HAL
open science

Introducing a new, robust galaxy-finder algorithm for simulations

Rodrigo Cañas, Pascal Elahi, Charlotte Welker, Claudia del P Lagos, Chris Power, Yohan Dubois, Christophe Pichon

To cite this version:

Rodrigo Cañas, Pascal Elahi, Charlotte Welker, Claudia del P Lagos, Chris Power, et al.. Introducing a new, robust galaxy-finder algorithm for simulations. Monthly Notices of the Royal Astronomical Society, 2019, 482 (2), pp.2039-2064. 10.1093/mnras/sty2725 . hal-02368711

HAL Id: hal-02368711

https://hal.science/hal-02368711v1

Submitted on 6 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **482,** 2039–2064 (2019)
Advance Access publication 2018 October 8

doi:10.1093/mnras/sty2725

Introducing a new, robust galaxy-finder algorithm for simulations

Rodrigo Cañas,[1,2★] Pascal J. Elahi,[1,2] Charlotte Welker,[1,2] Claudia del P. Lagos,[1,2] Chris Power,[1,2] Yohan Dubois[3,4,5] and Christophe Pichon[3,4,5]

[1] *International Centre for Radio Astronomy Research, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[2] *ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D)*
[3] *CNRS and UPMC Univ. Paris 06, UMR 7095, Institut d'Astrophysique de Paris, 98 bis Boulevard Arago, F-75014 Paris, France*
[4] *Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[5] *Korea Institute of Advanced Studies (KIAS), 85 Hoegiro, Dongdaemun-gu, Seoul 02455, Republic of Korea*

Accepted 2018 September 26. Received 2018 September 26; in original form 2018 April 27

ABSTRACT

Identifying galaxies in hydrodynamical simulations is a difficult task, particularly in regions of high density such as galaxy groups and clusters. We present a new scale-free, shape-independent algorithm to robustly and accurately identify galaxies in simulations, implemented within the phase-space halo-finder code VELOCIRAPTOR. This is achieved by using the full phase-space dispersion tensor for particle assignment and an iterative adjustment of search parameters, which help us overcome common structure-finding problems. We apply our improved method to the Horizon-AGN simulation and compare galaxy stellar masses (M_*), star formation rates (SFRs), and sizes with the elaborate configuration-space halo-finder HALOMAKER. Galaxies living in haloes with >1 galaxy are the most affected by the shortcomings of real-space finders, with their mass, SFR, and size being >2 times larger (smaller) in the case of host (satellite) galaxies. Thus, our ability to measure minor/major merger rates and disentangle environmental effects in simulations can be generally hindered if the identification of galaxies is not treated carefully. Though large systematic differences are obtained on a one-to-one basis, the overall galaxy stellar mass function, the SFR function, and the mass–size relations are not greatly affected. This is due to isolated galaxies being the most abundant population, dominating broad statistics.

Key words: methods: numerical – galaxies: evolution – dark matter – cosmology: theory.

1 INTRODUCTION

Galaxies are the result of a wide variety of physical processes. Their evolution and properties are determined by both their hierarchical assembly and the complex interplay between many multiscale non-linear processes, such as star formation, radiative cooling, and feedback loops (see Somerville & Davé 2015 for a recent review). Cosmological hydrodynamical simulations of galaxy formation are ideal laboratories to explore and isolate the effects of these physical processes on the evolution of galaxies in realistic environments (Dubois et al. 2014; Vogelsberger et al. 2014; Schaye et al. 2015). The advantage of these simulations over other numerical methods, such as abundance matching (e.g. Berlind, Narayanan & Weinberg 2001; Berlind & Weinberg 2002) and semi-analytic models of galaxy formation (Lacey & Cole 1993; Kauffmann & Charlot 1998; Cole et al. 2000), is the ability to predict the internal structure of galaxies, as the hydrodynamics that give rise to it are resolved through direct resolution of the equations of physics down to sub-galactic scales.

In recent years a major breakthrough in the capability of cosmological hydrodynamical simulations to produce realistic galaxy populations has taken place. This has been achieved thanks to the combined results of major improvements in numerical algorithms, availability of computing resources and improved subgrid models for unresolved feedback processes, and the calibration of subgrid feedback parameters to match key observables. Examples of this new generation of simulations include Horizon-AGN (Dubois et al. 2014), EAGLE (Schaye et al. 2015), Illustris (Vogelsberger et al. 2014), and IllustrisTNG (Pillepich et al. 2018). Simulated boxes of $\sim$(100 cMpc)^3 with sub-kpc resolution are becoming common. These simulations reproduce observables beyond those they were tuned for, with various degrees of success. For example, these simulations produce reasonable morphological diversity of galaxies, the colour bimodality of galaxies, the star formation rate (SFR)–stellar mass relation, the stellar mass function, and the cosmic SFR density

★ E-mail: rodrigo.canas@icrar.org

© 2018 The Author(s)
Published by Oxford University Press on behalf of the Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

evolution (e.g. Genel et al. 2014; Furlong et al. 2015; Snyder et al. 2015; Trayford et al. 2015, 2016; Dubois et al. 2016; Nelson et al. 2018).

In order to understand the physics involved in the formation of galaxies through simulations, we first need to understand and test the extent to which such results depend on numerical effects rather than on the physics (e.g. Klypin et al. 1999). This issue has been pointed out over the years by several studies which have shown that properties of galaxies and galaxy populations sensitively depend on the specific code used, the implemented subgrid physics and their respective tuning, as well as numerical resolution (see e.g. Frenk et al. 1999; Scannapieco et al. 2012; Kim et al. 2014; Power, Read & Hobbs 2014; Knebe et al. 2015; Schaye et al. 2015; Elahi et al. 2016; Sembolini et al. 2016a,b).

Often overlooked is the issue of the robustness with which we can measure galaxy properties in these simulations, which can affect the conclusions reached. The latter ultimately depend on how well we identify structures in the simulations (Knebe et al. 2011, 2013b). These issues are of particular interest for the new and coming generation of hydrodynamical simulations, which have taken the route of fine-tuning the free parameters of the subgrid physics modules (i.e. which describe the processes that are expected to take place at scales below the resolution limit) against a desired observable (e.g. the galaxy stellar mass function, GSMF, and the size–mass relation; Crain et al. 2015). Robustly measuring the desired galaxy property to perform the tuning in simulations is therefore crucial.

In the first studies of hierarchical formation, simple structure-finding algorithms, such as spherical overdensity (SO; Press & Schechter 1974) and friends-of-friends (FOF; Davis et al. 1985), were able to give a reasonable estimation of 'condensed' structures in simulations. However, with the ever-increasing size of simulations and the need for higher accuracy in measurements, such simple approaches are not necessarily optimal, and a large number of codes have appeared in the literature addressing the finding of structures in simulations (see Knebe et al. 2011, 2013b, and references therein). Early approaches were characterized by the use of solely configuration-space information (e.g. BDM, Klypin & Holtzman 1997; HOP, Eisenstein & Hut 1998; SUBFIND, Springel et al.2001; SKID, Stadel 2001; AHF, Gill, Knebe & Gibson 2004), while more recent sophisticated algorithms address the problem adding the velocity-space information (e.g. 6DFOF, Diemand, Kuhlen & Madau 2006; HSF, Maciejewski et al. 2009; VELOCIRAPTOR, Elahi, Thacker & Widrow 2011; ROCKSTAR, Behroozi, Wechsler & Wu 2013). Although all these algorithms attempt to solve the same problem, the specific details of each implementation can introduce artefacts in the final results. Other approaches tackle the problem by using temporal information by following (sub)haloes' bound particles through simulation snapshots to identify structures and *de-blend* systems in interaction, which can be done from either late to earlier times (e.g. SURV; Tormen, Moscardini & Yoshida 2004; Giocoli, Tormen & van den Bosch 2008; Giocoli et al. 2010) or vice-versa (e.g. HBT, HBT+; Han et al. 2012, 2018). Though powerful in principle, these methods rely heavily on identification at sufficiently early times and having at hand snapshots at a high cadence.

It is essential that we understand the reliability of measurements and the associated systematic uncertainties. This has been addressed by many comparison projects in which structure-finding codes are tested against the same data to study the similarities and differences between measurements of the properties of dark matter haloes (Knebe et al. 2011), subhaloes (Onions et al. 2012), galaxies (Knebe et al. 2013a), and tidal structures (Elahi et al. 2013). Such studies have found overall agreement when analysing dark matter halo populations (Knebe et al. 2011). However, large differences have been obtained in the overall mass recovered for dark matter subhaloes, satellite galaxies, and tidal streams (Onions et al. 2012; Elahi et al. 2013; Knebe et al. 2013a, 2013b). While the identification of substructures depends on the identification of density peaks, the major challenge is to assign the 'background' particles to statistically significant density peaks that can drastically affect the properties of the structures. For this reason, algorithms that only use configuration-space information, although fast, struggle to appropriately identify subhaloes in dense environments (e.g. galaxy groups and clusters and merging systems), while finders that also include velocity-space information obtain better results in these regimes (Knebe et al. 2011).

This paper presents a new galaxy-finding algorithm which makes use of the full configuration- and velocity-space information, and presents a thorough study of the effects that the identification method has on the properties of individual galaxies and galaxy populations. This implementation is an extension of the halo-finder code VELOCIRAPTOR (Elahi et al. 2011; Elahi et al., in preparation). We pay special attention to two regimes that have been traditionally challenging for galaxy-finding algorithms: (i) mergers and interactions and (ii) identification of substructures in high-density environments. The main problem in both of these regimes is that the outskirts of hosts and satellite structures can have similar densities, making it difficult to distinguish to which structure they belong. This is even harder if only configuration-space information is taken into account. These problems are equally valid for dark matter haloes and galaxies, and while there is a plethora of literature that addresses the former (see for reference Knebe et al. 2013b), the latter has not yet been thoroughly addressed. Galaxies have a range of morphologies that during interactions produce complex stellar structures that form on an already significant density peak. Thus, the problem of identifying galaxies cannot be solved using dark matter halo finding tools. We show that the undesirable consequences of poor identification affect radial mass profiles, sizes, and total masses. We apply our new galaxy-finding algorithm to the state-of-the-art cosmological hydrodynamical simulation Horizon-AGN (Dubois et al. 2014) and compare our results with the original galaxy catalogue, which was obtained by applying the configuration-space-finder HALOMAKER (Aubert, Pichon & Colombi 2004; Tweed et al. 2009).

This work is organised as follows: In Section 2 we provide a general and brief description of the code VELOCIRAPTOR and the Horizon-AGN simulation. In Section 3 we describe in detail the improved algorithm to identify galaxies in simulations and implemented in VELOCIRAPTOR. In Section 4 we present examples of the performance of our new algorithm on strongly interacting scenarios. In Section 5 we compare results obtained with VELOCIRAPTOR and the original Horizon-AGN galaxy catalogue on a galaxy-to-galaxy basis, as well as comparing the entire galaxy populations. Discussion is presented in Section 6, and summary and conclusions are presented in Section 7. Lastly, in Appendix A we show how configuration-space linking length affects galaxy delimitation, and in Appendix B we show how different weights affect particle assignment.

2 NUMERICAL METHODS

In this section, we briefly describe the Horizon-AGN simulation, and the structure-finding code VELOCIRAPTOR. For further details the interested reader is referred to Dubois et al. (2014), where the Horizon-AGN simulation was presented, and to Elahi et al. (2011) for a detailed description of VELOCIRAPTOR.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

2.1 Horizon-AGN simulation

Horizon-AGN is a state-of-the-art hydrodynamical simulation, presented in Dubois et al. (2014). It follows the formation and evolution of galaxies in a standard lambda cold dark matter (ΛCDM) cosmology, adopting values of total matter density $\Omega_{\rm m} = 0.272$, dark energy density $\Omega_\Lambda = 0.728$, amplitude of the linear power spectrum $\sigma_8 = 0.81$, baryon density $\Omega_{\rm b} = 0.045$, Hubble constant $H_0 = 70.4\,{\rm km\,s^{-1}\,Mpc^{-1}}$, and spectral index $n_{\rm s} = 0.967$, in concordance with results from the *Wilkinson Microwave Anisotropy Probe 7* (*WMAP7*; Komatsu et al. 2011).

The simulation was run using the adaptive mesh refinement (AMR) code RAMSES (Teyssier 2002), and it has a comoving box size of $L_{\rm box} = 100\,h^{-1}$ Mpc, a total of 1024^3 dark matter particles with mass $M_{\rm dm} = 8 \times 10^7\,{\rm M}_\odot$, and an initial number of 1024^3 gas cells, which are refined up to seven times to a maximum physical resolution of $\Delta x = 1$ kpc.

The implemented subgrid physics include gas cooling, heating from a uniform redshift-dependent ultraviolet (UV) background, star formation, stellar feedback driven by supernovae (SNe) Type Ia, II, and stellar winds, and black hole (BH) accretion and its associated active galactic nucleus (AGN) feedback.

Following Dubois et al. (2012), BHs are created with a seed mass of $M_{\rm BH} = 10^5\,{\rm M}_\odot$ and grown according to a Bondi–Hoyle–Lyttleon accretion scheme capped at the Eddington accretion rate. A two-mode AGN feedback is explicitly implemented as a bipolar outflow at accretion rates smaller than 1 per cent the Eddington accretion (Dubois et al. 2010), and as an isotropic thermal energy injection otherwise (see Dubois et al. 2014 and Volonteri et al. 2016 for further details).

Galaxies in Horizon-AGN were originally identified with the code HALOMAKER (Tweed et al. 2009). HALOMAKER uses the ADAPTAHOP (Aubert et al. 2004) algorithm (which is itself based on HOP; Eisenstein & Hut 1998) to identify structures and their corresponding substructures. The algorithm identifies high-density regions and the particles associated with those. This is done by estimating the density of all particles from $N_{\rm SPH}$ neighbours using a smoothed particle hydrodynamics (SPH) kernel. Then, starting at a reference particle, the density field gradient is followed by linking it to the densest particle within $N_{\rm HOP}$ neighbours and hopping it as the new reference particle. This process is done iteratively until the reference particle is the densest within its $N_{\rm HOP}$ neighbours. Particles with density above a density threshold $\rho_{\rm t}$ linked to the same peak constitute groups. A hierarchy is established by looking for saddle points in the density field between peaks and using them as boundaries to define the hierarchy levels. Groups whose saddle point is above $\rho_{\rm t}$ are linked as members of the same branch. The process is repeated iteratively for each level until no saddle points are found. The main structure (either dark matter halo or galaxy) is defined by following the branch to which the most massive or densest peak belongs. Groups from other branches will then become substructures, while those in branches within branches will be sub-substructures, and so on. Galaxies are identified using star particle information only, the local density for each particle is calculated using $N_{\rm SPH} = 20$ neighbours, and a local threshold of $\rho_{\rm t} = 178$ times the average total matter density is applied to select relevant densities. A minimum physical size above which substructures are considered relevant of $\sim$2 kpc is also applied. Only galactic structures with more than 50 star particles are considered.

Horizon-AGN has been used to study the alignments between the spin of galaxies and cosmic web filaments, and how mergers change the spin orientation of galaxies (Dubois et al. 2014; Welker et al. 2014). Its BH growth and AGN feedback implementations have succeeded in producing a BH population whose overall properties agree with observations (Volonteri et al. 2016), and have shown the importance of AGN feedback in helping the simulation to reproduce the observed morphology and kinematic properties of massive galaxies (Dubois et al. 2016). Additionally, the simulation has also been used to study the evolution of galaxy luminosity and stellar mass functions, star formation main sequence, and galaxy colours (Kaviraj et al. 2017).[1]

2.2 VELOCIRAPTOR

VELOCIRAPTOR (also known as STF; Elahi et al. 2011) is a structure-finding algorithm capable of identifying dark matter haloes, galaxies, and substructures such as satellite subhaloes and streams in simulations. Here we briefly summarize the algorithm presented in Elahi et al. (2011).

The standard VELOCIRAPTOR's algorithm is based on the assumption that the velocity distribution of a system composed of many objects can be split into a smooth background component with overdense features in it. The former would correspond to the main halo and the latter to the substructures embedded in it. Hence, substructures are found by identifying the particles whose local velocity density $f_{\rm l}(\boldsymbol{v})$ stands out from the expected background velocity density $f_{\rm bg}(\boldsymbol{v})$, effectively looking for clustering in orbit space.

In order to calculate these quantities for each particle, the main halo is split in volume cells using the KD-tree algorithm (Bentley 1975). This is done such that each cell contains enough particles to minimize statistical errors but not too many to avoid variations in the gravitational potential and velocity density in each cell. The expected background velocity density, $f_{\rm bg}$, is estimated as a multivariate Gaussian. Hence, for a particle i with velocity $\boldsymbol{v}_i$

$$f_{\rm bg}(\boldsymbol{v}_i) = \frac{\exp[-\frac{1}{2}(\boldsymbol{v}_i - \bar{\boldsymbol{v}}(\boldsymbol{x}_i))\,\boldsymbol{\Sigma}_v(\boldsymbol{x}_i)^{-1}\,(\boldsymbol{v}_i - \bar{\boldsymbol{v}}(\boldsymbol{x}_i))]}{(2\pi^{3/2}|\boldsymbol{\Sigma}_v(\boldsymbol{x}_i)|^{1/2})}\,, \tag{1}$$

where $\bar{\boldsymbol{v}}(\boldsymbol{x}_i)$ and $\boldsymbol{\Sigma}_v(\boldsymbol{x}_i)$ are respectively the local average velocity and velocity dispersion tensor about $\bar{\boldsymbol{v}}(\boldsymbol{x}_i)$, at the $i^{\rm th}$ particle's position $\boldsymbol{x}_i$. To accurately determine $\bar{\boldsymbol{v}}(\boldsymbol{x}_i)$ and $\boldsymbol{\Sigma}_v(\boldsymbol{x}_i)$, the $\bar{\boldsymbol{v}}_k$ and $\Sigma_{v,k}$ of each cell k are calculated, and these quantities are linearly interpolated to the $i^{\rm th}$ particle's position using the cell containing the particle and six neighbouring cells. For each cell k

$$\bar{\boldsymbol{v}}_k = \frac{1}{M_k}\sum_j^{N_k} m_j \boldsymbol{v}_j \tag{2}$$

and

$$\Sigma_{v,k} = \frac{1}{M_k}\sum_j^{N_k} m_j(\boldsymbol{v}_j - \bar{\boldsymbol{v}})(\boldsymbol{v}_j - \bar{\boldsymbol{v}})^T\,, \tag{3}$$

where m_j and $\boldsymbol{v}_j$ are the particle j's mass and velocity, respectively, and N_k[2]and M_k are the number of particles and mass of the cell

[1]Further research projects and publications can be found on the Horizon-AGN simulation website (http://www.horizon-simulation.org).

[2]VELOCIRAPTOR constructs KD-trees at several stages to calculate velocity density distribution, conduct FOF searches, and estimate gravitational potentials. The number of particles inside each cell N_k will vary depending on the purpose of the tree. To estimate $f_{\rm bg}(\boldsymbol{v})$, $N_k = 16$ is used when the $f(\boldsymbol{v})$ is estimated using 32 velocity-space nearest neighbours. For efficient FOF searches N_k is selected to be similar to the minimum number of particles threshold to define a structure. Finally, to calculate the gravitational potential $N_k = 8$ is used.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

k, respectively. Finally, the local velocity density $f_l(\boldsymbol{v}_i)$ is calculated using a smoothing kernel scheme from velocity-space nearest neighbours.

For each particle i, the logarithmic ratio of the local and background velocity distributions

$$\mathcal{R}_i = \ln \frac{f_l(\boldsymbol{v}_i)}{f_{bg}(\boldsymbol{v}_i)} \tag{4}$$

is calculated. Particles with $\mathcal{R}_{th}$ above a $\mathcal{R}_{th}$ threshold are kept and classified as potential substructures.

Once the outlying particles are found, they are clustered into groups using an FOF-motivated (Davis et al. 1985) algorithm. Particles i and j are grouped if

$$\frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^2}{l_x^2} < 1\,, \tag{5}$$

$$1/\mathcal{V}_r \leq v_i/v_j \leq \mathcal{V}_r\,, \tag{6}$$

$$\cos\Theta_{op} \leq \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{v_i v_j}\,, \tag{7}$$

where $\boldsymbol{x}$ and $\boldsymbol{v}$ are a particle's position and velocity, respectively, l_x is the configuration-space linking length, $\mathcal{V}_r$ is the velocity ratio threshold determining the range in which the norm of the particles' velocities are considered to be similar, and Θ_{op} is an opening angle threshold within which directions of the particles' velocity vector must align. This effectively means that particles in a group need to be not only physically close, but also close in orbital space.

VELOCIRAPTOR has been employed in several comparison projects that have confirmed its versatility and ability to accurately find structures and substructures in N-body and hydrodynamical simulations (e.g. Onions et al. 2012; Elahi et al. 2013; Knebe et al. 2013a; Behroozi et al. 2015). An updated version of the code along with new features and tools will be presented in Elahi et al. (in preparation).

3 ROBUST IDENTIFICATION OF GALAXIES

VELOCIRAPTOR was originally designed to find dark matter structures in simulations, including haloes, subhaloes, and dark matter streams. While it has also been used to identify galaxies in hydrodynamical simulations (Knebe et al. 2013a), the treatment of the baryonic component was limited to first identifying dark matter (sub)haloes and then linking gas and stellar particles to the nearest dark matter particle in phase space. Though this procedure in principle provides a phase-space assignment of baryons to dark matter haloes, there were two key aspects that needed improvement. First, the metric used for baryon assignment was quite simple, which could cause incorrect assignment of particles especially for non-spherical or complex geometries, which are particularly present in interacting galaxies. Secondly, for some interacting galaxies, the dark matter haloes might be indistinguishable, assigning the merging galaxies to a single halo.

These problems could be solved by running VELOCIRAPTOR independently over stellar particles to identify galaxies. However, the original VELOCIRAPTOR algorithm assumes the existence of a smooth, semi-virialized background. The code was not optimized to find substructures in any system where the background is sparsely sampled.

Here, we describe a new algorithm that uses the tools already implemented in VELOCIRAPTOR to perform fast and efficient phase-space FOF searches, but modifying several search and assignment criteria to get the desired robustness in the identification of galaxies.

3.1 An improved algorithm to identify galaxies

The exact definition of 'galaxy' is non-trivial in both simulations and observations. For hydrodynamical simulations a commonly adopted definition of a galaxy is all the baryonic mass bound to dark matter (sub)haloes. Hence, the identification of a galaxy relies on how well (sub)haloes are identified. Instead, our aim here is to be able to identify galaxies robustly independently of dark matter by using star particle information only. This is done first by identifying the regions where galaxies are expected to be and then separating kinematically distinct phase-space overdense structures. In this section we describe in detail the algorithm; a schematic representation is shown in Fig. 1.

3.1.1 Step 1 – 3DFOF

In the dark matter cosmological framework, galaxies reside inside large virialized dark matter haloes. Our 'first guess' of where galaxies are located will be the region delimited by the extent of its host dark matter halo. This is done by grouping particles that are close in physical space using a configuration-space FOF search (3DFOF), described by equation (5), on the star particles. Since its introduction in Davis et al. (1985), this first step has been commonly used by many finding algorithms (e.g. SUBFIND, HALOMAKER, ROCKSTAR; Springel et al. 2001; Aubert et al. 2004; Tweed et al. 2009; Behroozi et al. 2013) due to its simplicity and versatility.

For cosmological simulations, a widely adopted scheme is

$$l_{x(3D)} = b\,\Delta x\,, \tag{8}$$

where $l_{x(3D)}$ is the configuration-space linking length, Δx is the simulation's mean interparticle spacing, and $0 < b < 1$. We adopt the commonly used value of $b = 0.2$ (e.g. Boylan-Kolchin et al. 2009; Vogelsberger et al. 2014; Schaye et al. 2015), which will group star particles inside the dark matter halo.

3.1.2 Step 2 – 6DFOF

Galaxies are centrally concentrated distributions of stars in configuration and velocity space. In simulations, the positions and velocities of the constituent particles are expected to be found close in phase space. Galaxies are identified by performing a phase-space FOF (6DFOF) search separating each 3DFOF object into kinematically distinct substructures. Particles i and j are linked into 6DFOF groups if and only if

$$\frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^2}{l_{x(6D)}^2} + \frac{(\boldsymbol{v}_i - \boldsymbol{v}_j)^2}{l_{v(6D)}^2} \leq 1, \tag{9}$$

where $l_{x(6D)}$ and $l_{v(6D)}$ are the configuration-space and velocity linking lengths, respectively.

We stress that appropriate values of $l_{x(6D)}$ and $l_{v(6D)}$ have to be chosen in 6DFOF searches. If a very large value of $l_{x(6D)}$ is adopted, this would result in a velocity-only FOF search and vice-versa, while very small values of linking lengths would result in either splitting single structures into multiple components or missing structures.

At this point we are interested in separating structures that have been found in a common 3DFOF envelope. For this purpose $l_{x(6D)}$ is chosen to be a function of $l_{x(3D)}$ and $l_{v(6D)}$ is estimated from the

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

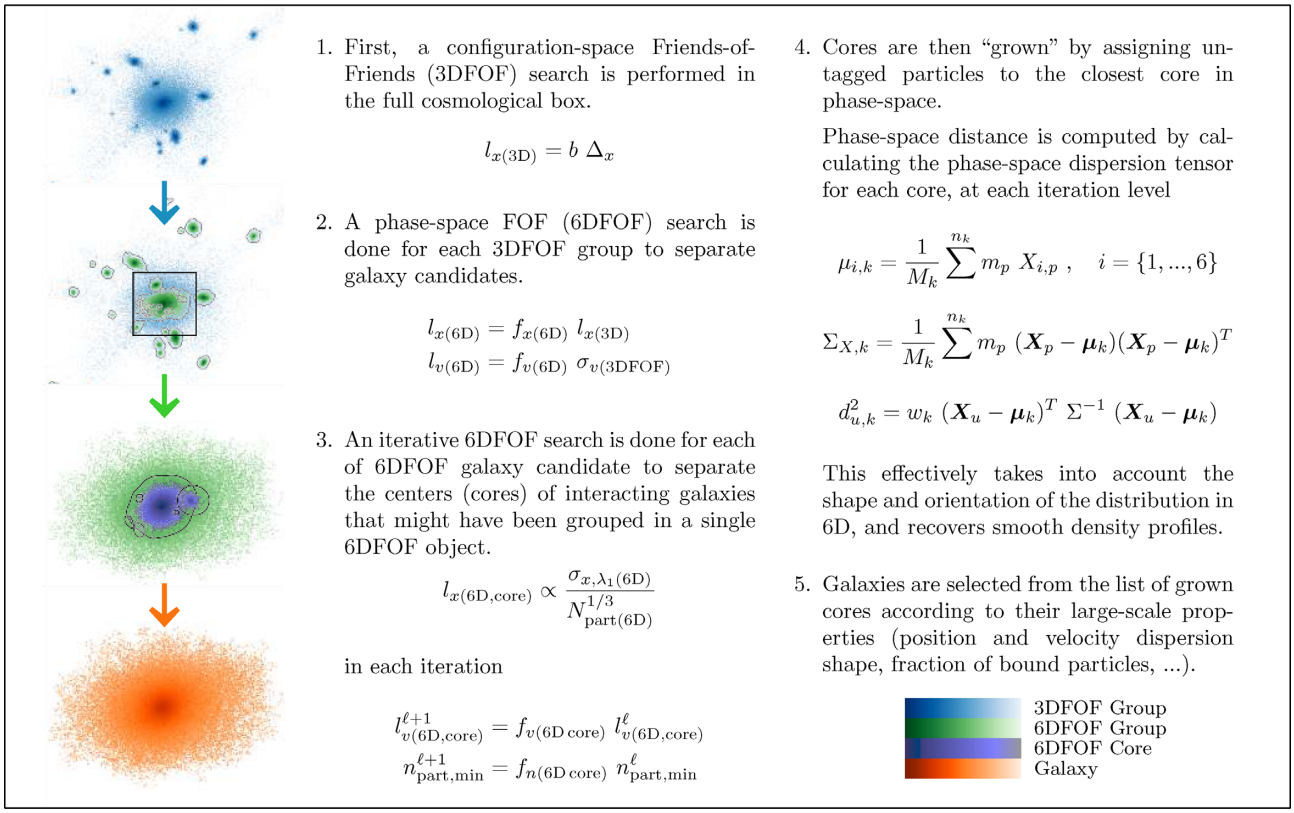

Figure 1. Summary of the algorithm to find galaxies with VELOCIRAPTOR introduced in Section 3. Structures are searched for by (1) separating particles in the simulation in 3DFOF objects and (2) posteriorly doing a 6DFOF search. (3) Then, an iterative 6DFOF search is done in each of these objects to look for dense cores of galaxies in close interactions or mergers. (4) Once cores are found, they are grown by assigning particles in the original 6DFOF object. (5) Finally, properties of all the objects found are calculated and galaxies are selected according to these properties. A key aspect of this algorithm is the particle assignment procedure (core growth), as even in the presence of satellites close to the host centre (6DFOF core objects in purple), we can obtain smooth profiles (see central galaxy in orange). See text for further details.

velocity dispersion of the full 3DFOF object

$$l_{x(6D)} = f_{x(6D)}\ l_{x(3D)} \tag{10}$$

and

$$l_{v(6D)} = f_{v(6D)}\ \sigma_v = f_{v(6D)}\sqrt{\sigma^2_{v,x} + \sigma^2_{v,y} + \sigma^2_{v,z}}\ . \tag{11}$$

Here $0 < f_{x(6D)} < 1$, σ_{v_j} is the velocity dispersion in the j direction, and $f_{v(6D)}$ is a user-defined parameter which should be of order unity. As local properties of each 3DFOF object are used for its 6DFOF search, we are effectively performing a 'tailored' 6DFOF search.[3]The above choice of parameters is motivated by the fact that galaxies reside in (sub)halo centres; hence, their overdensities are expected to be much higher than that of the dark matter halo. This condition is imposed by shrinking the configuration-space linking length. The velocity-space linking effectively removes particle bridges in the configuration space, resulting in the identification of kinematically distinct structures.

Intuitively it would be more consistent to compute $l_{x(6D)}$ using similar arguments as for $l_{v(6D)}$. However, due to the complexity of the environment in which some galaxies reside, measurements of the position dispersion of the particles would actually result in very large values of $l_{x(6D)}$. This is especially the case for galaxy groups and clusters where particle bridges between galaxies make 3DFOF structures too extended. A similar argument can be stated against using equation (11), as large 3DFOF objects are expected to have very large velocity dispersion and consequently very large values of $l_{v(6D)}$. However, in this case we do not have *a priori* knowledge of what the scale of the velocity linking length should be, as this is the first 6DFOF search; σ_v provides a good first estimation of $l_{v(6D)}$.

[3]Consider trying to link particles belonging to a Gaussian distribution. Its dispersion, σ, provides a good starting point for linking length.

3.1.3 Step 3 – iterative 6DFOF core search

Although the 6DFOF search should already have separated galaxies with distinct phase-space distributions, multiple galaxies can still be found in single 6DFOF groups. This is the case of merging galaxies whose outskirts have phase-mixed to some degree but whose cores (dense kinematically cold galactic centres) have not yet merged fully, or satellites that orbit close to the centre of a much bigger galaxy. Instead of trying to recover a group in its entirety, we adopt a different approach and attempt to isolate its cores. In order to separate galaxies in these structures we perform an iterative 6DFOF core search for each preliminary 6DFOF group. For this iterative 6DFOF core search we use the same criteria as equation (9) to link particles, but using a different choice of linking lengths, which for clarity will be identified with the subscript (6D,core). These linking lengths scale with the dispersion of the system being searched.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

FOF algorithms, particularly when used in an iterative fashion, are sensitive to the choice of linking parameters: Too large and separate structures can be joined; too small and structures can be fragmented. ROCKSTAR (Behroozi et al. 2013), which uses a 6DFOF to recover groups in full, addresses the latter problem by merging groups if their centres are closer than a phase-space distance threshold to clean for false positives. Although this is useful, our approach is oriented towards a robust search of the *densest portions of groups*, followed by carefully growing candidate cores, and does not rely solely on the effectiveness of cleaning procedures. Therefore, we first set the search parameters appropriately, which are then modified in each iteration.

For the initial velocity-space linking length we adopt

$$l_{v,(\mathrm{6D,core})} = \sigma_{v,\lambda_1} \,, \tag{12}$$

where σ^2_{v,λ_1} is the length of the largest principal axis of the velocity dispersion tensor, $\boldsymbol{\Sigma}_v$. As for the first 6DFOF search, equation (12) sets the scale for the initial velocity-space linking length. For the following iterations $l_{v(\mathrm{6D,core})}$ is iteratively shrunk, i.e.

$$l^{\ell+1}_{v(\mathrm{6D,core})} = f_{v(\mathrm{6D,core})}\, l^{\ell}_{v(\mathrm{6D,core})} \,, \tag{13}$$

where the superscript indicates the iteration level, and $0 < f_{v(\mathrm{6D,\,core})} < 1$ is a user-defined shrinking factor. By shrinking the velocity-space linking lengths this way, we remove the wings and bridges in the distribution, because in each iteration we truncate the original distribution towards the coldest regions, separating cores. For this study we adopt $f_{v(\mathrm{6D,core})} = 0.8$.

The adopted configuration-space linking length here is

$$l_{x(\mathrm{6D,core})} = 3\,\sigma_{x,\lambda_1} \left(\frac{4\pi}{3} \frac{1}{N_{\mathrm{part(6D)}}} \right)^{1/3} \,, \tag{14}$$

where σ^2_{x,λ_1} is the length of the largest principal axis of the configuration dispersion tensor, $\boldsymbol{\Sigma}_X$, and $N_{\mathrm{part(6D)}}$ is the number of particles in the 6DFOF group. Equation (14) is then the mean interparticle spacing in a $3\sigma_{x,\lambda_1(\mathrm{6D})}$ radius sphere. This linking length scales with configuration-space dispersion and the extent to which the distribution is well sampled. The logic of including a scaling that decreases the linking length with increasing number of particles is as follows. With a well-sampled distribution, the 3σ scaling used will link not only the central region but the outskirts as well, possibly joining this distribution with neighbouring ones. Decreasing the linking length, if well sampled, reduces the likelihood of artificially joining structures. Conversely, if poorly sampled, the measured dispersion will underestimate the true one. Therefore, relative to a well-sampled system, we scale up the linking length.

Although at this stage the iterative 6DFOF search is done to separate structures, the configuration-space linking length is kept fixed through iterations. We could in principle modify $l_{x(\mathrm{6D,core})}$ by some factor $f_{x,(\mathrm{6D,core})}$ at each iteration as is done for $l_{v(\mathrm{6D,core})}$. However, equation (14) already includes the information on how concentrated the distribution (6DFOF object) is in the configuration space. Reducing the $l_{x(\mathrm{6D,core})}$ value will likely cause us to either miss or fragment structures. Our approach requires a fixed $l_{x,(\mathrm{6D,core})}$ short enough to separate structures in the configuration space, and a $l_{v(\mathrm{6D,core})}$ long enough to gather statistically significant groups of particles. In each iteration $l_{v(\mathrm{6D,core})}$ is shrunk to separate structures that might be linked by their velocity-space outskirts.

For each FOF search, a minimum particle number, $n_{\mathrm{part,min}}$ has to be set to define statistically significant structures. For steps 1 and 2 (Sections 3.1.1 and 3.1.2) we suggest a $n_{\mathrm{part,min}} = 50$. For the iterative search, however, $n_{\mathrm{part,min}}$ is updated after each iteration as

$$n^{\ell+1}_{\mathrm{part,min}} = f_{n(\mathrm{6D,core})}\, n^{\ell}_{\mathrm{part,min}} \,, \tag{15}$$

where $n_{\mathrm{part,min}}$ is the minimum number of particles, $f_{n(\mathrm{6D,core})} > 1$, and the superscript ℓ indicates the iteration level. Increasing the minimum number of particles while shrinking linking lengths may sound non-intuitive at first as we expect to link *fewer* particles per group in each iteration. However, as the linking length $l_{v,(\mathrm{6D,core})}$ becomes smaller, it also becomes easier to identify small phase-space overdense (noisy) patches in the distribution, which can result in finding multiple spurious structures. Iteratively increasing $n_{\mathrm{part,min}}$ reduces the likelihood of finding noisy patches. For this study we adopt $f_{n(\mathrm{6D,core})} \sim 1.5$.

A more intuitive choice of $f_{n(\mathrm{6D,\,core})}$ would be one that scales with the number of particles in a given group or iteration level, instead of choosing a fixed $f_{n(\mathrm{6D,\,core})}$ for all searches. However, bearing in mind that the number of particles can differ by orders of magnitude between galaxies in the same system, even using a logarithmic scale of the number of particles can lead to $f_{n(\mathrm{6D,\,core})} \gg 1$, and consequently to very large $n_{\mathrm{part,\,min}}$ in a couple of iterations.

This iterative 6DFOF search starts with the entire 6DFOF object. For subsequent iterations the 6DFOF search is done only for the largest core. This prevents the loss of an already found structure due to the increment of $n_{\mathrm{part,\,min}}$. These cores are kept for particle assignment (core growth, Section 3.1.4) and are revisited later to look for possible mergers or close interactions. Iterations on the largest core stop when a user-defined maximum number of iterations, $N^{\mathrm{max}}_{\mathrm{iter}}$, has been reached, or when no more structures are found with the current iteration level search parameters.

3.1.4 Step 4 – core growth

The critical step once cores are identified is assigning particles to these cores, reconstructing the galaxies. We assign particles that belong to the original 6DFOF structure (step 2, Section 3.1.2) that are not member of a core. This process is crucial as the final product of structure searches (either galaxies or dark matter haloes) can be severely affected by how this is done.

Given the phase-space nature of the 6DFOF searches, the obvious criteria would be to assign a given particle to the closest core in phase-space. This concept has been previously used by other algorithms, but several implementations can exist. A naive 6D phase-space distance as implied by Behroozi et al. (2013), implicitly assumes a spherical morphology. This might work well for dark matter haloes but can lead to systematic effects due to the complex morphologies of galaxies.

Instead, starting at level ℓ, we characterize the phase-space distribution of each core k, by calculating its mean $\boldsymbol{\mu}$ (phase-space centre-of-mass vector), and phase-space dispersion tensor $\boldsymbol{\Sigma}_X$ (distribution's covariance matrix):

$$\boldsymbol{\mu}_k = \{\mu_{i,k}\} \quad i = \{1, ..., 6\} \,, \tag{16}$$

$$\mu_{i,k} = \frac{1}{M_k} \sum^{n_k} m_{\mathrm{p}}\, X_{i,\mathrm{p}} \,, \tag{17}$$

$$\Sigma_{X,k} = \frac{1}{M_k} \sum^{n_k} m_{\mathrm{p}}\, (\boldsymbol{X}_{\mathrm{p}} - \boldsymbol{\mu}_k)(\boldsymbol{X}_{\mathrm{p}} - \boldsymbol{\mu}_k)^T \,. \tag{18}$$

Here, M_k and n_k are the total mass and the total number of particles in the core k, respectively; $X_{i,\mathrm{p}}$ is the ith coordinate of the phase-

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

space coordinate vector $\boldsymbol{X}$ of particle p with mass $m_{\rm p}$ that belongs to core k. Then, for all the particles at $\ell - 1$ that were not assigned to any core at level ℓ, we calculate

$$d_{{\rm u},k}^2 = w_k \left(\boldsymbol{X}_{\rm u} - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_X^{-1} \left(\boldsymbol{X}_{\rm u} - \boldsymbol{\mu}_k\right) . \tag{19}$$

Here $d_{{\rm u},k}$ is the phase-space distance from untagged particle u to core k and w_k is a weighting constant. A weighting scheme is necessary to avoid assigning too many particles to tidal streams and shells. Without a weighting, this could happen as these structures can be quite extended and have large position and velocity dispersion compared to those of galaxies (compact centrally concentrated distributions). To compensate for this, we adopt

$$w_k = \frac{1}{M_k^{\alpha}} , \tag{20}$$

with α a free parameter. Taking $\alpha = 1$ can cause all particles to be assigned to the largest object, again, as galaxy masses in the same system can differ by orders of magnitude. Values of $1/3 \leq \alpha \leq 2/3$ give a w that scales with the tidal radius. We have found that $\alpha = 0.5$ leads to good results; we justify this choice of α in Appendix B.

After calculating these distances, particles are assigned to the closest core in phase space. When a single core is found at level ℓ, all untagged particles at the previous level, $\ell - 1$, are assigned to that single core. Then, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are recalculated for all the cores in the following levels and the process is repeated until all particles in the original 6DFOF group have been assigned to a core.

This approach is particularly powerful for many reasons: (i) It effectively takes into account the shape and orientation of the distribution; (ii) it allows the shape of the distribution to change from the inner to the outer parts; and (iii) this produces smooth density profiles for galaxies even when the galaxies are passing through the inner radii of larger galaxies. Hence, galaxies will not have missing holes or bubble-like structures . This is essential when measuring galaxy properties' radial profiles.

For each 6DFOF object (step 2, Section 3.1.2) the algorithm continues as follows. After performing step 3 (Section 3.1.3) on the largest core, particles are assigned to all cores inside following step 4. The top hierarchy level, i, is assigned to the largest core (candidate central galaxy). The rest of the cores will have hierarchy level $i + 1$. Steps 3 and 4 are then repeated for all $i + 1$ substructures. If any sub-substructures are found, they are assigned a hierarchy level $i + 2$, and so on. The algorithm finishes when all (sub)structures have been iteratively searched.

3.1.5 Step 5 – selecting galaxies

Once all (sub)structures have passed through the iterative core search and their respective core growth, bulk properties of the structures are calculated to determine whether they are galaxies or not. This is necessary because the versatility of the algorithm allows us to identify not only galaxies but also tidal features such as streams and shells. This catalogue can be cleaned if only galaxies are desired.

We classify objects as galaxies or streams following Elahi et al. (2013). We calculate the ratios $q \equiv \lambda_2/\lambda_1$ and $s \equiv \lambda_3/\lambda_1$ of the eigenvalues, λ_i, of the position and velocity dispersion tensors for all the structures, as well as the bound fraction of particles $f_{\rm b}$. A structure is not considered as a galaxy if

$$\begin{aligned} &(f_{\rm b} < 0.01) \quad \cup \\ &((q_x < 0.3 \cap s_x < 0.2) \cup (q_v < 0.5 \cap s_v < 0.2)) \cup \\ &(f_{\rm b} < 0.2 \cap ((q_x < 0.6 \cap s_x < 0.5) \cup (q_v < 0.5 \cap s_v < 0.4))) ; \end{aligned} \tag{21}$$

that is, galaxies are expected to be bound ellipsoidal distributions of stars. Structures with less than 1 per cent of bound particles are unlikely to be galaxies. Highly elongated structures either in configuration or in velocity space (i.e. low values of q_x, s_x, q_v, and s_v), which can be bound to some degree, are likely to be streams or shells. The fraction of bound particles is kept to such low thresholds as neither gas nor dark matter information is taken into account when computing the gravitational potential. The parameters and thresholds used in equation (21) are suggested values that were derived from calibration tests to give the desired results. At $z = 0$ this selection discards ∼30 per cent of structures with $10^8 < M_*/{\rm M}_\odot < 10^9$, ∼1.5 per cent of structures with $10^9 < M_*/{\rm M}_\odot < 10^{10}$, and ∼0.2 per cent of structures with $10^{10} < M_*/{\rm M}_\odot < 10^{11}$. If desired, other selection criteria can be used.

It is important to note that equation (21) was only tested for VELOCIRAPTOR outputs. Comparisons throughout this study between VELOCIRAPTOR and HALOMAKER are made using *raw* catalogues. We argue that selection of galaxies using equation (21) does not impact the results of this study as we focused on *well-resolved* structures with $M_* > 10^9\,{\rm M}_\odot$.

Table 1. Suggested values for the parameters used for galaxy identification with VELOCIRAPTOR.

Parameter	Value	Reference	
b	0.2	Equation	8
$f_{x({\rm 6D})}$	0.2	Equation	10
$f_{v({\rm 6D})}$	1.0	Equation	11
$f_{v({\rm 6D,core})}$	0.8	Equation	13
$f_{n({\rm 6D,core})}$	1.5	Equation	15
$N_{\rm iter}^{\rm max}$	8	Section	3.1.3
$n_{\rm part,min}$	≥50	Section	3.1.3
α	0.5	Equation	20

3.1.6 Intrahalo stellar component

Once galaxies have been identified inside a 3DFOF object, the remaining stellar particles are kept and labelled as intrahalo stellar component (IHSC). The extent, distribution, and shape of this component rely on the definition itself of galaxies (see Appendix A and Fig. A1). The IHSC is therefore all the material that is kinematically different enough from the distribution of any structure in the 3DFOF object. This diffuse component can be associated with either extended stellar haloes on Milky Way like systems or intracluster light in densely populated environments. In-depth analysis of the IHSC is beyond the scope of this work; thus, we address this in upcoming studies (Cañas et al., in preparation).

3.2 Adjustable parameters

Our new algorithm introduces a few tunable parameters, which determine key aspects of how the search is done. We show in Table 1 the values of the parameters used in this work. These values are, however, not fixed and can be modified to achieve different desired results. Here, we briefly describe how modifications to these values can change the identification.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

(i) b – step 1 (Section 3.1.1): As mentioned above our choice of b is the widely adopted $b = 0.2$, which is a good reference to define the extent of dark matter haloes in which we are interested in finding galaxies. This parameter can be changed if a different definition of the extent of FOF dark matter halo is adopted (e.g. $b = 0.28$; Behroozi et al. 2013).

(ii) $f_{x(6D)}$ – step 2 (Section 3.1.2): This parameter shrinks b in order to identify higher overdensities than those of dark matter haloes. From the tests and calibrations we have performed (Appendix A), we found that $f_{x(6D)} = 0.2$ separates most of the galaxies and satellites, leaving only strongly interacting systems linked as a single 6DFOF object. We further discuss the impact of $f_{x(6D)}$ in Appendix A.

(iii) $f_{v(6D)}$ – step 2 (Section 3.1.2): The velocity dispersion σ_v of a 3DFOF object can have different meanings for isolated and highly interacting systems due to the large dynamical range that is covered in cosmological simulations. As our aim is to have an automated algorithm to identify all the galaxies in such simulations, we suggest to keep it σ_v unchanged with $f_v(6D) = 1$. However, $f_{v(6D)} = 1$ is left as a free parameter for the possibility of tuning the initial 6DFOF for specific cases such as zoom simulations and non-cosmological models.

(iv) $f_{v(6D,core)}$ – step 3 (Section 3.1.3): This parameter sets how the velocity linking length scales in each iteration and can impact how many iterations are performed. Small values of $f_{v(6D,core)}$ will lead to fewer iterations, and therefore less use of computational resources; however, the identification of cores can be missed as an aggressively shrinking $l_{v6D,core}$ can cause particles to not be linked. A conservative choice would be values of $f_{v(6D,core)} \sim 1$, which in principle would be able to find all cores; however, this can lead to a very large number of iterations to separate cores, and consequently more use of computational resources, especially for major mergers. For such values a successful separation of all cores will then depend on $N_{\rm iter}^{\rm max}$. From calibration tests we found that values of $0.7 \leq f_{v(6D,core)} \leq 0.8$ successfully separate structures and minimize the total number of iterations.

(v) $f_{n(6D,core)}$ – step 3 (Section 3.1.3): This parameter dictates how the minimum number of particles threshold is modified between iterations. The purpose of this parameter is to avoid identifying small spurious structures, due to shrinking of $l_{v6D,core}$, which happen to be overdense patches in phase space. This parameter is particularly important for galaxy groups and clusters due to the amount of particle bridges caused by the large number of particles in the system and their interactions, and the large dynamical range of galaxy masses within them. The threshold $n_{\rm part,min}^{\ell}$ changes $\propto n_{\rm part,min} \times f_{n(6D,core)}^{m}$ at the $m^{\rm th}$ iteration. Values of $f_{n(6D,core)} \sim 1$ practically do not change $n_{\rm part,min}^{\ell}$, contradicting the purpose of this parameter. Values of $f_{n(6D,core)} \gg 1$ can lead to missing the identification of cores of small galaxies, specially for systems composed of a large number of particles. For example, for $f_{n(6D,core)} = 3$ and starting with $n_{\rm part,min} = 50$, we would require a *core* to have at least 4050 particles at the fourth iteration identified. From calibration tests we found $f_{n(6D,core)} = 1.5$ to give the desired results in a large simulation, such as Horizon-AGN. Deviations of ± 0.1 from the suggested value and starting with $n_{\rm part,min} = 50$ lead to differences of 40 per cent with particle thresholds of 192 for $f_{n(6D,core)} = 1.4$ and 327 for $f_{n(6D,core)} = 1.6$ at the *fifth* iteration, which are reasonable thresholds for the purpose of this parameter.

(vi) $N_{\rm iter}^{\rm max}$ – step 3 (Section 3.1.3): The iterative core search stops when no further cores are found with the parameters at a given iteration. Depending on the choices of $f_{v(6D,core)}$ and $f_{n,\,(6D,core)}$, it is possible that a large number of iterations are needed before the loop stops. The parameter $N_{\rm iter}^{\rm max}$ sets the maximum number of iterations in case the iterative core search has not stopped. Using the values in Table 1, the algorithm stops at the sixth iteration for the largest galaxy cluster in Horizon-AGN at $z = 0$. Choosing $N_{\rm iter}^{\rm max} = 8$ sets a reasonable threshold in case more iterations are needed.

(vii) $n_{\rm part,min}$ – steps 1, 2, and 3 (Sections 3.1.1, 3.1.2, and 3.1.3): This parameter sets a threshold over which structures are considered as relevant. This limit can be adjusted depending on the galaxies of interest. In our study we adopt a value of 50.

(viii) α – step 4 (Section 3.1.4): This parameter sets the strength of the mass-dependent weight to scale phase-space distances from untagged particles to cores. The purpose of this parameter is to compensate between tidal features with large dispersions and compact dense cores with small ones. The value of α can be adjusted depending on the scientific question to be addressed. For identification of galaxies and from our calibrations tests, we found $\alpha = 0.5$ to give the best results. A thorough discussion and comparison of different values of α as well as other choices of w for the core growth can be found in Appendix B.

3.3 Comments

This core-growth method has also been implemented in the VELOCIRAPTOR algorithm to find merging dark matter haloes.

We note that none of the finding algorithms are exempt from finding undesired (spurious) structures. Although for this study most such structures are removed from our galaxy catalogue with the criteria described in equation (21), some spurious structures can still be present if they happen to be not very elongated in phase space and are marginally bound. We leave methods and discussions on this matter for the upcoming VELOCIRAPTOR paper (Elahi et al., in preparation).

Many structure identification codes implement particle unbinding procedures to 'clean' substructures from particles that likely belong to a parent structure. This means that the algorithms are generally focused on finding density peaks (in configuration, velocity, or phase space), while the assignment of particles to these peaks is not well addressed and is generally overlooked (Knebe et al. 2011). VELOCIRAPTOR performs unbinding procedures for dark matter (sub)halo identification. In this study we only use the stellar particle information to identify galaxies; that is, we do not take into account any information from either the gas or dark matter distributions. Therefore, we cannot accurately estimate the true gravitational potential at each particle position to determine whether it is bound or not to a given (sub)structure. The latter is also true for the galaxy catalogues generated by HALOMAKER for the Horizon-AGN simulation (Dubois et al. 2014). We argue that for VELOCIRAPTOR, binding information is included to a certain degree by requiring that particles belonging to the same structure be close in phase space.[4] We stress though that it is crucial how particles with lower densities than the peaks are assigned to them. Even if particle unbinding is fully implemented, if the first guess of what a (sub)structure is is wrong,

[4]This can also be thought of the other way round. The way in which configuration-space-based finders include velocity-space information is by including unbinding procedures. Estimating the kinetic energy of each particle takes into account the information of the relative velocity of a particle with respect to the bulk velocity of a structure (either centred on mass, deepest potential, or highest density); bound particles would then need to be those which are close in phase space.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

no unbinding procedures will fix the problems, as particles would be assigned automatically to its direct host. In the structure-finding codes found in the literature (to the knowledge of the authors), particles are never reassigned from hosts to substructures, unless using temporal information (tracing) to decide where to reassign particles (e.g. HBT, HBT+; Han et al. 2012, 2018).

As it is shown in the following sections, this algorithm is quite efficient and powerful at finding galaxies at all simulation-resolved mass scales, in all environments. We note, however, that this is not the definitive method of finding simulated galaxies, because we do not include baryons in the form of gas. Hence, we may miss gas-dominated dwarf galaxies, which would have very few stellar particles or a bound fraction of particles below our adopted threshold. This is anyway solved by applying conservative particle number thresholds when selecting galaxies. In the future we plan to link gas to galaxies in a similar fashion as we do in the core growth, but to do this properly we need to take into account the thermal energy of the gas. This needs to be carefully implemented to include both particle-based and mesh-based algorithms. Further discussion on this matter is beyond the scope of this paper and is left for future studies.

4 CASE STUDIES

Here we present two case studies in which we compare the results of the improved algorithm of VELOCIRAPTOR and the galaxies from the original catalogue identified with HALOMAKER. With these case studies we address the most challenging cases for galaxy identification, which our new algorithm solves well: (i) strongly interacting and merging galaxies and (ii) robust identification and particle assignment in dense environments, such as galaxy groups and clusters.

4.1 Close interactions

Structure-finding algorithms have been known to struggle to produce robust results when trying to separate dark matter haloes and galaxies in the process of merging (Knebe et al. 2011; Behroozi et al. 2013, 2015). The reason behind this problem is that as structures start to get closer, the particle distributions that describe them start to mix, and separating them becomes a complicated task. For FOF finders, particle mixing creates bridges between the centres of the structures that link them together, while for density threshold algorithms, the mixture of the distribution reduces the contrast between peaks and saddle points in the density field, making it more difficult to correctly identify the components. As particle distributions also mix in phase space, even iterative procedures can struggle to find peaks and to assign particles correctly to structures; hence, host and substructure identities can be swapped between snapshots (see e.g. Behroozi et al. 2015; Poole et al. 2017). Here we show how our improved galaxy-finding algorithm performs in such cases.

We show an example of a close merger in Fig. 2. At a given snapshot, t, the galaxies are still separated, and have masses of $5.61 \times 10^{10}\,\mathrm{M}_\odot$ and $2.88 \times 10^{10}\,\mathrm{M}_\odot$ respectively, giving a merger ratio of $1:1.9$. In a subsequent snapshot, $t + \mathrm{d}t$, HALOMAKER identifies two galaxies with very different masses of 8.25×10^{10} and $4.49 \times 10^{9}\,\mathrm{M}_\odot$, respectively, corresponding to a merger ratio of $1:18$. During a merger, we expect some of the mass of one galaxy to be accreted by the other. However, from visual inspection we can tell that the galaxies have not been well separated by HALOMAKER, as it seems that only the core of one of them has been identified as an individual galaxy while its outer parts have been assigned to its companion. Although two galaxies are identified, the mass of the smallest galaxy is underestimated while the mass of the larger one is overestimated.

We ran VELOCIRAPTOR on the same merger and it can be seen from simple visual inspection that a better result is obtained, despite the complexity of the interaction. The recovered masses of the galaxies are 5.4×10^{10} and $3.08 \times 10^{10}\,\mathrm{M}_\odot$, giving a merger ratio of $1:1.75$. This is in much better agreement with what is measured at t, when galaxies were far enough as to be easily identified by a 3DFOF algorithm. It can be seen that not only are both the galaxy centres found, but the shapes of the galaxies are also well recovered thanks to the improved particle assignment (core-growth) implementation. In order to confirm the latter, we analysed different projections of the stellar mass maps of the galaxies, together with the velocity maps, and found that prior to the merger both galaxies have clear rotation-dominated kinematics, and flattened stellar discs, while during the merger the primary galaxy continues to have rotation-supported kinematics while the secondary galaxy becomes more disturbed. Correctly assigning particles to galaxy centres is crucial for an accurate estimation of the overall properties of galaxies. It affects the ratio of the merger, which in turn can affect the overall minor and major merger rate estimates, especially when only single snapshots are taken into account.

The (in)capability of disentangling structures in such complex interactions might not be considered as a relevant problem for finders, as it is easier to look for the progenitor structures at earlier times when they are still well separated, which ends up not affecting the merger ratio estimation in a major way. However, in general there is not always data available at a high-enough cadence to identify the galaxies at a mass that best represents the merger (e.g. maximum mass as is done by Rodriguez-Gomez et al. 2015) or simply snapshots may not be available. The capability of robustly identifying galaxies in these cases will become more important with the advent of even larger simulations for which storage of a large number of snapshots becomes undesirable and even implausible. The fact that VELOCIRAPTOR succeeds in this task without using any temporal information is a major success of our algorithm.

4.2 Groups and clusters of galaxies

Galaxy identification can be a complex task in galaxy groups and clusters. Stripped material from multiple interactions generates particle bridges and decreases the contrast in the density field, causing problems similar to the ones discussed in Section 4.1. Robust identification of galaxies in such systems is crucial as it can affect a very large number of galaxies. This can in principle affect environmental studies, as well as impacting galaxy population measurements.

We show in Fig. 3 two galaxy clusters in Horizon-AGN that host the two most massive galaxies identified by HALOMAKER. We show the projected stellar density of the full 3DFOF structure (step 1, Section 3.1.1), the central galaxy identified with HALOMAKER, and the VELOCIRAPTOR counterpart; a zoomed-out visualization of the objects is shown in the insets.

We can see that both codes are able to correctly identify a single peak in the central galaxy, meaning that there is no contamination from undetected satellite galaxies. In the zoomed-out images it can be seen that HALOMAKER tends to assign a large number of particles to the central galaxy that belong to other galaxies in the cluster. This leads to the odd bubble shapes observed for the second HALOMAKER galaxy on its top and to its bottom right in the zoomed-out inset. This problem causes the mass and size of the central galaxy to be overestimated. On the other hand, because it searches for structure in phase space, VELOCIRAPTOR is able to identify kinematically

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

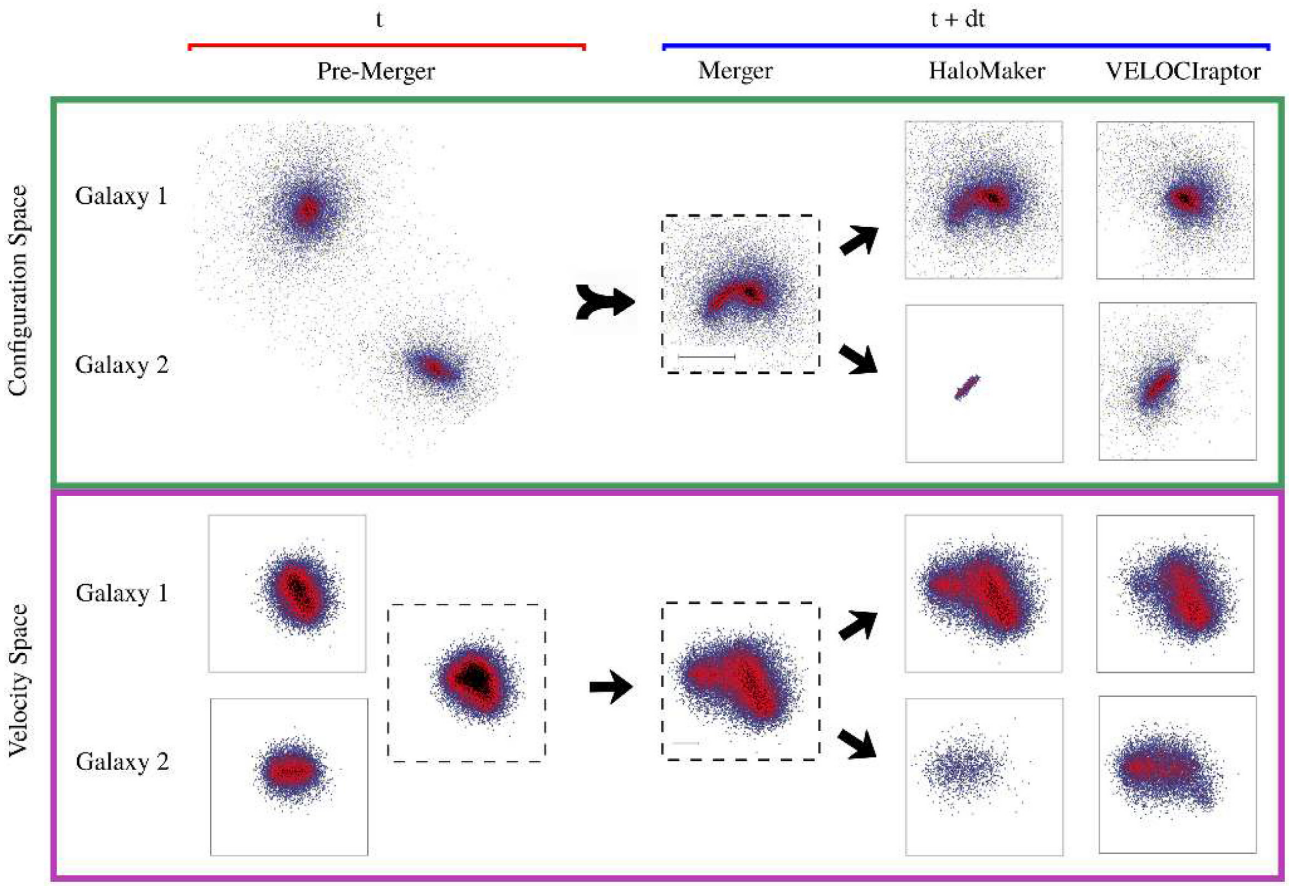

Figure 2. Projected stellar density for a major merger in the Horizon-AGN simulation in configuration space and velocity space, as labelled. t shows the galaxies before the merging occurs and $t + \mathrm{d}t$ during the merger. For clarity, for the velocity-space visualization at t, galaxies are shown both individually and as part of the same velocity space. Galaxies identified with HALOMAKER and VELOCIRAPTOR at $t + \mathrm{d}t$ are shown. We show for each space the projection in which particle distributions are most distinguishable. It can be seen that although HALOMAKER is able to identify two galaxies, it appears that for the small galaxy only the core of it is identified as an individual one, while its outer regions are assigned to its companion. Due to its phase-space implementation for search and core growth, VELOCIRAPTOR is able to find both galaxies and provide a better estimate of their mass and size. The horizontal line in the merger inset shows a length of 20 kpc (200 km s^{-1}), which is the same for all the configuration (velocity)-space insets.

distinct structures, resulting in a better delimitation of the galaxy's boundaries.

This example also demonstrates how crucial the particle assignment is for the robust identification of structures. Although both finders are capable of identifying the cores of the central and satellite galaxies, galaxies can be greatly different due to particle assignment procedures. This occurs for the galaxy in the first column of Fig. 3, where HALOMAKER assigns particles from an orbiting satellite to the central galaxy. Similarly as above, this is seen as a bubble-shaped feature corresponding to the outskirts of the satellite. On the other hand, due to the improvements of particle assignment using phase-space dispersion tensors, VELOCIRAPTOR is able to separate distinct components even if their distributions overlap. This not only produces a better estimation of the masses of the galaxies, but also allows us to recover smooth density profiles of the galaxies, which is important if we are interested in studying radial profiles of galaxy properties.

This problem is not unique to HALOMAKER, but affects structure-finding codes in general. This could in principle be tackled by reassigning procedures for which particles from central galaxies could be returned to any of the other substructures identified. However, as mentioned in Section 3.3 particles are never returned to substructures as particles that were not originally part of a substructure are expected to be bound to the central halo–galaxy system. VELOCIRAPTOR attempts to minimize this issue by carefully assigning particles to cores at each iteration level (Step 4 Section 3.1.4) without any prior assumption on whether cores will become central or satellite galaxies.

4.3 Temporal evolution of galaxy properties

We have shown how our new implementation to find galaxies with VELOCIRAPTOR is capable of identifying galaxies in complex environments. However, a robust algorithm requires that structures be identified consistently over time. This is necessary to ensure that studies focused on the evolution of single galaxies or systems are not affected by the finder. Temporal evolution of structures is either tracked by linking structures across catalogues using merger trees (see for reference Srisawat et al. 2013; Avila et al. 2014; Wang et al. 2016; Poole et al. 2017) or done *on the fly* during structure identification by tracing algorithms (e.g. HBT,HBT+; Han et al. 2012, 2018). It is well known that the evolution traced by merger trees can be severely affected by the specific implementation of structure-finding algorithms (e.g. Avila et al. 2014; Poole et al. 2017). Though

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

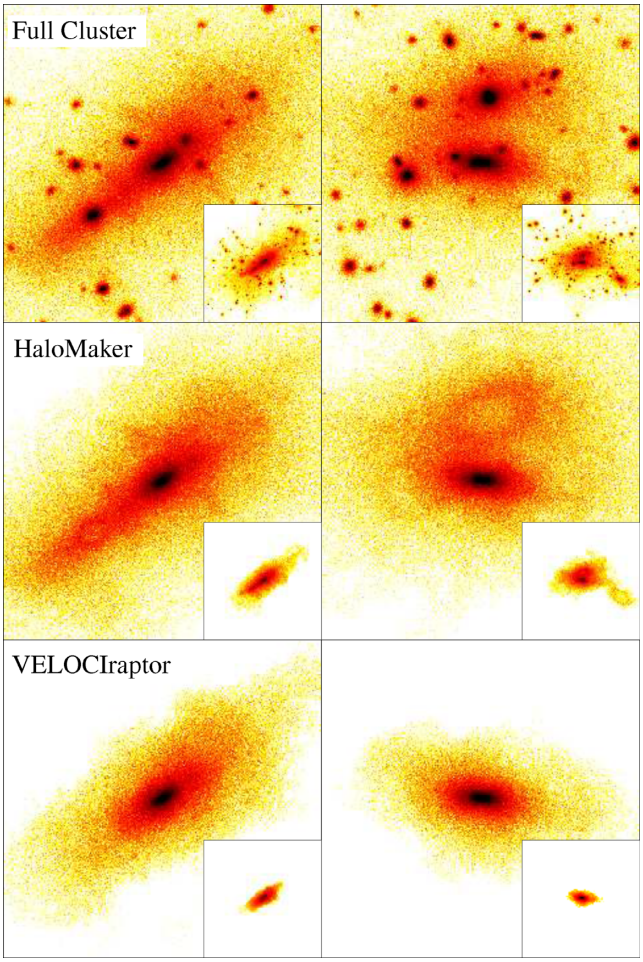

Figure 3. Projected stellar density of the two most massive galaxies found by HALOMAKER (middle row), their respective galaxy cluster (top row), and their VELOCIRAPTOR counterparts (bottom row). Although both codes are able to identify the central galaxy, HALOMAKER fails to separate stellar content that belongs to other galaxies. To emphasize the full extension of the galaxies, insets show a zoomed-out visualization of the same objects. Panels (insets) have a box size of 600 (2000) kpc.

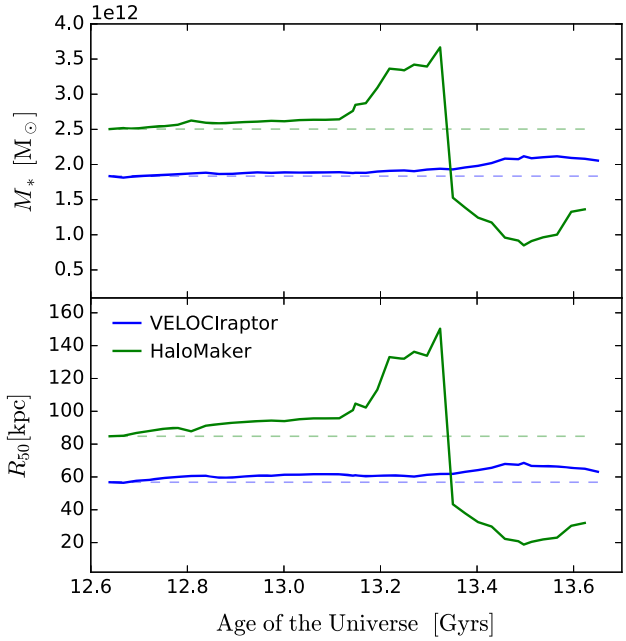

Figure 4. Evolution of the total stellar mass (M_*; upper panel) and spherical half-mass radius (R_{50}; bottom panel) for the central galaxy of the most massive cluster in Horizon-AGN at $z = 0$, as estimated by VELOCIRAPTOR (blue) and HALOMAKER (green). A dashed line is shown as reference for the initial estimated value of each property. VELOCIRAPTOR is capable of following consistently the evolution of bulk galaxy properties in complex environments without applying any temporal corrections. Further details for the other three massive members of the same cluster can be found in Appendix C.

the goal of this study is not focused on testing the consistency of merger trees for our galaxy catalogues, we show in this section how large-scale properties of galaxies as well as their radial distribution evolve for our catalogues.

For this purpose we generate galaxy catalogues on high-cadence snapshots produced for stellar particles only in Horizon-AGN. These catalogues are temporally spaced every ∼25 Myr, with a total of 778 snapshots being available. To test time consistency in the properties of the galaxies identified by our algorithm and HALOMAKER, we selected from the most massive galaxy cluster at $z = 0$ the four most massive galaxies (hereafter referred to as `Galaxy 1-4`, respectively) and followed their evolution backwards in time for 40 of the above-mentioned snapshots, corresponding to ∼1 Gyr of evolution.

We trace galaxies between snapshots using TREEFROG (Elahi et al., in preparation; Poulton et al. 2018), a tool associated with the VELOCIRAPTOR repository to construct merger trees for simulations. Galaxies in a reference snapshot are matched by finding the structure that shares the most particles in a subsequent snapshot. This is done by looking at the individual particle IDs that belong to the galaxies and computing a merit function:

$$\mathcal{M}_{ij} = \frac{N_{\rm sh}^2}{N_i\ N_j}\ . \tag{22}$$

Here, N_i and N_j are the total number of particles in structures i and j, respectively, and $N_{\rm sh}$ is the number of shared particles, thart is which exist in both i and j. This method ensures that galaxies in one snapshot are matched to the galaxy in the subsequent snapshot that is most similar in particle members and that shares a large fraction of those.

The upper panel of Fig. 4 shows the evolution of M_* for the most massive galaxy in the cluster, `Galaxy 1`, found by VELOCIRAPTOR (blue) and HALOMAKER (green). We calculate M_* simply by adding the stellar mass of all the particles in the galaxy. The bottom panel shows the evolution of R_{50}, which is the spherical radius that encloses half of M_*. Solid lines show the evolution of each quantity, and a dashed line shows, as reference, the initial amplitude of each quantity for each finder.

We demonstrate that for VELOCIRAPTOR the evolution of M_* and R_{50} is stable through time. Slight increments and decrements are expected due to the evolution of the galaxy through mergers and interactions. For HALOMAKER it is seen that the evolution for the first ∼500 Myr is quite stable; however, past that point there is huge increment of both M_* and R_{50} for ∼200 Myr, and then there is a sudden drop, which decreases to a minimum at $t = 13.5$ Gyr. During the last 100 Myr, the magnitude of the properties increases steadily with time. The sudden increment at $t = 13.1$ Gyr is consistent with the case studies presented above, which show that central galaxies identified by HALOMAKER include other galaxies' outskirts. In this case, as `Galaxy 1` gets closer to the other massive galaxies,

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

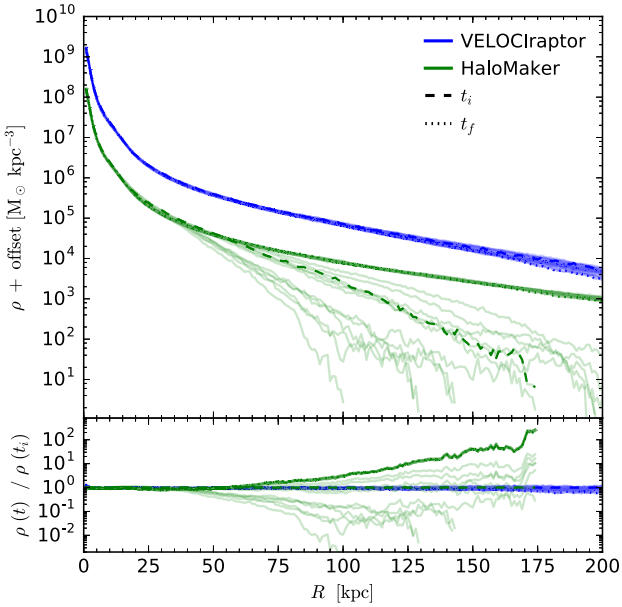

Figure 5. Evolution of the stellar density profile ρ for the same galaxy as that of of Fig. 4 for ∼1 Gyr. The profiles estimated by VELOCIRAPTOR and HALOMAKER are shown by the solid blue and green lines, respectively. HALOMAKER's profile is shifted by −1 dex for clarity. The dashed and dotted lines show the profile measured at the snapshot when we start (t_i) and stop (t_f) tracking the galaxies, respectively. This implementation of VELOCIRAPTOR is also capable of obtaining consistent density profiles to very large radii (>100 kpc) even for massive galaxies with multiple orbiting satellites interacting in a complex galaxy cluster. The evolution of the density profile for the other three most massive galaxies in the same cluster is shown in Appendix C.

HALOMAKER for a short period of time adds their outskirts as part of `Galaxy 1`; the abrupt decrement happens when `Galaxy 1` is not considered to be the central anymore, and galaxies' outskirts are assigned to a companion galaxy, `Galaxy 2`. A visualization of the evolution, as well as the evolution of the properties of the other three massive galaxies in the cluster, can be found in Appendix C.

We further test temporal consistency by measuring the radial stellar mass distribution of the galaxies. The top panel of Fig. 5 shows the stellar volume density profile ρ of `Galaxy 1` produced by VELOCIRAPTOR (blue lines) and HALOMAKER (green lines), from the snapshot where galaxies were first identified t_i ($z = 0$) to the last snapshot used t_f; for HALOMAKER we have offset the profile by −1 dex for clarity. The profile is calculated by adding the mass of all stellar particles inside fixed 1 kpc bins describing concentric spherical shells around the centre of mass of the galaxy and dividing over the volume of the shell. The bottom panel shows the ratio of the density of each bin at time t with respect to the density of the same bin at time t_i as solid lines. The density profiles at t_i and t_f for each finder are shown as dashed and dotted lines, respectively. We show that VELOCIRAPTOR produces not only stable large-scale properties, but also the mass profile of `Galaxy 1`. On the other hand, the HALOMAKER stellar mass profile is only stable for the inner 40 kpc, fluctuating by up to two orders of magnitude at large radii. This is also due to particle assignment, which truncates the outskirts of `Galaxy 1` when it is identified as a satellite rather than the central by HALOMAKER; this is seen as the decrement in the density profile, which corresponds to the 'valley' observed for M_* and R_{50} at $t \gtrsim 13.3$ Gyr. Between $12.6 \leq t/\mathrm{Gyr} \leq 13.1$, HALOMAKER's profile seems stable and smooth for three reasons: (i) The outskirts are not truncated; (ii) even if other galaxies' outskirts are added (producing the bumps in M_* and R_{50}), those particles have much larger radii than 200 kpc; and (iii) even if those particles are asymmetrically distributed with respect to `Galaxy 1`'s centre of mass, the profile looks smooth, as calculating ρ spherically averages the added outskirts (see Appendix C for further details).

5 RESULTS

In this section we study the differences between the HALOMAKER and VELOCIRAPTOR. For this purpose we generate a new galaxy catalogue for the Horizon-AGN simulation using our improved algorithm. In Section 5.1 we compare the catalogues on a galaxy-to-galaxy basis to study how much galaxy properties can be affected by the identification method. In Section 5.2 we investigate how differences in the identification can affect measurements of galaxy population properties.

5.1 Galaxy-to-galaxy comparison

We investigate differences between the finders by performing a galaxy-to-galaxy comparison. Matching structures between catalogues is a process similar to building merger trees. The best match of a galaxy is found by looking at the particle ID information only. Therefore, we use TREEFROG as a catalogue correlator and equation (22) to select the best match.

We compare the total stellar mass M_*, SFR, and sizes $R_{50,90}$ of matched galaxies by computing

$$f_Y = Y_{\mathrm{HaloMaker}}/Y_{\mathrm{VELOCIraptor}}\,, \tag{23}$$

which is the ratio between the above-mentioned quantities, Y, as measured for the HALOMAKER galaxy over the one measured for its VELOCIRAPTOR counterpart. In order to make a proper comparison and avoid resolution effects, we only show the f_Y of galaxies whose total stellar mass is greater than $10^9\,\mathrm{M}_\odot$ in both catalogues, and only matches with $\mathcal{M} > 0.1$ (galaxies sharing $\gtrsim 30$ per cent of particles) are shown.

To properly account for the cases shown in Section 4, we labelled galaxies depending on their degree of interaction as

(i) Isolated – the galaxy is the only structure found in the initial 3DFOF envelope.

(ii) Loosely interacting – the galaxy belongs to a 3DFOF object with multiple structures, and *no* structures were found in its iterative search, that is a single structure in a 6DFOF object.

(iii) Strongly interacting – the galaxy belongs to a 3DFOF object with multiple structures, and one or more additional structures were found in the 6DFOF object iterative search. The most massive galaxy in the 6DFOF object will be referred to as *host*, otherwise as *satellite*. Note that strongly interacting hosts are not necessarily central galaxies of the 3DFOF object (e.g. most massive galaxy of an interacting pair falling into a galaxy cluster).

A total of 81 583 matches fulfil the above criteria at $z = 0$. Of those, a total of 54 113 are isolated; from the remaining 27 470 interacting galaxies, 16 851 (10 619) are loosely (strongly) interacting. Approximately 3 per cent of all the structures in each of the catalogues do not have a counterpart in the other catalogue.

5.1.1 Total stellar mass

We measure the impact of identification on one of the most fundamental properties of a galaxy, the total stellar mass, M_*. We show in

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

Fig. 6 the distributions of the ratio f_M for galaxies in VELOCIRAPTOR in the mass ranges of $10^9 \leq M_*/\mathrm{M}_\odot \leq 10^{10}$, $10^{10} \leq M_*/\mathrm{M}_\odot \leq 10^{11}$, and $M_*/\mathrm{M}_\odot \geq 10^{11}$, which we will refer to as M09, M10, and M11, respectively. The solid line shows the f_M distribution of all the galaxies; the dashed lines show the contribution from both isolated and loosely interacting galaxies (i.e. not strongly interacting); and the dotted lines show the contribution from strongly interacting galaxies. For the latter we show the contribution from host and satellite galaxies as shaded blue and red regions, respectively. Vertical dashed lines are shown as reference at ± 0.2 dex with respect to an identical estimated mass, that is $f_M \equiv 1$.

At $z = 0$, the f_M distributions of the M09, M10, and M11 samples show that galaxies that are not strongly interacting are overall distributed around $f_M = 1$. This is something we would expect because any finder in the literature should not have any problem with the identification of isolated density peaks or particle distributions. Looking closely we see that the peak is slightly skewed towards $f_M > 1$, as a result of differences in the initial galaxy identification steps taken by the codes. HALOMAKER assigns all particles inside a 3DFOF object to identified structures, but VELOCIRAPTOR performs an additional 6DFOF search that delimits galaxies by grouping only phase-space close particles. This procedure effectively gets rid of the furthest phase-space particles, reducing the 'available' mass to distribute between galaxies inside the original 3DFOF object; the mass excess observed for HALOMAKER galaxies is the mass we consider to be part of the IHSC (Section 3.1.6). The reason why the M09 f_M distribution is not as narrow as that of the other sample is likely because all galaxies in M10 and M11 are well resolved, while some galaxies in M09 could still be affected by resolution effects. In Fig. 7 the f_M distributions for all not strongly interacting galaxies in the M09 sample are shown. The majority of isolated and loosely interacting hosts have $f_M > 1$, consistent with the above description. On the other hand, loosely interacting satellites describe a symmetric f_M distribution centred at $f_M = 1$, suggesting that their mass can be either over- or underestimated by HALOMAKER compared to VELOCIRAPTOR.

The greatest difference between the catalogues is seen in the strongly interacting population (dotted lines in Fig. 6). Their f_M distributions at all stellar masses are broader than those of the rest of the population. This shows that there are a non-negligible number of *resolved* galaxies that are affected by the artificial transfer of mass in interacting systems due to the particle assignment criteria of the finder (see Fig. 2 for an example). Host galaxies display a significant preference for $f_M > 1$, while the $f_M < 1$ part of the distribution is predominantly dominated by satellite galaxies, which is more evident for the M11 sample. This picture is consistent with the examples shown in Section 4 (Figs 2 and 3), meaning that the behaviour observed in those examples is not that of simple HALOMAKER outliers but is a recurrent phenomenon in the simulation.

Compared to VELOCIRAPTOR, HALOMAKER overestimates the mass of a considerable amount of strongly interacting satellites; that is, they have $f_M > 1$. As these are interacting systems, a fraction of these cases can be explained by host–satellite swapping, where a galaxy that is considered a satellite by VELOCIRAPTOR is in fact the host galaxy in HALOMAKER. This artificially increases their mass compared to what VELOCIRAPTOR estimates. It is not uncommon to see this phenomenon across catalogues from different finders, and it is even present for the same finder between different snapshots, especially in the case of major mergers (see e.g. Behroozi et al. 2015; Poole et al. 2017).

At higher redshifts, the f_M distributions display an amplified version of the behaviour observed at $z = 0$. Distributions for all galaxies at all masses become wider, and peak further from $f_M > 1$. The f_M distributions at $z = 2$, shown in the middle panels of Fig. 6, are wider, with more prominent wings compared to those at $z = 0$. In addition, the M09 f_M at $z = 2$ peaks at $f_M \approx 1.2$ for not strongly interacting galaxies. Strongly interacting galaxies show a similar behaviour to the $z = 0$ ones, with the $f_M > 1$ region being dominated by host galaxies and the $f_M < 1$ region by satellites, with some fraction of satellites also having $f_M > 1$, likely due to the host–satellite swapping. The f_M distributions at $z = 4$ (right-hand panels in Fig. 6) show an even wider distribution than that at $z = 2$ with a less prominent peak in the case of isolated galaxies.

In order to understand some of the differences at $z = 2$ and $z = 4$ we have to bear in mind the nature of the AMR calculation and the properties of high-redshift galaxies. Horizon-AGN was run using an AMR code for which the grid cells used to compute gravity and hydrodynamics change as the simulation evolves depending on the local density, affecting the effective resolution of the simulation. Cells are allowed to be refined when the universe has an expansion factor of $a = 0.1, 0.2, 0.4, 0.8$ ($z = 0.25, 1.5, 4.0, 9.0$), in order to keep the physical size of the cell somewhat constant. This affects the spatial and mass scales at which gas forms stars, hence the scales on which galaxies are resolved, impacting also their identification. Additionally, such large differences for the same galaxy can be explained by the fact that at high redshifts, galaxies are clumpier and more compact than at the present time. Bursts of star formation within the same galaxy could easily be identified as separate structures, by either of the finders. However, we expect that VELOCIRAPTOR is capable of joining structures that are kinematically similar which might appear as separate structures in the configuration space. Lastly, at high redshifts we also expect the number of mergers to increase (e.g. Fakhouri, Ma & Boylan-Kolchin 2010); hence, we expect that the example analysed in Section 4 becomes more frequent.

5.1.2 *Star formation rate*

Another fundamental quantity measured for galaxies is their star formation rate. We calculate the SFR for each galaxy by adding up the mass of all stellar particles with age smaller than a given Δt threshold and dividing the sum over that period of time. The results presented here were obtained adopting $\Delta t = 50$ Myr and were corrected for a recycling fraction of 0.44 for a Kroupa initial mass function following Courteau et al. (2014), implicitly assuming instantaneous recycling. We have also calculated SFRs using Δt windows of $\Delta t = 20$ and 100 Myr, and find that the results are robust.

We measure $f_{\mathrm{SFR}} = \mathrm{SFR}_{\mathrm{HaloMaker}}/\mathrm{SFR}_{\mathrm{VELOCIraptor}}$ to quantify the galaxy-to-galaxy difference in the estimated SFR. Fig. 8 shows the f_{SFR} distribution at $z = 2$ for samples M09, M10, and M11 as green, red, and blue lines, respectively. Similarly to Fig. 6, we show the total f_{SFR} distribution and the contribution from not strongly interacting galaxies and strongly interacting hosts and satellites, as labelled. f_{SFR} peaks close to $f_{\mathrm{SFR}} = 1$ for not strongly interacting galaxies in all mass ranges.

In this case the f_{SFR} distribution's peak is slightly more narrow compared to the one displayed by the f_M distribution (Fig. 6), and it is also slightly shifted towards $f_{\mathrm{SFR}} > 1$ due to the 6DFOF search of VELOCIRAPTOR that 'removes' the phase-space outermost particles of the initial 3DFOF object, as discussed in Section 5.1.1. The spread in the f_{SFR} distribution comes mostly from strongly interacting galaxies, with $f_{\mathrm{SFR}} > 1$ ($f_{\mathrm{SFR}} < 1$) corresponding to host (satellite)

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

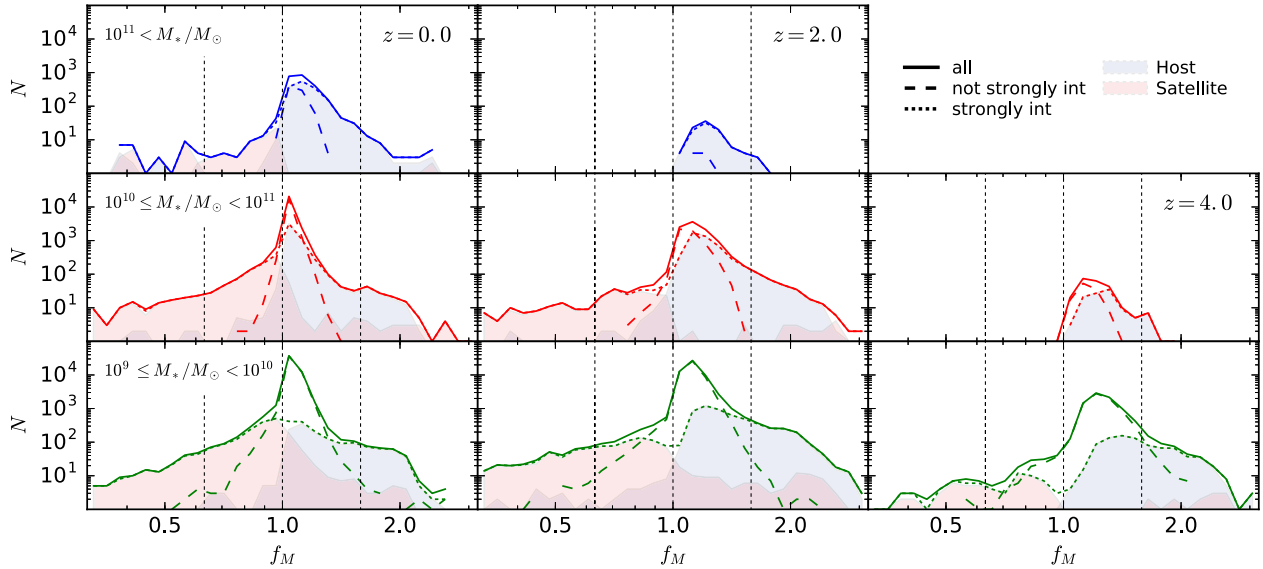

Figure 6. Distributions of the mass ratio $f_M = M_{*\mathrm{HaloMaker}}/M_{*\mathrm{VELOCIraptor}}$ at $z = 0$ (left-hand panels), $z = 2.0$ (middle panels), and $z = 4.0$ (right-hand panels) for different stellar mass ranges, as labelled. The contribution from isolated galaxies and loosely interacting galaxies is shown as a dashed line, from interacting galaxies as a dotted line, and the combined distribution as a solid line. For strongly interacting galaxies the contribution from hosts is shown as a shaded blue region, and that from satellites is shown as a shaded red region. Vertical dashed lines are shown as reference at ± 0.2 dex from an exact match ($f_M \equiv 1$).

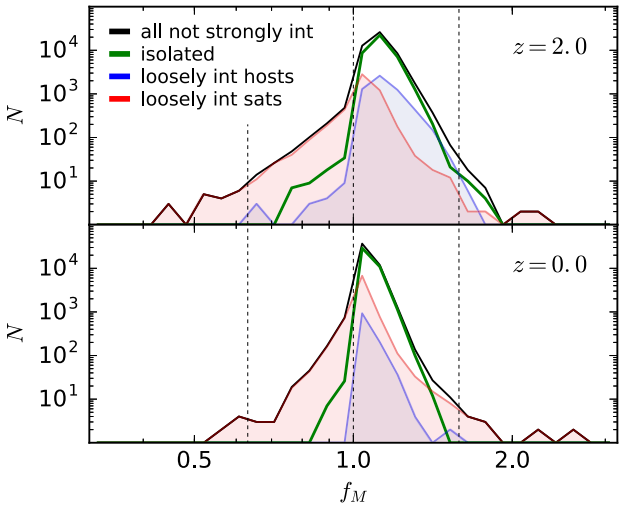

Figure 7. Distributions of the mass ratio f_M for not strongly interacting galaxies with $10^9 \leq M_*/\mathrm{M}_\odot \leq 10^{10}$ at $z = 2$ (top panel) and $z = 0$ (bottom panel). The contribution from isolated galaxies is shown as a solid green line and from loosely interacting hosts and satellites as shaded blue and red regions, respectively. Vertical dashed lines are shown as reference at ± 0.2 dex from an exact match ($f_M \equiv 1$).

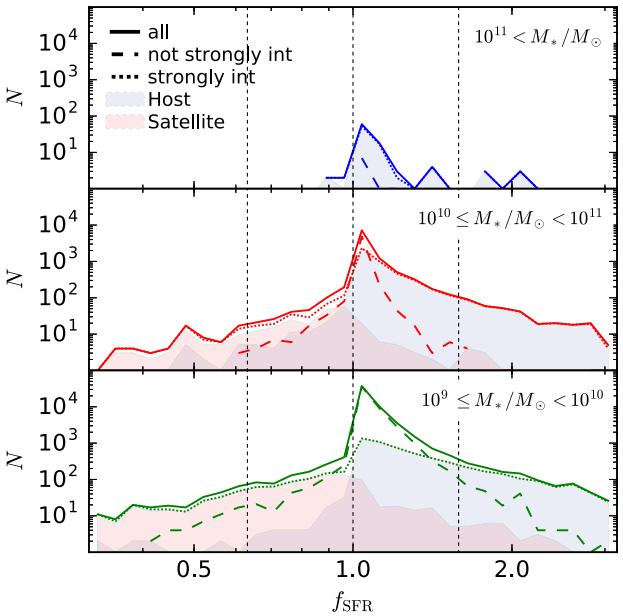

Figure 8. Distribution of the SFR ratio f_{SFR} at $z = 2$ for different mass ranges, as labelled. The contribution from isolated galaxies is shown as a dashed line, that from interacting galaxies as a dotted line, and the combined distribution as a solid line. For interacting galaxies the contribution from host and satellite galaxies is shown as shaded blue and red regions, respectively. For reference vertical dashed lines are shown at ± 0.2 dex from an exact match $f_{\mathrm{SFR}} \equiv 1$.

galaxies. It is interesting though that for host galaxies the f_{SFR} tail is quite prominent and extends to $f_{\mathrm{SFR}} > 2$ at all stellar masses. For satellites, on the other hand, $f_{\mathrm{SFR}} < 1$ tails are more prominent at lower stellar masses. Galaxies whose mass is overestimated via the spurious acquisition of outer material of an orbiting satellite will for instance increase their SFR, and vice versa. Moreover, the satellite galaxies affected by this are likely to have only the inner non-star-forming core as the galaxy (see for example Fig. 2). This will drastically reduce their estimated SFR, while for their hosts it will be enhanced by the incorrect assignment of the star-forming outskirts of the satellite.

5.1.3 *Sizes – enclosed mass radius*

Accurate estimation of galaxy sizes is crucial as they are used not only to test how well galaxy formation models agree with observations, but have been also used for calibration of subgrid physics parameters by some of the present-day hydrodynamical

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

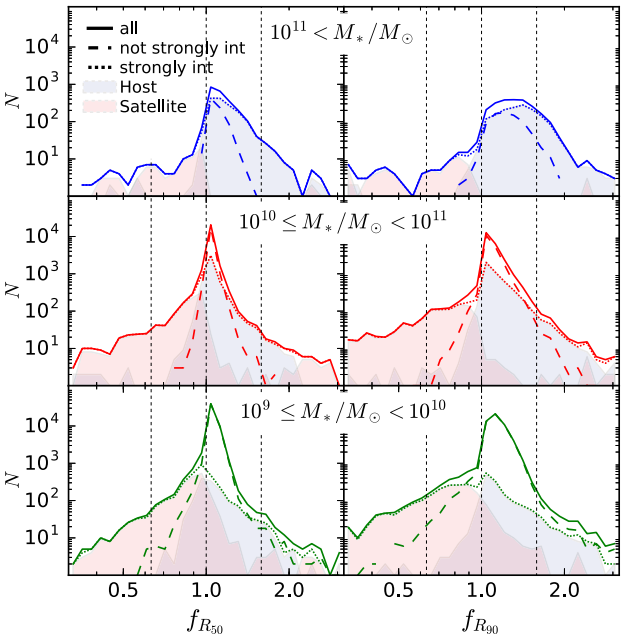

Figure 9. Distribution of the R_{50}, $f_{R_{50}}$, and R_{90}, $f_{R_{90}}$, ratios for different mass ranges at $z = 0$, as labelled. The contribution from not strongly interacting and strongly interacting host and satellite galaxies is shown, as labelled. For reference, vertical dashed lines show ±0.2 dex from $f_{R_X} \equiv 1$.

simulations (e.g Crain et al. 2015). We calculate spherical radii R_{50} and R_{90}, which enclose 50 per cent and 90 per cent, respectively, of the total stellar mass of a galaxy. We show in Fig. 9 the $f_{R_Y} = R_{Y,\mathrm{HaloMaker}}/R_{Y,\mathrm{VELOCIraptor}}$ distribution for R_{50} and R_{90} at $z = 0$. Galaxy samples are colour coded as in Figs 6 and 8.

At all stellar masses the total $f_{R_{50}}$ distribution peaks close to $f_{R_{50}} = 1$, and its tails extend beyond ± 0.3 dex from this value. At $10^9 \leq M_*/\mathrm{M}_\odot \leq 10^{11}$ the peak of $f_{R_{50}}$ comes from not strongly interacting galaxies, while at $M_* > 10^{11}\,\mathrm{M}_\odot$ isolated and interacting host galaxies contribute equally. Isolated and loosely interacting galaxies display a narrow distribution close to $f_{R_{50}} = 1$ at all stellar masses, with its peak being slightly shifted towards $f_{R_{50}} > 1$. Both of these behaviours are similar to those seen for f_M and obey the same reasons (see Section 5.1.1). Not strongly interacting galaxies, however, at $M_* < 10^{10}\,\mathrm{M}_\odot$ show a larger spread on $f_{R_{50}}$, as galaxies can have values from $f_{R_{50}} \lesssim 0.5$ up to $f_{R_{50}} \gtrsim 2.0$; while the former does not contribute largely to the low-$f_{R_{50}}$ tail, isolated and loosely interacting galaxies do contribute to the spread at the high-$f_{R_{50}}$ end.

Similarly to f_M and f_{SFR} (Figs 6 and 8, respectively), strongly interacting galaxies contribute the most to the spread in the $f_{R_{50}}$ distribution. Satellite galaxies comprise the majority of the low-$f_{R_{50}}$ population at all stellar masses, whereas the high-$f_{R_{50}}$ population tail arises from host galaxies, consistent with previous comparisons. This is most evident for galaxies with $M_* > 10^{11}\,\mathrm{M}_\odot$ (top panel), where strongly interacting host galaxies comprise the bulk of the total $f_{R_{50}}$ distribution, which is skewed towards $f_{R_{50}} > 1$. The latter can be seen in the bottom row of Fig. 3, where the zoomed-out insets clearly show that the extension of the galaxy is expanded by HALOMAKER as it also considers other galaxies' outskirts as part of the central.

The overall behaviour of $f_{R_{90}}$ (right-hand panels in Fig. 9) is similar to that of $f_{R_{50}}$; however, the spread is much larger at all stellar masses, for all the galaxy samples. For not strongly interacting galaxies a narrow peak is no longer visible and $f_{R_{90}}$ extends well above 0.2 dex from $f_{R_{90}} = 1$. Although R_{90} encloses almost all the mass of the galaxy, the distributions do not resemble those of f_M at the same redshift (see Fig. 6), especially for $M_* > 10^{11}\,\mathrm{M}_\odot$ galaxies, suggesting that sizes are more sensitive to finder systematics than the stellar mass is. Interacting host galaxies peak at $f_{R_{90}} = 1.41$, showing that the size of central galaxies of groups and clusters is greatly increased by HALOMAKER, which is consistent with the examples shown in Fig. 3.

For completeness purposes we repeated the above analysis for a radius enclosing 20 per cent and 100 per cent of the total stellar mass, and the results are consistent with those for R_{50} and R_{90}, respectively.

5.2 Galaxy population statistics

In this section we study the impact of the identification method on the statistical properties of the galaxy population of Horizon-AGN. We measure standard galaxy properties and compare VELOCIRAPTOR and HALOMAKER catalogues. Our main objective here is not to test how well Horizon-AGN reproduces the observed galaxy population, but to compare how statistical measurements of galaxy population can be affected by identification and the resulting consequences of the biases discussed in Section 5.1.

5.2.1 Galaxy stellar mass function

We start with the simplest measurement, the galaxy stellar mass function. Simulations are often tuned to reproduce this quantity (see review of Somerville & Davé 2015). Moreover, it has been demonstrated that a GSMF consistent with observations can be obtained by tuning subgrid physical model parameters (see e.g. Crain et al. 2015). Therefore, it is essential that we understand and control for all the systematic effects behind measuring the GSMF in our simulations.

We show in Fig. 10 the GSMF, $\Phi_M = \mathrm{d}N/\mathrm{d}\log M_*$, of Horizon-AGN measured with VELOCIRAPTOR (solid blue line) and HALOMAKER (solid green line) at redshifts $z = 0, 2,$ and 4. To quantify the agreement between the catalogues, we compute the relative difference

$$\epsilon = (\Phi_{M_*,\mathrm{HaloMaker}}/\Phi_{M_*,\mathrm{VELOCIraptor}}) - 1\,. \tag{24}$$

We also show the contribution to the GSMF from isolated (solid thin line) and strongly interacting hosts and satellites (dashed and dotted lines, respectively) from the matched galaxies as described in Section 5.1.

At $z = 0$ the overall shape of the GSMF measured by both catalogues is similar. However, differences can be seen at both the low- and high-mass ends. At $M_* < 10^9\,\mathrm{M}_\odot$ HALOMAKER finds fewer galaxies than VELOCIRAPTOR. The GSMF predicted by HALOMAKER displays a declining curve towards lower stellar masses, while VELOCIRAPTOR's GSMF shows a plateau. This can be attributed to two factors. First, at low number of particles, the density field used by HALOMAKER is likely to be poorly sampled, making it possible that structures are not dense enough in the configuration space to be identified. VELOCIRAPTOR is better at picking up these structures as they are dense in the velocity-space as well. This is consistent with comparison studies that have found that in general 6D-based finders tend to perform better at identifying structures with a low number of particles (Knebe et al. 2011, 2013b). The second reason is attributed to the specific particle and density thresholds used by HALOMAKER to define relevant structures. At this mass range,

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

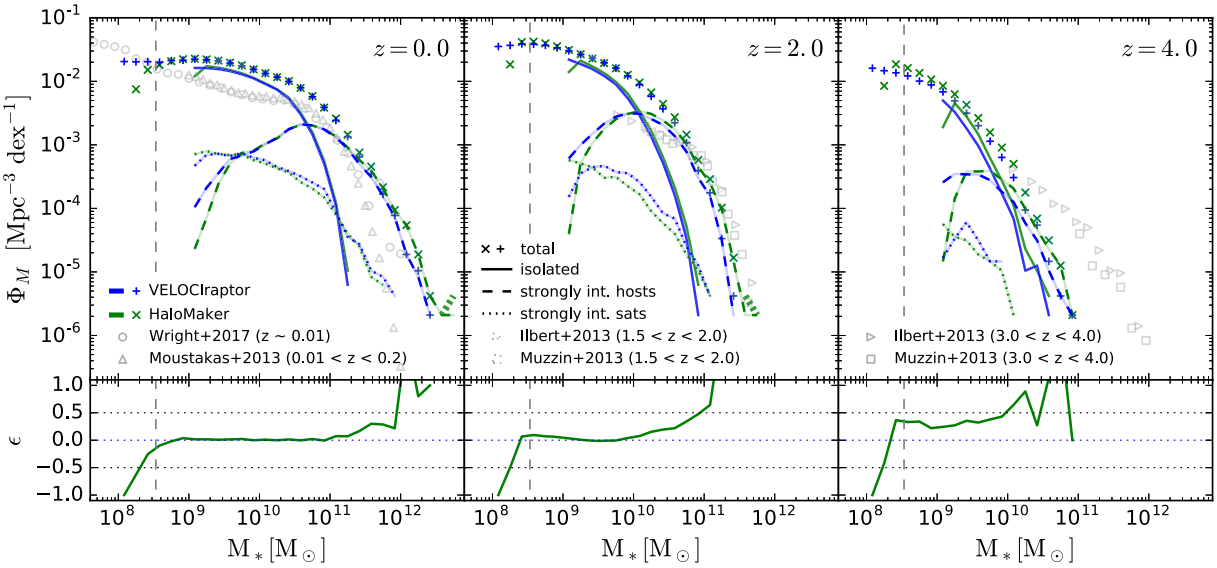

Figure 10. Galaxy stellar mass function of the Horizon-AGN simulation calculated using VELOCIRAPTOR (blue) and HALOMAKER (green), measured at redshifts $z = 0, 2$, and 4, as labelled in each panel. The GSMF of all galaxies is shown as plus and cross symbols for VELOCIRAPTOR and HALOMAKER, respectively. The contribution from not strongly interacting galaxies is shown as a solid thin line and that from strongly interacting hosts and satellites as dashed and dotted lines, respectively. The bottom panels show the relative difference, $\varepsilon = (\Phi_{\rm HaloMaker})/\Phi_{\rm VELOCI} - 1$ of the total GSMF. A dotted thick line is shown for HALOMAKER's total GSMF at stellar masses where VELOCIRAPTOR does not find any structures. A vertical dashed grey line is shown at the mass equivalent to structures composed of 100 particles for reference. Observations of the GSMF from Wright et al. (2017), Moustakas et al. (2013), Muzzin et al. (2013), and Ilbert et al. (2013) are shown as symbols, as labelled.

galaxies are composed of $\lesssim$300 particles, close to the resolution limit of the simulation, making the identification of its peaks and saddle points challenging.

It has been pointed out by several studies that structures need to be composed of at least hundreds or thousands of particles in order to have reliable measurements of their internal properties, as well as resolved merger histories, (e.g Knebe et al. 2013b; Chisari et al. 2015; van den Bosch 2017; Elahi et al. 2018; van den Bosch et al. 2018). We argue that VELOCIRAPTOR is capable of robustly identifying structures at very low particle numbers, as has been shown in other studies (Elahi et al. 2018). However, we leave further analysis of structures with a small number of particles to an upcoming study (Elahi et al., in preparation). Finally, it is worth mentioning that for Horizon-AGN, only galaxies with $M_* > 10^9\,{\rm M}_\odot$ are considered as resolved structures due to resolution.

At $M_* > 10^{11}\,{\rm M}_\odot$, the GSMF of HALOMAKER predicts between 20 per cent and 100 per cent more galaxies than VELOCIRAPTOR. This difference is a result of the IHSC being assigned to the central galaxy in HALOMAKER. Therefore, it is not that VELOCIRAPTOR is unable to find such big galaxies, but the fact that the mass of central galaxies in HALOMAKER is systematically increased by the finder. Note that HALOMAKER's GSMF, shown as a dotted thick green line in Fig. 10, extends to higher masses than that of the most massive galaxy obtained by VELOCIRAPTOR.

Despite the wide f_M distributions shown in Fig. 6, the GSMFs practically overlap in the mass range $10^9 \leq {\rm M}_\odot \leq 10^{11}$. This is partially explained by the peak of the total f_M distribution (Fig. 6), located close to $f_M = 1$, which mostly comes from isolated galaxies (see Fig. 7). The latter are the galaxies that contribute the most to the GSMF at $M_* < 10^{11}\,{\rm M}_\odot$. Another factor is that the over- and underestimation of the stellar mass by HALOMAKER is compensated for between systems of different masses. This effect can be seen at $M_* \gtrsim 10^9{\rm M}_\odot$ from the mass functions of different galaxy populations, where VELOCIRAPTOR predicts more isolated and strongly interacting host galaxies than HALOMAKER. The opposite happens for strongly interacting satellites, giving a total GSMF that agrees at those stellar masses. This is even more evident at $z = 2$ at $10^9 \leq M_*/{\rm M}_\odot \leq 10^{10.5}$, where catalogues predict different numbers of galaxies for different populations, and still the total GSMFs agree relatively well.

The observed difference in the estimation of M_* (as seen in Section 5.1.1) leads to a shift in mass for the GSMF. Such a difference can only be distinguished beyond the break of the GSMF, as the flat slope at lower masses has the effect of making the shift of the GSMF indistinguishable.

At higher redshifts the GSMF has a different behaviour than at $z = 0$. Although the overall shape of the GSMF is roughly similar, there is a clear offset in the normalization. At $z = 2$ (middle panel of Fig. 10) the GSMFs at $M_* < 10^9\,{\rm M}_\odot$ behave similarly to the $z = 0$ ones, except for a slightly higher number density obtained by HALOMAKER compared to VELOCIRAPTOR at $2.5 \times 10^8 \lesssim M_*/{\rm M}_\odot \lesssim 9 \times 10^8$. At $M_* \sim 10^9\,{\rm M}_\odot$, both GSMFs agree well; however, as we go to higher masses, HALOMAKER's GSMF starts to deviate from the one measured by VELOCIRAPTOR, with up to ~50 per cent more galaxies at $M_* \sim 10^{11}\,{\rm M}_\odot$. At $z = 4.0$ (right-hand panel in Fig. 10) the GSMFs of VELOCIRAPTOR and HALOMAKER are completely offset at all masses. For galaxies with $10^{8.5} \lesssim M_*/{\rm M}_\odot \lesssim 10^{9.5}$, HALOMAKER predicts between 30 per cent and 40 per cent more galaxies than VELOCIRAPTOR. This difference increases to 50 per cent up to more than 100 per cent at $M_* \gtrsim 10^{10}\,{\rm M}_\odot$. We can see from the mass function of isolated galaxies and strongly interacting hosts that HALOMAKER predicts more galaxies at all stellar masses. This difference, however, to some degree can be explained by the IHSC that VELOCIRAPTOR is able to separate but which HALOMAKER includes as part of the galaxy. By adding the IHSC mass to their respective central galaxies we could in prin-

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

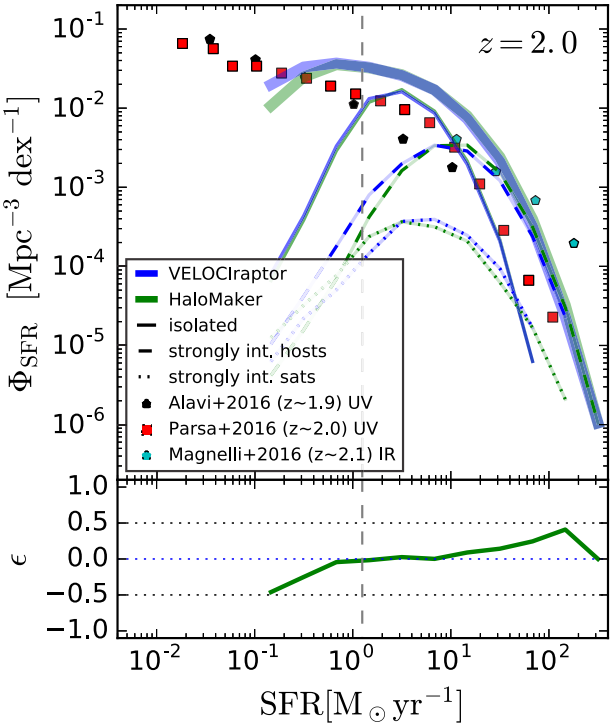

Figure 11. Star formation rate function of Horizon-AGN using VELOCIRAPTOR (blue line) and HALOMAKER (green line) at $z = 2$. The SFRF from all galaxies is shown as a solid thick line; the contribution from isolated galaxies is shown as a solid thin line; and the contribution from strongly interacting hosts and satellites is shown as dashed and dotted lines, respectively. The bottom panels show the $\epsilon = \Phi_{\rm SFR,HaloMaker}/\Phi_{\rm SFR,VELOCIraptor}$. The vertical dashed grey line shows an SFR equivalent to ∼10 star particles formed in the last 50 Myr for reference. From the compilation of observations presented by Katsianis et al. (2017), we show estimations of the SFRF derived from infrared LF from Magnelli et al. (2011), and UV LF from Alavi et al. (2016) and Parsa et al. (2016) are shown as symbols, as labelled.

ciple shift these mass functions to the right, matching those obtained by HALOMAKER. As discussed for Fig. 6, differences between the catalogues at high redshift can also be attributed to (i) higher merger rates, (ii) bursty star formation, and (iii) AMR resolution implementation.

5.2.2 *Star formation rate function (SFRF)*

In Fig. 11 we show the estimated SFRF, $\Phi_{\rm SFR}$, of Horizon-AGN using VELOCIRAPTOR (blue line) and HALOMAKER (green line) at $z = 2$. SFRFs from isolated galaxies as well as strongly interacting hosts and satellites are shown, as labelled. The bottom panel shows the relative difference $\varepsilon = (\Phi_{\rm HaloMaker}/\Phi_{\rm VELOCI}) - 1$, for the total SFRF. The dashed vertical line shows a SFR equivalent to ∼10 new star particles formed in the last 50 Myr for reference.

The overall shape of the total SFRF is in good agreement between the finders. At 0.1 M$_\odot$ yr^{-1}, HALOMAKER predicts 50 per cent less galaxies than VELOCIRAPTOR. These are, however, SFR values close to $\gtrsim 1$ star particle formed in the last 50 Myr. At higher SFRs, the values of SFRFs start to become more similar, reaching a negligible difference at $0.6 \leq$ SFR/ M$_\odot$ yr^{-1} ≤ 6. At these SFRs the total SFRF is principally dominated by isolated galaxies whose SFRFs agree between the catalogues; however, similarly to the GSMF, there is also 'compensation' from different galaxy samples, as VELOCIRAPTOR predicts more strongly interacting hosts than HALOMAKER, but less strongly interacting satellites at SFR $\lesssim 3$ M$_\odot$ yr^{-1}, and vice versa at $3 \lesssim$ SFR / M$_\odot$ yr^{-1} $\lesssim 30$. At SFRs > 30 M$_\odot$ yr^{-1}, HALOMAKER predicts more galaxies than VELOCIRAPTOR, reaching a maximum difference of 50 per cent at ∼100 M$_\odot$ yr^{-1}. This excess on number density predicted by HALOMAKER compared to VELOCIRAPTOR is caused by the addition of mass from other galaxies to the central (see Fig. 3 and the IHSC that is separated in VELOCIRAPTOR). As we showed in Section 5.1.1, systematics affect most of the systems composed of multiple galaxies, and not only contact mergers, hence the differences observed in the estimated SFRFs for different galaxy samples.

5.2.3 *Galaxy mass–size relation*

We show in Fig. 12 the estimated galaxy mass–size relation using R_{50} (circles) and R_{90} (triangles) for the galaxies found by VELOCIRAPTOR (blue) and HALOMAKER (green) in the sample of matched galaxies (as described in Section 5.1), at $z = 0$. Symbols show the median calculated in equal size bins in logarithmic scale, using the same bins for both catalogues. The four panels show the mass–size relation for all galaxies (first panel), isolated galaxies (second panel), and strongly interacting host and satellite galaxies (third and fourth panel, respectively). Vertical dashed lines are shown at the high-mass end where bins have less than 10 galaxies. For reference, the *J*-band mass–size relation linear fit from Lange et al. (2015) is shown as a solid red (magenta) line for early (late)-type galaxies. We apply an average correction to the observations due to the fact that we are measuring 3D sizes in the simulation, while observations measure projected sizes. The latter is a simple scaling of 1.35 applied to the observations (which comes from the fact that galaxies have minor-to-major axial ratios of ≈2 and are inclined by 60 degrees, on average).

The overall shape for the R_{50} and R_{90} mass–size relation of the whole sample of matched galaxies is roughly similar for both finders. At $M_* \gtrsim 10^9$ M$_\odot$, the R_{50} (R_{90}) values of all the galaxies in the sample are on average 10 per cent (20 per cent) larger in HALOMAKER than in VELOCIRAPTOR. At $M_* \sim 10^{10}$ M$_\odot$ this difference reaches a minimum for both radii, being almost negligible for R_{50} and $\gtrsim 5$ per cent for R_{90}. At $M_* > 10^{10.5}$ M$_\odot$, the difference in the sizes of galaxies between the catalogues starts to increase. The difference in the estimated radii of galaxies peaks at $M_* \sim 10^{12}$ M$_\odot$, where on average HALOMAKER galaxies have R_{50} (R_{90}) values up to $\gtrsim 20$ per cent (∼50 per cent) larger than do those of VELOCIRAPTOR. Although there is agreement between both finders at $M_* > 2 \times 10^{12}$ M$_\odot$, the number of galaxies is very low. As discussed in Sections 5.1.1 and 5.2.1, high-mass galaxies are more massive in HALOMAKER than in VELOCIRAPTOR, extending the mass–size relation to larger values. It is interesting though that despite the individual differences seen in Figs 6 and 9, a simple extrapolation of the VELOCIRAPTOR relation would agree with the one described by HALOMAKER.

The mass–size relation for isolated and loosely interacting galaxies, as well as for strongly interacting host galaxies (second and third panel of Fig. 12, respectively), has a similar behaviour as the complete sample at all stellar masses. Smaller R_{50} and R_{90} in VELOCIRAPTOR are expected for isolated galaxies due to its 6DOF implementation, which, as discussed in Section 5.1.1, reduces the stellar particle budget for galaxies and is kept as the IHSC. Although the latter also affects the sizes of interacting host galaxies,

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

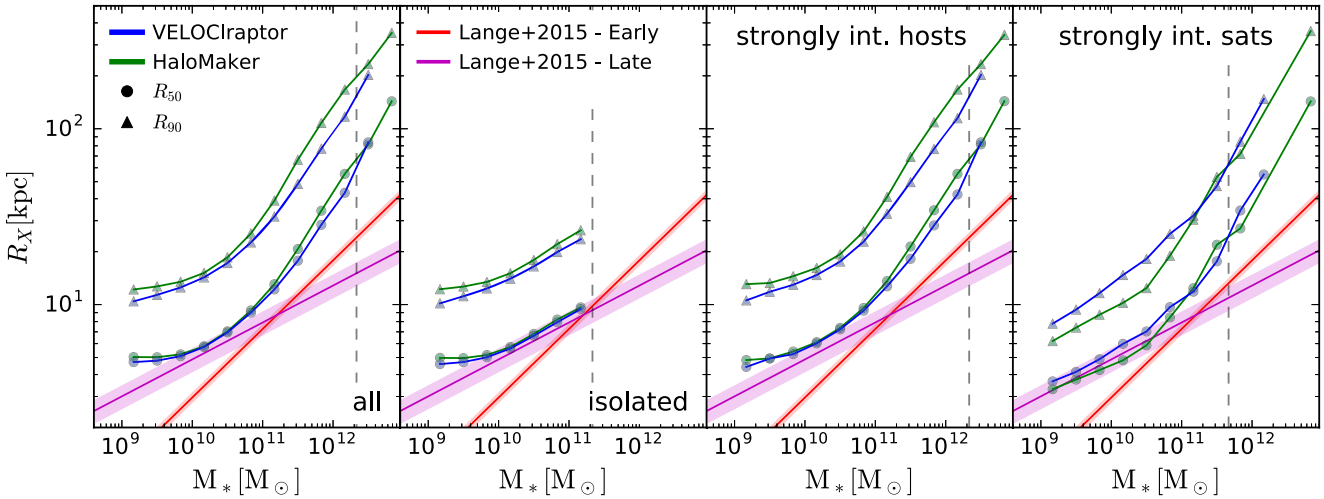

Figure 12. Galaxy size–mass relation as measured by VELOCIRAPTOR (green) and HALOMAKER (blue). The relation for R_{50} and R_{90} is shown as circles and triangles, respectively. The mass–size relations are shown for all the matched galaxies (first panel), isolated galaxies (second panel), and strongly interacting host and satellite galaxies (third and fourth panel, respectively). The dashed grey vertical lines delimit the left edge of mass bins that have less than 10 galaxies. The *J*-band linear fit from Lange et al. (2015) (corrected for the observational 2D projection) for early (late)-type galaxies is shown in red (magenta).

their sizes are again artificially increased because of how particles in a common 3DFOF object are distributed, as was shown in Fig. 3. Both effects are evident at $M_* > 10^{11}\,\mathrm{M}_\odot$ for R_{50}, and to a greater extent for R_{90}. Although at the very high-mass end there seems to be agreement, as discussed above, the number of galaxies for both finders is very low, preventing us from reaching any conclusion.

The difference in the mass–size relation for strongly interacting satellites, however, has a different behaviour compared to other samples. At $M_* > 10^{11}\,\mathrm{M}_\odot$, interacting satellite galaxies are on average more compact in the HALOMAKER catalogue; both R_{50} and R_{90} have on average lower values than their VELOCIRAPTOR counterparts. Similarly to other samples, differences are larger for R_{90} than for R_{50} in the same stellar mass bin. At $M_* \gtrsim 10^9\,\mathrm{M}_\odot$, R_{50} (R_{90}) is on average ~ 10 per cent (~ 20 per cent) smaller in HALOMAKER than in VELOCIRAPTOR. At higher stellar masses the difference increases, reaching a maximum at $\sim 10^{10.5}\,\mathrm{M}_\odot$ with galaxies being on average ~ 20 per cent and ~ 30 per cent smaller for R_{50} and R_{90}, respectively, in HALOMAKER compared to VELOCIRAPTOR. At $10^{11} \leq M_*/\mathrm{M}_\odot \leq 5 \times 10^{11}\,\mathrm{M}_\odot$ there are two mass bins where the R_{50} and R_{90} of interacting satellites are on average similar and even larger in HALOMAKER, contrary to what would be expected. This is likely to be caused by host–satellite swapping, and can be seen in Fig. 9 as a small bump at $f_{R_X} > 2$. At $M_* > 10^{12}\,\mathrm{M}_\odot$ R_{50} (R_{90}) is on average $\sim$15 per cent (~ 20 per cent) smaller for HALOMAKER satellites.

6 DISCUSSION

We have presented an improved algorithm for identifying galaxies in simulations and showed how galaxy properties are affected by the finding algorithm. In this section we discuss implications of our algorithm, as well as possible consequences that non-robust identification of galaxies can have in cosmological hydrodynamical simulations.

6.1 Identifying galaxies versus dark matter haloes

Many structure-finding codes are capable of finding galaxies in simulations (see for reference Knebe et al. 2013a). The vast majority of them are generally limited to either taking all bound baryons inside a dark matter (sub)halo and labelling them as the galaxy or using the same algorithm and parameters adopted for dark matter haloes to identify galaxies. Although both approaches are valid for the identification of galaxies, there are important differences between dark matter haloes and galaxies to keep in mind: (i) Stars that make up galaxies are formed from gas elements at the bottom of potential wells, and hence galaxies are expected to be more compact than dark matter haloes; such gas elements can also cool into discs, the geometry of which is very different from the geometry of dark matter haloes. Consequently, for a large fraction of the galaxies, the stellar density profiles do not resemble those of their dark matter counterpart. (ii) Taking into account this variety of shapes and distributions is extremely important for the identification of merging and interacting galaxies, as such morphologies can distort and become quite complex, making their identification a non-trivial task. (iii) During mergers, the outer dark matter component will at some point phase-mix, but the stars in its centre do that on a different time-scale, with some features being long-living (such as streams and shells). This makes it important to analyse them separately. Even if stars and dark matter are both collisionless in simulations and interact solely through gravity, we should not use the same approach if codes were designed under assumptions that are valid only for dark matter haloes. Although for some galaxies the above approaches might work, that is not expected to be the case for the entire galaxy population in cosmological simulations, as we have shown in this study.

The algorithm presented here is *a* solution to tackle this problem. It is particularly powerful as it was designed to work without any *a priori* assumption on shape or distribution, and is capable of handling the large dynamical range covered in cosmological simulations. It is therefore a generalized solution that can also be easily

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

applied to other components in simulations, which we explore in upcoming studies (Elahi et al., in preparation).

6.2 Impact of identification

6.2.1 Simulation results

We have shown in this study how the total mass, size, and star formation rate of galaxies can be affected by the assumptions and sometimes oversimplification of the finder. Additionally, as seen in the case studies (Section 4), misestimation of masses of merging galaxies impacts the estimated merger ratio. This has several consequences as galaxy mergers are essential for the growth of massive galaxies (Robotham et al. 2014). Inability to resolve galaxies in interactions can affect estimated merger ratios and therefore estimated minor and major merger rates, and the impact they have on the build-up of galaxies. A related area of great interest is whether interactions enhance/suppress the star formation activity in galaxies (Ellison et al. 2008; Davies et al. 2015; Kaviraj et al. 2015; Martin et al. 2017; Davies et al., submitted). As shown in this work galaxies that are strongly interacting have their SFRs affected within a fraction of 2 to 3 mostly due to how particles in the outskirts of galaxies are assigned to density peaks. The latter is comparable to the enhancements inferred observationally, showing that it is critical to robustly measure SFRs in simulated galaxies if we want to use them to offer physical interpretations at these environmental trends.

Misestimation of masses and sizes can impact the interpretation that we can give to galaxies in dense environments such as groups and clusters, where both the central and satellites can be largely affected. This implies that we could get misleading results when studying environmental effects on galaxy quenching in hydrodynamical simulations.

These are not the only possible consequences. Our case studies also showed that radial profiles can be affected, such as angular momentum and inertia tensors, both used for alignment studies. Regarding angular momentum, it has been recently shown by for example Cortese et al. (2016) in observations and Lagos et al. (2017) in simulations that the estimated specific angular momentum can be up to ∼2.5 times (∼0.4 dex) higher if measured at two effective radii rather than one. It is therefore important for related studies in simulations to account for systematic effects that can severely affect the estimated sizes (e.g R_{50}) of interacting galaxies. Taking into account the offset we found when comparing HALOMAKER with our new algorithm in Section 5.1.3, we would expect 3D finders to bias the specific angular momentum of satellites (centrals) towards low (high) values.

6.2.2 The effects on our understanding of galaxy formation

We showed in this study that despite the large differences seen in individual galaxies, especially the interacting ones, the overall galaxy population statistics are not severely affected by finder systematics. This has important implications for the way the galaxy formation is modelled and understood. We have already stated that population statistics are used to tune free parameters of subgrid physics models in simulations. To a certain extent, through tuning we can learn how recipes affect galaxies as well as the impact of different models, for example star formation and stellar and AGN feedback. However, we have shown that we can obtain the right amplitude of a relation or function (i.e. stellar mass function or mass–size relation) using vastly different finders but for different reasons. We argue that the study of subsamples of the galaxy population (e.g. satellite galaxies and galaxy groups/clusters) can unveil such differences, and therefore provide key information to estimate the systematic effects introduced by the choice of finder. Subgrid physics often model unresolved and generally not-so-well understood physical processes that can affect the large-scale properties of galaxies. This is the case of BH growth and its corresponding AGN feedback. A major growth channel of BHs is mergers, and we have shown that the choice of finder affects the derived merger ratio. This in turn affects our estimates of BH-merging time-scales, possibly causing the existence of multiple BHs in merger remnants, and thus changing the associated effect of AGN feedback on the galaxy properties.

Overall, there are many unknowns in simulations, and the exact way in which one decides to compare with observations or even among simulations is a non-trivial task. In this paper, we focus on the effect the galaxy identification has on the derived galaxy properties, and in many cases those differences will be smaller than other uncertainties, such as the exact way one measures a property (Stevens et al. 2014), or the systematic effects the physical modelling itself has on the predicted population. However, in some cases (such as galaxy mergers and satellite galaxies in dense environments), the bias introduced by the chosen algorithm could be a dominant effect.

All this shows that perfecting our ability to identify galaxies and measure their properties in simulations is a key task that cannot be overlooked. Our new algorithm offers a new, robust, and accurate way of doing this, yielding smoother stellar profiles (see Figs 2 and 3) and more robust stellar mass estimates than do widely used 3D finders. This implementation of our algorithm in VELOCIRAPTOR will be made public in the next release of the code (Elahi et al., in preparation).

7 SUMMARY AND CONCLUSIONS

We have extended the halo-finder code VELOCIRAPTOR (Elahi et al. 2011; Elahi et al., in preparation) to robustly identify galaxies in state-of-the-art simulations of galaxy formation. This new implementation overcomes many common problems that even state-of-the-art structure-finding codes struggle with, such as particle assignment and accurate identification of strongly interacting systems. We have paid special attention to the appropriate selection and iterative adjustment of search parameters, to account for the wide dynamical range that simulations can have. Particle assignment (core growth) was improved by using the full phase-space dispersion tensor, allowing us not only to recover arbitrary galaxy shapes, but also to obtain smooth density profiles even for galaxies with satellites embedded within them.

With our improved code, we built an additional galaxy catalogue for the state-of-the-art cosmological hydrodynamical simulation Horizon-AGN and compared its outcomes with those of the complex configuration-space-based finder HALOMAKER. Case studies confirmed the versatility and robustness of our algorithm, and provided insight into how identification tools can affect galaxy properties (e.g. mass and sizes) as well as the estimates of merger ratios. Below we summarize our main results.

Galaxy-to-galaxy comparison. We matched the galaxy catalogues to quantify how the total M_*, SFR, and sizes (R_{50} and R_{90}) can be affected by the chosen finder. We built distributions of

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

$f_Y = Y_{\rm HaloMaker}/Y_{\rm VELOCIraptor}$, where Y corresponds to each of the properties above, and separated the contributions from isolated and interacting galaxies. Interacting galaxies are those hosted by haloes with more than one substructure; otherwise galaxies are considered as isolated.

(i) Isolated galaxies are in general narrowly distributed close to $f_Y = 1$ for M_*, SFR, and R_{50}. Such similarities between catalogues are expected as the identification of isolated galaxies should be straightforward. For R_{90}, however, the peak is not narrow and a considerable amount of isolated galaxies have $f_{R_{90}}$ values around $\pm$ 0.3 dex from $f_{R_{90}} = 1$. This suggests that R_{90} is highly dependent on the finding algorithm.

(ii) Interacting galaxies show a very wide f_Y distribution for all quantities studied. There is an evident difference between host and satellite galaxies, which peak at $f_Y > 1$ and $f_Y < 1$, respectively. These differences are mainly caused by inadequate particle assignment in HALOMAKER, which we show our improved version of VELOCIRAPTOR handles better. HALOMAKER artificially increases (decreases) the estimated values of M_*, SFR, and $R_{50,90}$ for host (satellite) galaxies in interacting systems.

(iii) Differences between the catalogues are amplified at higher redshifts, where the f_Y distributions of interacting and isolated galaxies widen.

Galaxy population statistics. We investigate how the choice of finder affects the overall galaxy population statistics. We explore the GSMF, SFRF, and mass–size relation for R_{50} and R_{90}.

(i) At $M_* < 10^{11}\,{\rm M}_\odot$ the $z = 0$ GSMFs of HALOMAKER and VELOCIRAPTOR agree well, while at higher stellar masses the former predicts from 20 per cent to 100 per cent more galaxies than the latter. At higher redshifts differences are amplified. At $z = 4$ HALOMAKER's GSMF predicts a number density of galaxies at least 30 per cent higher than that of VELOCIRAPTOR over the whole mass range, increasing to 100 per cent at $M_* \gtrsim 10^{10}\,{\rm M}_\odot$.

(ii) The SFRF at $z = 0$ is also fairly similar between the finders, with differences increasing with redshift. At $z = 2$, the peak of the cosmic star formation history, HALOMAKER predicts up to 50 per cent more galaxies at $3 \leq {\rm SFR}/\,{\rm M}_\odot\,{\rm yr}^{-1} \leq 100$ than VELOCIRAPTOR. This is important as these galaxies are expected to dominate the cosmic SFR.

(iii) We compare the R_{50} and R_{90} size–mass relation predicted by both finders. We find that the R_{50} mass–size relations resulting from the two finders are similar, except at the high-mass end, $M_* \simeq 10^{12}\,{\rm M}_\odot$, where HALOMAKER'S galaxies are 20 per cent larger than VELOCIRAPTOR'S. These differences increase by 30 per cent when we study R_{90}. This results from the fact that the stellar content and structure in the outskirts of galaxies are very sensitive to the choice of finder.

Although we see that the overall $z = 0$ galaxy statistics are not greatly impacted by the choice of finder, individual galaxies can display differences in mass and size of more than a factor of 3 between the two finders studied here. We suggest that the tuning of simulations of galaxy formation is relatively robust as it has consistently focused on population statistics. However, comparisons of galaxy subpopulations with observations, specifically in the context of pairs, groups, and clusters, can be greatly affected by the choice of finder. We showed that our new algorithm outperforms 3D finders and provides extensive evidence of this.

One of our key findings is that the stellar outskirts of galaxies is greatly affected by the choice of finder. In upcoming studies we will explore in detail the diffuse IHSC, the stellar streams, and the outer stellar profiles of galaxies. Another important area of investigation will be comparing our theoretical measurements of the diffuse stellar halo with observations, by mimicking the observational effects, such as selection, surface brightness biases, among others.

ACKNOWLEDGEMENTS

The authors thank Aaron Robotham, Matthieu Schaller, and the theory and computing group at ICRAR for helpful discussions. RC is supported by the MERAC foundation postdoctoral grant awarded to CL and by the Consejo Nacional de Ciencia y Tecnología CONACYT CVU 520137 Scholar 290609 Overseas Scholarship 438594. CL is funded by a Discovery Early Career Researcher Award (DE150100618). Parts of this research were conducted by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. This research is part of Spin(e) (ANR-13-BS05-0005; http://cosmicorigin.org). This work has made use of the Horizon Cluster hosted by Institut d'Astrophysique de Paris. We thank Stephane Rouberol for running this cluster smoothly for us.

REFERENCES

Alavi A. et al., 2016, ApJ, 832, 56
Aubert D., Pichon C., Colombi S., 2004, MNRAS, 352, 376
Avila S. et al., 2014, MNRAS, 441, 3488
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109
Behroozi P. et al., 2015, MNRAS, 454, 3020
Bentley J. L., 1975, Commun. ACM, 18, 509
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Berlind A. A., Narayanan V. K., Weinberg D. H., 2001, ApJ, 549, 688
Binney J., Tremaine S., 2008, Galactic Dynamics, 2nd edn. Princeton Univ. Press, Princeton, NJ
Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., Lemson G., 2009, MNRAS, 398, 1150
Chisari N. et al., 2015, MNRAS, 454, 2736
Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168
Cortese L. et al., 2016, MNRAS, 463, 170
Courteau S. et al., 2014, Rev. Mod. Phys., 86, 47
Crain R. A. et al., 2015, MNRAS, 450, 1937
Davies L. J. M. et al., 2015, MNRAS, 452, 616
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Diemand J., Kuhlen M., Madau P., 2006, ApJ, 649, 1
Dubois Y., Devriendt J., Slyz A., Teyssier R., 2010, MNRAS, 409, 985
Dubois Y., Devriendt J., Slyz A., Teyssier R., 2012, MNRAS, 420, 2662
Dubois Y. et al., 2014, MNRAS, 444, 1453
Dubois Y., Peirani S., Pichon C., Devriendt J., Gavazzi R., Welker C., Volonteri M., 2016, MNRAS, 463, 3948
Eisenstein D. J., Hut P., 1998, ApJ, 498, 137
Elahi P. J., Thacker R. J., Widrow L. M., 2011, MNRAS, 418, 320
Elahi P. J. et al., 2013, MNRAS, 433, 1537
Elahi P. J. et al., 2016, MNRAS, 458, 1096
Elahi P. J., Welker C., Power C., Lagos C. d. P., Robotham A. S. G., Cañas R., Poulton R., 2018, MNRAS, 475, 5338
Ellison S. L., Patton D. R., Simard L., McConnachie A. W., 2008, AJ, 135, 1877
Fakhouri O., Ma C.-P., Boylan-Kolchin M., 2010, MNRAS, 406, 2267
Frenk C. S. et al., 1999, ApJ, 525, 554

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

Furlong M. et al., 2015, MNRAS, 450, 4486
Genel S. et al., 2014, MNRAS, 445, 175
Gill S. P. D., Knebe A., Gibson B. K., 2004, MNRAS, 351, 399
Giocoli C., Tormen G., van den Bosch F. C., 2008, MNRAS, 386, 2135
Giocoli C., Tormen G., Sheth R. K., van den Bosch F. C., 2010, MNRAS, 404, 502
Han J., Jing Y. P., Wang H., Wang W., 2012, MNRAS, 427, 2437
Han J., Cole S., Frenk C. S., Benitez-Llambay A., Helly J., 2018, MNRAS, 474, 604
Ilbert O. et al., 2013, A&A, 556, A55
Katsianis A. et al., 2017, MNRAS, 472, 919
Kauffmann G., Charlot S., 1998, MNRAS, 294, 705
Kaviraj S., Devriendt J., Dubois Y., Slyz A., Welker C., Pichon C., Peirani S., Le Borgne D., 2015, MNRAS, 452, 2845
Kaviraj S. et al., 2017, MNRAS, 467, 4739
Kim J.-h. et al., 2014, ApJS, 210, 14
Klypin A., Holtzman J., 1997, preprint (arXiv:astro-ph/9712217)
Klypin A., Gottlöber S., Kravtsov A. V., Khokhlov A. M., 1999, ApJ, 516, 530
Knebe A. et al., 2011, MNRAS, 415, 2293
Knebe A. et al., 2013a, MNRAS, 428, 2039
Knebe A. et al., 2013b, MNRAS, 435, 1618
Knebe A. et al., 2015, MNRAS, 451, 4029
Komatsu E. et al., 2011, ApJS, 192, 18
Lacey C., Cole S., 1993, MNRAS, 262, 627
Lagos C. d. P., Theuns T., Stevens A. R. H., Cortese L., Padilla N. D., Davis T. A., Contreras S., Croton D., 2017, MNRAS, 464, 3850
Lange R. et al., 2015, MNRAS, 447, 2603
Maciejewski M., Colombi S., Springel V., Alard C., Bouchet F. R., 2009, MNRAS, 396, 1329
Magnelli B., Elbaz D., Chary R. R., Dickinson M., Le Borgne D., Frayer D. T., Willmer C. N. A., 2011, A&A, 528, A35
Martin G., Kaviraj S., Devriendt J. E. G., Dubois Y., Laigle C., Pichon C., 2017, MNRAS, 472, L50
Moustakas J. et al., 2013, ApJ, 767, 50
Muzzin A. et al., 2013, ApJ, 777, 18
Nelson D. et al., 2018, MNRAS, 475, 624
Onions J. et al., 2012, MNRAS, 423, 1200
Parsa S., Dunlop J. S., McLure R. J., Mortlock A., 2016, MNRAS, 456, 3194
Pillepich A. et al., 2018, MNRAS, 473, 4077
Poole G. B., Mutch S. J., Croton D. J., Wyithe S., 2017, MNRAS, 472, 3659
Poulton R. J. J., Robotham A. S. G., Power C., Elahi P. J., 2018, preprint (arXiv:1809.06043)
Power C., Read J. I., Hobbs A., 2014, MNRAS, 440, 3243
Press W. H., Schechter P., 1974, ApJ, 187, 425
Robotham A. S. G. et al., 2014, MNRAS, 444, 3986
Rodriguez-Gomez V. et al., 2015, MNRAS, 449, 49
Scannapieco C. et al., 2012, MNRAS, 423, 1726
Schaye J. et al., 2015, MNRAS, 446, 521
Sembolini F. et al., 2016a, MNRAS, 457, 4063
Sembolini F. et al., 2016b, MNRAS, 459, 2973
Snyder G. F. et al., 2015, MNRAS, 454, 1886
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
Srisawat C. et al., 2013, MNRAS, 436, 150
Stadel J. G., 2001, PhD thesis, Univ. Washington
Stevens A. R. H., Martig M., Croton D. J., Feng Y., 2014, MNRAS, 445, 239
Teyssier R., 2002, A&A, 385, 337
Tormen G., Moscardini L., Yoshida N., 2004, MNRAS, 350, 1397
Trayford J. W. et al., 2015, MNRAS, 452, 2879
Trayford J. W., Theuns T., Bower R. G., Crain R. A., Lagos C. d. P., Schaller M., Schaye J., 2016, MNRAS, 460, 3925
Tweed D., Devriendt J., Blaizot J., Colombi S., Slyz A., 2009, A&A, 506, 647
van den Bosch F. C., 2017, MNRAS, 468, 885
van den Bosch F. C., Ogiya G., Hahn O., Burkert A., 2018, MNRAS, 474, 3043
Vogelsberger M. et al., 2014, MNRAS, 444, 1518
Volonteri M., Dubois Y., Pichon C., Devriendt J., 2016, MNRAS, 460, 2979
Wang Y. et al., 2016, MNRAS, 459, 1554
Welker C., Devriendt J., Dubois Y., Pichon C., Peirani S., 2014, MNRAS, 445, L46
Wright A. H. et al., 2017, MNRAS, 470, 283

APPENDIX A: 6DFOF LINKING LENGTH

As described in Section 3.1.2, we use a fraction $f_{x,(6D)}$ of b for the configuration-space linking length, described in equation (10). We show in Fig. A1 how the particular choice of $f_{x,(6D)}$ can affect how galaxies are delimited in VELOCIRAPTOR, and how its IHSC changes with it.

Values of $0.6 \lesssim f_{x,(6D)} \lesssim 1$ link a large number of substructures to the 6DFOF object. Though smaller values get rid of some of the substructures in the outskirts of the central/group, they can also shrink the size of the central galaxy to only the inner parts as is seen for $f_{x,(6D)} = 0.1$. The latter is appreciated in the IHSC, where bubble-like features can be seen at the position of galaxies, as if the central parts of the galaxies were carved leaving the outer parts unassigned to any galaxy. Mid-range values of $0.2 \lesssim f_{x,(6D)} \lesssim 0.4$ leave out some of the outskirts of the group without leaving bubble-like structures in the IHSC. For this study our preferred choice is $f_{x,(6D)} = 0.2$ as it leaves out most of the satellites while preserving structures that can only be separated using the iterative search in the 6DFOF object.[5]

Though it can be seen that the estimated IHSC depends on the $f_{x,(6D)}$ used, this example is illustrative of upcoming studies focused on the IHSC (Cañas et al., in preparation).

[5]It is important to notice that even with $f_{x,(6D)} = 0.1$ a small satellite still remains as part of the 6DFOF object (this can be seen by zooming the figure in the electronic version), and that independently of the $f_{x,(6D)}$ value used, the inner profile of the galaxy is smooth, showing the power of the iterative search and core growth.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

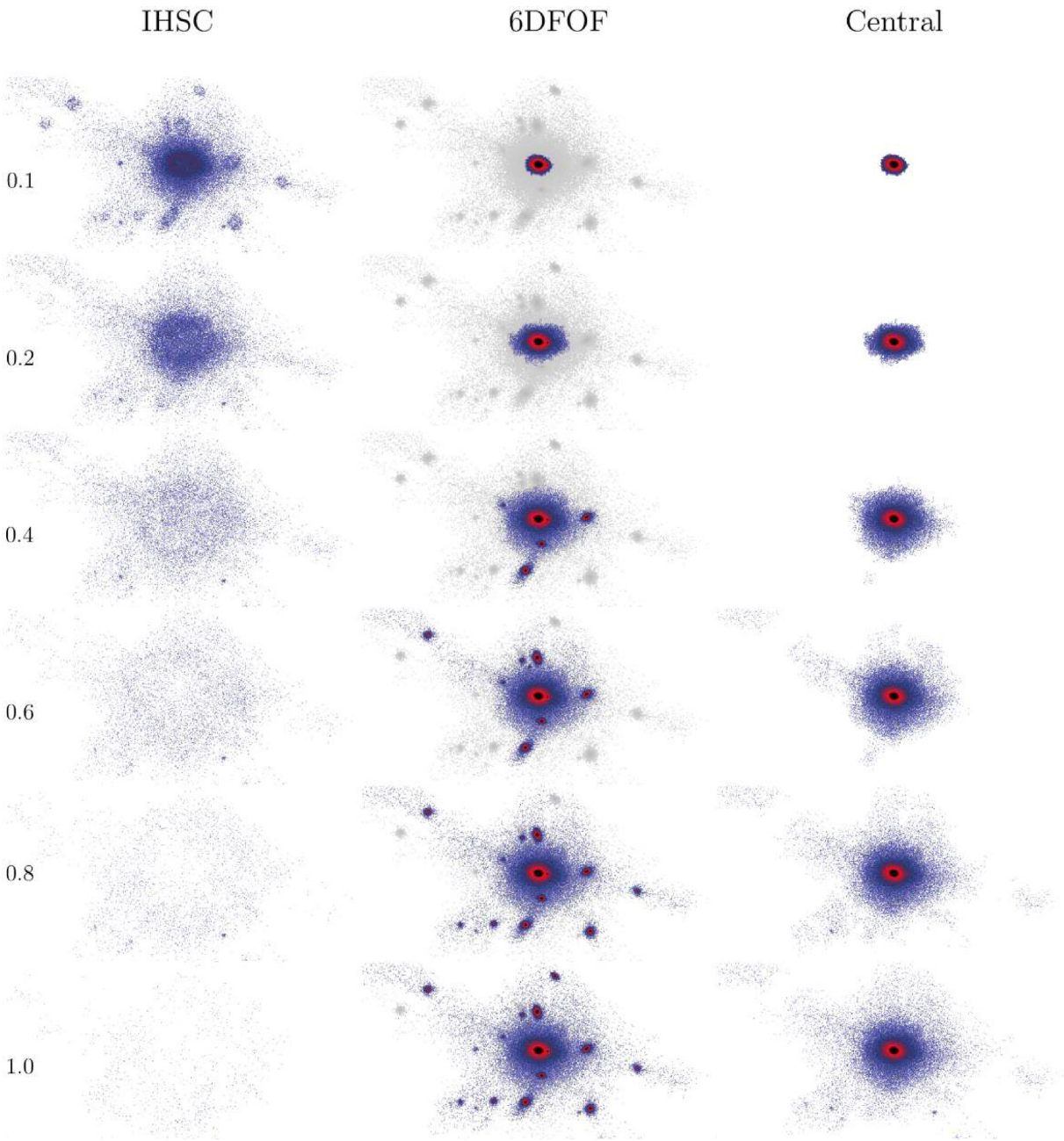

Figure A1. Projected stellar density of the IHSC, the 6DFOF object (step 2), and the final central galaxy, for different choices of $f_{x,(6D)}$ (row labels). The 3DFOF object (step 1) is shown in grey in the middle panels. For this work our standard choice is $f_{x,(6D)} = 0.2$.

APPENDIX B: CORE-GROWTH WEIGHTING

In Section 3.1.4 we stated that when calculating phase-space distances from cores, equation (19), a weight w_k was needed to compensate for the compactness of galaxies and extension of streams and shells. For this we used a mass-dependent weight, as this would reduce the phase-space distance from particles to phase-space compact high-mass objects (i.e. galaxy candidates), compared to low-density extended objects (streams and shells). We show in Fig. B1 surface density projections of how the growth of phase-space cores changes for $w = \{1, 1/\log M, 1/M, 1/\sqrt{M}\}$, where M is the total mass of the core at a given iteration level.

We show how the central galaxy is affected by w in the first row. For $w = 1$ and $w = 1/\log M$ it can be seen that some of the outer parts of the galaxy are missing. The reason behind this is the fact that some streams can have very large dispersion so that 'weak' weights do not compensate, as is seen in the last row. On the other hand 'strong' weights can in fact make the central galaxy absorb the vast majority of the particles, as they would have to be close to any structure by the same orders of magnitude difference of their masses. A weight of $1/\sqrt{M}$ does a slightly better job than $1/M$, as for the largest satellite (second row) it is able to correctly return its outskirts. For satellites with leading or trailing tails, and some streams (rows 3 to 5), it can be seen that weak weights recover a

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

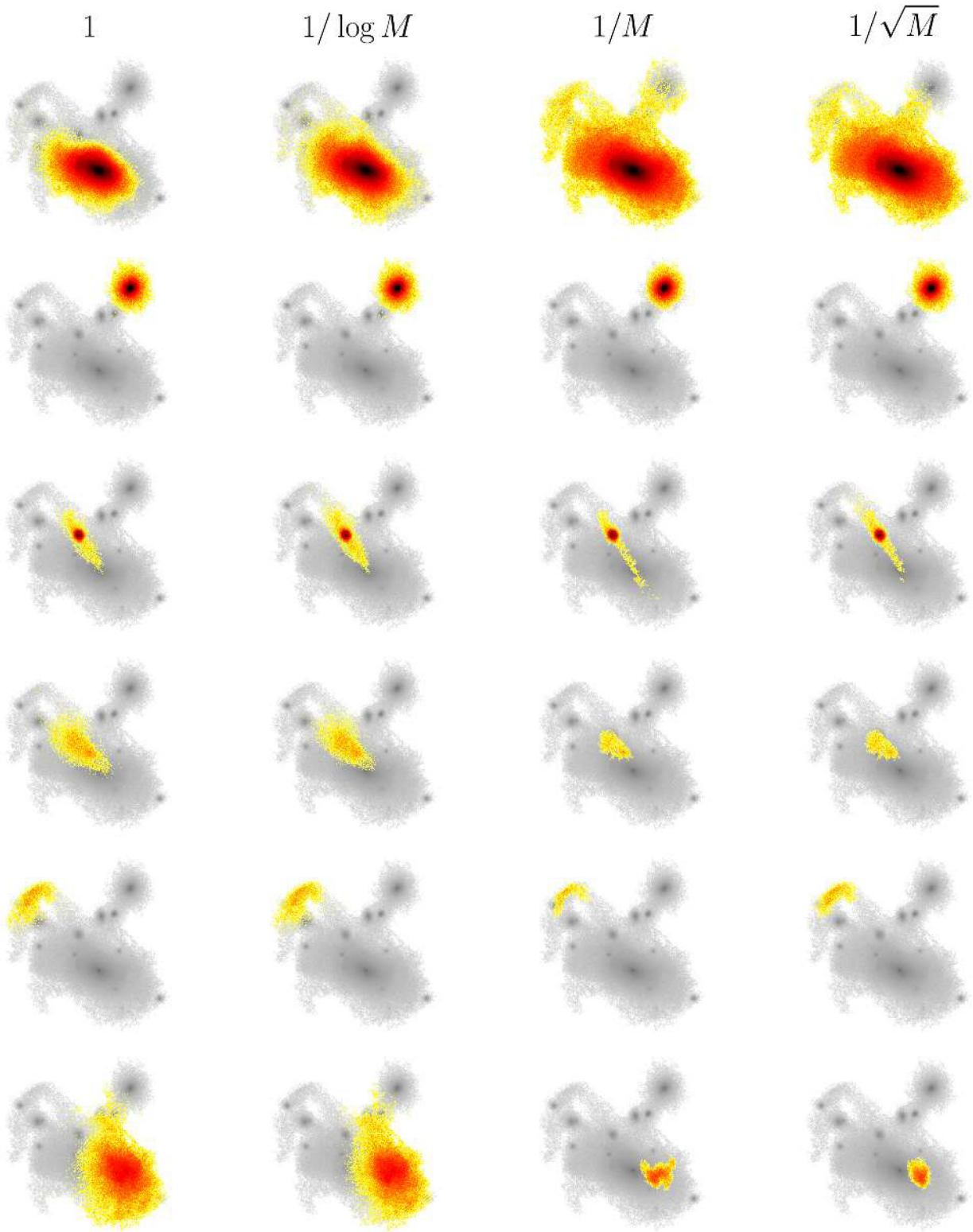

Figure B1. Projected stellar density of structures found by VELOCIRAPTOR inside a galaxy cluster in Horizon-AGN using different weights for the core growth.

little better outer material than strong ones, as for the latter streams seem to look a little fragmented. The last row shows an example of a tidal structure that grows drastically when weak weights are used, due to their large dispersions and the metric used to assign particles.

VELOCIRAPTOR can also identify tidal features with the same algorithm. Though the ability of identifying streams is important for many studies, our priority is first to identify galaxies as cleanly as possible. In order to avoid the extreme growth of streams, we

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

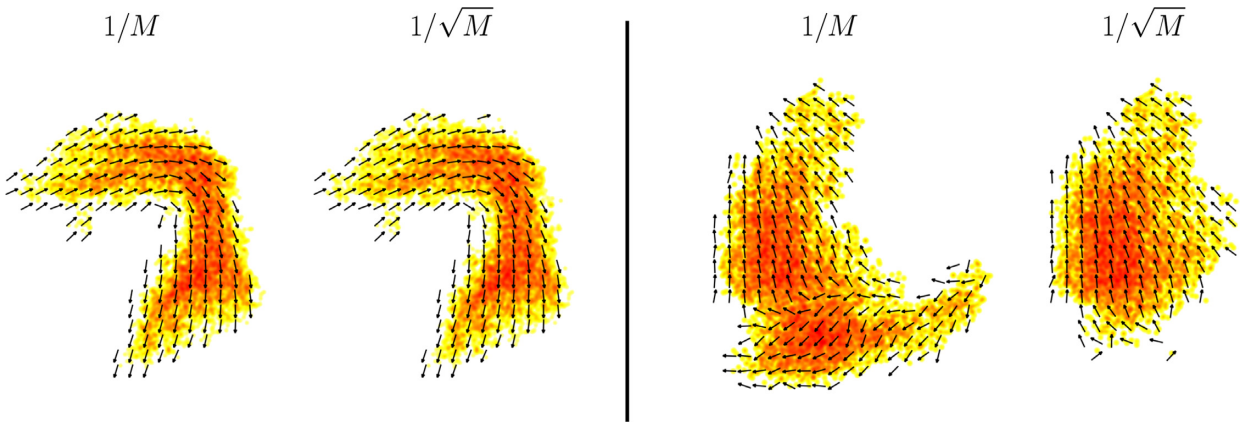

Figure B2. Projected stellar density (colour) and velocity field (arrows) for two streams identified by VELOCIRAPTOR and grown with weights proportional to $1/M$ and $1/\sqrt{M}$, as labelled. For the left-hand one it can be seen that the velocity field changes smoothly along the structure for both weights. For the right-hand one it can be seen that $1/M$ has assigned particles of a distinct structure to the stream, while for $1/\sqrt{M}$ the velocity field is much smoother.

incline for stronger weights. Both $1/M$ and $1/\sqrt{M}$ give similar results; however, the latter does a better job in preventing all particles from being assigned to the central, and recovers better outskirts for satellites. We further compare these two weights by taking a closer look at the streams from the last rows of Fig. B1. In Fig. B2 we show zoomed-in projected surface densities and velocity fields for both structures. The left-hand panel shows how for an arc stream the velocity field changes smoothly along it. The same is recovered for both weights. However, for the last stream it can be seen that for a weight of $1/M$, the velocity field is not continuous along the structure, and there are in fact distinct structures, while for a weight of $1/\sqrt{M}$ the stream (or tidal feature) has a smooth velocity field.

Weights proportional to other powers of M, or to any arbitrary quantity, can be easily applied. However, a weight of $1/\sqrt{M}$ is used as it gives the desired results for the purposes of this study.

Physically, M^{α} for $\alpha < 1$ is proportional to a galaxy's tidal radius, $r_{\rm t}$. In the case of point masses, the tidal radii can be approximated as the Roche lobe radii, that is $r_{\rm t} \propto M^{1/3}$. For a King's profile the tidal radii are approximately $\propto M^{0.4}$ (see e.g. Binney & Tremaine 2008). For more realistic profiles $r_{\rm t} \propto R\,M^{1/3}$, which for the spherical-collapse model gives $r_{\rm t} \propto M^{2/3}$. Therefore, $w \propto 1/r_{\rm t}$ gives us $w \propto 1/M^{\alpha}$, with $1/3 \leq \alpha \leq 2/3$. Our choice of $\alpha = 0.5$ is then appropriate to properly account for the size of galaxies for particle assignment.

APPENDIX C: EVOLUTION OF MASSIVE GALAXIES IN GALAXY CLUSTER

In Section 4.3 we show how the properties of the most massive galaxy in the most massive cluster in HORIZON-AGN evolve over the last Gyr. In this appendix, we show the evolution of the subsequent three most massive galaxies in the same cluster. In Fig. C1 we show the projected mass density of the cluster (left-hand column) at different times across the 40 snapshots analysed. Galaxies identified by HALOMAKER (central column) and VELOCIRAPTOR are coloured as blue, orange, green, and red for `Galaxy 1-4`, respectively. When galaxies are well separated (first row), both finders are able to identify them independently without any contamination of other galaxies' outskirts. At subsequent snapshots, as galaxies get closer, HALOMAKER starts to assign the outskirts of other galaxies to what is considered the central, as seen in the second row for `Galaxy 2` (orange) and `3` (green). This problem gets worse at later times as the two most massive galaxies experience a flyby (rows three to five): First, `Galaxy 1` (blue) gets assigned the outskirts of galaxies `2` (orange) and `3` (green). Later, they are assigned to `Galaxy 2` (orange). It can be seen in rows three to five that the central galaxy can extend to very far regions. Due to the 6DFOF implementation of VELOCIRAPTOR and its phase-space dispersion tensor based particle assignment, the time-independent identification of galaxies is much more consistent over time even during the flyby of `Galaxy 1` and `2`.

We further test the consistency of the properties of these galaxies. In Fig. C2, we show the evolution of M_* and R_{50} (top panels) and the stellar volume density profile, ρ (bottom panels), of `Galaxy 2-4` (left, centre, and right column, respectively). As seen in Fig. C1 and consistent with Figs 4 and 5, the properties of `Galaxy 2` change abruptly in time due to how particles are assigned to what is considered the central galaxy. A dip is seen when `Galaxy 1` is the central and an abrupt increment when `Galaxy 2` is. For `Galaxy 3` and `4` the evolution of M_* and R_{50} is more consistent over time, and is similar to what VELOCIRAPTOR estimates. The properties of HALOMAKER's `Galaxy 3` fluctuate significantly compared to its counterpart in VELOCIRAPTOR, due to its outskirts being assigned to the central. These are, however, quite stable compared to the evolution of those of `Galaxy 1` and `2`. Consistent with what is shown throughout the paper, strongly interacting galaxies can be affected significantly due to the finder.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

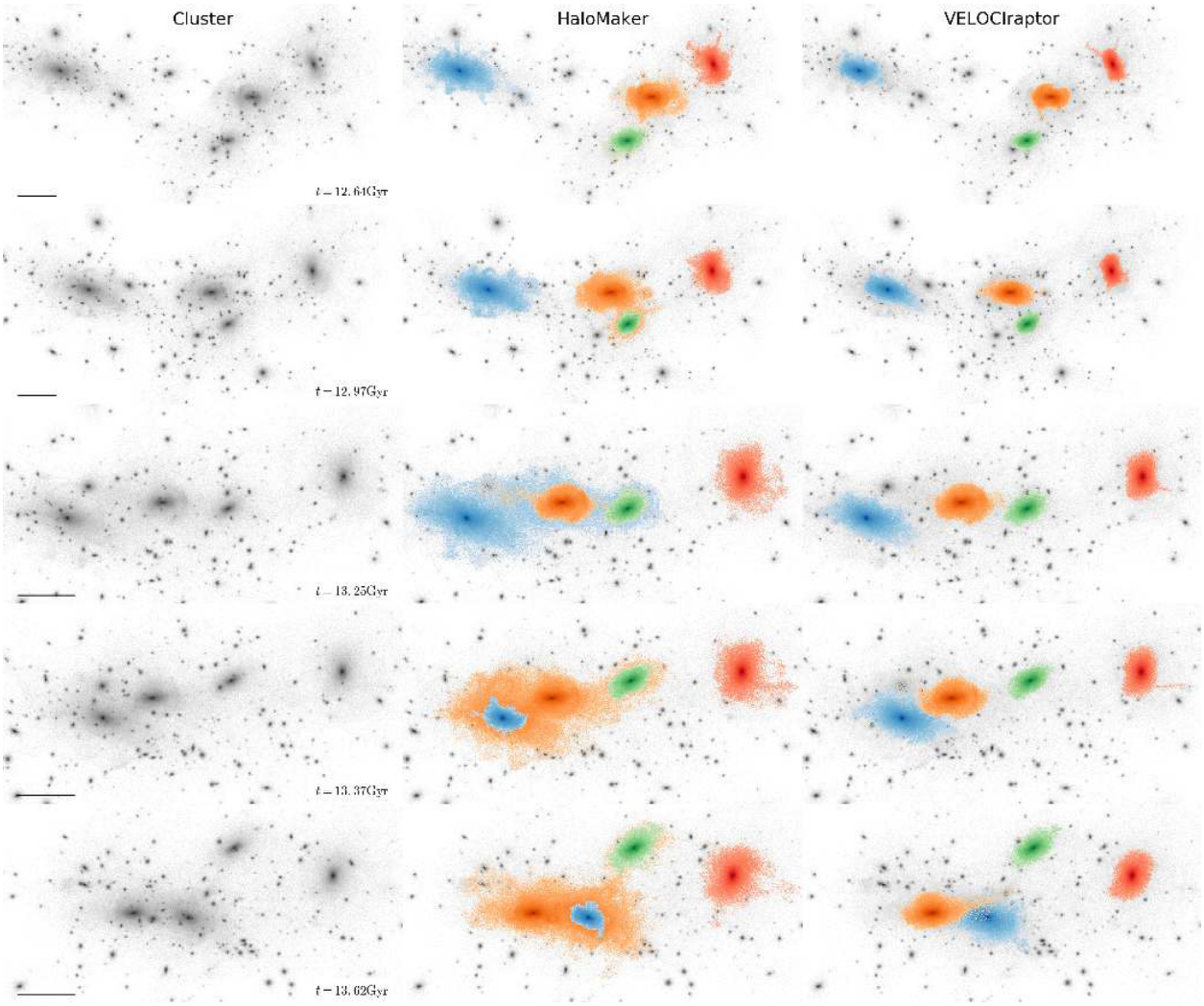

Figure C1. Evolution of the most massive galaxy cluster at $z = 0$ in the Horizon-AGN simulation (left) and how the four most massive galaxies (`Galaxy 1-4`, coloured blue, orange, green and red, respectively) are identified by HALOMAKER (centre) and VELOCIRAPTOR (right). The lower right corner of the first column displays the age of the universe at that snapshot, and a horizontal bar in the lower left corner corresponds to a length of 400 kpc.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022

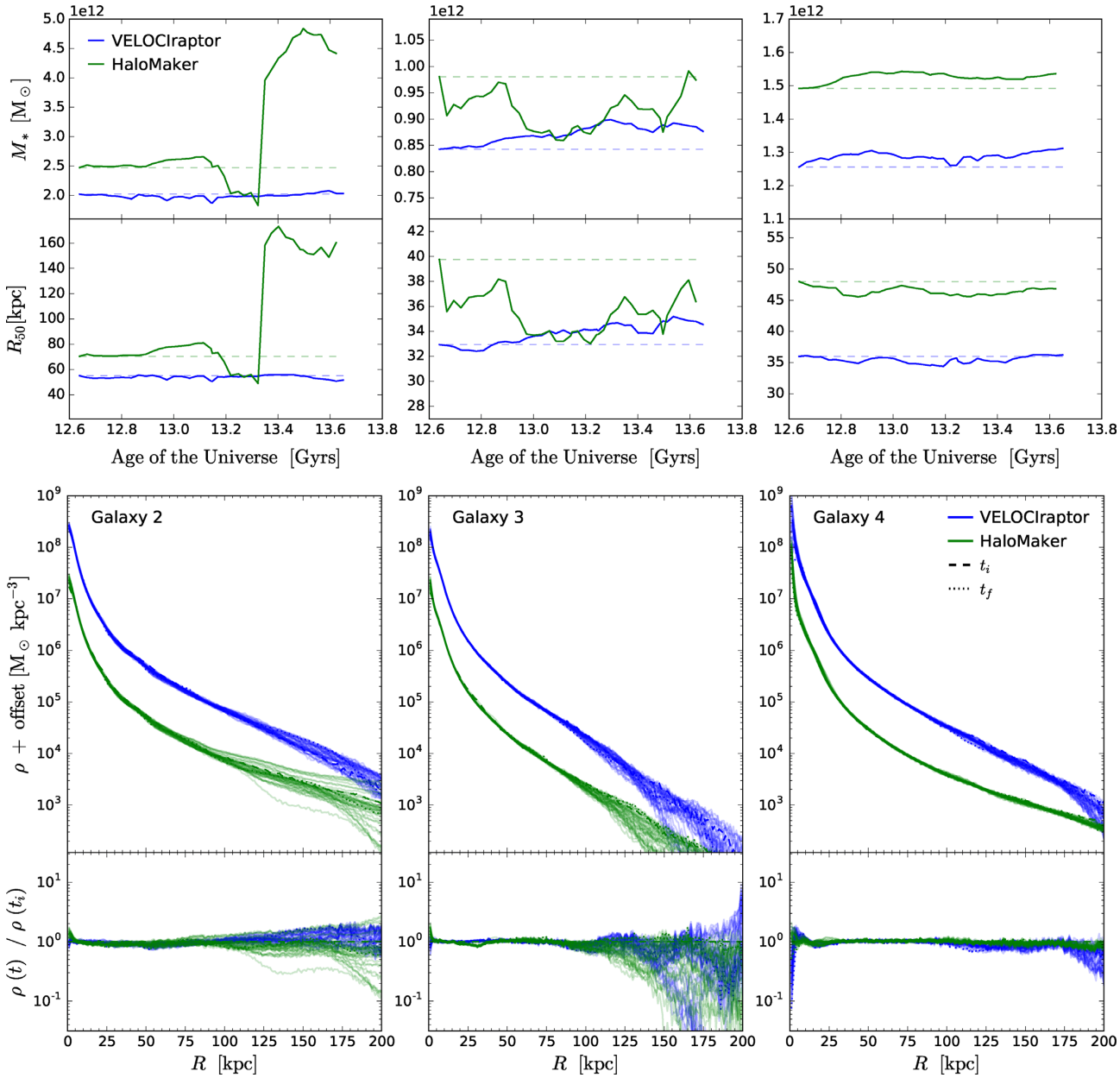

Figure C2. Evolution of M_* and R_{50} (top panels) and the stellar volume density profiles, ρ (bottom panels), of `Galaxy 2-4` (left, centre, and right columns, as labelled). Profiles of HALOMAKER's galaxies are shifted by -1 dex for clarity.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/mnras/article/482/2/2039/5123724 by CNRS - ISTO user on 06 July 2022