decreases with the number of BHs. However, the effects are insignificant as the GCs with 30 and 60 BHs, respectively still fit the data reasonably well.



**Figure 13.** The dispersion profiles of simulated clusters with varying numbers of stellar-mass BHs. Compared to a cluster without BHs, the effects of BHs on the dispersion profile are small. None of the simulated GCs is able to reproduce the large observed velocity dispersion in the outer cluster region.

performed using the dynamics package `gala` (Price-Whelan 2017). However, we found that with a significantly larger escape rate, the final dispersion is still about  $1.5 \text{ km s}^{-1}$ , which is roughly consistent with the  $N$ -body model, but still lower than the observations. In the tail model, we find 405 tail stars projected within  $r_{\text{Jacobi}}$ . Because the escape rate in this model is a factor of 10–30 too high, and only a small

fraction of the escaping stars are bright enough for our observations, we conclude that tail stars have a negligible contribution to the kinematics within  $r_{\text{Jacobi}}$ .

#### 4 CONCLUSIONS

In this paper, we described our GC survey with 2dF/AAOmega and the first results for the GC NGC 3201. Aiming at constraining their evolutionary history, we acquired spectra of stars in the outer part of five GCs (see Table 1). Those observations are designed to be divided into different blocks separated by at least one month so that we could detect short-period binaries.

We select stars near the turn-off, sub-giant branch and in the RGB phase as our targets, as those stars have three strong calcium absorption lines (CaT) at near-infrared wavelengths. Templates based on the three lines are fitted to the spectrum to extract information from our observations, from which we can determine the redshift and stellar parameters, and hence the LOS velocity and stellar properties like metallicity. The detailed study and comparison of the five GCs will be presented in future work.

As the first result of the survey, we discussed the dynamics of NGC 3201 from our observations. A dark matter free  $N$ -body simulation, that includes the effect from the MW potential, is built and compared with the observations. We confirm the LOS velocity gradient observed in the GC comes mainly from perspective rotation effects. In addition, we found a weak rotational signal in the inner part of the GC with amplitude of  $\sim 1 \text{ km s}^{-1}$ . The PA of this signal is different to the tidal tails and the potential escapers, which suggests that it comes from the internal rotation of the GC. Besides, we found a rotational signal at the outer part of the GC that has the same rotational direction compared to the tidal tails and potential escapers. However, within the field of view, the contribution from tidal tails are limited, suggesting that the stars at the outskirts are likely potential escapers.

We also discussed the dispersion profile of the GC. Compared to the simulation, the observed dispersion profile is lower beyond the King tidal radius. We discussed the potential source of this discrepancy. Effects due to the interaction of the cluster with the MW potential and potential escapers are included in the  $N$ -body model; hence, we conclude that both the MW potential and potential escapers cannot solve the discrepancy. With the simulations that include BHs, we found adding BHs in simulation can increase the escape rate of the cluster, but the change in dispersion is insignificant. Mock tidal tails produced with large escape rates also have a small dispersion compared to the data, consistent with the  $N$ -body model. In addition, we performed an analysis of the effect of binaries with multi-epoch observations, finding that they should be taken into account, but are unlikely to fully explain the difference in dispersion.

Since the dispersion relates to the dynamical mass, the presence of dark matter (e.g. Peebles 1984; Trenti et al. 2015) at larger radii would naturally lead to a flattened dispersion. B19 discuss the possibility that NGC 3201 is embedded in a dark matter halo. Although this can naturally explain an increased dispersion, there is still no direct evidence for the existence of dark matter in GCs. In addition, the signal of unbounded stars at the outer part of the GC, as well as the evidence of tidal streams associated with NGC 3201 (Chen & Chen 2010; Kunder et al. 2014; Anguiano et al. 2016; Ibata et al. 2020; Palau & Miralda-Escudé 2020), argue against the presence of dark matter. Hence the existence of the dark matter needs further confirmation. One could argue that we see the final phases of the stripping of the dark matter halo, e.g. if NGC 3201 was in the nuclear cluster of the dwarf galaxy, which could explain why the stars are also

affected by tides. However, NGC 3201 is association with the Gaia–Enceladus/Sequoia accretion (Massari et al. 2019; Myeong et al. 2019), which suggests that this cluster was accreted  $\gtrsim 9$  Gyr ago as part of a dwarf galaxy with multiple-star clusters, among which  $\omega$  Centauri. The high mass, and multiple metallicities of  $\omega$  Centauri make this cluster a much more plausible ‘former nuclear cluster’ compared to NGC 3201. We note that there are some additional effects that can potentially influence the dispersion. The orbital phase of this cluster is less constrained due to the uncertainties in proper motion, distance and the galactic potential, even though the position of the NGC 3201 on the sky is well known. Also, an initially rotating cluster in the simulation might result in a different dispersion profile. Finally, interactions with sub-structure in the MW – either baryonic or non-baryonic – may have heated the stars in the cluster and in the tails (e.g. Erkal, Koposov & Belokurov 2017). This discrepancy suggests that there is more we can learn from the dynamics of the outer part of the GC on its evolutionary history. Analyses of the dark matter content/distribution, as well as the unexplored effects on the velocity gradient, dispersion and tidal tails, will be presented in our future work.

#### ACKNOWLEDGEMENTS

ZW is supported by a Dean’s International Postgraduate Research Scholarship at the University of Sydney. WHO gratefully acknowledges financial support through the Hunstead Student Support Scholarship from the Dick Hunstead Fund in the University of Sydney’s School of Physics. MG, TdB, and EB acknowledge financial support from the European Research Council (ERC StG-335936, CLUSTERS). MG acknowledges support from the Ministry of Science and Innovation through a Europa Excelencia grant (EUR2020-112157). VHB acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant RGPIN-2020-05990. EB acknowledges financial support from a Vici grant from the Netherlands Organisation for Scientific Research (NWO). We thank Paolo Bianchini for providing the dispersion profile from his proper motion studies. Based in part on data acquired through the Australian Astronomical Observatory. We acknowledge the traditional owners of the land on which the AAT stands, the Gamilaraay people, and pay our respects to elders past, present and emerging.

Parts of this work were performed on the OzSTAR national facility at Swinburne University of Technology. The OzSTAR program receives funding in part from the National Collaborative Research Infrastructure Strategy (NCRIS) Astronomy allocation provided by the Australian Government.

#### DATA AVAILABILITY

The data underlying this paper may be made available on a reasonable request to the corresponding author.

#### REFERENCES

- AAO Software Team, 2015, *Astrophysics Source Code Library*, record ascl:1505.015
- Anguiano B. et al., 2016, *MNRAS*, 457, 2078
- Askar A., Arca Sedda M., Giersz M., 2018, *MNRAS*, 478, 1844
- Balbinot E., Gieles M., 2018, *MNRAS*, 474, 2479
- Baumgardt H., 2001, *MNRAS*, 325, 1323
- Baumgardt H., 2017, *MNRAS*, 464, 2174
- Baumgardt H., Hilker M., 2018, *MNRAS*, 478, 1520

- Baumgardt H., Makino J., 2003, *MNRAS*, 340, 227
- Baumgardt H., Mieske S., 2008, *MNRAS*, 391, 942
- Baumgardt H., Hilker M., Sollima A., Bellini A., 2019, *MNRAS*, 482, 5138
- Bellazzini M., Bragaglia A., Carretta E., Gratton R. G., Lucatello S., Catanzaro G., Leone F., 2012, *A&A*, 538, A18
- Bianchini P., van der Marel R. P., del Pino A., Watkins L. L., Bellini A., Fardal M. A., Libralato M., Sills A., 2018, *MNRAS*, 481, 2125
- Bianchini P., Ibata R., Famaey B., 2019, *ApJ*, 887, L12 (B19)
- Bovy J., 2015, *ApJS*, 216, 29
- Cai M. X., Gieles M., Heggie D. C., Varri A. L., 2016, *MNRAS*, 455, 596
- Carballo-Bello J. A., Gieles M., Sollima A., Koposov S., Martínez-Delgado D., Peñarrubia J., 2012, *MNRAS*, 419, 14
- Chen C. W., Chen W. P., 2010, *ApJ*, 721, 1790
- Chun S.-H. et al., 2010, *AJ*, 139, 606
- Claydon I., Gieles M., Zocchi A., 2017, *MNRAS*, 466, 3937
- Claydon I., Gieles M., Varri A. L., Heggie D. C., Zocchi A., 2019, *MNRAS*, 487, 147
- Conroy C., Loeb A., Spergel D. N., 2011, *ApJ*, 741, 72
- Cote P., Welch D. L., Fischer P., Gebhardt K., 1995, *ApJ*, 454, 788
- Cottaar M., Hénault-Brunet V., 2014, *A&A*, 562, A20
- Cottaar M., Meyer M. R., Parker R. J., 2012, *A&A*, 547, A35
- Da Costa G. S., 2012, *ApJ*, 751, 6
- Daniel K. J., Heggie D. C., Varri A. L., 2017, *MNRAS*, 468, 1453
- de Boer T. J. L., Gieles M., Balbinot E., Hénault-Brunet V., Sollima A., Watkins L. L., Claydon I., 2019, *MNRAS*, 485, 4906
- Edlén B., Risberg P., 1956, *Ark. Fys*, 10, 553
- Erkal D., Koposov S. E., Belokurov V., 2017, *MNRAS*, 470, 60
- Fardal M. A., Huang S., Weinberg M. D., 2015, *MNRAS*, 452, 301
- Ferraro F. R. et al., 2018, *ApJ*, 860, 50
- Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, *PASP*, 125, 306
- Fukushige T., Heggie D. C., 2000, *MNRAS*, 318, 753
- Gaia Collaboration et al., 2018a, *A&A*, 616, A1
- Gaia Collaboration et al., 2018b, *A&A*, 616, A10
- Gaia Collaboration et al., 2018c, *A&A*, 616, A12
- Gieles M., Zocchi A., 2015, *MNRAS*, 454, 576
- Gieles M., Balbinot E., Yaaqib R. I. S. M., Hénault-Brunet V., Zocchi A., Peuten M., Jonker P. G., 2018, *MNRAS*, 473, 4832
- Giersz M., Askar A., Wang L., Hypki A., Leveque A., Spurzem R., 2019, *MNRAS*, 487, 2412
- Giesers B. et al., 2018, *MNRAS*, 475, L15
- Giesers B. et al., 2019, *A&A*, 632, A3
- Hansen T. T., Riley A. H., Strigari L. E., Marshall J. L., Ferguson P. S., Zepeda J., Sneden C., 2020, *ApJ*, preprint ([arXiv:2007.12165](https://arxiv.org/abs/2007.12165))
- Harris W. E., 1996, *AJ*, 112, 1487
- Harris W. E., 2010, preprint ([arXiv:1012.3224](https://arxiv.org/abs/1012.3224))
- Helmi A. et al., 2018, *A&A*, 616, A12
- Hénault-Brunet V., Gieles M., Sollima A., Watkins L. L., Zocchi A., Claydon I., Pancino E., Baumgardt H., 2019, *MNRAS*, 483, 1400
- Henon M., 1970, *A&A*, 9, 24
- Ibata R. et al., 2020, preprint ([arXiv:2012.05245](https://arxiv.org/abs/2012.05245))
- Irrgang A., Wilcox B., Tucker E., Schiefelbein L., 2013, *A&A*, 549, A137
- Jordi K., Grebel E. K., 2010, *A&A*, 522, A71
- Kamann S. et al., 2018, *MNRAS*, 473, 5591
- Kimmig B., Seth A., Ivans I. I., Strader J., Caldwell N., Anderton T., Gregersen D., 2015, *AJ*, 149, 53
- King I., 1962, *AJ*, 67, 471
- King I. R., 1966, *AJ*, 71, 64
- Kremer K., Chatterjee S., Ye C. S., Rodriguez C. L., Rasio F. A., 2019, *ApJ*, 871, 38
- Kroupa P., 2001, *MNRAS*, 322, 231
- Kunder A. et al., 2014, *A&A*, 572, A30
- Küpper A. H. W., Kroupa P., Baumgardt H., Heggie D. C., 2010, *MNRAS*, 407, 2241
- Küpper A. H. W., Balbinot E., Bonaca A., Johnston K. V., Hogg D. W., Kroupa P., Santiago B. X., 2015, *ApJ*, 803, 80
- Kuzma P. B., Da Costa G. S., Mackey A. D., 2018, *MNRAS*, 473, 2881
- Lanzoni B. et al., 2018a, *ApJ*, 861, 16
- Lanzoni B. et al., 2018b, *ApJ*, 865, 11
- Mackey A. D., Wilkinson M. I., Davies M. B., Gilmore G. F., 2007, *MNRAS*, 379, L40
- Mackey A. D., Wilkinson M. I., Davies M. B., Gilmore G. F., 2008, *MNRAS*, 386, 65
- Marigo P. et al., 2017, *ApJ*, 835, 77
- Marín-Franch A. et al., 2009, *ApJ*, 694, 1498
- Marino A. F. et al., 2014, *MNRAS*, 442, 3044
- Mashchenko S., Sills A., 2005a, *ApJ*, 619, 243
- Mashchenko S., Sills A., 2005b, *ApJ*, 619, 258
- Massari D., Koppelman H. H., Helmi A., 2019, *A&A*, 630, L4
- Merritt D., Piatek S., Portegies Zwart S., Hensendorf M., 2004, *ApJ*, 608, L25
- Milone A. P. et al., 2016, *MNRAS*, 455, 3009
- Miszalski B., Shorridge K., Saunders W., Parker Q. A., Croom S. M., 2006, *MNRAS*, 371, 1537
- Moore B., 1996, *ApJ*, 461, L13
- Myeong G. C., Vasiliev E., Iorio G., Evans N. W., Belokurov V., 2019, *MNRAS*, 488, 1235
- Nitadori K., Aarseth S. J., 2012, *MNRAS*, 424, 545
- Odenkirchen M. et al., 2001, *ApJ*, 548, L165
- Palau C. G., Miralda-Escudé J., 2020, *MNRAS*, preprint ([arXiv:2010.14381](https://arxiv.org/abs/2010.14381))
- Peebles P. J. E., 1984, *ApJ*, 277, 470
- Peñarrubia J., Varri A. L., Breen P. G., Ferguson A. M. N., Sánchez-Janssen R., 2017, *MNRAS*, 471, L31
- Peuten M., Zocchi A., Gieles M., Gualandris A., Hénault-Brunet V., 2016, *MNRAS*, 462, 2333
- Price-Whelan A. M., 2017, *J Open Source Software*, 2, 388
- Raghavan D. et al., 2010, *ApJS*, 190, 1
- Scarpa R., Marconi G., Gilmozzi R., 2003, *A&A*, 405, L15
- Scarpa R., Marconi G., Gilmozzi R., Carraro G., 2007, *A&A*, 462, L9
- Simioni M. et al., 2018, *MNRAS*, 476, 271
- Sollima A., Martínez-Delgado D., Valls-Gabaud D., Peñarrubia J., 2011, *ApJ*, 726, 47
- Sollima A., Baumgardt H., Hilker M., 2019, *MNRAS*, 485, 1460
- Tiongco M. A., Vesperini E., Varri A. L., 2016, *MNRAS*, 461, 402
- Tiongco M. A., Vesperini E., Varri A. L., 2018, *MNRAS*, 475, L86
- Trager S. C., King I. R., Djorgovski S., 1995, *AJ*, 109, 218
- Trenti M., Padoan P., Jimenez R., 2015, *ApJ*, 808, L35
- van de Ven G., van den Bosch R. C. E., Verolme E. K., de Zeeuw P. T., 2006, *A&A*, 445, 513
- VandenBerg D. A., Brogaard K., Leaman R., Casagrande L., 2013, *ApJ*, 775, 134
- Vasiliev E., 2019a, *MNRAS*, 484, 2832
- Vasiliev E., 2019b, *MNRAS*, 489, 623 (V19)
- Wang L., 2020, *MNRAS*, 491, 2413
- Watkins L. L., van der Marel R. P., Bellini A., Anderson J., 2015, *ApJ*, 803, 29

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.

# Appendix B

## Supplementary Codes

Each of the papers reproduced in Chapters 3, 4, and 5 present a novel astrophysical clustering algorithm. In this appendix I provide the `PYTHON3` code for these algorithms. The line wrapping has been artificially altered within each of these codes in order to maintain the readability and indentations. As such, some reformatting will be needed in order to reproduce them as working algorithms.

### B.1 The Halo-OPTICS algorithm

Featured in paper 1, the `HALO-OPTICS` algorithm contains three classes; `Point`, `Cluster`, and `Halo_OPTICS`. While the code is far from the most succinct way to write this algorithm, the core of the algorithm – the process of ordering points and constructing the reachability plot – is highly efficient for a `PYTHON3` implementation.

Running this code first requires creating a list of `Point` classes containing the attributes of each point within the input data. The values of `eps` and `minpts` ( $N_{\min}$ ) must then also be chosen. The method for finding `eps` so that `FOF` haloes are produced is not provided, although this is discussed in detail within Sec. 3.1 of paper 1. Once these three inputs are constructed, the `HALO-OPTICS` algorithm may be run with default settings by calling the following lines:

```
1 hoptics = Halo_OPTICS(list_of_point_objects, eps, minpts)
2 hoptics.run()
```

Once finished, the reachability plot can be created by plotting the `rd` attribute of each `Point` against the index of each `Point` within the `hoptics.ordered` list. The clusters can also be obtained using the list of `Cluster` objects `hoptics.clusters`. The hierarchy can be understood by viewing the `id` attribute of each cluster i.e. '1', '2' etc. are root level clusters and '1-1', '1-2', '2-1', '2-1-1' etc. are their child (and grandchild) clusters.

```

1 import numpy as np
2 from sklearn.utils import gen_batches, get_chunk_n_rows
3 from sklearn.neighbors import NearestNeighbors
4 from sklearn.metrics import pairwise_distances
5 from scipy.signal import find_peaks
6
7 class Point:
8     """
9     Point objects given to Halo_OPTICS.
10
11     Parameters
12     -----
13     x, y, z: float
14         Cartesian coordinates of the 3D spatial positions of
15         simulation particles.
16
17     Attributes
18     -----
19     idx: int >= 0 or None
20         Index within the ordered list from Halo_OPTICS.
21     rd: float > 0 or None
22         This point's reachability distance.
23     processed: bool
24         A flag indicating the processing status of this point.
25     clusterid: str or None
26         The id of the cluster this point belongs to.
27     """
28
29     def __init__(self, x, y, z):
30         # Initialise object
31         self.x = x
32         self.y = y
33         self.z = z
34         self.idx = None
35         self.rd = None
36         self.processed = False
37         self.clusterid = None
38
39     def coords(self):
40         # Returns list of Cartesian coordinates
41         return [self.x, self.y, self.z]
42
43
44
45

```

```

46 class Cluster:
47     """
48     Cluster objects created by OPTICS.
49
50     Parameters
51     -----
52     id: str
53         Identification of Cluster object and its relation to
54         the cluster hierarchy, i.e. id = '1' indicates the
55         first root-level cluster and id = '1-1' indicates its
56         first child cluster.
57     points: list of Point objects
58         The points that belong to this cluster.
59
60     Attributes
61     -----
62     parentid: str or None
63         Id of the parent cluster of this cluster. Used to
64         construct the hierarchy of clusters.
65     minIdx: int > 0 or None
66         Lower bound of cluster in the ordered list.
67     maxIdx: int > 0 or None
68         Upper bound of cluster in the ordered list.
69     lone: bool or None
70         Flag indicating whether this cluster has no parent
71         or child clusters associated with it.
72     """
73
74     def __init__(self, id, points):
75         # Initialise object
76         self.id = id
77         self.points = points
78         if '-' not in self.id: self.parentid = None
79         else: self.parentid = '-'.join(self.id.split('-')[:-1])
80         self.minIdx = None
81         self.maxIdx = None
82         self.lone = None
83
84     def _index_bounds(self):
85         # Find and return index bounds
86         if self.maxIdx is None or self.minIdx is None:
87             i = [p.idx for p in self.points]
88             self.minIdx, self.maxIdx = min(i), max(i)
89         return [self.minIdx, self.maxIdx]
90

```



```

91 class Halo_OPTICS:
92     """
93     Astrophysical clustering algorithm based on the
94     'Ordering Points To Identify Clustering Structure' algorithm.
95
96     Parameters
97     -----
98     points: list of Point objects
99         The input data to be clustered.
100    eps: float > 0
101        Search radius and maximum reachability distance.
102    minpts: int > 1
103        Size of core neighbourhoods and the minimum number of
104        points in a cluster.
105    rho_threshold: float >= 1, default = 2
106        The minimum overdensity that clusters can have.
107    f_reject: float between 0 and 1, default = 0.9
108        The maximum fraction of points that a child cluster can
109        share with its parent.
110    s_outlier: float, default = 2
111        The maximum local-outlier-factor that points in
112        clusters can have.
113
114    Attributes
115    -----
116    n_samples: int > 0
117        Number of points in the input data.
118    processed: ndarray of shape (n_samples,) and dtype bool
119        Flags indicating which points have been processed.
120    ordering: ndarray of shape (n_samples,) and dtype int
121        Index of points in the ordered list.
122    reachability_: ndarray of shape (n_samples,) and dtype float
123        Reachability distances of points.
124    cds_: ndarray of shape (n_samples,) and dtype float
125        Core distances of points.
126    nbrs: NearestNeighbors object
127        Used to construct a kd-tree and find the nearest
128        neighbours of points in the input data.
129    clusters: list of Cluster objects
130        The predicted clusters.
131    ordered: list of Point objects
132        The ordered list of points.
133    """
134
135

```

```

136     def __init__(self, points, eps, minpts, rho_threshold = 2,
137                 f_reject = 0.9, s_outlier = 2):
138         # Initialise object
139         self.points = points
140         self.eps = eps
141         self.minpts = minpts
142         self.rho_threshold = rho_threshold
143         self.f_reject = f_reject
144         self.s_outlier = s_outlier
145         self.n_samples = len(self.points)
146         self.processed = np.zeros(self.n_samples, dtype = bool)
147         self.ordering = np.zeros(self.n_samples, dtype = int)
148         self.reachability_ = np.full(self.n_samples, np.inf)
149         self.cds_ = np.full(self.n_samples, np.nan)
150         self.nbrs = NearestNeighbors(n_neighbors = self.minpts)
151         self.clusters = []
152         self.ordered = []
153
154     def run(self, detectClusters = True):
155         # Run Halo-OPTICS.
156         self._compute_core_distances()
157         self._compute_ordered_list()
158         self._save_ordered_list()
159         if detectClusters: self.detect_clusters()
160
161     def _compute_core_distances(self):
162         # Find core distances.
163         self.X = np.array([p.coords() for p in self.points])
164         self.nbrs.fit(self.X)
165         chunks = get_chunk_n_rows(row_bytes = 16*self.minpts,
166                                   max_n_rows = self.n_samples)
167         for sl in gen_batches(self.n_samples, chunks):
168             d, i = self.nbrs.kneighbors(self.X[sl], self.minpts)
169             self.cds_[sl] = d[:, -1]
170         not_core = self.cds_ > self.eps
171         self.cds_[not_core] = np.inf
172
173     def _compute_ordered_list(self):
174         # Find ordered list.
175         for ordering_idx in range(self.n_samples):
176             idx = np.where(self.processed == 0)[0]
177             self.point = idx[np.argmin(self.reachability_[idx])]
178             self.processed[self.point] = True
179             self.ordering[ordering_idx] = self.point
180             if self.cds_[self.point] != np.inf: self._set_rd()

```

```

181
182     def _set_rd(self):
183         # Update reachability distance.
184         P = self.X[self.point:self.point + 1]
185         i = self.nbrs.radius_neighbors(P, radius = self.eps,
186                                     return_distance = False)[0]
187         unproc = np.compress(~np.take(self.processed, i), i)
188         if not unproc.size: return
189         unproc_X = np.take(self.X, unproc, axis = 0)
190         dists = pairwise_distances(P, unproc_X).ravel()
191         rdists = np.maximum(dists, self.cds_[self.point])
192         unproc_rd = np.take(self.reachability_, unproc)
193         improved = np.where(rdists < unproc_rd)
194         self.reachability_[unproc[improved]] = rdists[improved]
195
196     def _save_ordered_list(self):
197         # Recreate ordered list as a list of point objects.
198         if self.ordered: self.ordered = []
199         for i, idx in enumerate(self.ordering):
200             self.ordered.append(self.points[idx])
201             self.points[idx].idx = i
202             if self.reachability_[idx] <= self.eps:
203                 self.points[idx].rd = self.reachability_[idx]
204             else:
205                 self.points[idx].rd = None
206             self._progress(i, self.n_samples)
207
208     def detect_clusters(self):
209         # Finds clusters from reachability plot.
210         if self.clusters: self.clusters = []
211         rd = np.array([point.rd for point in self.ordered])
212         # Altrd is used to invoke find_peaks since this
213         # cannot compare Nonetype and Float.
214         altrd = np.array([d if d is not None else self.eps
215                          for d in rd])
216         maxima = [0] + list(find_peaks(altrd)[0])
217         bounds = []
218         totalMaxima = len(maxima)
219         for i, max in enumerate(maxima):
220             self._progress(i, totalMaxima)
221             if rd[max] is None:
222                 lower = next((max + j + 1 for j, v in
223                             enumerate(rd[max + 1:])
224                             if v is not None), None)
225             if lower is not None:

```



```

271     innerrd = np.median(altrd[lower: upper + 1])
272     if (outerrd/innerrd)**3 > self.rho_threshold: return True
273     else: return False
274
275 def _cluster(self, separators):
276     # Creates the clustering hierarchy of cluster objects.
277     separators.sort(key = lambda pair: [pair[0], -pair[1]])
278     ids = []
279     children = []
280     for i, sep in enumerate(separators):
281         children.append(0)
282         parent = next((-j - 1 for j, v in
283                       enumerate(separators[:i][::-1])
284                       if sep[0] < v[1] and sep[1] <= v[1]),
285                       None)
286         if parent is not None:
287             children[parent] += 1
288             ids.append(f"{ids[parent]}-{children[parent]}")
289         elif len(ids) > 0:
290             ids.append(str(max([int(id.split('-')[0])
291                               for id in ids]) + 1))
292         else:
293             ids.append('1')
294         self.clusters.append(Cluster(ids[i],
295                                     self.ordered[sep[0]:sep[1] + 1]))
296
297 def _single_child(self):
298     # Removes single child clusters.
299     ids = [clst.id for clst in self.clusters]
300     delete = [i + 1 for i, v in enumerate(self.clusters)
301              if i != len(self.clusters) - 1 and
302              v.id + '-1' in ids and v.id + '-2' not in ids]
303     self.clusters = [v for i, v in enumerate(self.clusters)
304                    if i not in delete]
305     self._new_ids()
306
307 def _new_ids(self):
308     # Updates clustering hierarchy.
309     self.clusters.sort(key = lambda c:
310                       list(np.multiply(c._index_bounds(),
311                                         [1, -1])))
312     separators = [cluster._index_bounds()
313                  for cluster in self.clusters]
314     ids = []
315     children = []

```

```

316     for i, sep in enumerate(separators):
317         children.append(0)
318         parent = next((-j - 1 for j, v in
319                       enumerate(separators[:i][::-1])
320                       if sep[0] < v[1] and sep[1] <= v[1]),
321                       None)
322         if parent is not None:
323             children[parent] += 1
324             ids.append(f"{ids[parent]}-{children[parent]}")
325         elif len(ids) > 0:
326             ids.append(str(max([int(id.split('-')[0])
327                               for id in ids] + 1))
328                       else: ids.append('1'))
329         self.clusters[i].id = ids[i]
330
331     def _particle_similarity(self):
332         # Assesses the similarity of potential clusters.
333         self.clusters.sort(key = lambda c:
334                             list(np.multiply(c._index_bounds(),
335                                                [1, -1])))
336         bounds = [clst._index_bounds() for clst in self.clusters]
337         condition1 = lambda j: next((True for v in
338                                     self.clusters[j + 1:] if
339                                     v.id.startswith(self.clusters[j].id + '-')),
340                                     False)
341         condition2 = lambda i: '-' in self.clusters[i].id
342         delete = []
343         for i, b1 in enumerate(bounds):
344             for j, b2 in enumerate(bounds[i + 1:]):
345                 if b2[0] > b1[1]: break
346                 f_shared = (b2[1] - b2[0] + 1)/(b1[1] - b1[0] + 1)
347                 if f_shared > self.f_reject:
348                     if condition1(j + i + 1):
349                         delete.append(j + i + 1)
350                     elif condition2(i):
351                         delete.append(i)
352                     break
353         self.clusters = [v for i, v in enumerate(self.clusters)
354                          if i not in delete]
355         self._new_ids()
356
357     def _outlier_factors(self):
358         # Removes local outliers from clusters.
359         def _get_lofs(points):
360             X = np.array([p.coords() for p in points])

```

```

361         c_nbrs = NearestNeighbors(n_neighbors = self.minpts)
362         c_nbrs.fit(X)
363         d, i = c_nbrs.kneighbors(X, self.minpts)
364         cds = d[i, self.minpts - 1]
365         rds = np.maximum(d, cds)
366         lrds = 1/np.mean(rds, axis = 1)
367         return np.mean(lrds[i]/lrds[:, np.newaxis], axis = 1)
368
369     numClusters = len(self.clusters)
370     removed = 0
371     clusterdelete = []
372     for index, cluster in enumerate(self.clusters[::-1]):
373         self._progress(index, numClusters)
374         cluster.lone = ('-' not in cluster.id and
375                        next((False for v in
376                             self.clusters[numClusters - index:]
377                             if v.id.startswith(cluster.id + '-') and
378                             len(v.points) >= self.minpts), True))
379         if '-' in cluster.id or cluster.lone:
380             lofs = _get_lofs(cluster.points)
381             pointsdelete = np.compress(lofs > self.s_outlier,
382                                     np.array(range(len(cluster.points))))
383             cluster.points = [point for i, point in
384                              enumerate(cluster.points)
385                              if i not in pointsdelete]
386             removed += pointsdelete.shape[0]
387             if len(cluster.points) < self.minpts:
388                 clusterdelete.append(numClusters - index)
389     self.clusters = [cluster for i, cluster in
390                     enumerate(self.clusters)
391                     if i not in clusterdelete]
392     self._new_ids()
393
394     def _bounds_exceeded(self):
395         # Checks that parent clusters completely envelope
396         # their child clusters.
397         self.clusters.sort(key = lambda c:
398                           list(np.multiply(c._index_bounds(),
399                                           [1, -1])))
400         bounds = [clst._index_bounds() for clst in self.clusters]
401         delete = []
402         for i, b1 in enumerate(bounds):
403             for j, b2 in enumerate(bounds[i + 1:]):
404                 if b2[0] > b1[1]: break
405                 if b2[1] > b1[1]:

```

```

406         if b1[1] - b1[0] > b2[1] - b2[0]:
407             delete.append(i)
408         else:
409             delete.append(j + i + 1)
410     self.clusters = [v for i, v in enumerate(self.clusters)
411                     if i not in delete]
412     self._new_ids()
413
414     def _weakly_enclosed(self):
415         # Checks that a cluster does not contain a point
416         # that is less dense than its surrounds.
417         delete = []
418         for i, v in enumerate(self.clusters):
419             if '-' in v.id:
420                 bounds = v._index_bounds()
421                 maxInternal = max([p.rd for p in
422                                 self.ordered[bounds[0]:bounds[1] + 1]])
423                 maxBoundary = max([self.ordered[bounds[0]].rd,
424                                   self.ordered[bounds[1]].rd])
425                 if maxInternal > maxBoundary:
426                     hasChild = next((True for clst in
427                                     self.clusters[i + 1:] if
428                                     clst.id.startswith(v.id + '-')),
429                                     False)
430                 if hasChild:
431                     delete.append(i)
432                 else:
433                     maxima = np.argmax([p.rd for p in
434                                         self.ordered[bounds[0]:bounds[1] + 1]])
435                     if maxima < self.minpts:
436                         v.points = [p for p in v.points
437                                     if p.idx > bounds[0] + maxima]
438                     else:
439                         v.points = [p for p in v.points
440                                     if p.idx < bounds[0] + maxima]
441     self.clusters = [v for i, v in enumerate(self.clusters)
442                     if i not in delete]
443     self._new_ids()
444
445     def _rename_cluster_point_ids(self):
446         # Updates clusterid attributes within Point objects.
447         for cluster in self.clusters:
448             for point in cluster.points:
449                 point.clusterid = cluster.id

```



## B.2 The CluSTAR-ND algorithm

The CLUSTAR-ND algorithm is the focus of paper 2. In order to run this algorithm, a `np.ndarray` of floating point values (`P`) with shape `(n, d)` must be created to represent the  $n$  points of the input data and each of their  $d$  features. A choice also needs to be made about whether the user wishes to find 3D FOF field haloes within the data or not – if so the value of the linking length (`l_x`) must be chosen accordingly. The values of `k_den` and `adaptive` should also be considered, however the default values of 20 and 1 were shown to provide consistent and robust results regardless of the size or dimensionality of the input data. Running the algorithm with its default settings can be achieved by calling the following lines:

```
1 cstar = CluSTAR_ND(P)
2 cstar.run()
```

After the algorithm has finished running, the clusters can be created using the arrays `cstar.clusters` and `cstar.ids`. The `cstar.clusters` array indicates for each point the index of the child-most cluster id within the `cstar.ids` array that it belongs to. Using these arrays, the catalogue of the complete clusters can be constructed by calling the following lines:

```
1 # Find the descendants of each cluster.
2 ids_rs = cstar.ids.reshape(1, -1)
3 descArr = np.char.startswith(ids_rs, np.char.add(ids_rs.T, '-'))
4 # Add the cluster itself to the record of each of its own
5 # descendants.
6 rng = np.arange(cstar.ids.size)
7 descArr[rng, rng] = True
8 # Create the list of arrays containing the index of each point
9 # (as it appears in cstar.P) belonging to each cluster.
10 clst_catalogue = []
11 for idx in rng:
12     which_idx = np.where(descArr[idx])[0]
13     clst = np.where(np.isin(cstar.clusters, which_idx))[0]
14     clst_catalogue.append(clst)
```

Now `clst_catalogue`, although a list of arrays and not a list of `Cluster` objects, provides the same format of the returned clustering as with HALO-OPTICS. The `cstar.ids` also provides the same format for the cluster ids (`'1'`, `'1-1'`, `'1-1-1'`, `'1-2'`, etc.) as can be retrieved from each `Cluster` object through the `id` attribute when using HALO-OPTICS.

```

1 import numpy as np
2 import pyfof
3 from sklearn.decomposition import PCA
4 from scipy.spatial import cKDTree
5 from sklearn.utils import gen_batches, get_chunk_n_rows
6 from scipy.spatial.distance import cdist
7 from numba import njit
8
9 class CluSTAR_ND:
10     """
11     The CluSTAR-ND algorithm.
12
13     Parameters
14     -----
15     P: ndarray of shape (n_samples, features) and dtype float
16         The input data to be clustered over.
17     l_x: float > 0, default = np.inf
18         The linking length used to find 3D-FOF field haloes.
19     k_den: int >= 7, default = 20
20         The number of nearest neighbours used to estimate
21         the local density at each point in P.
22     adaptive: int in [0, 1, 2], default = 1
23         The setting controlling the adaptivity of the metric.
24         0 specifies no transformation, 1 specifies a single
25         global PCA transformation, 2 specifies an iterative
26         PCA transformation.
27     k_link: int >= 7 (for reliable behaviour), default = 'auto'
28         The number of nearest neighbours used to densely
29         connect the points in P.
30     rho_threshold: float >= 1, default = 'auto'
31         The minimum overdensity that clusters can have.
32     f_reject: float between 0 and 1, default = 0.85
33         The maximum fraction of points that a child cluster can
34         share with its parent.
35     s_outlier: float, default = 2.5
36         Used to define the cut-off density for each cluster
37         such that all points within any given cluster with a
38         density less than the cut-off density of that cluster
39         are removed from that cluster. The cut-off density is
40         the minimum density of points that are not local
41         outliers. A local outlier is any point with a
42         local-outlier-factor greater than s_outlier. Decreasing
43         s_outlier has the effect of shedding the outer layers
44         off the cluster.
45     workers: int <= number of cpus, default = -1 (uses all cpus)

```

```

46         The number of core processing units used to perform
47         some parallelised calculations.
48
49     Attributes
50     -----
51     n_samples: int > 0
52         The number of points in P, i.e. n_samples = P.shape[0].
53     features: int > 0
54         The number of features/dimensions of P,
55         i.e. features = P.shape[1].
56     transform: PCA object
57         The PCA transformation object. Initialising this now
58         saves computation time that would otherwise be incurred
59         from repeated initialising.
60     origin_for_cdist: ndarray of shape (1, k_den) and dtype float
61         Allows the use of cdist to compute the sum of squared
62         distances instead of numpy (which is slower).
63     clusters: ndarray of shape (n_samples,) and dtype int
64         The index of each point indicating the smallest cluster
65         it belongs to. A index of -1 implies that the point is
66         not clustered and is instead treated as noise.
67     ids: ndarray of shape (np.unique(clusters).size,) and
68         dtype <UX
69         Contains the cluster identification string and its
70         relation to the cluster hierarchy, i.e. id = '1'
71         indicates the first root-level cluster and id = '1-1'
72         indicates its first child cluster.
73     """
74
75     def __init__(self, P, l_x = np.inf, k_den = 20, adaptive = 1,
76                 k_link = 'auto', rho_threshold = 'auto',
77                 f_reject = 0.85, s_outlier = 2.5, workers = -1):
78         # Initialise
79         self.P = P
80         self.n_samples, self.features = self.P.shape
81         self.l_x = l_x
82         self.k_den = k_den
83         self.adaptive = adaptive
84         if k_link == 'auto':
85             contin_klink = 11.97*self.features**(-2.23)\
86                 - 22.97*self.k_den**(-0.57)\
87                 + 10.03
88             self.k_link = max(int(np.ceil(contin_klink)), 7)
89         else: self.k_link = k_link
90         if rho_threshold == 'auto':

```

```

91         nmrtr = 0.8076*(self.features**0.8099)
92         self.rho_threshold = nmrtr/np.log(self.k_den) + 1
93     else: self.rho_threshold = rho_threshold
94     self.f_reject = f_reject
95     self.s_outlier = s_outlier
96     self.workers = workers
97     self.transform = PCA(whiten = True)
98     self.origin_for_cdist = np.zeros((1, self.k_den))
99
100     def run(self):
101         # Find 3D-F0F root-level haloes.
102         if np.isfinite(self.l_x):
103             groups = pyfof.friends_of_friends(self.P[:, :3],
104                                               self.l_x)
105         else: groups = [range(self.n_samples)]
106         # Find substructure.
107         self.clusters = -np.ones(self.n_samples, dtype = np.int64)
108         self.ids = []
109         for i, g in enumerate(groups):
110             clstIdx = np.array(g)
111             clusters[clstIdx] = len(self.ids)
112             nextID = f"{i}"
113             self.ids.append(nextID)
114             self._find_substructure(clstIdx, parentID = nextID)
115         self.ids = np.array(self.ids)
116
117     def _find_substructure(self, node, parentID):
118         if node.size >= self.k_den:
119             # Transform the data.
120             if self.adaptive:
121                 transNode = np.ascontiguousarray(
122                     self.transform.fit_transform(self.P[node]))
123             else: transNode = self.P[node]
124
125             # Compute density and nearest neighbours.
126             indices = np.empty((node.size, self.k_den),
127                               dtype = np.int64)
128             density = np.empty(node.size)
129             nbrs = cKDTree(transNode)
130             chunks = get_chunk_n_rows(row_bytes = 16*self.k_den,
131                                       max_n_rows = node.size)
132             for sl in gen_batches(node.size, chunks):
133                 dist, indices[sl] = nbrs.query(transNode[sl],
134                                               k = self.k_den, workers = self.workers)
135             coreDist = dist[:, -1]

```

```

136         kernal = self.k_den - cdist(self.origin_for_cdist,
137                                     dist, 'sqeuclidean').ravel()/coreDist**2
138         density[sl] = kernal/coreDist**(self.features)
139
140         # Densely connect points separated by saddle points
141         # in the density field and extract relevant clusters.
142         ind = indices[:, :self.k_link]
143         clusters, ids = self._densely_aggregate(density, ind,
144                                               density[ind].argmax(axis = 1) == 0, indices,
145                                               self.rho_threshold, self.f_reject,
146                                               self.s_outlier, self.adaptive, self.features)
147
148         # Save clusters to list and if adaptive == 2 then
149         # search within them.
150         if clusters:
151             sortedClstID = sorted(zip(clusters, ids),
152                                   key = lambda c_id: c_id[1])
153             for clst, id in sortedClstID:
154                 nextCluster = node[clst]
155                 self.clusters[nextCluster] = len(self.ids)
156                 childNum = '-'.join([str(i) for i in id])
157                 nextID = f"{parentID}-{childNum}"
158                 self.ids.append(nextID)
159                 if self.adaptive == 2:
160                     self._find_substructure(nextCluster,
161                                             nextID)
162
163     @staticmethod
164     @njit(fastmath = True)
165     def _densely_aggregate(density, indices, localMaxima,
166                           indices_full, rho_threshold, f_reject,
167                           s_outlier, adaptive, features):
168         # Preparation.
169         procOrder = density.argsort()[::-1]
170         procOrder = procOrder[np.invert(localMaxima[procOrder])]
171         n_samples, k_link = indices.shape
172         localMaxima = np.where(localMaxima)[0]
173         densityConnect = [[locMax for locMax in localMaxima]
174                           for i in range(n_samples)]
175         whichDC = np.empty(n_samples, dtype = np.int64)
176         whichDC[localMaxima] = np.arange(localMaxima.size)
177         sizesDC = [1]*localMaxima.size
178         emptyIntList = [0 for i in range(0)]
179         children = [emptyIntList[:] for i in
180                    range(localMaxima.size)]
181         emptyIntArr = np.empty(0, dtype = np.int64)

```



```

226         sigCount += 1
227         whichDC[clst] = bigCnct
228     else:
229         for j in extDC:
230             whichDC[j] = bigCnct
231     # Check largest group last to save time.
232     if sigCount > 0:
233         # Has child clusters.
234         if children[bigCnct]:
235             cLst = densityConnect[bigCnct][:bigSizeDC]
236             cArr = np.array(cLst)
237             newClsts.append(cArr)
238             significant.append(bigCnct)
239             sigCount += 1
240         elif bigSizeDC >= k_link:
241             # Groups must be larger than k_link.
242             cLst = densityConnect[bigCnct][:bigSizeDC]
243             cArr = np.array(cLst)
244             # Density ratio condition.
245             medThr = rho_threshold*currDensity
246             if np.median(density[clst]) > medThr:
247                 newClsts.append(clst)
248                 significant.append(bigCnct)
249                 sigCount += 1
250     # If more than one significant group exists
251     # create new clusters.
252     if sigCount > 1:
253         clusters.extend(newClsts)
254         for i, cnct in enumerate(significant):
255             for child in children[cnct]:
256                 parents[child] = lenClsts + i
257                 children[cnct] = [lenClsts + i]
258                 parents.append(-1)
259         lenClsts += sigCount
260         densityConnect[bigCnct].append(idx)
261         whichDC[idx] = bigCnct
262         sizesDC[bigCnct] += addSize
263         for cnct in connections:
264             if cnct != bigCnct:
265                 children[bigCnct].extend(children[cnct])
266     connectionsArr = np.unique(whichDC)
267     # If points were not all aggregated together check
268     # if remaining groups are significant
269     if connectionsArr.size > 1:
270         significantList = [sizesDC[cnct] >= k_link

```

```

271         for cnct in connectionsArr]
272     if significantList.count(True) > 1:
273         for i, cnct in enumerate(connectionsArr):
274             if significantList[i]:
275                 cArr = np.array(densityConnect[cnct])
276                 clusters.append(cArr)
277                 for child in children[cnct]:
278                     parents[child] = lenClsts
279                 parents.append(-1)
280                 lenClsts += 1
281
282     # Check similarity between parent-child pairs
283     keep = [True]*lenClsts
284     remove = 0
285     for i, (clst, parent) in enumerate(zip(clusters, parents)):
286         if parent != -1:
287             similarity = clst.size/clusters[parent].size
288         else: similarity = clst.size/n_samples
289         if similarity > f_reject:
290             if parent == -1 or i in parents[:i]:
291                 if keep[i]:
292                     keep[i] = False
293                     remove += 1
294             else:
295                 keep[parent] = False
296                 remove += 1
297
298     if lenClsts - remove > 1:
299         # Remove outliers
300         if s_outlier < np.inf:
301             invDensity = density**(-1/features)
302             lrds = np.empty_like(density)
303             for i in range(n_samples):
304                 kdenInvDensity = invDensity[indices_full[i]]
305                 lrds[i] = np.maximum(kdenInvDensity,
306                                     invDensity[i]).sum()
307             lrds = 1/lrds
308             lofs = np.empty_like(density)
309             for i in range(n_samples):
310                 lofs[i] = lrds[indices_full[i]].mean()
311             lofs = lofs/lrds
312             inliers = lofs < s_outlier
313             for i, clst in enumerate(clusters):
314                 if keep[i]:
315                     cutOff = inliers[clst]

```



```

316         if cutOff.sum() >= k_link:
317             clstDensity = density[clst]
318             cutoffDnsty = clstDnsty[cutOff].min()
319             inlierBool = clstDnsty >= cutoffDnsty
320             if inlierBool.sum() < k_link:
321                 keep[i] = False
322                 remove += 1
323             else: clusters[i] = clst[inlierBool]
324         else:
325             keep[i] = False
326             remove += 1
327     if lenClsts - remove > 1:
328         # Assign correct parents
329         for i, parent in enumerate(parents):
330             if keep[i]:
331                 while not (keep[parent] or parent == -1):
332                     parent = parents[parent]
333                 parents[i] = parent
334         # Assign correct ids
335         ids = [emptIntList]*lenClsts
336         activeParents = [-1]
337         for pIdx in activeParents:
338             j = 1
339             for child, p_iter in enumerate(parents):
340                 if p_iter == pIdx and keep[child]:
341                     ids[child] = ids[pIdx] + [j]
342                     j += 1
343                 activeParents.append(child)
344         # Remove insignificant clusters
345         finalClsts = []
346         finalIDs = []
347         for i, (clst, id) in enumerate(zip(clusters, ids)):
348             if keep[i]:
349                 if adaptive == 2:
350                     if len(id) == 1:
351                         finalClsts.append(clst)
352                         finalIDs.append(id)
353                 else:
354                     finalClsts.append(clst)
355                     finalIDs.append(id)
356         return finalClsts, finalIDs
357     finalClsts = [emptIntArr for i in range(0)]
358     finalIDs = [[1] for i in range(0)]
359     return finalClsts, finalIDs

```

## B.3 The CluSTARR-ND algorithm

As described in paper 3, this section presents the CLUSTARR-ND algorithm. Similarly to CLUSTAR-ND, to run the algorithm the user first needs to construct `P`, the `np.ndarray` of shape  $(n, d)$  that represents the  $n$  points and each of their  $d$  features that exist within the data-to-be-clustered. The functionality of producing 3D FOF field haloes is not yet written into this code – however, this will appear at a later date. The values of `k_den` and `adaptive` should again also be considered, however as with CLUSTAR-ND, the default values of 20 and 1 work well in nearly all cases.

Unlike CLUSTAR-ND and owing to the distinct cluster extraction method within CLUSTARR-ND, the values of `S` and `h_style` should be considered. Strictly speaking, there is no one *best* value for either of these parameters – although their default values will give good quality clusterings but may classify groups too many or too few clusters for the user’s needs. The algorithm can be applied to the input data with its default settings by running the following lines:

```
1 cstarr = CluSTARR_ND(P)
2 cstarr.run()
```

Then the CLUSTARR-ND equivalent of the OPTICS reachability plot, the ordered-density plot, can be constructed by plotting the following `y` vs `x`:

```
1 y = cstarr.logRho[cstarr.ordering]
2 x = np.arange(cstarr.n_samples)
```

Each cluster that has been found can be retrieved by running the following:

```
1 clst_catalogue = []
2 for start, end in cstarr.clusters:
3     clst = cstarr.ordering[start:end]
4     clst_catalogue.append(clst)
```

In a similar format to both HALO-OPTICS and CLUSTAR-ND, `clst_catalogue` is now a list of arrays that each represent the indices of the points (with respect to `cstarr.P`) that belong to each cluster.

```

1 import numpy as np
2 from scipy.spatial import cKDTree
3 from scipy.spatial.distance import cdist
4 from scipy.stats import norm, beta
5 from scipy.optimize import minimize
6 from sklearn.decomposition import PCA
7 from sklearn.utils import gen_batches, get_chunk_n_rows
8 from numba import njit, prange
9
10 class CluSTARR_ND:
11     """
12     The CluSTARR-ND algorithm.
13
14     Parameters
15     -----
16     P: ndarray of shape (n_samples, features) and dtype float
17         The input data to be clustered over.
18     l_x: float > 0, default = np.inf
19         The linking length used to find 3D-FOF field haloes.
20     k_den: int >= 7, default = 20
21         The number of nearest neighbours used to estimate
22         the local density at each point in P.
23     adaptive: int in [0, 1], default = 1
24         The setting controlling the adaptivity of the metric.
25         0 specifies no transformation and 1 specifies a single
26         global PCA transformation.
27     S: float, default = 5
28         The statistical significance that clusters must have
29         when compared to noisy density fluctuations.
30     k_link: int >= 7 (for reliable behaviour), default = 'auto'
31         The number of nearest neighbours used to densely
32         connect the points in P.
33     h_style: int = 0 or 1
34         A flag indicating the style of hierarchy that is returned.
35         A value of 1 is closer to the Halo-OPTICS and CluSTAR-ND
36         style of hierarchy.
37     workers: int <= num_cpus, default = -1
38         The number of core processing units (cpus) used to perform
39         some parallelised calculations. A value of -1 uses all
40         available cpus.
41
42     Attributes
43     -----
44     n_samples: int > 0
45         The number of points in P, i.e. n_samples = P.shape[0].

```

```

46     features: int > 0
47         The number of features/dimensions of P,
48         i.e. features = P.shape[1].
49     transform: PCA object
50         The PCA transformation object. Initialising this now
51         saves computation time that would otherwise be incurred
52         from repeated initialising.
53     origin_for_cdists: ndarray of shape (1, k_den) and dtype float
54         Allows the use of cdists to compute the sum of squared
55         distances instead of numpy (which is slower).
56     ordering: ndarray of shape (n_samples,) and dtype float
57         The ordered list that can be used to create the
58         ordered-density plot (an analogue of the reachability
59         plot that OPTICS produces).
60     logRho: ndarray of shape (n_samples,) and dtype float
61         The logarithm of the local density (scaled between 0 and
62         1) of each point in P.
63     groups: ndarray of shape (num_groups, 2) and dtype int
64         The start and end positions of each aggregated group
65         within the ordered list.
66     prominences: ndarray of shape (num_groups,) and dtype float
67         The prominences of each aggregated group.
68     significances: ndarray of shape (num_clusters,) and
69     dtype float
70         The statistical significances of each returned cluster.
71     clusters: ndarray of shape (n_samples,) and dtype int
72         The index of each point indicating the smallest cluster
73         it belongs to. A index of -1 implies that the point is
74         not clustered and is instead treated as noise.
75     ids: ndarray of shape (num_clusters,) and dtype <UX (X is
76     twice the max level of the hierarchy + 1)
77         Contains the cluster identification string and its
78         relation to the cluster hierarchy, i.e. id = '1'
79         indicates the first root-level cluster and id = '1-1'
80         indicates its first child cluster.
81     """
82
83     def __init__(self, P, k_den = 20, adaptive = 1, S = 5,
84                 k_link = 'auto', h_style = 1, workers = -1):
85         # Initialise.
86         self.P = P
87         self.n_samples, self.features = self.P.shape
88         self.k_den = k_den
89         self.adaptive = adaptive
90         self.S = S

```

```

91     if k_link == 'auto':
92         contin_klink = 11.97*self.features**(-2.23)\
93             - 22.97*self.k_den**(-0.57)\
94             + 10.03
95         self.k_link = max(int(np.ceil(contin_klink)), 7)
96     else: self.k_link = k_link
97     self.h_style = h_style
98     self.workers = workers
99     self.transform = PCA(whiten = True)
100    self.origin_for_cdists = np.zeros((1, self.k_den))
101
102    def run(self):
103        self._find_substructure(np.arange(self.n_samples), '1')
104
105    def _find_substructure(self, node, parentID):
106        # Find the substructure of each field halo.
107        node = self._transform_data(node)
108        indices, self.logRho = self._find_rho_and_kNN(node)
109        del node
110        # Order points and find groups.
111        localMaxima = self.logRho[indices].argmax(axis = 1) == 0
112        self.ordering, self.groups, self.prominences,
113            childCheck, self.blrs = self._aggregate(self.logRho,
114            indices, localMaxima)
115
116        self.prominences = self._adjust_prominences(
117            self.logRho[self.ordering], self.groups,
118            self.prominences, childCheck)
119        self._find_significances()
120        self.find_clusters(parentID)
121
122    def _transform_data(self, node):
123        # Transform the data.
124        if self.adaptive:
125            return np.ascontiguousarray(
126                self.transform.fit_transform(self.P[node]))
127        else: return self.P[node]
128
129    def _find_rho_and_kNN(self, transNode):
130        # Compute density and nearest neighbours.
131        d = np.empty((transNode.shape[0], self.k_den))
132        i = np.empty((transNode.shape[0], self.k_den),
133                    dtype = np.int64)
134        nbrs = cKDTree(transNode)
135        chunks = get_chunk_n_rows(row_bytes = 16*self.k_den,

```

```

136         max_n_rows = transNode.shape[0])
137     for sl in gen_batches(transNode.shape[0], chunks):
138         d[sl], i[sl] = nbrs.query(transNode[sl],
139                                 k = self.k_den, workers = self.workers)
140     coreDist = dist[:, -1]
141     uSqr = cdist(self.origin_for_cdist, d,
142                 'sqeuclidean').ravel()
143     kernal = self.k_den - uSqr/coreDist**2
144     logRho = np.log(kernal) - self.features*np.log(coreDist)
145     minVal, maxVal = self._find_minmax(logRho)
146     scaledLogRho = (logRho - minVal)/(maxVal - minVal)
147     return i[:, :self.k_link], scaledLogRho
148
149     @staticmethod
150     @njit(fastmath = True)
151     def _find_minmax(x):
152         # Find min and max of x quickly.
153         minVal = x[0]
154         maxVal = x[0]
155         for xi in x[1:]:
156             if xi < minVal: minVal = xi
157             elif xi > maxVal: maxVal = xi
158         return minVal, maxVal
159
160     @staticmethod
161     @njit(fastmath = True)
162     def _aggregate(logRho, indices, localMaxima):
163         # Preparation.
164         procOrder = logRho.argsort()[::-1]
165         procOrder = procOrder[np.invert(localMaxima[procOrder])]
166         n_samples, k_link = indices.shape
167         localMaxima = np.where(localMaxima)[0]
168         aggregations = [[locMax] for locMax in localMaxima]
169         whichAgg = np.empty(n_samples, dtype = np.int64)
170         whichAgg[localMaxima] = np.arange(localMaxima.size)
171         sizesAgg = np.ones(localMaxima.size, dtype = np.int64)
172         prominences = logRho[localMaxima]
173         blrs = np.zeros(localMaxima.size)
174         groupStarts = np.zeros(localMaxima.size, dtype = np.int64)
175         emptIntList = [0 for i in range(0)]
176         children = [[0 for i in range(0)] for i in
177                    range(localMaxima.size)]
178
179         # Connect densely into significant hierarchy.
180         zippedIdx = zip(procOrder, indices[procOrder],

```

```

181         logRho[procOrder])
182     for idx, currNN, currLogRho in zippedIdx:
183         unproc = logRho[currNN] > currLogRho
184         connections = set(whichAgg[currNN[unproc]])
185         # Connect to existing group iff one connection exists.
186         if len(connections) == 1:
187             mainConnect = connections.pop()
188             addSize = 1
189         else: # Otherwise join multiple groups and append idx.
190             addSize = 0
191             sortedConnects = sorted(zip([sizesAgg[connect]
192                                     for connect in connections],
193                                     connections))
194             mainSizeAgg, mainConnect = sortedConnects[-1]
195             for sizeAgg, connect in sortedConnects[-2::-1]:
196                 extendAgg = aggregations[connect]
197                 aggregations[mainConnect].extend(extendAgg)
198                 aggregations[connect] = emptyIntList
199                 for jdx in extendAgg:
200                     whichAgg[jdx] = mainConnect
201                 prominences[mainConnect] = np.maximum(
202                     prominences[mainConnect],
203                     prominences[connect])
204                 if sizeAgg >= k_link:
205                     prominences[connect] -= currLogRho
206                     blrs[connect] = currLogRho
207                     groupStarts[connect] = mainSizeAgg + \
208                                             addSize
209                     children[mainConnect].append(connect)
210                 addSize += sizeAgg
211             addSize += 1
212             aggregations[mainConnect].append(idx)
213             whichAgg[idx] = mainConnect
214             sizesAgg[mainConnect] += addSize
215
216     connectionsArr = np.unique(whichAgg)
217     # If not all points were aggregated together, make it so.
218     if connectionsArr.size == 1:
219         mainConnect = connectionsArr[0]
220     else:
221         addSize = 0
222         sortedConnects = sorted(zip([sizesAgg[connect]
223                                     for connect in connectionsArr],
224                                     connectionsArr))
225         mainSizeAgg, mainConnect = sortedConnects[-1]

```

```

226         for sizeAgg, connect in sortedConnects[-2::-1]:
227             extendAgg = aggregations[connect]
228             aggregations[connect] = emptyIntList
229             aggregations[mainConnect].extend(extendAgg)
230             prominences[mainConnect] = np.maximum(
231                                     prominences[mainConnect],
232                                     prominences[connect])
233             if sizeAgg >= k_link:
234                 prominences[connect] -= currLogRho
235                 blrs[connect] = currLogRho
236                 groupStarts[connect] = mainSizeAgg + addSize
237                 children[mainConnect].append(connect)
238             addSize += sizeAgg
239
240     # Finalise ordering.
241     ordering = np.array(aggregations[mainConnect])
242
243     # Finalise groups.
244     childCheck = np.zeros(localMaxima.size, dtype = np.bool_)
245     activeGroups = [childConnect for childConnect in
246                    children[mainConnect]]
247     while activeGroups:
248         connect = activeGroups.pop()
249         startAdjust = groupStarts[connect]
250         childConnects = children[connect]
251         if childConnects:
252             for childConnect in childConnects:
253                 groupStarts[childConnect] += startAdjust
254             activeGroups.extend(childConnects)
255             childCheck[connect] = True
256     keepCheck = sizesAgg >= k_link
257     keepCheck[mainConnect] = False
258     groups = np.concatenate((groupStarts.reshape(-1, 1),
259                             sizesAgg.reshape(-1, 1)),
260                             axis = 1)[keepCheck]
261     groups[:, 1] += groups[:, 0]
262     prominences = prominences[keepCheck]
263     childCheck = childCheck[keepCheck]
264     blrs = blrs[keepCheck]
265     reorder = groups[:, 0].argsort()
266
267     # Reorder according to ordered list.
268     groups = groups[reorder]
269     prominences = prominences[reorder]
270     childCheck = childCheck[reorder]

```



```

271     blrs = blrs[reorder]
272     return ordering, groups, prominences, childCheck, blrs
273
274 @staticmethod
275 @njit(fastmath = True, parallel = True)
276 def _adjust_prominences(lrOrdIdx, grps, proms, chldChk):
277     # Adjust group prominences for noise.
278     for i in prange(grps.shape[0]):
279         if chldChk[i]:
280             start, end = grps[i]
281             grpLR = lrOrdIdx[start:end]
282             ord = grpLR.argsort()[::-1]
283             idx_diff_1 = ord - np.arange(ord.size)
284             w = np.abs(idx_diff_1).astype(np.float64)
285             if np.any(w != 0):
286                 w /= w.sum()
287                 idx_diff_2 = np.abs(grpLR[ord] - grpLR)
288                 proms[i] -= np.sum(w*idx_diff_2)
289     return proms
290
291 def _find_significances(self):
292     # Fit beta distribution to prominences and return
293     # group significance.
294     if self.prominences.size > 1:
295         # Model pdf.
296         def model_pdf(p):
297             return p[0]*beta.pdf(self.prominences,
298                                 p[1], p[2]) + 1 - p[0]
299         # Negative log-likelihood.
300         def negLL(p):
301             return -np.sum(np.log(np.maximum(
302                 model_pdf(p), 10**(-323.6))))
303         # Boundary restrictions.
304         bnds = ((0, 1), (1e-15, None), (1e-15, None))
305         # Initial guess.
306         mu = self.prominences.mean()
307         var = self.prominences.var()
308         p = [1, mu*(mu*(1 - mu)/var - 1),
309             (1 - mu)*(mu*(1 - mu)/var - 1)]
310         # Fitting.
311         sol = minimize(negLL, p, jac = '3-point',
312                       bounds = bnds)
313         if sol.success: self.pFit = sol.x
314         else: self.pFit = p
315         self.group_sigs = norm.isf(beta.sf(self.prominences,

```

```

316             self.pFit[1], self.pFit[2]))
317     else:
318         self.pFit = [1, 1, 1]
319         self.group_sigs = np.full(self.prominences.size,
320                                 np.nan)
321
322     def find_clusters(self, parentID = '1'):
323         # Classify clusters as significant groups.
324         sl = self.group_sigs >= self.S
325         self.significances = self.group_sigs[sl]
326         self.clusters = self.groups[sl]
327         clst_blrs = self.blrs[sl]
328
329         # Add on root-level cluster.
330         self.significances = np.concatenate((np.array([np.inf]),
331                                             self.significances))
332         self.clusters = np.concatenate((
333             np.array([[0, self.n_samples]]),
334             self.clusters), axis = 0)
335         clst_blrs = np.concatenate((np.array([0.0]), clst_blrs))
336
337         # Label clusters according to their hierarchy.
338         self.ids = []
339         children = []
340         for i, clst in enumerate(self.clusters):
341             children.append(0)
342             parent = next((-j - 1 for j, v in
343                          enumerate(self.clusters[:i][::-1])
344                          if clst[0] < v[1] and clst[1] <= v[1]),
345                          None)
346             if parent is not None: # Child cluster of parent.
347                 children[parent] += 1
348                 nextID = f"{self.ids[parent]}-{children[parent]}"
349                 self.ids.append(nextID)
350             elif len(self.ids) > 0: # Multiple root clusters.
351                 self.ids.append(str(max([int(id.split('-')[0])
352                                         for id in self.ids]) + 1))
353             else: # First root cluster.
354                 self.ids.append(parentID)
355         self.ids = np.array(self.ids)
356
357         # Optionally reduce hierarchy.
358         if self.h_style and self.ids.size > 1:
359             # Find children.
360             hLevel = np.char.count(self.ids, '-')

```

```

361     oneLevelSep = hLevel.reshape(1, -1) == \
362                 hLevel.reshape(-1, 1) + 1
363     subClst = np.char.startswith(self.ids.reshape(1, -1),
364                                 np.char.add(self.ids, '-').reshape(-1, 1))
365     children_bool = np.logical_and(subClst, oneLevelSep)
366     # Hierarchy correction.
367     compC, compP, pTrack = self._hierarchy_correction(
368         self.clusters, children_bool, clst_blrs,
369         self.logRho[self.ordering],
370         beta.isf(norm.sf(self.S), self.pFit[1],
371                 self.pFit[2]))
372     compS = norm.isf(beta.sf(compP,
373                             self.pFit[1], self.pFit[2]))
374     keep = np.ones(self.ids.size, dtype = np.bool_)
375     compS_bool = compS >= self.significances[pTrack]
376     keep[pTrack][compS_bool] = False
377     keep[0] = True
378     self.significances = np.concatenate((
379         self.significances[keep], compS))
380     self.clusters = np.concatenate((self.clusters[keep],
381                                     compC), axis = 0)
382
383     # Reorder the list of clusters according increasing
384     # start and decreasing end indices.
385     reorder = sorted(np.arange(self.clusters.shape[0]),
386                     key = lambda i:
387                         [self.clusters[i, 0], -self.clusters[i, 1]])
388     self.significances = self.significances[reorder]
389     self.clusters = self.clusters[reorder]
390
391     # Re-label clusters.
392     self.ids = []
393     children = []
394     for i, clst in enumerate(self.clusters):
395         children.append(0)
396         parent = next((-j - 1 for j, v in
397                       enumerate(self.clusters[:i][::-1])
398                       if clst[0] < v[1] and clst[1] <= v[1]),
399                       None)
400         if parent is not None: # Child cluster of parent.
401             children[parent] += 1
402             nextID = f"{self.ids[parent]}-\
403                     {children[parent]}"
404             self.ids.append(nextID)
405     elif len(self.ids) > 0: # Multiple root clusters.

```

```

406         self.ids.append(str(max([int(id.split('-')[0])
407                                for id in self.ids]) + 1))
408     else: # First root cluster
409         self.ids.append(parentID)
410     self.ids = np.array(self.ids)
411
412     @staticmethod
413     @njit(fastmath = True)
414     def _hierarchy_correction(clusters, children_bool, blrs,
415                              lrOrdIdx, cutOff):
416         compC = [[0 for j in range(0)] for i in range(0)]
417         compP = [0.0 for i in range(0)]
418         pTrack = [0 for i in range(0)]
419         # Cycle through each parent with children.
420         for p in range(clusters.shape[0]):
421             if children_bool[p].any():
422                 newClst = [clusters[p, 0], 0]
423                 blr_catch = 1.0
424                 # Check each child for adjoining complementary
425                 # clusters.
426                 for c in np.where(children_bool[p])[0]:
427                     newClst[1] = clusters[c, 0]
428                     if blrs[c] != blr_catch:
429                         blr_catch = blrs[c]
430                     # Find prominence.
431                     grpLR = lrOrdIdx[newClst[0]:newClst[1]]
432                     ord = grpLR.argsort()[::-1]
433                     grpLROrd = grpLR[ord]
434                     newProm = grpLROrd[0] - blrs[c]
435                     if newProm >= cutOff:
436                         # Adjust for noise.
437                         w = np.abs(ord - np.arange(ord.size)
438                                     ).astype(np.float64)
439                         if np.any(w != 0.0):
440                             w /= w.sum()
441                             idx_diff_2 = np.abs(grpLROrd -\
442                                                     grpLR)
443                             newProm -= np.sum(w*idx_diff_2)
444                         if newProm >= cutOff:
445                             compC.append(newClst)
446                             compP.append(newProm)
447                             pTrack.append(p)
448                             break
449         return np.array(compC), np.array(compP), np.array(pTrack)

```

# Bibliography

- (1) S. S. Kumar, “The Structure of Stars of Very Low Mass.”, *The Astrophysical Journal*, 1963, **137**, 1121.
- (2) C. Hayashi and T. Nakano, “Evolution of Stars of Small Masses in the Pre-Main-Sequence Stages”, *Progress of Theoretical Physics*, 1963, **30**, 460–474.
- (3) T. Nakano, in *50 Years of Brown Dwarfs: From Prediction to Discovery to Forefront of Research*, ed. V. Joergens, Springer International Publishing, Cham, 2014, pp. 5–17.
- (4) F. LeBlanc, *An Introduction to Stellar Astrophysics*, Wiley, 2010.
- (5) I. Horváth, Z. Bagoly, J. Hakkila and L. V. Tóth, Proceedings of Swift: 10 Years of Discovery, 2014, 78, p. 78.
- (6) Horváth, István, Hakkila, Jon and Bagoly, Zsolt, “Possible structure in the GRB sky distribution at redshift two”, *A&A*, 2014, **561**, L12.
- (7) Horváth, István, Bagoly, Zsolt, Hakkila, Jon and Tóth, L. V., “New data support the existence of the Hercules-Corona Borealis Great Wall”, *A&A*, 2015, **584**, A48.
- (8) G. Efstathiou, W. J. Sutherland and S. Maddox, “The cosmological constant and cold dark matter”, *Nature*, 1990, **348**, 705–707.
- (9) S. M. Carroll, W. H. Press and E. L. Turner, “The cosmological constant”, *Annual review of astronomy and astrophysics*, 1992, **30**, 499–542.
- (10) M. Fukugita, F. Takahara, K. Yamashita and Y. Yoshii, “Test for the cosmological constant with the number count of faint galaxies”, *The Astrophysical Journal*, 1990, **361**, L1–L4.
- (11) S. D. M. White, J. F. Navarro, A. E. Evrard and C. S. Frenk, “The baryon content of galaxy clusters: a challenge to cosmological orthodoxy”, *Nature*, 1993, **366**, 429–433.

- (12) J. P. Ostriker and P. J. Steinhardt, “The observational case for a low-density Universe with a non-zero cosmological constant”, *Nature*, 1995, **377**, 600–602.
- (13) S. D. M. White and M. J. Rees, “Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering”, *Monthly Notices of the Royal Astronomical Society*, 1978, **183**, 341–358.
- (14) G. Kauffmann, S. D. M. White and B. Guiderdoni, “The formation and evolution of galaxies within merging dark matter haloes”, *Monthly Notices of the Royal Astronomical Society*, 1993, **264**, 201–218.
- (15) S. Ghigna, B. Moore, F. Governato, G. Lake, T. Quinn and J. Stadel, “Dark matter haloes within clusters”, *Monthly Notices of the Royal Astronomical Society*, 1998, **300**, 146–162.
- (16) V. Springel, J. Wang, M. Vogelsberger, A. Ludlow, A. Jenkins, A. Helmi, J. F. Navarro, C. S. Frenk and S. D. M. White, “The Aquarius Project: the subhaloes of galactic haloes”, *Monthly Notices of the Royal Astronomical Society*, 2008, **391**, 1685–1711.
- (17) V. M. Slipher, “The radial velocity of the Andromeda Nebula”, *Lowell Observatory Bulletin*, 1913, **1**, 56–57.
- (18) V. M. Slipher, “Spectrographic Observations of Nebulae”, *Popular Astronomy*, 1915, **23**, 21–24.
- (19) E. Hubble, “A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae”, *Proceedings of the National Academy of Science*, 1929, **15**, 168–173.
- (20) G. Lemaître, “Expansion of the universe, A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae”, *Monthly Notices of the Royal Astronomical Society*, 1931, **91**, 483–490.
- (21) S. W. Hawking and G. F. R. Ellis, “The Cosmic Black-Body Radiation and the Existence of Singularities in Our Universe”, *The Astrophysical Journal*, 1968, **152**, 25.
- (22) S. W. Hawking and R. Penrose, “The Singularities of Gravitational Collapse and Cosmology”, *Proceedings of the Royal Society of London Series A*, 1970, **314**, 529–548.
- (23) A. H. Guth, “Inflationary universe: A possible solution to the horizon and flatness problems”, *Physical Review D*, 1981, **23**, 347–356.

- (24) A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner et al., “Observational evidence from supernovae for an accelerating universe and a cosmological constant”, *The Astronomical Journal*, 1998, **116**, 1009.
- (25) S. Perlmutter, G. Aldering, G. Goldhaber, R. Knop, P. Nugent, P. G. Castro, S. Deustua, S. Fabbro, A. Goobar, D. E. Groom et al., “Measurements of  $\Omega$  and  $\Lambda$  from 42 high-redshift supernovae”, *The Astrophysical Journal*, 1999, **517**, 565.
- (26) D. G. York, J. Adelman, J. E. Anderson Jr, S. F. Anderson, J. Annis, N. A. Bahcall, J. Bakken, R. Barkhouser, S. Bastian, E. Berman et al., “The sloan digital sky survey: Technical summary”, *The Astronomical Journal*, 2000, **120**, 1579.
- (27) M. Colless, G. Dalton, S. Maddox, W. Sutherland, P. Norberg, S. Cole, J. Bland-Hawthorn, T. Bridges, R. Cannon, C. Collins et al., “The 2df galaxy redshift survey: spectra and redshifts”, *Monthly Notices of the Royal Astronomical Society*, 2001, **328**, 1039–1063.
- (28) P. J. E. Peebles, “Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations”, *The Astrophysical Journal Letters*, 1982, **263**, L1–L5.
- (29) G. R. Blumenthal, S. M. Faber, J. R. Primack and M. J. Rees, “Formation of galaxies and large-scale structure with cold dark matter.”, *Nature*, 1984, **311**, 517–525.
- (30) P. de Bernardis, P. A. Ade, J. J. Bock, J. Bond, J. Borrill, A. Boscaleri, K. Coble, B. Crill, G. De Gasperis, P. Farese et al., “A flat Universe from high-resolution maps of the cosmic microwave background radiation”, *Nature*, 2000, **404**, 955–959.
- (31) S. Hanany, P. Ade, A. Balbi, J. Bock, J. Borrill, A. Boscaleri, P. De Bernardis, P. Ferreira, V. Hristov, A. Jaffe et al., “MAXIMA-1: a measurement of the cosmic microwave background anisotropy on angular scales of  $10^{\circ}$ – $5^{\circ}$ ”, *The Astrophysical Journal*, 2000, **545**, L5.
- (32) N. Aghanim, Y. Akrami, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. Barreiro, N. Bartolo et al., “Planck 2018 results-I. Overview and the cosmological legacy of Planck”, *Astronomy & Astrophysics*, 2020, **641**, A1.

- (33) N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. Banday, R. Barreiro, N. Bartolo, S. Basak et al., “Planck 2018 results-VI. Cosmological parameters”, *Astronomy & Astrophysics*, 2020, **641**, A6.
- (34) A. Einstein, “Die feldgleichungen der gravitation”, *Sitzung der physikalisch-mathematischen Klasse*, 1915, **25**, 844–847.
- (35) A. Einstein, “Die Grundlage der allgemeinen Relativitätstheorie”, *Annalen der Physik*, 1916, **354**, 769–822.
- (36) A. Friedmann, “Über die Krümmung des Raumes”, *Zeitschrift für Physik*, 1922, **10**, 377–386.
- (37) A. Friedmann, “Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes”, *Zeitschrift für Physik*, 1924, **21**, 326–332.
- (38) G. Lemaître, “L’Univers en expansion”, *Annales de la Société Scientifique de Bruxelles*, 1933, **53**, 51.
- (39) H. P. Robertson, “Kinematics and World-Structure”, *The Astrophysical Journal*, 1935, **82**, 284.
- (40) H. P. Robertson, “Kinematics and World-Structure II.”, *The Astrophysical Journal*, 1936, **83**, 187.
- (41) H. P. Robertson, “Kinematics and World-Structure III.”, *The Astrophysical Journal*, 1936, **83**, 257.
- (42) A. G. Walker, “On Milne’s Theory of World-Structure”, *Proceedings of the London Mathematical Society*, 1937, **42**, 90–127.
- (43) *Timeline of the Universe*, <https://map.gsfc.nasa.gov/media/060915/index.html>, Accessed: 13-06-2022.
- (44) N. Jarosik, C. L. Bennett, J. Dunkley, B. Gold, M. R. Greason, M. Halpern, R. S. Hill, G. Hinshaw, A. Kogut, E. Komatsu, D. Larson, M. Limon, S. S. Meyer, M. R. Nolte, N. Odegard, L. Page, K. M. Smith, D. N. Spergel, G. S. Tucker, J. L. Weiland, E. Wollack and E. L. Wright, “Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors, and Basic Results”, *Astrophysical Journal Supplement*, 2011, **192**, 14.



- (45) Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, N. Bartolo, E. Battaner, R. Battye, K. Benabed, A. Benoît, A. Benoit-Lévy, J. -. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, A. Bonaldi, L. Bonavera, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. -. Cardoso, A. Catalano, A. Challinor, A. Chamballu, R. -. Chary, H. C. Chiang, J. Chluba, P. R. Christensen, S. Church, D. L. Clements, S. Colombi, L. P. L. Colombo, C. Combet, A. Coulais, B. P. Crill, A. Curto, F. Cuttaia, L. Danese, R. D. Davies, R. J. Davis, P. de Bernardis, A. de Rosa, G. de Zotti, J. Delabrouille, F. -. Désert, E. Di Valentino, C. Dickinson, J. M. Diego, K. Dolag, H. Dole, S. Donzelli, O. Doré, M. Douspis, A. Ducout, J. Dunkley, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, M. Farhang, J. Fergusson, F. Finelli, O. Forni, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frejsel, S. Galeotta, S. Galli, K. Ganga, C. Gauthier, M. Gerbino, T. Ghosh, M. Giard, Y. Giraud-Héraud, E. Giusarma, E. Gjerlow, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gregorio, A. Gruppuso, J. E. Gudmundsson, J. Hamann, F. K. Hansen, D. Hanson, D. L. Harrison, G. Helou, S. Henrot-Versillé, C. Hernández-Monteagudo, D. Herranz, S. R. Hildebrandt, E. Hivon, M. Hobson, W. A. Holmes, A. Hornstrup, W. Hovest, Z. Huang, K. M. Huffenberger, G. Hurier, A. H. Jaffe, T. R. Jaffe, W. C. Jones, M. Juvela, E. Keihänen, R. Keskitalo, T. S. Kisner, R. Kneissl, J. Knoche, L. Knox, M. Kunz, H. Kurki-Suonio, G. Lagache, A. Lähteenmäki, J. -. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, J. P. Leahy, R. Leonardi, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Linden-Vørnle, M. López-Cañiego, P. M. Lubin, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marchini, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Masi, S. Matarrese, P. McGehee, P. R. Meinhold, A. Melchiorri, J. -. Melin, L. Mendes, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. -. Miville-Deschênes, A. Moneti, L. Montier, G. Morgante, D. Mortlock, A. Moss, D. Munshi, J. A. Murphy, P. Naselsky, F. Nati, P. Natoli, C. B. Netterfield, H. U. Norgaard-Nielsen, F. Noviello, D. Novikov, I. Novikov, C. A. Oxborrow, F. Paci, L. Pagano, F. Pajot, R. Paladini, D. Paoletti, B. Partridge, F. Pasian, G. Patanchon, T. J. Pearson, O. Perdereau, L. Perotto, F. Perrotta, V. Pettorino, F. Piacentini, M. Piat, E. Pierpaoli, D. Pietrobon, S. Plaszczynski, E. Pointecouteau, G. Polenta, L. Popa, G. W. Pratt, G. Prézeau, S. Prunet, J. -. Puget, J. P. Rachen, W. T. Reach, R. Rebolo, M. Reinecke, M. Remazeilles, C. Renault,

- A. Renzi, I. Ristorcelli, G. Rocha, C. Rosset, M. Rossetti, G. Roudier, B. Rouillé d'Orfeuill, M. Rowan-Robinson, J. A. Rubiño-Martín, B. Rusholme, N. Said, V. Salvatelli, L. Salvati, M. Sandri, D. Santos, M. Savelainen, G. Savini, D. Scott, M. D. Seiffert, P. Serra, E. P. S. Shellard, L. D. Spencer, M. Spinelli, V. Stolyarov, R. Stompor, R. Sudiwala, R. Sunyaev, D. Sutton, A. -. Suur-Uski, J. -. Sygnet, J. A. Tauber, L. Terenzi, L. Toffolatti, M. Tomasi, M. Tristram, T. Trombetti, M. Tucci, J. Tuovinen, M. Türlér, G. Umama, L. Valenziano, J. Valiviita, F. Van Tent, P. Vielva, F. Villa, L. A. Wade, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Wilkinson, D. Yvon, A. Zacchei and A. Zonca, "Planck 2015 results. XIII. Cosmological parameters", *Astronomy and Astrophysics*, 2016, **594**, A13.
- (46) A. R. Liddle and D. H. Lyth, *Cosmological inflation and large-scale structure*, Cambridge university press, 2000.
- (47) R. A. Alpher and R. Herman, "Evolution of the Universe", *Nature*, 1948, **162**, 774–775.
- (48) A. A. Penzias and R. W. Wilson, "A Measurement of Excess Antenna Temperature at 4080 Mc/s.", *The Astrophysical Journal*, 1965, **142**, 419–421.
- (49) N. Halverson, E. Leitch, C. Pryke, J. Kovac, J. Carlstrom, W. Holzapfel, M. Dragovan, J. Cartwright, B. Mason, S. Padin et al., "Degree angular scale interferometer first results: a measurement of the cosmic microwave background angular power spectrum", *The Astrophysical Journal*, 2002, **568**, 38.
- (50) D. N. Spergel, L. Verde, H. V. Peiris, E. Komatsu, M. Nolta, C. L. Bennett, M. Halpern, G. Hinshaw, N. Jarosik, A. Kogut et al., "First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: determination of cosmological parameters", *The Astrophysical Journal Supplement Series*, 2003, **148**, 175.
- (51) H. Mo, F. Van den Bosch and S. White, *Galaxy formation and evolution*, Cambridge University Press, 2010.
- (52) R. H. Cyburt, B. D. Fields, K. A. Olive and T.-H. Yeh, "Big bang nucleosynthesis: Present status", *Reviews of Modern Physics*, 2016, **88**, 015004.
- (53) M. Tanabashi, K. Hagiwara, K. Hikasa, K. Nakamura, Y. Sumino, F. Takahashi, J. Tanaka, K. Agashe, G. Aielli, C. Amsler, M. Antonelli, D. M. Asner, H. Baer, S. Banerjee, R. M. Barnett, T. Basaglia, C. W. Bauer, J. J. Beatty, V. I. Belousov, J. Beringer, S. Bethke, A. Bettini, H. Bichsel, O. Biebel,

K. M. Black, E. Blucher, O. Buchmuller, V. Burkert, M. A. Bychkov, R. N. Cahn, M. Carena, A. Ceccucci, A. Cerri, D. Chakraborty, M. -. Chen, R. S. Chivukula, G. Cowan, O. Dahl, G. D'Ambrosio, T. Damour, D. de Florian, A. de Gouvêa, T. DeGrand, P. de Jong, G. Dissertori, B. A. Dobrescu, M. D'Onofrio, M. Doser, M. Drees, H. K. Dreiner, D. A. Dwyer, P. Eerola, S. Eidelman, J. Ellis, J. Erler, V. V. Ezhela, W. Fetscher, B. D. Fields, R. Firestone, B. Foster, A. Freitas, H. Gallagher, L. Garren, H. -. Gerber, G. Gerbier, T. Gershon, Y. Gershtein, T. Gherghetta, A. A. Godizov, M. Goodman, C. Grab, A. V. Gribsan, C. Grojean, D. E. Groom, M. Grünewald, A. Gurtu, T. Gutsche, H. E. Haber, C. Hanhart, S. Hashimoto, Y. Hayato, K. G. Hayes, A. Hebecker, S. Heinemeyer, B. Heltsley, J. J. Hernández-Rey, J. Hisano, A. Höcker, J. Holder, A. Holtkamp, T. Hyodo, K. D. Irwin, K. F. Johnson, M. Kado, M. Karliner, U. F. Katz, S. R. Klein, E. Klempt, R. V. Kowalewski, F. Krauss, M. Kreps, B. Krusche, Y. V. Kuyanov, Y. Kwon, O. Lahav, J. Laiho, J. Lesgourgues, A. Liddle, Z. Ligeti, C. -. Lin, C. Lippmann, T. M. Liss, L. Littenberg, K. S. Lugovsky, S. B. Lugovsky, A. Lusiani, Y. Makida, F. Maltoni, T. Mannel, A. V. Manohar, W. J. Marciano, A. D. Martin, A. Masoni, J. Matthews, U. -. Meißner, D. Milstead, R. E. Mitchell, K. Mönig, P. Molaro, F. Moortgat, M. Moskovic, H. Murayama, M. Narain, P. Nason, S. Navas, M. Neubert, P. Nevski, Y. Nir, K. A. Olive, S. Pagan Griso, J. Parsons, C. Patrignani, J. A. Peacock, M. Pennington, S. T. Petcov, V. A. Petrov, E. Pianori, A. Piepke, A. Pomarol, A. Quadt, J. Rademacker, G. Raffelt, B. N. Ratcliff, P. Richardson, A. Ringwald, S. Roesler, S. Rolli, A. Romaniouk, L. J. Rosenberg, J. L. Rosner, G. Rybka, R. A. Ryutin, C. T. Sachrajda, Y. Sakai, G. P. Salam, S. Sarkar, F. Sauli, O. Schneider, K. Scholberg, A. J. Schwartz, D. Scott, V. Sharma, S. R. Sharpe, T. Shutt, M. Silari, T. Sjöstrand, P. Skands, T. Skwarnicki, J. G. Smith, G. F. Smoot, S. Spanier, H. Spieler, C. Spiering, A. Stahl, S. L. Stone, T. Sumiyoshi, M. J. Syphers, K. Terashi, J. Terning, U. Thoma, R. S. Thorne, L. Tiator, M. Titov, N. P. Tkachenko, N. A. Törnqvist, D. R. Tovey, G. Valencia, R. Van de Water, N. Varelas, G. Venanzoni, L. Verde, M. G. Vinciter, P. Vogel, A. Vogt, S. P. Wakely, W. Walkowiak, C. W. Walter, D. Wands, D. R. Ward, M. O. Wascko, G. Weiglein, D. H. Weinberg, E. J. Weinberg, M. White, L. R. Wiencke, S. Willocq, C. G. Wohl, J. Womersley, C. L. Woody, R. L. Workman, W. -. Yao, G. P. Zeller, O. V. Zenin, R. -. Zhu, S. -. Zhu, F. Zimmermann, P. A. Zyla, J. Anderson, L. Fuller, V. S. Lugovsky, P. Schaffner and Particle Data Group, "Review of Particle Physics\*", *Physical Review D*, 2018, **98**, 030001.

- (54) C. Hayashi, “Evolution of Protostars”, *Annual Review of Astronomy and Astrophysics*, 1966, **4**, 171.
- (55) F. van Leeuwen, “Validation of the new Hipparcos reduction”, *Astronomy and Astrophysics*, 2007, **474**, 653–664.
- (56) W. Gliese and H. Jahreiß, *Preliminary Version of the Third Catalogue of Nearby Stars*, On: The Astronomical Data Center CD-ROM: Selected Astronomical Catalogs, Vol. I; L.E. Brodzmann, S.E. Gesser (eds.), NASA/Astronomical Data Center, Goddard Space Flight Center, Greenbelt, MD, 1991.
- (57) F. Mignard, “Astronomical distance scales”, *Comptes Rendus Physique*, 2019, **20**, DOI: [10.1016/j.crhy.2019.02.001](https://doi.org/10.1016/j.crhy.2019.02.001).
- (58) R. B. Larson, “Numerical calculations of the dynamics of collapsing protostar”, *Monthly Notices of the Royal Astronomical Society*, 1969, **145**, 271.
- (59) K. .-. A. Winkler and M. J. Newman, “Formation of solar-type stars in spherical symmetry. I - The key role of the accretion shock”, *The Astrophysical Journal*, 1980, **236**, 201–211.
- (60) S. W. Stahler, F. H. Shu and R. E. Taam, “The evolution of protostars. I - Global formulation and results”, *The Astrophysical Journal*, 1980, **241**, 637–654.
- (61) M. M. Dunham, A. M. Stutz, L. E. Allen, N. J. Evans II, W. J. Fischer, S. T. Megeath, P. C. Myers, S. S. Offner, C. A. Poteet, J. J. Tobin et al., “The evolution of protostars: Insights from ten years of infrared surveys with Spitzer and Herschel”, *Protostars and Planets VI*, 2014, **195**.
- (62) W. W. Morgan, P. C. Keenan and E. Kellman, “An atlas of stellar spectra, with an outline of spectral classification”, *Chicago*, 1943.
- (63) W. W. Morgan and P. C. Keenan, “Spectral Classification”, *Annual Review of Astronomy and Astrophysics*, 1973, **11**, 29.
- (64) W. W. Campbell, “The Wolf-Rayet stars.”, *Astronomy and Astro-Physics (formerly The Sidereal Messenger)*, 1894, **13**, 448–476.
- (65) In *Encyclopedia of Astronomy and Astrophysics*, ed. P. Murdin, 2000, 4101, p. 4101.
- (66) J. D. Kirkpatrick, I. N. Reid, J. Liebert, R. M. Cutri, B. Nelson, C. A. Beichman, C. C. Dahn, D. G. Monet, J. E. Gizis and M. F. Skrutskie, “Dwarfs Cooler than “M“: The Definition of Spectral Type “L” Using Discoveries from the 2 Micron All-Sky Survey (2MASS)”, *The Astrophysical Journal*, 1999, **519**, 802–833.

- (67) J. D. Kirkpatrick, “New Spectral Types L and T”, *Annual Review of Astronomy and Astrophysics*, 2005, **43**, 195–245.
- (68) J. D. Kirkpatrick, T. S. Barman, A. J. Burgasser, M. R. McGovern, I. S. McLean, C. G. Tinney and P. J. Lowrance, “Discovery of a Very Young Field L Dwarf, 2MASS J01415823-4633574”, *The Astrophysical Journal*, 2006, **639**, 1120–1128.
- (69) N. R. Deacon and N. C. Hambly, “The possibility of detection of ultracool dwarfs with the UKIRT Infrared Deep Sky Survey”, *Monthly Notices of the Royal Astronomical Society*, 2006, **371**, 1722–1730.
- (70) B. Zuckerman and I. Song, “The minimum Jeans mass, brown dwarf companion IMF, and predictions for detection of Y-type dwarfs”, *Astronomy and Astrophysics*, 2009, **493**, 1149–1154.
- (71) J. D. Kirkpatrick, M. C. Cushing, C. R. Gelino, C. A. Beichman, C. G. Tinney, J. K. Faherty, A. Schneider and G. N. Mace, “Discovery of the Y1 Dwarf WISE J064723.23-623235.5”, *The Astrophysical Journal*, 2013, **776**, 128.
- (72) T. J. Dupuy and A. L. Kraus, “Distances, Luminosities, and Temperatures of the Coldest Known Substellar Objects”, *Science*, 2013, **341**, 1492–1495.
- (73) W. Baade, “The Resolution of Messier 32, NGC 205, and the Central Region of the Andromeda Nebula.”, *The Astrophysical Journal*, 1944, **100**, 137.
- (74) M. J. Rees, “Origin of pregalactic microwave background”, *Nature*, 1978, **275**, 35–37.
- (75) J. L. Puget and J. Heyvaerts, “Population III stars and the shape of the cosmological black body radiation”, *Astronomy and Astrophysics*, 1980, **83**, L10–L12.
- (76) A. Heger and S. E. Woosley, “The Nucleosynthetic Signature of Population III”, *The Astrophysical Journal*, 2002, **567**, 532–543.
- (77) R. A. E. Fosbury, M. Villar-Martín, A. Humphrey, M. Lombardi, P. Rosati, D. Stern, R. N. Hook, B. P. Holden, S. A. Stanford, G. K. Squires, M. Rauch and W. L. W. Sargent, “Massive Star Formation in a Gravitationally Lensed H II Galaxy at  $z = 3.357$ ”, *The Astrophysical Journal*, 2003, **596**, 797–809.
- (78) V. Bromm, N. Yoshida, L. Hernquist and C. F. McKee, “The formation of the first stars and galaxies”, *Nature*, 2009, **459**, 49–54.

- (79) D. Sobral, J. Matthee, B. Darvish, D. Schaerer, B. Mobasher, H. J. A. Röttgering, S. Santos and S. Hemmati, “Evidence for PopIII-like Stellar Populations in the Most Luminous Lyman- $\alpha$  Emitters at the Epoch of Reionization: Spectroscopic Confirmation”, *The Astrophysical Journal*, 2015, **808**, 139.
- (80) B. K. Gibson, Y. Fenner, A. Renda, D. Kawata and H.-c. Lee, “Galactic Chemical Evolution”, *Publications of the Astronomical Society of Australia*, 2003, **20**, 401–415.
- (81) T. S. van Albada and N. Baker, “On the Two Oosterhoff Groups of Globular Clusters”, *The Astrophysical Journal*, 1973, **185**, 477–498.
- (82) W. Dias, B. Alessi, A. Moitinho and J. Lépine, “New catalogue of optically visible open clusters and candidates”, *Astronomy & Astrophysics*, 2002, **389**, 871–873.
- (83) N. V. Kharchenko, A. E. Piskunov, E. Schilbach, S. Röser and R. -. Scholz, “Global survey of star clusters in the Milky Way. II. The catalogue of basic parameters”, *Astronomy and Astrophysics*, 2013, **558**, A53.
- (84) Cantat-Gaudin, T., Jordi, C., Vallenari, A., Bragaglia, A., Balaguer-Núñez, L., Soubiran, C., Bossini, D., Moitinho, A., Castro-Ginard, A., Krone-Martins, A., Casamiquela, L., Sordo, R. and Carrera, R., “A Gaia DR2 view of the open cluster population in the Milky Way”, *A&A*, 2018, **618**, A93.
- (85) Hao, C. J., Xu, Y., Hou, L. G., Bian, S. B., Li, J. J., Wu, Z. Y., He, Z. H., Li, Y. J. and Liu, D. J., “Evolution of the local spiral structure of the Milky Way revealed by open clusters”, *A&A*, 2021, **652**, A102.
- (86) C. J. Lada and E. A. Lada, “Embedded Clusters in Molecular Clouds”, *Annual Review of Astronomy and Astrophysics*, 2003, **41**, 57–115.
- (87) J. H. Jeans, “On the law of distribution in star-clusters”, *Monthly Notices of the Royal Astronomical Society*, 1916, **76**, 567.
- (88) J. H. Oort, “Luminosity distribution and shape of the Hyades cluster.”, *Astronomy and Astrophysics*, 1979, **78**, 312–317.
- (89) G. Bergond, S. Leon and J. Guibert, “Gravitational tidal effects on galactic open clusters”, *Astronomy and Astrophysics*, 2001, **377**, 462–472.
- (90) C. E. Jones and S. Basu, “The Intrinsic Shapes of Molecular Cloud Fragments over a Range of Length Scales”, *The Astrophysical Journal*, 2002, **569**, 280–287.

- (91) C. L. Curry, “Shapes of Molecular Cloud Cores and the Filamentary Mode of Star Formation”, *The Astrophysical Journal*, 2002, **576**, 849–859.
- (92) W. P. Chen, C. W. Chen and C. G. Shu, “Morphology of Galactic Open Clusters”, *The Astronomical Journal*, 2004, **128**, 2306–2315.
- (93) T. Jerabkova, H. M. J. Boffin, G. Beccari and R. I. Anderson, “A stellar relic filament in the Orion star-forming region”, *Monthly Notices of the Royal Astronomical Society*, 2019, **489**, 4418–4428.
- (94) A. Ballone, M. Mapelli, U. N. Di Carlo, S. Torniamenti, M. Spera and S. Rastello, “Evolution of fractality and rotation in embedded star clusters”, *Monthly Notices of the Royal Astronomical Society*, 2020, **496**, 49–59.
- (95) C. Gaia, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, C. Babusiaux, C. A. L. Bailer-Jones, M. Biermann, D. W. Evans, L. Eyer, F. Jansen, C. Jordi, S. A. Klioner, U. Lammers, L. Lindegren, X. Luri, F. Mignard, C. Panem, D. Pourbaix, S. Randich, P. Sartoretti, H. I. Siddiqui, C. Soubiran, F. van Leeuwen, N. A. Walton, F. Arenou, U. Bastian, M. Cropper, R. Drimmel, D. Katz, M. G. Lattanzi, J. Bakker, C. Cacciari, J. Castañeda, L. Chaoul, N. Cheek, F. De Angeli, C. Fabricius, R. Guerra, B. Holl, E. Masana, R. Messineo, N. Mowlavi, K. Nienartowicz, P. Panuzzo, J. Portell, M. Riello, G. M. Seabroke, P. Tanga, F. Thévenin, G. Gracia-Abril, G. Comoretto, M. Garcia-Reinaldos, D. Teyssier, M. Altmann, R. Andrae, M. Audard, I. Bellas-Velidis, K. Benson, J. Berthier, R. Blomme, P. Burgess, G. Busso, B. Carry, A. Cellino, G. Clementini, M. Clotet, O. Creevey, M. Davidson, J. De Ridder, L. Delchambre, A. Dell’Oro, C. Ducourant, J. Fernández-Hernández, M. Fouesneau, Y. Frémat, L. Galluccio, M. García-Torres, J. González-Núñez, J. J. González-Vidal, E. Gosset, L. P. Guy, J. L. Halbwachs, N. C. Hambly, D. L. Harrison, J. Hernández, D. Hestroffer, S. T. Hodgkin, A. Hutton, G. Jasniewicz, A. Jean-Antoine-Piccolo, S. Jordan, A. J. Korn, A. Krone-Martins, A. C. Lanzafame, T. Lebzelter, W. Löffler, M. Manteiga, P. M. Marrese, J. M. Martín-Fleitas et al., “Gaia Data Release 2”, *A&A*, 2018, **616**.
- (96) Gaia Collaboration, Babusiaux, C., van Leeuwen, F., Barstow, M. A., Jordi, C., Vallenari, A., Bossini, D., Bressan, A., Cantat-Gaudin, T., van Leeuwen, M., Brown, A. G. A., Prusti, T., de Bruijne, J. H. J., Bailer-Jones, C. A. L., Biermann, M., Evans, D. W., Eyer, L., Jansen, F., Klioner, S. A., Lammers, U., Lindegren, L., Luri, X., Mignard, F., Panem, C., Pourbaix, D., Randich, S., Sartoretti, P., Siddiqui, H. I., Soubiran, C., Walton, N. A.,

Arenou, F., Bastian, U., Cropper, M., Drimmel, R., Katz, D., Lattanzi, M. G., Bakker, J., Cacciari, C., Castañeda, J., Chaoul, L., Cheek, N., De Angeli, F., Fabricius, C., Guerra, R., Holl, B., Masana, E., Messineo, R., Mowlavi, N., Nienartowicz, K., Panuzzo, P., Portell, J., Riello, M., Seabroke, G. M., Tanga, P., Thévenin, F., Gracia-Abril, G., Comoretto, G., Garcia-Reinaldos, M., Teyssier, D., Altmann, M., Andrae, R., Audard, M., Bellas-Velidis, I., Benson, K., Berthier, J., Blomme, R., Burgess, P., Busso, G., Carry, B., Cellino, A., Clementini, G., Clotet, M., Creevey, O., Davidson, M., De Ridder, J., Delchambre, L., Dell'Oro, A., Ducourant, C., Fernández-Hernández, J., Fouesneau, M., Frémat, Y., Galluccio, L., García-Torres, M., González-Núñez, J., González-Vidal, J. J., Gosset, E., Guy, L. P., Halbwachs, J.-L., Hambly, N. C., Harrison, D. L., Hernández, J., Hestroffer, D., Hodgkin, S. T., Hutton, A., Jasniewicz, G., Jean-Antoine-Piccolo, A., Jordan, S., Korn, A. J., Krone-Martins, A., Lanzafame, A. C., Lebzelter, T., Löffler, W., Manteiga, M., Marrese, P. M., Martín-Fleitas, J. M., Moitinho, A., Mora, A., Muinonen, K., Osinde, J., Pancino, E., Pauwels, T., Petit, J.-M., Recio-Blanco, A., Richards, P. J., Rimoldini, L., Robin, A. C., Sarro, L. M., Siopis, C., Smith, M., Sozzetti, A., Süveges, M., Torra, J., van Reeve, W., Abbas, U., Abreu Aramburu, A., Accart, S., Aerts, C., Altavilla, G., Álvarez, M. A., Alvarez, R., Alves, J., Anderson, R. I., Andrei, A. H., Anglada Varela, E., Antiche, E., Antoja, T., Arcay, B., Astraatmadja, T. L., Bach, N., Baker, S. G., Balaguer-Núñez, L., Balm, P., Barache, C., Barata, C., Barbato, D., Barblan, F., Barklem, P. S., Barrado, D., Barros, M., Bartholomé Muñoz, L., Bassilana, J.-L., Becciani, U., Bellazzini, M., Berihuete, A., Bertone, S., Bianchi, L., Bienaymé, O., Blanco-Cuaresma, S., Boch, T., Boeche, C., Bombrun, A., Borrachero, R., Bouquillon, S., Bourda, G., Bragaglia, A., Bramante, L., Breddels, M. A., Brouillet, N., Brüsemeister, T., Brugaletta, E., Bucciarelli, B., Burlacu, A., Busonero, D., Butkevich, A. G., Buzzi, R., Caffau, E., Cancelliere, R., Cannizzaro, G., Carballo, R., Carlucci, T., Carrasco, J. M., Casamiquela, L., Castellani, M., Castro-Ginard, A., Charlot, P., Chemin, L., Chiavassa, A., Cocozza, G., Costigan, G., Cowell, S., Crifo, F., Crosta, M., Crowley, C., Cuypers, J., Dafonte, C., Damerdji, Y., Dapergolas, A., David, P., David, M., de Laverny, P., De Luise, F., De March, R., de Martino, D., de Souza, R., de Torres, A., Debosscher, J., del Pozo, E., Delbo, M., Delgado, A., Delgado, H. E., Diakite, S., Diener, C., Distefano, E., Dolding, C., Drazinos, P., Durán, J., Edvardsson, B., Enke, H., Eriksson, K., Esquej, P., Eynard Bontemps, G., Fabre, C., Fabrizio, M., Faigler, S., Falcão, A. J., Farràs Casas, M., Federici,



L., Fedorets, G., Fernique, P., Figueras, F., Filippi, F., Findeisen, K., Fonti, A., Fraile, E., Fraser, M., Frézouls, B., Gai, M., Galleti, S., Garabato, D., García-Sedano, F., Garofalo, A., Garralda, N., Gavel, A., Gavras, P., Gerssen, J., Geyer, R., Giacobbe, P., Gilmore, G., Girona, S., Giuffrida, G., Glass, F., Gomes, M., Granvik, M., Gueguen, A., Guerrier, A., Guiraud, J., Gutiérrez, R., Haigron, R., Hatzidimitriou, D., Hauser, M., Haywood, M., Heiter, U., Helmi, A., Heu, J., Hilger, T., Hobbs, D., Hofmann, W., Holland, G., Huckle, H. E., Hypki, A., Icardi, V., Janßen, K., Jevardat de Fombelle, G., Jonker, P. G., Juhász, Á. L., Julbe, F., Karampelas, A., Kewley, A., Klar, J., Kochoska, A., Kohley, R., Kolenberg, K., Kontizas, M., Kontizas, E., Koposov, S. E., Kordopatis, G., Kostrzewa-Rutkowska, Z., Koubsky, P., Lambert, S., Lanza, A. F., Lasne, Y., Lavigne, J.-B., Le Fustec, Y., Le Poncin-Lafitte, C., Lebreton, Y., Leccia, S., Leclerc, N., Lecoœur-Taibi, I., Lenhardt, H., Leroux, F., Liao, S., Licata, E., Lindstrøm, H. E. P., Lister, T. A., Livanou, E., Lobel, A., López, M., Managau, S., Mann, R. G., Mantelet, G., Marchal, O., Marchant, J. M., Marconi, M., Marinoni, S., Marschalkó, G., Marshall, D. J., Martino, M., Marton, G., Mary, N., Massari, D., Matijevic, G., Mazeh, T., McMillan, P. J., Messina, S., Michalik, D., Millar, N. R., Molina, D., Molinaro, R., Molnár, L., Montegriffo, P., Mor, R., Morbidelli, R., Morel, T., Morris, D., Mulone, A. F., Muraveva, T., Musella, I., Nelemans, G., Nicastro, L., Noval, L., O'Mullane, W., Ordénovic, C., Ordóñez-Blanco, D., Osborne, P., Pagani, C., Pagano, I., Pailler, F., Palacin, H., Palaversa, L., Panahi, A., Pawlak, M., Piersimoni, A. M., Pineau, F.-X., Plachy, E., Plum, G., Poggio, E., Poujoulet, E., Prsa, A., Pulone, L., Racero, E., Ragaini, S., Rambaux, N., Ramos-Lerate, M., Regibo, S., Reylé, C., Riclet, F., Ripepi, V., Riva, A., Rivard, A., Rixon, G., Roegiers, T., Roelens, M., Romero-Gómez, M., Rowell, N., Royer, F., Ruiz-Dern, L., Sadowski, G., Sagristà Sellés, T., Sahlmann, J., Salgado, J., Salguero, E., Sanna, N., Santana-Ros, T., Sarasso, M., Savietto, H., Schultheis, M., Sciacca, E., Segol, M., Segovia, J. C., Ségransan, D., Shih, I-C., Siltala, L., Silva, A. F., Smart, R. L., Smith, K. W., Solano, E., Solitro, F., Sordo, R., Soria Nieto, S., Souchay, J., Spagna, A., Spoto, F., Stampa, U., Steele, I. A., Steidelmüller, H., Stephenson, C. A., Stoev, H., Suess, F. F., Surdej, J., Szabados, L., Szegedi-Elek, E., Tapiador, D., Taris, F., Tauran, G., Taylor, M. B., Teixeira, R., Terrett, D., Teyssandier, P., Thuillot, W., Titarenko, A., Torra Clotet, F., Turon, C., Ulla, A., Utrilla, E., Uzzi, S., Vaillant, M., Valentini, G., Valette, V., van Elteren, A., Van Hemelryck, E., Vaschetto, M., Vecchiato, A., Veljanoski, J., Viala, Y., Vicente, D., Vogt, S., von Essen, C.,

- Voss, H., Votruba, V., Voutsinas, S., Walmsley, G., Weiler, M., Wertz, O., Wevers, T., Wyrzykowski, L., Yoldas, A., Zerjal, M., Ziaeeepour, H., Zorec, J., Zschocke, S., Zucker, S., Zurbach, C. and Zwitter, and T., “Gaia Data Release 2 - Observational Hertzsprung-Russell diagrams”, *A&A*, 2018, **616**, A10.
- (97) R. Gratton, A. Bragaglia, E. Carretta, V. D’Orazi, S. Lucatello and A. Sollima, “What is a globular cluster? An observational perspective”, *Astronomy and Astrophysics Review*, 2019, **27**, 8.
- (98) K. Ashman and S. Zepf, *Globular Cluster Systems*, Cambridge University Press, 1998.
- (99) H. Shapley and H. B. Sawyer, *Harvard Observatory Bulletin*, 1927, **848**.
- (100) H. Shapley and H. B. Sawyer, *Harvard Observatory Bulletin*, 1927, **849**.
- (101) H. Shapley and H. B. Sawyer, *Harvard Observatory Bulletin*, 1927, **852**.
- (102) H. Shapley and H. B. Sawyer, *Harvard Observatory Bulletin*, 1929, **869**.
- (103) H. S. Hogg, “Harlow Shapley and Globular Clusters”, *Publications of the Astronomical Society of the Pacific*, 1965, **77**, 336.
- (104) C. Firmani and V. Avila-Reese, *Revista Mexicana de Astronomia y Astrofisica Conference Series*, ed. V. Avila-Reese, C. Firmani, C. S. Frenk and C. Allen, 2003, vol. 17, pp. 107–120.
- (105) P. Guhathakurta, J. Tyson and S. Majewski, “A redshift limit for the faint blue galaxy population from deep U band imaging”, *The Astrophysical Journal*, 1990, **357**, L9–L12.
- (106) R. J. Bouwens, P. A. Oesch, G. D. Illingworth, I. Labbé, P. G. van Dokkum, G. Brammer, D. Magee, L. R. Spitler, M. Franx, R. Smit, M. Trenti, V. Gonzalez and C. M. Carollo, “Photometric Constraints on the Redshift of  $z \sim 10$  Candidate UDFj-39546284 from Deeper WFC3/IR+ACS+IRAC Observations over the HUDF”, *The Astrophysical Journal Letters*, 2013, **765**, L16.
- (107) P. A. Oesch, P. G. van Dokkum, G. D. Illingworth, R. J. Bouwens, I. Momcheva, B. Holden, G. W. Roberts-Borsani, R. Smit, M. Franx, I. Labbé, V. González and D. Magee, “A Spectroscopic Redshift Measurement for a Luminous Lyman Break Galaxy at  $z = 7.730$  Using Keck/MOSFIRE”, *The Astrophysical Journal Letters*, 2015, **804**, L30.

- (108) P. A. Oesch, G. Brammer, P. G. van Dokkum, G. D. Illingworth, R. J. Bouwens, I. Labbé, M. Franx, I. Momcheva, M. L. N. Ashby, G. G. Fazio, V. Gonzalez, B. Holden, D. Magee, R. E. Skelton, R. Smit, L. R. Spitler, M. Trenti and S. P. Willner, “A Remarkably Luminous Galaxy at  $z=11.1$  Measured with Hubble Space Telescope Grism Spectroscopy”, *The Astrophysical Journal*, 2016, **819**, 129.
- (109) T. Hashimoto, N. Laporte, K. Mawatari, R. S. Ellis, A. K. Inoue, E. Zackrisson, G. Roberts-Borsani, W. Zheng, Y. Tamura, F. E. Bauer, T. Fletcher, Y. Harikane, B. Hatsukade, N. H. Hayatsu, Y. Matsuda, H. Matsuo, T. Okamoto, M. Ouchi, R. Pelló, C.-E. Rydberg, I. Shimizu, Y. Taniguchi, H. Umehata and N. Yoshida, “The onset of star formation 250 million years after the Big Bang”, *Nature*, 2018, **557**, 392–395.
- (110) Y. Harikane, A. K. Inoue, K. Mawatari, T. Hashimoto, S. Yamanaka, Y. Fudamoto, H. Matsuo, Y. Tamura, P. Dayal, L. Y. A. Yung, A. Hutter, F. Pacucci, Y. Sugahara and A. M. Koekemoer, “A Search for H-Dropout Lyman Break Galaxies at  $z \sim 12\text{--}16$ ”, *The Astrophysical Journal*, 2022, **929**, 1.
- (111) E. Hubble, “No. 324. Extra-galactic nebulae.”, *Contributions from the Mount Wilson Observatory / Carnegie Institution of Washington*, 1926, **324**, 1–49.
- (112) E. Hubble, “Realm of the Nebulae, ed”, *Hubble, EP*, 1936, **2**.
- (113) A. W. Graham, in *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, ed. T. D. Oswalt and W. C. Keel, 2013, vol. 6, pp. 91–140.
- (114) M. H. Liller, “The Distribution of Intensity in Elliptical Galaxies of the Virgo Cluster. II”, *The Astrophysical Journal*, 1966, **146**, 28.
- (115) A. W. Graham, M. M. Colless, G. Busarello, S. Zaggia and G. Longo, “Extended stellar kinematics of elliptical galaxies in the Fornax cluster”, *Astronomy and Astrophysics Supplement*, 1998, **133**, 325–336.
- (116) E. Emsellem, M. Cappellari, D. Krajnović, K. Alatalo, L. Blitz, M. Bois, F. Bournaud, M. Bureau, R. L. Davies, T. A. Davis, P. T. de Zeeuw, S. Khochfar, H. Kuntschner, P.-Y. Lablanche, R. M. McDermid, R. Morganti, T. Naab, T. Oosterloo, M. Sarzi, N. Scott, P. Serra, G. van de Ven, A.-M. Weijmans and L. M. Young, “The ATLAS<sup>3D</sup> project - III. A census of the stellar angular momentum within the effective radius of early-type galaxies: unveiling the distribution of fast and slow rotators”, *Monthly Notices of the Royal Astronomical Society*, 2011, **414**, 888–912.

- (117) J. Binney, M. Michael and M. Merrifield, *Galactic astronomy*, Princeton University Press, 1998.
- (118) D. Mihalas and J. Binney, *Galactic astronomy*, W.H. Freeman & Co., 1968.
- (119) A. Dekel, Y. Birnboim, G. Engel, J. Freundlich, T. Goerdt, M. Mumcuoglu, E. Neistein, C. Pichon, R. Teyssier and E. Zinger, “Cold streams in early massive hot haloes as the main mode of galaxy formation”, *Nature*, 2009, **457**, 451–454.
- (120) K. R. Stewart, A. M. Brooks, J. S. Bullock, A. H. Maller, J. Diemand, J. Wadsley and L. A. Moustakas, “Angular Momentum Acquisition in Galaxy Halos”, *The Astrophysical Journal*, 2013, **769**, 74.
- (121) G. de Vaucouleurs, “Classification and Morphology of External Galaxies.”, *Handbuch der Physik*, 1959, **53**, 275.
- (122) G. de Vaucouleurs, “Revised Classification of 1500 Bright Galaxies.”, *Astrophysical Journal Supplement*, 1963, **8**, 31.
- (123) G. De Vaucouleurs, Symposium-International Astronomical Union, 1974, vol. 58, pp. 1–53.
- (124) C. Mateu, “galstreams: A Library of Milky Way Stellar Stream Footprints and Tracks”, *arXiv preprint arXiv:2204.10326*, 2022.
- (125) G. Tormen, A. Diaferio and D. Syer, “Survival of substructure within dark matter haloes”, *Monthly Notices of the Royal Astronomical Society*, 1998, **299**, 728–742.
- (126) O. Y. Gnedin, “Tidal Effects in Clusters of Galaxies”, *The Astrophysical Journal*, 2003, **582**, 141–161.
- (127) J. S. Bullock and K. V. Johnston, “Tracing Galaxy Formation with Stellar Halos. I. Methods”, *The Astrophysical Journal*, 2005, **635**, 931–949.
- (128) K. V. Johnston, J. S. Bullock, S. Sharma, A. Font, B. E. Robertson and S. N. Leitner, “Tracing Galaxy Formation with Stellar Halos. II. Relating Substructure in Phase and Abundance Space to Accretion Histories”, *The Astrophysical Journal*, 2008, **689**, 936–957.
- (129) K. V. Johnston, L. Hernquist and M. Bolte, “Fossil Signatures of Ancient Accretion Events in the Halo”, *The Astrophysical Journal*, 1996, **465**, 278.
- (130) A. Helmi and S. D. M. White, “Building up the stellar halo of the Galaxy”, *Monthly Notices of the Royal Astronomical Society*, 1999, **307**, 495–517.

- (131) R. Ibata, G. Lewis, M. Irwin and T. Quinn, “Uncovering cold dark matter halo substructure with tidal streams”, *Monthly Notices of the Royal Astronomical Society*, 2002, **332**, 915–920.
- (132) J. Diemand, M. Kuhlen and P. Madau, “Formation and Evolution of Galaxy Dark Matter Halos and Their Substructure”, *The Astrophysical Journal*, 2007, **667**, 859–877.
- (133) F. Schweizer, “An optical study of the giant radio galaxy NGC 1316 (Fornax A).”, *The Astrophysical Journal*, 1980, **237**, 303–318.
- (134) P. J. Quinn, “On the formation and dynamics of shells around elliptical galaxies.”, *The Astrophysical Journal*, 1984, **279**, 596–609.
- (135) D. Martínez-Delgado, J. Peñarrubia, R. J. Gabany, I. Trujillo, S. R. Majewski and M. Pohlen, “The Ghost of a Dwarf Galaxy: Fossils of the Hierarchical Formation of the Nearby Spiral Galaxy NGC 5907”, *The Astrophysical Journal*, 2008, **689**, 184–193.
- (136) D. Martínez-Delgado, M. Pohlen, R. J. Gabany, S. R. Majewski, J. Peñarrubia and C. Palma, “Discovery of a Giant Stellar Tidal Stream Around The Disk Galaxy NGC 4013”, *The Astrophysical Journal*, 2009, **692**, 955–963.
- (137) D. Martínez-Delgado, R. J. Gabany, K. Crawford, S. Zibetti, S. R. Majewski, H.-W. Rix, J. Fliri, J. A. Carballo-Bello, D. C. Bardalez-Gagliuffi, J. Peñarrubia, T. S. Chonis, B. Madore, I. Trujillo, M. Schirmer and D. A. McDavid, “Stellar Tidal Streams in Spiral Galaxies of the Local Volume: A Pilot Survey with Modest Aperture Telescopes”, *The Astronomical Journal*, 2010, **140**, 962–967.
- (138) A. W. McConnachie, R. Ibata, N. Martin, A. M. N. Ferguson, M. Collins, S. Gwyn, M. Irwin, G. F. Lewis, A. D. Mackey, T. Davidge, V. Arias, A. Conn, P. Côté, D. Crnojevic, A. Huxor, J. Penarrubia, C. Spengler, N. Tanvir, D. Valls-Gabaud, A. Babul, P. Barmby, N. F. Bate, E. Bernard, S. Chapman, A. Dotter, W. Harris, B. McMonigal, J. Navarro, T. H. Puzia, R. M. Rich, G. Thomas and L. 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*, 2018, **868**, 55.
- (139) R. Ibata, K. Malhan, N. Martin, D. Aubert, B. Famaey, P. Bianchini, G. Monari, A. Siebert, G. F. Thomas, M. Bellazzini et al., “Charting the Galactic acceleration field. I. A search for stellar streams with Gaia DR2 and EDR3 with follow-up from ESPaDOnS and UVES”, *The Astrophysical Journal*, 2021, **914**, 123.

- (140) D. Malin and D. Carter, “A catalog of elliptical galaxies with shells”, *The Astrophysical Journal*, 1983, **274**, 534–540.
- (141) A. M. Atkinson, R. G. Abraham and A. M. Ferguson, “Faint tidal features in galaxies within the Canada–France–Hawaii Telescope legacy survey wide fields”, *The Astrophysical Journal*, 2013, **765**, 28.
- (142) M. Bílek, P.-A. Duc, J.-C. Cuillandre, S. Gwyn, M. Cappellari, D. V. Bekaert, P. Bonfini, T. Bitsakis, S. Paudel, D. Krajnović et al., “Census and classification of low-surface-brightness structures in nearby early-type galaxies from the MATLAS survey”, *Monthly Notices of the Royal Astronomical Society*, 2020, **498**, 2138–2166.
- (143) N. Amorisco, “On feathers, bifurcations and shells: the dynamics of tidal streams across the mass scale”, *Monthly Notices of the Royal Astronomical Society*, 2015, **450**, 575–591.
- (144) G. S. Karademir, R.-S. Remus, A. Burkert, K. Dolag, T. L. Hoffmann, B. P. Moster, U. P. Steinwandel and J. Zhang, “The outer stellar halos of galaxies: how radial merger mass deposition, shells, and streams depend on infall-orbit configurations”, *Monthly Notices of the Royal Astronomical Society*, 2019, **487**, 318–332.
- (145) A. Klypin, A. V. Kravtsov, O. Valenzuela and F. Prada, “Where Are the Missing Galactic Satellites?”, *The Astrophysical Journal*, 1999, **522**, 82–92.
- (146) B. Moore, S. Ghigna, F. Governato, G. Lake, T. Quinn, J. Stadel and P. Tozzi, “Dark Matter Substructure within Galactic Halos”, *The Astrophysical Journal*, 1999, **524**, L19–L22.
- (147) D. Reed, F. Governato, T. Quinn, J. Gardner, J. Stadel and G. Lake, “Dark matter subhaloes in numerical simulations”, *Monthly Notices of the Royal Astronomical Society*, 2005, **359**, 1537–1548.
- (148) E. J. Tollerud, J. S. Bullock, L. E. Strigari and B. Willman, “Hundreds of Milky Way Satellites? Luminosity Bias in the Satellite Luminosity Function”, *The Astrophysical Journal*, 2008, **688**, 277–289.
- (149) T. Ishiyama, S. Rieder, J. Makino, S. Portegies Zwart, D. Groen, K. Nitadori, C. de Laat, S. McMillan, K. Hiraki and S. Harfst, “The Cosmogrid Simulation: Statistical properties of small dark matter halos”, *The Astrophysical Journal*, 2013, **767**, 146.

- (150) V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg and F. Pearce, “Simulations of the formation, evolution and clustering of galaxies and quasars”, *Nature*, 2005, **435**, 629–636.
- (151) V. Springel, C. S. Frenk and S. D. M. White, “The large-scale structure of the Universe”, *Nature*, 2006, **440**, 1137–1144.
- (152) D. E. S. Collaboration, “The dark energy survey”, *International Journal of Modern Physics A*, 2005, **20**, 3121–3123.
- (153) J. Amiaux, R. Scaramella, Y. Mellier, B. Altieri, C. Burigana, A. D. Silva, P. Gomez, J. Hoar, R. Laureijs, E. Maiorano, D. M. Oliveira, F. Renk, G. S. Criado, I. Tereno, J. L. Auguères, J. Brinchmann, M. Cropper, L. Duvet, A. Ealet, P. Franzetti, B. Garilli, P. Gondoin, L. Guzzo, H. Hoekstra, R. Holmes, K. Jahnke, T. Kitching, M. Meneghetti, W. Percival and S. Warren, *Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*, ed. M. C. Clampin, G. G. Fazio, H. A. MacEwen and J. M. O. Jr., SPIE, 2012, vol. 8442, 84420Z.
- (154) R. S. de Jong, O. Agertz, A. A. Berbel, J. Aird, D. A. Alexander, A. Amarsi, F. Anders, R. Andrae, B. Ansarinejad, W. Ansorge et al., “4MOST: Project overview and information for the First Call for Proposals”, *arXiv preprint arXiv:1903.02464*, 2019.
- (155) J. Forero–Romero, Y. Hoffman, S. Gottlöber, A. Klypin and G. Yepes, “A dynamical classification of the cosmic web”, *Monthly Notices of the Royal Astronomical Society*, 2009, **396**, 1815–1824.
- (156) M. Cautun, R. Van De Weygaert, B. J. Jones and C. S. Frenk, “Evolution of the cosmic web”, *Monthly Notices of the Royal Astronomical Society*, 2014, **441**, 2923–2973.
- (157) N. K. Hayles, in *The Cosmic Web*, Cornell University Press, 2018.
- (158) N. I. Libeskind, R. Van De Weygaert, M. Cautun, B. Falck, E. Tempel, T. Abel, M. Alpaslan, M. A. Aragón-Calvo, J. E. Forero-Romero, R. Gonzalez et al., “Tracing the cosmic web”, *Monthly Notices of the Royal Astronomical Society*, 2018, **473**, 1195–1217.
- (159) D. J. Eisenstein, W. Hu and M. Tegmark, “Cosmic Complementarity:  $H_0$  and  $\Omega_m$  from Combining Cosmic Microwave Background Experiments and Redshift Surveys”, *The Astrophysical Journal Letters*, 1998, **504**, L57–L60.

- (160) A. Meiksin, M. White and J. Peacock, “Baryonic signatures in large-scale structure”, *Monthly Notices of the Royal Astronomical Society*, 1999, **304**, 851–864.
- (161) D. J. Eisenstein, “Dark energy and cosmic sound [review article]”, *New Astronomy Reviews*, 2005, **49**, 360–365.
- (162) D. J. Eisenstein, I. Zehavi, D. W. Hogg, R. Scoccimarro, M. R. Blanton, R. C. Nichol, R. Scranton, H.-J. Seo, M. Tegmark, Z. Zheng, S. F. Anderson, J. Annis, N. Bahcall, J. Brinkmann, S. Burles, F. J. Castander, A. Connolly, I. Csabai, M. Doi, M. Fukugita, J. A. Frieman, K. Glazebrook, J. E. Gunn, J. S. Hendry, G. Hennessy, Z. Ivezić, S. Kent, G. R. Knapp, H. Lin, Y.-S. Loh, R. H. Lupton, B. Margon, T. A. McKay, A. Meiksin, J. A. Munn, A. Pope, M. W. Richmond, D. Schlegel, D. P. Schneider, K. Shimasaku, C. Stoughton, M. A. Strauss, M. SubbaRao, A. S. Szalay, I. Szapudi, D. L. Tucker, B. Yanny and D. G. York, “Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies”, *The Astrophysical Journal*, 2005, **633**, 560–574.
- (163) M. Plionis and S. Basilakos, “The size and shape of local voids”, *Monthly Notices of the Royal Astronomical Society*, 2002, **330**, 399–404.
- (164) R. K. Sheth and R. Van De Weygaert, “A hierarchy of voids: much ado about nothing”, *Monthly Notices of the Royal Astronomical Society*, 2004, **350**, 517–538.
- (165) A. P. Fairall, “Large-Scale Structures in the Distribution of Galaxies”, *Astrophysics and Space Science*, 1995, **230**, 225–235.
- (166) S. D. Landy, “Mapping the Universe.”, *Scientific American*, 1999, **280**, 30–37.
- (167) X.-F. Deng, Y.-Q. Chen, Q. Zhang and J.-Z. He, “Super-large-scale structures in the Sloan Digital Sky Survey”, *Chinese Journal of Astronomy and Astrophysics*, 2006, **6**, 35.
- (168) J. D. Barrow, S. P. Bhavsar and D. Sonoda, “Minimal spanning trees, filaments and galaxy clustering”, *Monthly Notices of the Royal Astronomical Society*, 1985, **216**, 17–35.
- (169) E. Tempel, R. Kipper, E. Saar, M. Bussov, A. Hektor and J. Pelt, “Galaxy filaments as pearl necklaces”, *Astronomy & Astrophysics*, 2014, **572**, A8.
- (170) E. Tempel and N. I. Libeskind, “Galaxy spin alignment in filaments and sheets: observational evidence”, *The Astrophysical Journal Letters*, 2013, **775**, L42.



- (171) E. Tempel and A. Tamm, “Galaxy pairs align with Galactic filaments”, *Astronomy & Astrophysics*, 2015, **576**, L5.
- (172) A. R. Zentner, A. V. Kravtsov, O. Y. Gnedin and A. A. Klypin, “The Anisotropic Distribution of Galactic Satellites”, *The Astrophysical Journal*, 2005, **629**, 219–232.
- (173) Y.-S. Li and A. Helmi, “Group infall of substructures on to a Milky Way-like dark halo”, *Proceedings of the International Astronomical Union*, 2008, **4**, 263–268.
- (174) J. I. Read, G. Lake, O. Agertz and V. P. Debattista, “Thin, thick and dark discs in  $\Lambda$ CDM”, *Monthly Notices of the Royal Astronomical Society*, 2008, **389**, 1041–1057.
- (175) E. D’Onghia and G. Lake, “Small Dwarf Galaxies within Larger Dwarfs: Why Some Are Luminous while Most Go Dark”, *The Astrophysical Journal Letters*, 2008, **686**, L61.
- (176) N. I. Libeskind, A. Knebe, Y. Hoffman, S. Gottlöber, G. Yepes and M. Steinmetz, “The preferred direction of infalling satellite galaxies in the Local Group”, *Monthly Notices of the Royal Astronomical Society*, 2011, **411**, 1525–1535.
- (177) N. I. Libeskind, Y. Hoffman, R. B. Tully, H. M. Courtois, D. Pomarède, S. Gottlöber and M. Steinmetz, “Planes of satellite galaxies and the cosmic web”, *Monthly Notices of the Royal Astronomical Society*, 2015, **452**, 1052–1059.
- (178) G. O. Abell, “The distribution of rich clusters of galaxies.”, *The Astrophysical Journal Supplement Series*, 1958, **3**, 211.
- (179) R. Carlberg, H. Yee and E. Ellingson, “The average mass and light profiles of galaxy clusters”, *The Astrophysical Journal*, 1997, **478**, 462.
- (180) R. Carlberg, H. Yee, E. Ellingson, S. Morris, R. Abraham, P. Gravel, C. Pritchet, T. Smecker-Hane, F. Hartwick, J. Hesser et al., “The average mass profile of galaxy clusters”, *The Astrophysical Journal*, 1997, **485**, L13.
- (181) A. V. Kravtsov and S. Borgani, “Formation of galaxy clusters”, *Annual Review of Astronomy and Astrophysics*, 2012, **50**, 353–409.
- (182) L. P. Bautz and W. W. Morgan, “On the Classification of the Forms of Clusters of Galaxies”, *The Astrophysical Journal Letters*, 1970, **162**, L149.
- (183) L. P. Bautz and W. W. Morgan, *Bulletin of the American Astronomical Society*, 1970, vol. 2, p. 294.

- (184) I. Karachentsev, “The local group and other neighboring galaxy groups”, *The Astronomical Journal*, 2005, **129**, 178.
- (185) X. Yang, H. Mo, F. C. Van den Bosch, A. Pasquali, C. Li and M. Barden, “Galaxy groups in the SDSS DR4. I. The catalog and basic properties”, *The Astrophysical Journal*, 2007, **671**, 153.
- (186) R. B. Tully, “Galaxy Groups”, *The Astronomical Journal*, 2015, **149**, 54.
- (187) S. Mitra, *The World of Galaxies*, ed. H. G. Corwin and L. Bottinelli, Springer US, New York, NY, 1989, pp. 426–427.
- (188) T. Hasegawa, K.-i. Wakamatsu, M. Malkan, K. Sekiguchi, J. W. Menzies, Q. A. Parker, J. Jugaku, H. Karoji and S. Okamura, “Large-scale structure of galaxies in the Ophiuchus region”, *Monthly Notices of the Royal Astronomical Society*, 2000, **316**, 326–344.
- (189) F. X. Hu, G. X. Wu, G. X. Song, Q. R. Yuan and S. Okamura, “Orientation of Galaxies in the Local Supercluster: A Review”, *Astrophysics and Space Science*, 2006, **302**, 43–59.
- (190) R. B. Tully, H. Courtois, Y. Hoffman and D. Pomarède, “The Laniakea supercluster of galaxies”, *Nature*, 2014, **513**, 71–73.
- (191) J. Czekanowski, *Forschungen im Nil-Kongo-Zwischengebiet: Ethnographisch-antropologischer Atlas, Zwischenseen-Bantu; Pygmäen und Pygmoiden; Urwaldstämme*, Klinkhardt, 1911, vol. 3.
- (192) H. E. Driver and A. L. Kroeber, *Quantitative expression of cultural relationships*, Berkeley: University of California Press, 1932, vol. 31.
- (193) W. Stephenson, “The inverted factor technique”, *British Journal of Psychology*, 1936, **26**, 344.
- (194) J. Zubin, “A technique for measuring like-mindedness.”, *The Journal of Abnormal and Social Psychology*, 1938, **33**, 508.
- (195) R. C. Tryon, “Cluster analysis: correlation profile and orthometric analysis for the isolation of unities in mind and personality”, *Ann Arbor: Edward Brothers*, 1939.
- (196) R. B. Cattell, “The description of personality: Basic traits resolved into clusters.”, *The journal of abnormal and social psychology*, 1943, **38**, 476.
- (197) R. R. Sokal and P. Sneath, “Principles of numerical taxonomy”, *CA and London: WH Freeman*, 1963.
- (198) P. H. Sneath and R. R. Sokal, “Numerical taxonomy”, 1973.

- (199) L. Stark, J. F. Dickson, G. H. Whipple and H. Horibe, "REMOTE REAL-TIME DIAGNOSIS OF CLINICAL ELECTROCARDIOGRAMS BY A DIGITAL COMPUTER SYSTEM", *Annals of the New York Academy of Sciences*, 1965, **126**, 851–872.
- (200) R. T. Manning and L. Watson, "Signs, symptoms, and systematics", *JAMA*, 1966, **198**, 1180–1184.
- (201) D. Baron and P. M. Fraser, "Medical applications of taxonomic methods", *British medical bulletin*, 1968, **24**, 236–240.
- (202) R. KNUSMAN and M. TOELLER, *DIABETOLOGIA*, 1972, vol. 8, p. 53.
- (203) M. Lorr, C. J. Klett and D. M. McNair, "Syndromes of psychosis.", 1963.
- (204) E. S. Paykel, G. L. Klerman and B. A. Prusoff, "Treatment setting and clinical depression", *Archives of General Psychiatry*, 1970, **22**, 11–21.
- (205) B. Everitt, A. Gourlay and R. Kendell, "An attempt at validation of traditional psychiatric syndromes by cluster analysis", *The British Journal of Psychiatry*, 1971, **119**, 399–412.
- (206) J. Hautaluoma, "Syndromes, antecedents, and outcomes of psychosis: A cluster-analytic study.", *Journal of Consulting and Clinical Psychology*, 1971, **37**, 332.
- (207) F. R. Hodson, "Searching for structure within multivariate archaeological data", *World Archaeology*, 1969, **1**, 90–105.
- (208) F. R. Hodson, "Cluster analysis and archaeology: some new developments and applications", *World archaeology*, 1970, **1**, 299–320.
- (209) B. F. King, "Market and industry factors in stock price behavior", *the Journal of Business*, 1966, **39**, 139–190.
- (210) F. Goronzy, "A numerical taxonomy of business enterprises", *Numerical taxonomy*, 1969, 42–52.
- (211) I. Dyen, A. James and J. Cole, "Language divergence and estimated word retention rate", *Language*, 1967, 150–171.
- (212) J. B. Weaver and S. W. Hess, "A procedure for nonpartisan districting: Development of computer techniques", *Yale LJ*, 1963, **73**, 288.
- (213) H. F. Kaiser, "An objective method for establishing legislative districts", *Midwest Journal of Political Science*, 1966, **10**, 200–213.
- (214) J. A. Hartigan, *Clustering algorithms*, John Wiley & Sons, Inc., 1975.

- (215) E. Čech and M. Katětov, “Chapter II: General metric spaces”, *Point Sets*, 1969, 42–91.
- (216) R. Hamming, “Error detecting and error correcting codes”, *Bell System Technical Journal*, 1950, **29**.
- (217) C. D. Pilcher, J. K. Wong and S. K. Pillai, “Inferring HIV transmission dynamics from phylogenetic sequence relationships”, *PLoS medicine*, 2008, **5**, e69.
- (218) C. Lee, “Some properties of nonbinary error-correcting codes”, *IRE Transactions on Information Theory*, 1958, **4**, 77–82.
- (219) V. I. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions and Reversals”, *Soviet Physics Doklady*, 1966, **10**, 707.
- (220) P. C. Mahalanobis, 1936.
- (221) L. N. Vaserstein, “Markov processes over denumerable products of spaces, describing large systems of automata”, *Problemy Peredachi Informatsii*, 1969, **5**, 64–72.
- (222) S. Kullback and R. A. Leibler, “On Information and Sufficiency”, *The Annals of Mathematical Statistics*, 1951, **22**, 79–86.
- (223) S. Kullback, “Information theory and statistics. John Riley and sons”, *Inc. New York*, 1959.
- (224) E. Hellinger, “Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen.”, *Journal für die reine und angewandte Mathematik*, 1909, **1909**, 210–271.
- (225) K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus and S. Zubrzycki, *Colloquium mathematicum*, 1951, vol. 2, pp. 282–285.
- (226) K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus and S. Zubrzycki, “Taksonomia wrocławska”, *Przegląd Antropologiczny*, 1951, **17**, 193–211.
- (227) R. Sibson, “SLINK: An optimally efficient algorithm for the single-link cluster method”, *The Computer Journal*, 1973, **16**, 30–34.
- (228) D. Defays, “An efficient algorithm for a complete link method”, *The Computer Journal*, 1977, **20**, 364–366.
- (229) R. R. Sokal, “A statistical method for evaluating systematic relationships.”, *Univ. Kansas, Sci. Bull.*, 1958, **38**, 1409–1438.

- (230) *Hierarchical Clustering / Dendrogram: Simple Definition, Examples*, <https://www.statisticshowto.com/hierarchical-clustering/>, Accessed: 13-06-2022.
- (231) J. MacQueen et al., Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967, vol. 1, pp. 281–297.
- (232) S. Lloyd, “Least squares quantization in PCM”, *IEEE transactions on information theory*, 1982, **28**, 129–137.
- (233) S. Vassilvitskii and D. Arthur, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2006, pp. 1027–1035.
- (234) L. Kaufman and P. J. Rousseeuw, “Clustering large applications (Program CLARA)”, *Finding groups in data: an introduction to cluster analysis*, 2008, 126–146.
- (235) A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice-Hall, Inc., 1988.
- (236) P. Bradley, O. Mangasarian and W. Street, “Clustering via concave minimization”, *Advances in neural information processing systems*, 1996, **9**.
- (237) J. C. Dunn, “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters”, *Journal of Cybernetics*, 1973, **3**, 32–57.
- (238) A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977, **39**, 1–22.
- (239) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.
- (240) *Comparing different clustering algorithms on toy datasets*, [https://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_cluster\\_comparison.html](https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html), Accessed: 13-06-2022.
- (241) M. Ester, H.-P. Kriegel, J. Sander and X. Xu, *Kdd*, 1996, vol. 96, pp. 226–231.
- (242) M. Ankerst, M. M. Breunig, H.-P. Kriegel and J. Sander, *ACM Sigmod record*, ACM, 1999, vol. 28, pp. 49–60.

- (243) E. Achtert, C. Böhm and P. Kröger, Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2006, pp. 119–128.
- (244) R. J. G. B. Campello, D. Moulavi, A. Zimek and J. Sander, “Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection”, *ACM Trans. Knowl. Discov. Data*, 2015, **10**, DOI: [10.1145/2733381](https://doi.org/10.1145/2733381).
- (245) L. McInnes, J. Healy and S. Astels, “hdbscan: Hierarchical density based clustering”, *Journal of Open Source Software*, 2017, **2**, 205.
- (246) K. Fukunaga and L. Hostetler, “The estimation of the gradient of a density function, with applications in pattern recognition”, *IEEE Transactions on information theory*, 1975, **21**, 32–40.
- (247) M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, Springer Berlin Heidelberg, 1999, pp. 262–270.
- (248) O. Alghushairy, R. Alsini, T. Soule and X. Ma, “A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams”, *Big Data and Cognitive Computing*, 2021, **5**, DOI: [10.3390/bdcc5010001](https://doi.org/10.3390/bdcc5010001).
- (249) K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space”, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901, **2**, 559–572.
- (250) H. Hotelling, “Analysis of a complex of statistical variables with principal components”, *J. Educ. Psy.*, 1933, **24**, 498–520.
- (251) J. Héroult, “Réseaux de neurones à synapses modifiables: décodage de messages sensoriels composites par un apprentissage non supervisé et permanent”, *CR Acad. Sci. Paris*, 1984, 525–528.
- (252) B. Ans, J. Héroult and C. Jutten, “Architectures neuromimétiques adaptatives: Détection de primitives”, *Proceedings of Cognitiva*, 1985, **85**, 593–597.
- (253) J. Héroult, C. Jutten and B. Ans, 10 Colloque sur le traitement du signal et des images, FRA, 1985, 1985.
- (254) J. Héroult and C. Jutten, AIP conference proceedings, 1986, vol. 151, pp. 206–211.
- (255) P. Comon, “Independent component analysis, a new concept?”, *Signal processing*, 1994, **36**, 287–314.

- (256) S. Sharma and M. Steinmetz, “Multidimensional density estimation and phase-space structure of dark matter haloes”, *Monthly Notices of the Royal Astronomical Society*, 2006, **373**, 1293–1307.
- (257) L. McInnes, J. Healy and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction”, *arXiv preprint arXiv:1802.03426*, 2018.
- (258) L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE.”, *Journal of machine learning research*, 2008, **9**.
- (259) J. L. Bentley, “Multidimensional Binary Search Trees Used for Associative Searching”, *Commun. ACM*, 1975, **18**, 509–517.
- (260) S. M. Omohundro, *Five balltree construction algorithms*, International Computer Science Institute Berkeley, 1989.
- (261) A. Guttman, “R-Trees: A Dynamic Index Structure for Spatial Searching”, *SIGMOD Rec.*, 1984, **14**, 47–57.
- (262) J. Johnson, M. Douze and H. Jégou, “Billion-scale similarity search with GPUs”, *IEEE Transactions on Big Data*, 2019, **7**, 535–547.
- (263) J. Nocedal and S. J. Wright, *Numerical optimization*, Springer, 1999.
- (264) C. E. Shannon, “A mathematical theory of communication”, *The Bell System Technical Journal*, 1948, **27**, 379–423.
- (265) J. A. Nelder and R. Mead, “A Simplex Method for Function Minimization”, *The Computer Journal*, 1965, **7**, 308–313.
- (266) A. Cauchy et al., “Méthode générale pour la résolution des systemes d’équations simultanées”, *Comp. Rend. Sci. Paris*, 1847, **25**, 536–538.
- (267) J. Hadamard, *Mémoire sur le problème d’analyse relatif à l’équilibre des plaques élastiques encastrées*, Imprimerie nationale, 1908, vol. 33.
- (268) C. G. Broyden, “The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations”, *IMA Journal of Applied Mathematics*, 1970, **6**, 76–90.
- (269) R. Fletcher, “A new approach to variable metric algorithms”, *The Computer Journal*, 1970, **13**, 317–322.
- (270) D. Goldfarb, “A family of variable-metric methods derived by variational means”, *Mathematics of computation*, 1970, **24**, 23–26.
- (271) D. F. Shanno, “Conditioning of quasi-Newton methods for function minimization”, *Mathematics of computation*, 1970, **24**, 647–656.

- (272) B. Olson, I. Hashmi, K. Molloy and A. Shehu, “Basin Hopping as a General and Versatile Optimization Framework for the Characterization of Biological Macromolecules.”, *Advances in Artificial Intelligence (16877470)*, 2012.
- (273) M. Pincus, “A Monte Carlo method for the approximate solution of certain types of constrained optimization problems”, *Operations research*, 1970, **18**, 1225–1228.
- (274) A. Khachaturyan, S. Semenovskaya and B. Vainstein, “Statistical thermodynamic approach to determination of structure amplitude phases”, *Sov. Phys. Crystallography*, 1979, **24**, 519–524.
- (275) A. Khachaturyan, S. Semenovskaya and B. Vainshtein, “The thermodynamic approach to the structure analysis of crystals”, *Acta Crystallographica Section A*, 1981, **37**, 742–754.
- (276) S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, “Optimization by Simulated Annealing”, *Science*, 1983, **220**, 671–680.
- (277) E. Fix and J. L. Hodges, “Nonparametric discrimination: consistency properties”, *Randolph Field, Texas, Project*, 1951, 21–49.
- (278) T. Cover and P. Hart, “Nearest neighbor pattern classification”, *IEEE transactions on information theory*, 1967, **13**, 21–27.
- (279) L. Devroye, L. Györfi and G. Lugosi, *A probabilistic theory of pattern recognition*, Springer Science & Business Media, 2013, vol. 31.
- (280) S. Guha, N. Mishra, R. Motwani and L. O’Callaghan, Proceedings 41st Annual Symposium on Foundations of Computer Science, 2000, pp. 359–366.
- (281) T. Zhang, R. Ramakrishnan and M. Livny, Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Association for Computing Machinery, Montreal, Quebec, Canada, 1996, pp. 103–114.
- (282) D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979, **PAMI-1**, 224–227.
- (283) J. C. Dunn, “Well-Separated Clusters and Optimal Fuzzy Partitions”, *Journal of Cybernetics*, 1974, **4**, 95–104.
- (284) P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”, *Journal of Computational and Applied Mathematics*, 1987, **20**, 53–65.



- (285) M. Hassani and T. Seidl, “Using internal evaluation measures to validate the quality of diverse stream clustering algorithms”, *Vietnam Journal of Computer Science*, 2017, **4**, 171–183.
- (286) D. L. Olson and D. Delen, *Advanced data mining techniques*, Springer Science & Business Media, 2008.
- (287) P. Jaccard, “The distribution of the flora in the alpine zone. 1”, *New phytologist*, 1912, **11**, 37–50.
- (288) N. X. Vinh, J. Epps and J. Bailey, Proceedings of the 26th Annual International Conference on Machine Learning, Association for Computing Machinery, Montreal, Quebec, Canada, 2009, pp. 1073–1080.
- (289) M. Meilă, “Comparing clusterings—an information based distance”, *Journal of Multivariate Analysis*, 2007, **98**, 873–895.
- (290) M. I. Arifyanto and B. Fuchs, “Fine structure in the phase space distribution of nearby subdwarfs”, *Astronomy and Astrophysics*, 2006, **449**, 533–538.
- (291) S. Duffau, R. Zinn, A. K. Vivas, G. Carraro, R. A. Méndez, R. Winnick and C. Gallart, “Spectroscopy of QUEST RR Lyrae Variables: The New Virgo Stellar Stream”, *The Astrophysical Journal Letters*, 2006, **636**, L97–L100.
- (292) M. E. K. Williams, M. Steinmetz, S. Sharma, J. Bland-Hawthorn, R. S. de Jong, G. M. Seabroke, A. Helmi, K. C. Freeman, J. Binney, I. Minchev, O. Bienaymé, R. Campbell, J. P. Fulbright, B. K. Gibson, G. F. Gilmore, E. K. Grebel, U. Munari, J. F. Navarro, Q. A. Parker, W. Reid, A. Siebert, A. Siviero, F. G. Watson, R. F. G. Wyse and T. Zwitter, “The dawning of the stream of Aquarius in RAVE”, *The Astrophysical Journal*, 2011, **728**, 102.
- (293) Helmi, Amina, Veljanoski, Jovan, Breddels, Maarten A., Tian, Hao and Sales, Laura V., “A box full of chocolates: The rich structure of the nearby stellar halo revealed by Gaia and RAVE”, *A&A*, 2017, **598**, A58.
- (294) W. H. Press and P. Schechter, “Formation of galaxies and clusters of galaxies by self-similar gravitational condensation”, *The Astrophysical Journal*, 1974, **187**, 425–438.
- (295) M. Davis, G. Efstathiou, C. S. Frenk and S. D. White, “The evolution of large-scale structure in a universe dominated by cold dark matter”, *The Astrophysical Journal*, 1985, **292**, 371–394.
- (296) E. Bertschinger and J. M. Gelb, “Cosmological N-body simulations”, *Computers in Physics*, 1991, **5**, 164–175.

- (297) J. M. Gelb and E. Bertschinger, “Cold Dark Matter. I. The Formation of Dark Halos”, *The Astrophysical Journal*, 1994, **436**, 467.
- (298) E. van Kampen, “Improved numerical modelling of clusters of galaxies”, *Monthly Notices of the Royal Astronomical Society*, 1995, **273**, 295–327.
- (299) D. W. Pfitzner and J. K. Salmon, KDD, 1996, pp. 26–31.
- (300) A. Klypin and J. Holtzman, “Particle-Mesh code for cosmological simulations”, *arXiv preprint astro-ph/9712217*, 1997.
- (301) D. J. Eisenstein and P. Hut, “HOP: a new group-finding algorithm for N-body simulations”, *The Astrophysical Journal*, 1998, **498**, 137.
- (302) A. Klypin, S. Gottlober, A. V. Kravtsov and A. M. Khokhlov, “Galaxies in N-Body Simulations: Overcoming the Overmerging Problem”, 1999, **516**, 530–551.
- (303) F. Governato, B. Moore, R. Cen, J. Stadel, G. Lake and T. Quinn, “The Local Group as a test of cosmological models”, *New Astronomy*, 1997, **2**, 91–106.
- (304) J. Stadel, Ph.D. Thesis, 2001.
- (305) J. S. Bullock, T. S. Kolatt, Y. Sigad, R. S. Somerville, A. V. Kravtsov, A. A. Klypin, J. R. Primack and A. Dekel, “Profiles of dark haloes: evolution, scatter and environment”, *Monthly Notices of the Royal Astronomical Society*, 2001, **321**, 559–575.
- (306) V. Springel, S. D. M. White, G. Tormen and G. Kauffmann, “Populating a cluster of galaxies – I. Results at  $z = 0$ ”, *Monthly Notices of the Royal Astronomical Society*, 2001, **328**, 726–750.
- (307) S. P. D. Gill, A. Knebe and B. K. Gibson, “The evolution of substructure - I. A new identification method”, *Monthly Notices of the Royal Astronomical Society*, 2004, **351**, 399–409.
- (308) D. Aubert, C. Pichon and S. Colombi, “The origin and implications of dark matter anisotropic cosmic infall on L haloes”, *Monthly Notices of the Royal Astronomical Society*, 2004, **352**, 376–398.
- (309) M. C. Neyrinck, A. J. S. Hamilton and N. Y. Gnedin, “Understanding the PSCz galaxy power spectrum with N-body simulations”, *Monthly Notices of the Royal Astronomical Society*, 2004, **348**, 1–11.

- (310) G. Tormen, L. Moscardini and N. Yoshida, “Properties of cluster satellites in hydrodynamical simulations”, *Monthly Notices of the Royal Astronomical Society*, 2004, **350**, 1397–1408.
- (311) C. Giocoli, G. Tormen and F. C. van den Bosch, “The population of dark matter subhaloes: mass functions and average mass-loss rates”, *Monthly Notices of the Royal Astronomical Society*, 2008, **386**, 2135–2144.
- (312) J. Weller, J. P. Ostriker, P. Bode and L. Shaw, “Fast identification of bound structures in large N-body simulations”, *Monthly Notices of the Royal Astronomical Society*, 2005, **364**, 823–832.
- (313) M. C. Neyrinck, N. Y. Gnedin and A. J. Hamilton, “VOBOZ: an almost-parameter-free halo-finding algorithm”, *Monthly Notices of the Royal Astronomical Society*, 2005, **356**, 1222–1232.
- (314) J. Kim and C. Park, “A new halo-finding method for n-body simulations”, *The Astrophysical Journal*, 2006, **639**, 600.
- (315) J. Diemand, M. Kuhlen and P. Madau, “Early supersymmetric cold dark matter substructure”, *The Astrophysical Journal*, 2006, **649**, 1.
- (316) L. D. Shaw, J. Weller, J. P. Ostriker and P. Bode, “The bound mass of substructures in dark matter halos”, *The Astrophysical Journal*, 2007, **659**, 1082.
- (317) J. Gardner, A. Connolly and C. McBride, *Broadening Participation in the TeraGrid Enabling Knowledge Discovery in a Virtual Universe*, 2007.
- (318) J. P. Gardner, A. Connolly and C. McBride, Proceedings of the 5th IEEE workshop on Challenges of large applications in distributed environments, 2007, pp. 1–10.
- (319) M. Maciejewski, S. Colombi, V. Springel, C. Alard and F. R. Bouchet, “Phase-space structures - II. Hierarchical Structure Finder”, *Monthly Notices of the Royal Astronomical Society*, 2009, **396**, 1329–1348.
- (320) S. Habib, A. Pope, Z. Lukić, D. Daniel, P. Fasel, N. Desai, K. Heitmann, C.-H. Hsu, L. Ankeny, G. Mark, S. Bhattacharya and J. Ahrens, *Journal of Physics Conference Series*, 2009, vol. 180, 012019, p. 012019.
- (321) S. R. Knollmann and A. Knebe, “AHF: Amiga’s Halo Finder”, *Astrophysical Journal Supplement*, 2009, **182**, 608–624.
- (322) S. Skory, M. J. Turk, M. L. Norman and A. L. Coil, “Parallel HOP: A Scalable Halo Finder for Massive Cosmological Data Sets”, *Astrophysical Journal Supplement*, 2010, **191**, 43–57.

- (323) S. Planelles and V. Quilis, “ASOHF: a new adaptive spherical overdensity halo finder”, *Astronomy and Astrophysics*, 2010, **519**, A94.
- (324) P. M. Sutter and P. M. Ricker, “Examining Subgrid Models of Supermassive Black Holes in Cosmological Simulation”, *The Astrophysical Journal*, 2010, **723**, 1308–1318.
- (325) Y. Rasera, J. .-. Alimi, J. Courtin, F. Roy, P. .-. Corasaniti, A. Füzfa and V. Boucher, *Invisible Universe*, ed. J.-M. Alimi and A. Fuözfa, 2010, vol. 1241, pp. 1134–1139.
- (326) J. Courtin, Y. Rasera, J. .-. Alimi, P. .-. Corasaniti, V. Boucher and A. Füzfa, “Imprints of dark energy on cosmic structure formation - II. Non-universality of the halo mass function”, *Monthly Notices of the Royal Astronomical Society*, 2011, **410**, 1911–1931.
- (327) B. L. Falck, M. C. Neyrinck and A. S. Szalay, “ORIGAMI: Delineating Halos Using Phase-space Folds”, *The Astrophysical Journal*, 2012, **754**, 126.
- (328) M. Sgró, A. Ruiz and M. Merchán, “Hierarchical Friend-of-Friend algorithm to extract substructures from dark matter halos”, *Boletín de la Asociación Argentina de Astronomía La Plata Argentina*, 2010, **53**, 43–46.
- (329) C. Giocoli, G. Tormen, R. K. Sheth and F. C. van den Bosch, “The substructure hierarchy in dark matter haloes”, *Monthly Notices of the Royal Astronomical Society*, 2010, **404**, 502–517.
- (330) P. J. Elahi, R. J. Thacker and L. M. Widrow, “Peaks above the Maxwellian Sea: a new approach to finding substructures in N-body haloes”, *Monthly Notices of the Royal Astronomical Society*, 2011, **418**, 320–335.
- (331) P. S. Behroozi, R. H. Wechsler and H.-Y. Wu, “The rockstar phase-space temporal halo finder and the velocity offsets of cluster cores”, *The Astrophysical Journal*, 2012, **762**, 109.
- (332) J. Han, Y. P. Jing, H. Wang and W. Wang, “Resolving subhaloes’ lives with the Hierarchical Bound-Tracing algorithm”, *Monthly Notices of the Royal Astronomical Society*, 2012, **427**, 2437–2449.
- (333) A. Knebe, S. R. Knollmann, S. I. Muldrew, F. R. Pearce, M. A. Aragon-Calvo, Y. Ascasibar, P. S. Behroozi, D. Ceverino, S. Colombi, J. Diemand, K. Dolag, B. L. Falck, P. Fasel, J. Gardner, S. Gottlöber, C.-H. Hsu, F. Iannuzzi, A. Klypin, Z. Lukić, M. Maciejewski, C. McBride, M. C. Neyrinck, S. Planelles, D. Potter, V. Quilis, Y. Rasera, J. I. Read, P. M. Ricker, F. Roy, V. Springel, J. Stadel, G. Stinson, P. M. Sutter, V. Turchaninov, D. Tweed, G. Yepes

- and M. Zemp, “Haloes gone MAD14: The Halo-Finder Comparison Project”, *Monthly Notices of the Royal Astronomical Society*, 2011, **415**, 2293–2318.
- (334) A. Knebe, N. I. Libeskind, F. Pearce, P. Behroozi, J. Casado, K. Dolag, R. Dominguez-Tenreiro, P. Elahi, H. Lux, S. I. Muldrew et al., “Galaxies going MAD: the galaxy-finder comparison project”, *Monthly Notices of the Royal Astronomical Society*, 2013, **428**, 2039–2052.
- (335) P. J. Elahi, J. Han, H. Lux, Y. Ascasibar, P. Behroozi, A. Knebe, S. I. Muldrew, J. Onions and F. Pearce, “Streams going Notts: the tidal debris finder comparison project”, *Monthly Notices of the Royal Astronomical Society*, 2013, **433**, 1537–1555.
- (336) A. Knebe, F. R. Pearce, H. Lux, Y. Ascasibar, P. Behroozi, J. Casado, C. C. Moran, J. Diemand, K. Dolag, R. Dominguez-Tenreiro, P. Elahi, B. Falck, S. Gottlöber, J. Han, A. Klypin, Z. Lukić, M. Maciejewski, C. K. McBride, M. E. Merchán, S. I. Muldrew, M. Neyrinck, J. Onions, S. Planelles, D. Potter, V. Quilis, Y. Rasera, P. M. Ricker, F. Roy, A. N. Ruiz, M. A. Sgró, V. Springel, J. Stadel, P. M. Sutter, D. Tweed and M. Zemp, “Structure finding in cosmological simulations: the state of affairs”, *Monthly Notices of the Royal Astronomical Society*, 2013, **435**, 1618–1658.
- (337) J. Han, S. Cole, C. S. Frenk, A. Benitez-Llambay and J. Helly, “hbt+: an improved code for finding subhaloes and building merger trees in cosmological simulations”, *Monthly Notices of the Royal Astronomical Society*, 2017, **474**, 604–617.
- (338) P. J. Elahi, R. Canas, R. J. J. Poulton, R. J. Tobar, J. S. Willis, C. d. P. Lagos, C. Power and A. S. G. Robotham, “Hunting for galaxies and halos in simulations with VELOCIRaptor”, *Publications of the Astronomical Society of Australia*, 2019, **36**, e021.
- (339) S.-P. Sun, S.-H. Liao, Q. Guo, Q. Wang and L. Gao, “HIKER: a halo-finding method based on kernel-shift algorithm”, *Research in Astronomy and Astrophysics*, 2020, **20**, 046.
- (340) R. Mondal, S. Bharadwaj, S. Majumdar, A. Bera and A. Acharyya, *FoF-Halo-finder: Halo location and size*, Astrophysics Source Code Library, record ascl:2107.004, 2021.
- (341) B. Hadzhiyska, D. Eisenstein, S. Bose, L. H. Garrison and N. Maksimova, “compaso: A new halo finder for competitive assignment to spherical overdensities”, *Monthly Notices of the Royal Astronomical Society*, 2021, **509**, 501–521.

- (342) J. Onions, A. Knebe, F. R. Pearce, S. I. Muldrew, H. Lux, S. R. Knollmann, Y. Ascasibar, P. Behroozi, P. Elahi, J. Han, M. Maciejewski, M. E. Merchán, M. Neyrinck, A. N. Ruiz, M. A. Sgró, V. Springel and D. Tweed, “Subhaloes going Notts: the subhalo-finder comparison project”, *Monthly Notices of the Royal Astronomical Society*, 2012, **423**, 1200–1214.
- (343) J. Onions, Y. Ascasibar, P. Behroozi, J. Casado, P. Elahi, J. Han, A. Knebe, H. Lux, M. E. Merchán, S. I. Muldrew, M. Neyrinck, L. Old, F. R. Pearce, D. Potter, A. N. Ruiz, M. A. Sgró, D. Tweed and T. Yue, “Subhaloes gone Notts: spin across subhaloes and finders”, *Monthly Notices of the Royal Astronomical Society*, 2013, **429**, 2739–2747.
- (344) S. Avila, A. Knebe, F. R. Pearce, A. Schneider, C. Srisawat, P. A. Thomas, P. Behroozi, P. J. Elahi, J. Han, Y.-Y. Mao, J. Onions, V. Rodriguez-Gomez and D. Tweed, “SUSSING MERGER TREES: the influence of the halo finder”, *Monthly Notices of the Royal Astronomical Society*, 2014, **441**, 3488–3501.
- (345) J. Lee, S. K. Yi, P. J. Elahi, P. A. Thomas, F. R. Pearce, P. Behroozi, J. Han, J. Helly, I. Jung, A. Knebe, Y.-Y. Mao, J. Onions, V. Rodriguez-Gomez, A. Schneider, C. Srisawat and D. 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*, 2014, **445**, 4197–4210.
- (346) P. Behroozi, A. Knebe, F. R. Pearce, P. Elahi, J. Han, H. Lux, Y.-Y. Mao, S. I. Muldrew, D. Potter and C. Srisawat, “Major mergers going Notts: challenges for modern halo finders”, *Monthly Notices of the Royal Astronomical Society*, 2015, **454**, 3020–3029.
- (347) C. M. Rockosi, M. Odenkirchen, E. K. Grebel, W. Dehnen, K. M. Cudworth, J. E. Gunn, D. G. York, J. Brinkmann, G. S. Hennessy and Ž. Ivezić, “A Matched-Filter Analysis of the Tidal Tails of the Globular Cluster Palomar 5”, *The Astronomical Journal*, 2002, **124**, 349–363.
- (348) N. Wiener, N. Wiener, C. Mathematician, N. Wiener, N. Wiener and C. Mathématicien, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, MIT press Cambridge, MA, 1949, vol. 113.
- (349) J. Kepner, X. Fan, N. Bahcall, J. Gunn, R. Lupton and G. Xu, “An automated cluster finder: the adaptive matched filter”, *The Astrophysical Journal*, 1999, **517**, 78.

- (350) L. R. Doyle, H. J. Deeg, V. P. Kozhevnikov, B. Oetiker, E. L. Martin, J. E. Blue, L. Rottler, R. P. S. Stone, Z. Ninkov, J. M. Jenkins, J. Schneider, E. W. Dunham, M. F. Doyle and E. Paleologou, “Observational Limits on Terrestrial-sized Inner Planets around the CM Draconis System Using the Photometric Transit Method with a Matched-Filter Algorithm”, *The Astrophysical Journal*, 2000, **535**, 338–349.
- (351) C. Mateu, G. Bruzual, L. Aguilar, A. G. A. Brown, O. Valenzuela, L. Carigi, H. Velázquez and F. Hernández, “Detection of satellite remnants in the Galactic Halo with Gaia – II. A modified great circle cell method”, *Monthly Notices of the Royal Astronomical Society*, 2011, **415**, 214–224.
- (352) E. Balbinot, B. X. Santiago, L. N. da Costa, M. Makler and M. A. G. Maia, “The tidal tails of NGC 2298”, *Monthly Notices of the Royal Astronomical Society*, 2011, **416**, 393–402.
- (353) C. Mateu, J. I. Read and D. Kawata, “Fourteen candidate RR Lyrae star streams in the inner Galaxy”, *Monthly Notices of the Royal Astronomical Society*, 2017, **474**, 4112–4129.
- (354) K. Malhan and R. A. Ibata, “STREAMFINDER – I. A new algorithm for detecting stellar streams”, *Monthly Notices of the Royal Astronomical Society*, 2018, **477**, 4063–4076.
- (355) Z. Yuan, J. Chang, P. Banerjee, J. Han, X. Kang and M. C. Smith, “StarGO: A New Method to Identify the Galactic Origins of Halo Stars”, *The Astrophysical Journal*, 2018, **863**, 26.
- (356) T. Kohonen, *Self-Organizing Maps*, 2001.
- (357) S. Pearson, S. E. Clark, A. J. Demirjian, K. V. Johnston, M. K. Ness, T. K. Starkenburg, B. F. Williams and R. A. Ibata, “The Hough Stream Spotter: A New Method for Detecting Linear Structure in Resolved Stars and Application to the Stellar Halo of M31”, *The Astrophysical Journal*, 2022, **926**, 166.
- (358) D. Shih, M. R. Buckley, L. Necib and J. Tamasas, “via machinae: Searching for stellar streams using unsupervised machine learning”, *Monthly Notices of the Royal Astronomical Society*, 2021, **509**, 5992–6007.
- (359) S. S. Lövdal, T. Ruiz-Lara, H. H. Koppelman, T. Matsuno, E. Dodd and A. Helmi, “Substructure in the stellar halo near the Sun. I. Data-driven clustering in Integrals of Motion space”, *arXiv preprint arXiv:2201.02404*, 2022.

- (360) T. Prusti, J. De Bruijne, A. G. Brown, A. Vallenari, C. Babusiaux, C. Bailer-Jones, U. Bastian, M. Biermann, D. W. Evans, L. Eyer et al., “The Gaia mission”, *Astronomy & Astrophysics*, 2016, **595**, A1.
- (361) J. A. Kollmeier, G. Zasowski, H.-W. Rix, M. Johns, S. F. Anderson, N. Drory, J. A. Johnson, R. W. Pogge, J. C. Bird, G. A. Blanc et al., “SDSS-V: pioneering panoptic spectroscopy”, *arXiv preprint arXiv:1711.03234*, 2017.
- (362) A. W. McConnachie, M. J. Irwin, R. A. Ibata, J. Dubinski, L. M. Widrow, N. F. Martin, P. Côté, A. L. Dotter, J. F. Navarro, A. M. N. Ferguson, T. H. Puzia, G. F. Lewis, A. Babul, P. Barmby, O. Bienaymé, S. C. Chapman, R. Cockcroft, M. L. M. Collins, M. A. Fardal, W. E. Harris, A. Huxor, A. D. Mackey, J. Peñarrubia, R. M. Rich, H. B. Richer, A. Siebert, N. Tanvir, D. Valls-Gabaud and K. A. Venn, “The remnants of galaxy formation from a panoramic survey of the region around M31”, *Nature*, 2009, **461**, 66–69.
- (363) S. Buder, S. Sharma, J. Kos, A. M. Amarsi, T. Nordlander, K. Lind, S. L. Martell, M. Asplund, J. Bland-Hawthorn, A. R. Casey et al., “The GALAH+ survey: Third data release”, *Monthly Notices of the Royal Astronomical Society*, 2021, **506**, 150–201.
- (364) M. A. El Aziz, I. Selim and A. Essam, “Open cluster membership probability based on K-means clustering algorithm”, *Experimental Astronomy*, 2016, **42**, 49–59.
- (365) S. Hasselquist, J. L. Carlin, J. A. Holtzman, M. Shetrone, C. R. Hayes, K. Cunha, V. Smith, R. L. Beaton, J. Sobeck, C. A. Prieto et al., “Identifying Sagittarius stream stars by their APOGEE chemical abundance signatures”, *The Astrophysical Journal*, 2019, **872**, 58.
- (366) A. Castro-Ginard, C. Jordi, X. Luri, T. Cantat-Gaudin and L. Balaguer-Núñez, “Hunting for open clusters in Gaia DR2: the Galactic anticentre”, *Astronomy and Astrophysics*, 2019, **627**, A35.
- (367) N. Price-Jones and J. Bovy, “Blind chemical tagging with DBSCAN: prospects for spectroscopic surveys”, *Monthly Notices of the Royal Astronomical Society*, 2019, **487**, 871–886.
- (368) M. Kounkel and K. Covey, “Untangling the Galaxy. I. Local Structure and Star Formation History of the Milky Way”, *The Astronomical Journal*, 2019, **158**, 122.



- (369) E. L. Hunt and S. Reffert, “Improving the open cluster census-I. Comparison of clustering algorithms applied to Gaia DR2 data”, *Astronomy & Astrophysics*, 2021, **646**, A104.
- (370) K. Malhan, R. A. Ibata, B. Goldman, N. F. Martin, E. Magnier and K. Chambers, “STREAMFINDER II: A possible fanning structure parallel to the GD-1 stream in Pan-STARRS1”, *Monthly Notices of the Royal Astronomical Society*, 2018, **478**, 3862–3870.
- (371) K. Malhan, R. A. Ibata and N. F. Martin, “Ghostly tributaries to the Milky Way: charting the halo’s stellar streams with the Gaia DR2 catalogue”, *Monthly Notices of the Royal Astronomical Society*, 2018, **481**, 3442–3455.
- (372) R. A. Ibata, K. Malhan and N. F. Martin, “The Streams of the Gaping Abyss: A Population of Entangled Stellar Streams Surrounding the Inner Galaxy”, *The Astrophysical Journal*, 2019, **872**, 152.
- (373) K. Malhan, R. A. Ibata, R. G. Carlberg, M. Bellazzini, B. Famaey and N. F. Martin, “Phase-space Correlation in Stellar Streams of the Milky Way Halo: The Clash of Kshir and GD-1”, *The Astrophysical Journal*, 2019, **886**, L7.
- (374) K. Malhan, R. A. Ibata, S. Sharma, B. Famaey, M. Bellazzini, R. G. Carlberg, R. D’Souza, Z. Yuan, N. F. Martin and G. 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*, 2022, **926**, 107.
- (375) P. V. Hough, Proc. of the International Conference on High Energy Accelerators and Instrumentation, Sept. 1959, 1959, pp. 554–556.
- (376) S. Sharma and K. V. Johnston, “A GROUP FINDING ALGORITHM FOR MULTIDIMENSIONAL DATA SETS”, *The Astrophysical Journal*, 2009, **703**, 1061–1077.
- (377) J. B. Kruskal, “On the shortest spanning subtree of a graph and the traveling salesman problem”, *Proceedings of the American Mathematical society*, 1956, **7**, 48–50.
- (378) S. S. Fuentes, J. De Ridder and J. Debosscher, “Stellar halo hierarchical density structure identification using (F) OPTICS”, *Astronomy & Astrophysics*, 2017, **599**, A143.
- (379) M. T. Costado, E. J. Alfaro, M. González and L. Sampedro, “Analysis of the kinematic structure of the Cygnus OB1 association”, *Monthly Notices of the Royal Astronomical Society*, 2016, **465**, 3879–3888.

- (380) H. Canovas, C. Cantero, L. Cieza, A. Bombrun, U. Lammers, B. Merin, A. Mora, Á. Ribas and D. Ruiz-Rodríguez, “Census of  $\rho$  Ophiuchi candidate members from Gaia Data Release 2”, *Astronomy and Astrophysics*, 2019, **626**, DOI: [10.1051/0004-6361/201935321](https://doi.org/10.1051/0004-6361/201935321).
- (381) F. Massaro, N. Alvarez-Crespo, A. Capetti, R. Baldi, I. Pillitteri, R. Campana and A. Paggi, “Deciphering the Large-scale Environment of Radio Galaxies in the Local Universe: Where Are They Born? Where Do They Grow? Where Do They Die?”, *The Astrophysical Journal Supplement Series*, 2019, **240**, 20.
- (382) J. L. Ward, J. D. Kruijssen and H.-W. Rix, “Not all stars form in clusters—Gaia-DR2 uncovers the origin of OB associations”, *Monthly Notices of the Royal Astronomical Society*, 2020, **495**, 663–685.
- (383) C. Higgs, A. McConnachie, N. Annau, M. Irwin, G. Battaglia, P. Côté, G. Lewis and K. Venn, “Solo dwarfs II: the stellar structure of isolated Local Group dwarf galaxies”, *Monthly Notices of the Royal Astronomical Society*, 2021, **503**, 176–199.
- (384) J. Jensen, G. Thomas, A. W. McConnachie, E. Starkenburg, K. Malhan, J. Navarro, N. Martin, B. Famaey, R. Ibata, S. Chapman et al., “Uncovering fossils of the distant Milky Way with UNIONS: NGC 5466 and its stellar stream”, *Monthly Notices of the Royal Astronomical Society*, 2021, **507**, 1923–1936.
- (385) M. Soto, M. A. Sgró, L. D. Baravalle, M. V. Alonso, J. L. Nilo Castellón, C. Valotto, A. Taverna, E. Díaz-Giménez, C. Villalón and D. Minniti, “Galaxy clustering in the VVV near-IR galaxy catalogue”, *Monthly Notices of the Royal Astronomical Society*, 2022, **513**, 2747–2760.
- (386) V. A. Epanechnikov, “Non-parametric estimation of a multivariate probability density”, *Theory of Probability & Its Applications*, 1969, **14**, 153–158.
- (387) S. R. Sain, “Multivariate locally adaptive density estimation”, *Computational Statistics & Data Analysis*, 2002, **39**, 165–186.
- (388) S. Sharma, J. Bland-Hawthorn, K. V. Johnston and J. Binney, “GALAXIA: A CODE TO GENERATE A SYNTHETIC SURVEY OF THE MILKY WAY”, *The Astrophysical Journal*, 2011, **730**, 3.
- (389) J. D. Simpson, S. L. Martell, G. Da Costa, J. Horner, R. F. Wyse, Y.-S. Ting, M. Asplund, J. Bland-Hawthorn, S. Buder, G. M. De Silva et al., “The GALAH Survey: Chemically tagging the Fimbulthul stream to the globular

cluster  $\omega$  Centauri”, *Monthly Notices of the Royal Astronomical Society*, 2020, **491**, 3374–3384.