# Durham E-Theses

## *The tidal evolution of dark galactic substructures*

### CURRAN, AOIFE,MARIE

**How to cite:**

**Use policy**

# The tidal evolution of dark galactic substructures
## Aoife Marie Curran

### Abstract

In this thesis we study the tidal evolution of dark matter subhaloes. We first look at reproducing the work of van den Bosch and Ogiya (2018), who suggested that modern cosmological simulations still suffer from excessive disruption of subhaloes due to gravitational tides. We were able to reproduce the results qualitatively, but not exactly, and found that our subhaloes were slightly more robust and resistant to disruption.

We examine substructures in state-of-the-art cosmological simulations. We develop a technique to study substructures of a Milky Way-like halo from the Aquarius project (Springel et al., 2008) using the HEX technique (Lowing et al., 2011). HEX allows us to realistically model the potential of a halo in a computationally efficient fashion, which means that it is possible to run a large number of simulations of individual subhaloes.

We find that the softening length does not seem to have a significant effect on the survival of substructure in realistic conditions. We find that with sufficient resolution, subhaloes which were lost in the original simulation do survive until the end. However, these subhaloes are relatively rare. We thus confirm that while there is artificial disruption, this does not appear to affect the substructure population as a whole in a realistic simulation.

# The tidal evolution of dark galactic substructures

## Aoife Marie Curran

A thesis presented for the degree of
Doctor of Philosophy

Institute for Computational Cosmology

Department of Physics

Durham University

United Kingdom

September 2022

# Declaration

The work in this thesis is based on research carried out at the Institute for Computational Cosmology, Department of Physics, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification, and it is the sole work of the author unless referenced to the contrary in the text.

As my PhD was part of the Centre for Doctoral Training in Data Intensive Science, I undertook a three-month placement in industry. Chapter 5 covers the work that I did on this placement - creating a REST API for R package analysis.

Following the outbreak of the SARS-CoV-2 virus, a multidisciplinary team of researchers from Durham University worked to develop JUNE, an open-source model for the simulation of epidemics. JUNE is described in Chapter 5. My contribution to this project was in the area of data analysis and visualisation. I analysed survey data and generated visualisations of the output to help improve the accuracy of the model, as well as creating plots for the publication.

This work comprises Appendix A and has been published in

J. Aylett-Bullock, C. Cuesta-Lazaro, A. Quera-Bofarull, M. Icaza-Lizaola, A. Sedgewick, H. Truong, **A. Curran**, et al. *June: open-source individual-based epidemiology simulation.* Royal Society Open Science 8 210506 (2021)

# Acknowledgements

I would like to thank my supervisory team, without whose support and guidance this thesis would not be here. Adrian, Carlos and Azi, I am incredibly grateful for both the academic and personal support that you have provided throughout the years. You have all been amazing. Adrian, thank you for being a great teacher and for always being available whenever I needed help. Carlos, thank you for your guidance and for keeping me entertained with your stories during the dark days of the pandemic. Azi, thank you for always being there to answer my many questions.

Many thanks to everyone in the ICC. Particular thanks go to John Helly for always being generous with his time and expertise, the COSMA support team, and everyone involved in the CDT programme. Thanks also to Dorothy Jenkins for the chats during lockdown.

I am grateful to Stelios Kazantzidis for the use of his initial conditions code and to Frank van den Bosch for his assistance and the use of his codes.

Thank you to everyone at Jumping Rivers for being so welcoming and enthusiastic during my time there.

Thank you to the Finder-Outers for providing a puzzling and much-needed distraction during lockdown.

Thanks to my parents for all their love and support throughout the years.

Finally, thank you to Jonathan, who has always been there for me and has always supported me through good times and bad. Thank you for being my rubber duck and for all of the proofreading and personal RSE support. Tá mo chroí istigh ionat.

# Contents

# List of Figures

# List of Tables

# Introduction

## 1.1 The ΛCDM Model

In the standard ΛCDM cosmological paradigm the Universe is composed of three main components: the cosmological constant, $\Lambda$, which is associated with dark energy, cold dark matter (CDM) and ordinary matter or baryons. Dark energy is estimated to comprise $\sim 68\%$ of the mass-energy density of the Universe, while dark matter comprises $\sim 27\%$. Ordinary matter makes up just $\sim 5\%$. Neutrinos and photons also contribute a small percentage.

Cold dark matter is so called because it has negative thermal velocities at early times and it does not interact with the electromagnetic force. As dark matter is not visible, evidence for its existence comes entirely from its gravitational interactions. In 1933, while studying the redshifts of galaxy clusters, Fritz Zwicky noticed a large scatter in the apparent velocities of eight galaxies within the Coma cluster (Zwicky, 1933), indicating a total mass that exceeded the visible, baryonic mass. More evidence for dark matter came from the work of Rubin and Ford Jr (1970), who studied the rotation curve of the Andromeda Galaxy. While there have been numerous experiments to detect dark matter both on Earth (e.g. LUX-ZEPLIN (Akerib et al., 2020), XMASS (Abe et al., 2013)) and in space (e.g. Fermi-LAT (Albert et al., 2017)) they have so far failed to identify a dark matter particle.

The cosmological constant, $\Lambda$, is thought to drive the accelerating expansion of the Universe. Evidence for dark energy comes from measurements of Type Ia supernovae. These type of supernovae are considered "standardizable candles" and through measurement of both their redshift and their apparent brightness, it was discovered that the expansion of the Universe was accelerating (Riess et al., 1998).

The ΛCDM model is often referred to as the standard Big Bang model of cosmology, as it is the simplest model that can explain what we see in the Universe today. In this model, the universe began at the Big Bang approximately 13.7 billion years ago. This was then followed by a period of rapid expansion known as inflation. While this inflation is responsible for smoothing out initial inhomogeneities, it is also the cause of the small fluctuations in the cosmological density. These fluctuations lead to a gravitational instability that leads to the formation of structure in the Universe (Coles and Lucchin, 2003). Regions of overdensity collapse to form dark matter haloes, once a critical density is reached. These dark matter haloes continue to grow, through both the accretion of nearby material and by hierarchically merging with other dark matter haloes.

The discovery of the cosmic microwave background (CMB) by Penzias and Wilson (1965), provided one of the most important pieces of evidence for the ΛCDM model. The power spectrum of the CMB was mapped by WMAP (Bennett et al., 2013) and later by Planck (Aghanim et al., 2020) and good consistency was found with the ΛCDM model.

While the ΛCDM model has been successful at explaining the large-scale structure of the Universe, it has been more difficult to test at smaller scales. Testing the consequences of small scale power is key to falsifying CDM, as small-scale structure is very sensitive to the exact nature of dark matter. Dark matter substructure is important, not just for understanding galaxy formation, but also for modelling galaxy clustering and constraining the nature of dark matter. The smallest haloes which are too small to host galaxies are still interesting from a dark matter detection perspective. A quantitative understanding of the small-scale dark matter

distribution within our galaxy is necessary if we are to detect dark matter, either through direct methods or from annihilation products.

## 1.2 Dark matter substructure

$N$-body simulations can be used to model the non-linear growth of this dark matter structure, as both the initial conditions and the equations of evolution are known. However, early attempts to simulate this process suffered from overmerging due to low resolution, of both force and mass, which meant that large amounts of dark matter substructure was missing (Moore et al., 1998). As $N$-body simulation software and computer hardware have improved, new simulations began to show the substructure of these dark matter haloes more clearly.

### 1.2.1 The Overmerging Problem

Early simulations exhibited very little substructure, especially in high density regions. This was known as the "overmerging problem", because bound, virialised systems of multiple galaxies do exist in the real Universe. This problem was first noted by White et al. (1987) and Frenk et al. (1988), who both found that the substructure of dark matter haloes was erased after they collapsed into regions of high density in $N$-body simulations. Developments in code and computer hardware led to improvements in simulations, with increasing resolutions of both force and mass. However, these simulations still produced smooth dark matter haloes, with very little substructure.

Modern cosmological $N$-body simulations predict that 5-10% of the mass of a dark matter halo is composed of substructure (Gao et al., 2004), but it was not until the late 1990s that cosmological simulations were able to resolve dark matter substructure in Milky Way-sized haloes. Moore et al. (1996) determined that softening lengths which were too large were to blame for this overmerging problem. Large

softening lengths lead to haloes with large low-density cores, which are susceptible to disruption. Still, overmerging was recognised as a problem caused largely by insufficient numerical resolution (Moore et al., 1998; Klypin et al., 1999).

In the 2000's, two dark matter simulations were run, both with more than one billion particles, which have not yet been superseded in terms of resolution - the Aquarius Project (Springel et al., 2008) and the GHALO Simulation (Stadel et al., 2009). The GHALO Simulation (Stadel et al., 2009) performed a series of simulations of a Galactic mass halo at a number of resolutions. The highest resolution simulation contained over three billion particles and had a particle mass of $10^3$ M$_\odot$, allowing for thousands of subhaloes to be resolved. The Aquarius Project will be described in more detail below.

## 1.2.2   The Aquarius Simulations

In the highest resolution simulation substructures are found within substructures. Subhalo density profile is independent of subhalo mass. Most subhaloes are found in the outermost regions of the halo.

In the Aquarius Project, dark matter only zoom-in simulations were performed on six Milky Way-sized haloes. The six haloes were selected at random from isolated haloes of a similar mass to the Milky Way within a lower-resolution parent simulation with $900^3$ particles in a box with sides $100h^{-1}$ Mpc. In this context, 'isolated' means that each halo had no neighbours greater than half its mass within $1h^{-1}$ Mpc.

The six haloes were named 'Aq-A' through 'Aq-F'. These haloes were simulated in their full cosmological context, at a number of different resolutions in order to test for convergence. Once results are robust to an increase in resolution they are said to be converged. The resolution levels were named 1-5, where the suffix '1' indicates that that is the highest resolution, while '5' indicates that that is the lowest level of resolution. Only one halo, Aq-A, was simulated at the highest resolution. This

simulation contains over one billion particles and almost 300,000 gravitationally bound subhaloes within the virialized region of the main halo.

The Aq-A is shown in five different resolutions in Figure 1.1. These images clearly show that the halo is a complex object, which is rich in substructure, even at the lowest resolution. The small substructures which have grown the halo through merging are visible, dotted throughout. These substructures repeat closely between the simulations at different resolutions, albeit in slightly different positions, thus demonstrating that there is good convergence across the different resolutions. A number of convergence tests were performed on the haloes and good convergence across different resolutions on both the amount of substructure in the Aquarius simulations and the internal properties of said substructure was found.

Figure 1.2 shows the differential subhalo abundance by mass for the different resolutions in the Aq-A halo. There is good convergence across the different resolutions. The bottom panel compresses the vertical range, by multiplying the data by $M_{\text{sub}}^2$. This shows a slight divergence in the number of objects at masses below a few hundred particles, but for properly resolved subhaloes, there is good convergence. Springel et al. (2008) found that there was a strong dependence on radius for the abundance of substructure, with the majority of substructure found in the outer regions of the halo.

The Aquarius simulations are incredibly detailed and at the highest resolution show several levels of substructure within substructure. Figure 1.3 shows closeups of six of the largest subhaloes in Aq-A-1. It is clear that each subhalo has its own associated substructures, some of which even possess their own substructures. The bottom row of images shows four levels of substructure within subhalo F. Springel et al. (2008) found that the substructure fraction found within subhaloes was much smaller than that found in the main haloes, particularly for those in the inner regions of the halo. They determined that this was due to the effects of tidal stripping and the lack of replacement substructure infalling into older objects.
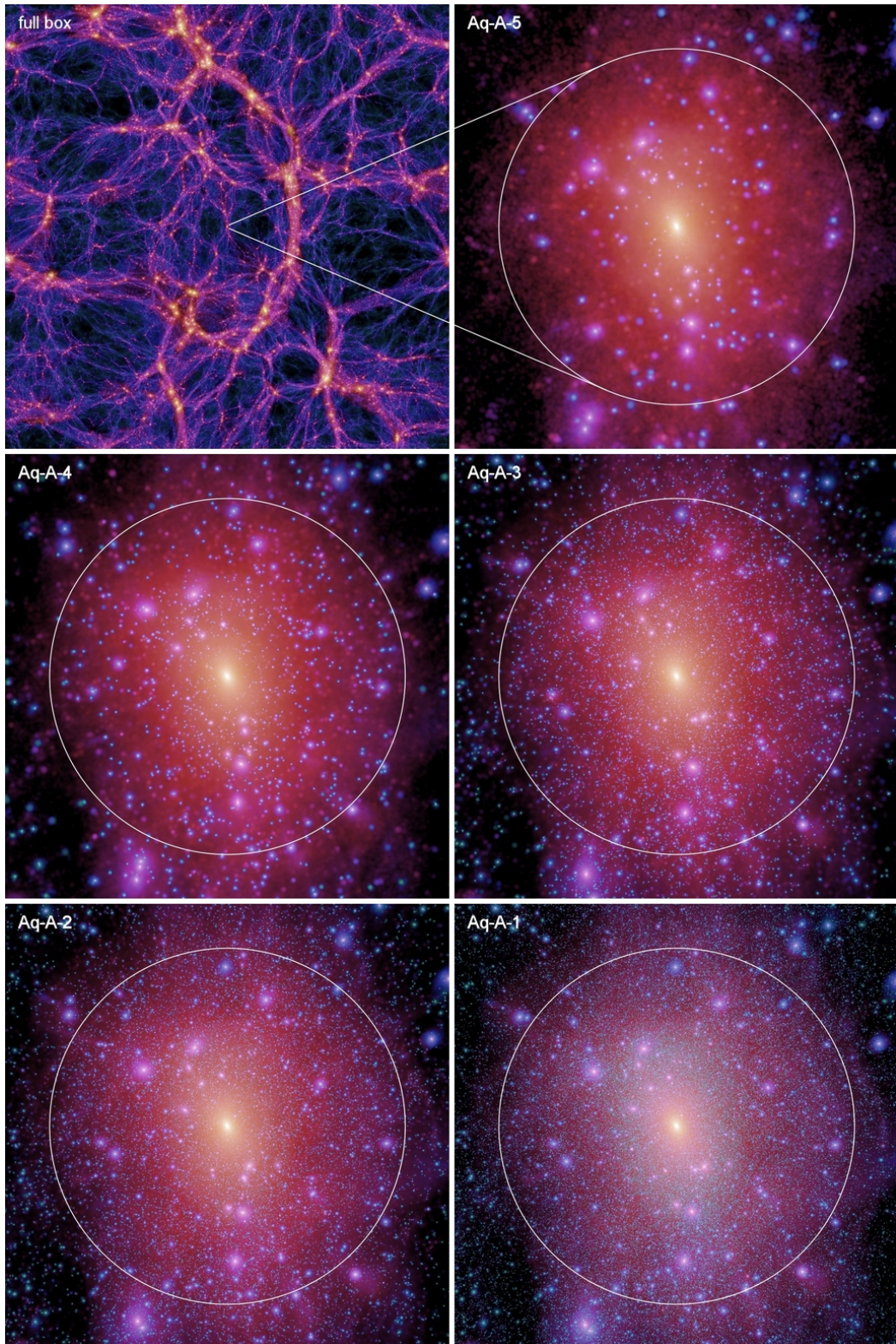
Figure 1.1:  The top left panel shows the projected dark matter density at $z = 0$ of a 13.7 Mpc slice through the parent simulation. The other panels show the Aquarius A halo at different resolutions. The brightness is proportional to the log of the squared dark matter density projected along the line of sight, and the colour hue shows the local velocity dispersion weighted by the squared density along the line of sight. The white circles show $r_{50}$. *Figure 2 from Springel et al. (2008).*
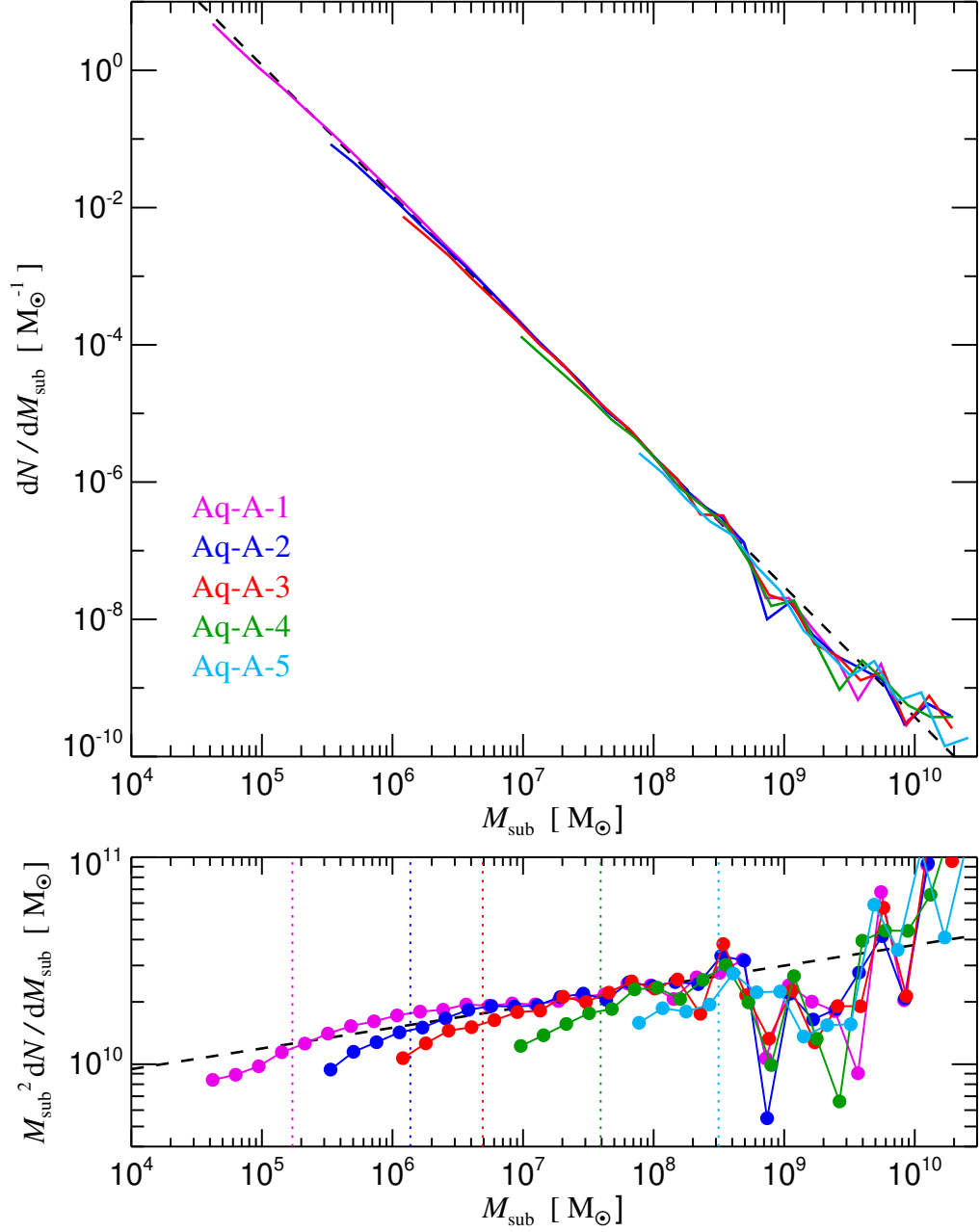
Figure 1.2: Differential subhalo abundance by mass for different resolution levels in Aq-A. The bottom panel shows the data multiplied by $M_{sub}^2$ to compress the vertical dynamic range. The dashed lines show the power law $\mathrm{d}N/\mathrm{d}M \propto M^{-1.9}$. *Figure 6 from Springel et al. (2008).*
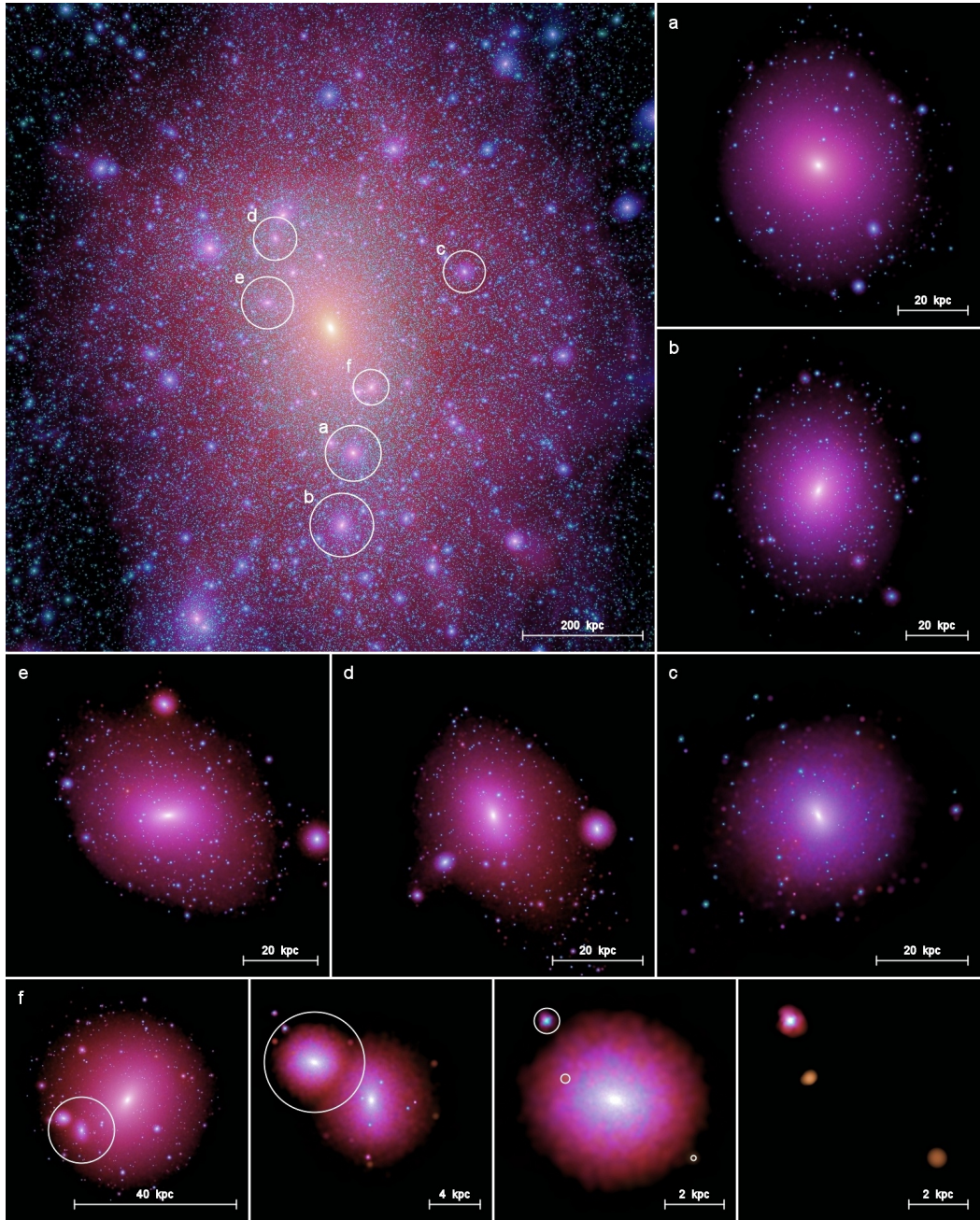
Figure 1.3:   Image showing substructure within substructure. The top left panel shows the dark matter distribution within the main halo in the Aq-A-1 simulation. The six subhaloes marked by circles are shown in detail in the surrounding panels. It is clear that these subhaloes possess their own substructures. In the case of the subhalo shown in the bottom panels, four levels of substructure are shown. *Figure 13 from Springel et al. (2008).*

## 1.3 Recent Work on Substructure

Springel et al. (2008) found an abundance of substructure and significantly more substructure than was found in the Via Lactea I and II simulations (Diemand et al., 2007, 2008; Kuhlen et al., 2008; Madau et al., 2008). Detailed convergence studies were also performed on these Aquarius subhaloes at different resolutions and good convergence was found. A number of other resolution studies, such as Onions et al. (2012), Knebe et al. (2013) and van den Bosch and Jiang (2016), have also found that subhalo mass functions are converged down to 50-100 particles per subhalo.

However, recent work suggests that current cosmological simulations, such as Aquarius, may still suffer from overmerging and that the majority of disruption is in fact artificial, rather than due to physical processes (Penarrubia et al., 2010; van den Bosch et al., 2018; van den Bosch and Ogiya, 2018). For example, van den Bosch (2017) estimated that 80% of disruption in the Bolshoi simulation (Klypin et al., 2011) was probably due to numerical effects. Analytical work by van den Bosch et al. (2018) demonstrated that neither tidal stripping nor tidal heating should lead to complete disruption of cold dark matter substructure. The authors concluded that it was almost impossible to completely physically disrupt a CDM subhalo and that the inner remnant of an NFW (Navarro et al., 1996) subhalo should survive, even after 99.9% of its mass is lost due to tidal stripping. This work was followed up by a detailed convergence study (van den Bosch and Ogiya, 2018) which found that inadequate force-softening was driving disruption of substructure in $N$-body simulations and that subhaloes with NFW profiles almost never become fully disrupted.

Following this work, van den Bosch and Ogiya (2018) used a large suite of idealised numerical simulations to examine the tidal evolution and disruption of subhaloes. They attempted to determine whether disruption seen in subhaloes in simulations is physical or merely due to numerical effects. The authors claim that most state-of-the-art cosmological simulations, such as Aquarius, do not have the necessary

force softening or a sufficient number of particles to properly resolve the dynamical evolution of subhaloes. They concluded that most disruption of substructure in $N$-body simulations is numerical in origin, rather than physical, and that cosmological simulations suffer from significant overmerging, mainly due to inadequate softening and discreteness noise. They argued that this overmerging is overlooked in standard convergence tests.

Building on this work, (Ogiya et al., 2019) developed the DASH library. This is a database of over 2,000 idealised, high-resolution ($N_p = N = 1,048,576$) $N$-body simulations following the tidal evolution of individual subhaloes orbiting around larger haloes. Two orbital parameters and the concentration of both the subhalo and the main halo are modified in the library, covering much of the parameter space. The simulations are of a high enough resolution such that they are largely unaffected by discreteness noise and insufficient force softening. This library is intended for use with semi-analytical models, in order to improve their calibration. The host is modelled as a fixed, analytical, spherical potential.

Green and van den Bosch (2019) used the DASH library to look at substructure and found that modern cosmological simulations may have been systematically under-predicting the amount of substructure by as much as a factor of two. However, further work in Green et al. (2021) determined that this was an overestimate and instead found that artificial disruption only results in an $\sim 8 - 12\%$ suppression of the fraction of substructure. They also found that artificial disruption was more pronounced in objects which orbit close to the centre of the main halo.

Errani and Navarro (2021) looked at the evolution of cuspy dark matter haloes in a static, spherical potential. Their subhaloes are modelled as NFW $N$-body objects, most of which have $10^7$ particles and have masses of $10^6$ M$_\odot$. They looked at a range of pericentre radii and eccentricities and also found that the orbits with smaller pericentres suffer from the greatest mass loss. They found that although circular orbits lead to a faster mass loss, the remnant is indistinguishable from that on an eccentric orbit after a sufficient mass loss. They also found that mass loss is

continuous, but it slows down as the simulation progresses and that tidal evolution on highly eccentric orbits is delayed, but otherwise similar to less eccentric orbits.

Earlier this year, another paper looking at the evolution of subhaloes was published by Aguirre-Santaella et al. (2023). This paper used an improved version of DASH to look at the evolution of subhaloes on orbits with pericentres in the solar vicinity. They examined dark matter subhaloes orbiting in a Milky Way-type potential. The subhaloes are represented by $N$-body systems, while the potential is a time-evolving analytical, spherical potential. For most of the simulations they examined subhaloes with masses of $10^6$ M$_\odot$, but they did test down to 1 M$_\odot$ in mass.

They found that pericentre distance has the greatest effect on mass loss, compared to all other orbital parameters. Subhaloes whose pericentric passages come very close to the centre of the main halo (less than 15 kpc) do survive the simulation, but lose up to 90% of their mass. Subhaloes which have a lower concentration and closer pericentre lose the most mass. They confirm that for dark matter subhaloes lighter than $10^8$ M$_\odot$, the results are virtually independent of mass. They also find that the case is the same when baryons are included. The time of accretion was set to $z = 2$ for most cases and the concentration of subhaloes from 5 to 50 was tested.

## 1.4   Tidal Disruption of Subhaloes

Subhaloes can become disrupted due to tidal forces. Tidal stripping occurs when a smaller halo orbits a significantly larger halo. As the subhalo orbits the halo, all particles which lie beyond the tidal radius, $r_t$, and so are less gravitationally bound, are stripped from the subhalo due to tidal forces (van den Bosch and Ogiya, 2018). As the particles are stripped, the subhalo tries to virialize again. This causes the subhalo to expand and the tidal radius to increase. Once again, the particles which lie beyond $r_t$ are stripped and the subhalo tries to reach virial equilibrium. This is a continuous process.

Tidal shocking occurs when there is a sudden change in the external gravitational field. van den Bosch and Ogiya (2018) look at mainly circular orbits in order to avoid having to account for this. However, Errani and Navarro (2021) found that subhaloes on elliptical orbits take longer to get stripped compared to those on circular orbits. The difference in time is related to the ellipticity of their orbits. For example, it takes approximately five times as many orbits for a subhalo on a 1:5 orbit to reach the same final stage as an object on a circular orbit. This is due to the fact that although the subhalo is tidally shocked when it reaches pericentre, it spends the vast majority of its time away from the centre of the main halo, where the tidal forces are strongest.

Dynamical friction can also contribute to subhalo disruption, as it can drag subhaloes closer to the centre of the main halo, thus hastening its tidal stripping. However, as the subhaloes which we will be examining are several orders of magnitude smaller than their hosts, it is not necessary to account for dynamical friction in our simulations. The deceleration caused by dynamical friction is proportional to the mass of the subhalo.

## 1.5 Thesis Structure

In this thesis, we will examine the tidal evolution of dark matter subhaloes on galactic scales. As described above, on this scale there are discrepancies between the results of state-of-the-art cosmological simulations and work which looks at subhaloes on their own in idealistic conditions.

In Chapter 2 we explain some of the techniques used in this thesis.

In Chapter 3, we take a close look at van den Bosch and Ogiya (2018). We attempt to reproduce some of the key plots from this work, first using GADGET-4 and SUBFIND, and then using code provided by Frank van den Bosch. We perform a convergence study, looking at several parameters which can affect the accuracy of

numerical simulations, such as force softening, force accuracy and the number of particles which are used to simulate the subhalo.

In Chapter 4, we perform a similar study, but instead of the idealised conditions used by van den Bosch and Ogiya (2018), we now use more realistic conditions by extracting subhaloes from the Aquarius simulation and resimulating their orbits using the HEX code developed by Ben Lowing (Lowing et al., 2011).

In Chapter 5, we discuss work done as part of the CDT placement - the Diffify project, a tool which provides a comparison between different versions of any R package stored on CRAN and the JUNE project, an open-source model for the simulation of epidemics.

# Methods

## 2.1   Introduction

In this chapter we will discuss the methods and techniques which are used throughout this thesis. We will discuss $N$-body codes and some of their parameters in Section 2.2. In Section 2.3 we talk about how to identify substructure in $N$-body simulations using SUBFIND. We look at the procedure which we followed in order to fit subhaloes with NFW profiles in Section 2.4 and the generation of initial conditions in Section 2.5. Finally, we discuss the halo expansion (HEX) method in Section 2.6.

## 2.2   $N$-body Codes

### 2.2.1   Tree Codes

The Barnes-Hut Approximation scheme was devised by Josh Barnes and Piet Hut in order to make large-scale, highly-clustered $N$-body simulations computationally tractable (Barnes and Hut, 1986). It is a tree code, where the main idea is that long-range forces are approximated by replacing groups of distant points with the centre of mass of the cluster. This introduces a certain amount of error, but also

speeds up the calculation massively, from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, where $N$ is the total number of particles in the simulation.

There are three main stages in the Barnes-Hut algorithm. First, the computational space is recursively subdivided into smaller and smaller regions, until each region contains at most one particle. The data structure that is used to do this in three dimensions is an octree. The entire computational domain represents the root of the octree. This is broken down into eight cubes, each with a volume that is one eighth of the total domain. These eight cubes are the eight children of the root. Each child can be subsequently subdivided into eight smaller cubes and so on.

The second step in the algorithm is to calculate the centre of mass positions of the particles within the cubic cells. Each node of the tree must be made aware of the total mass and the centre of mass of the particles in its cell.

The final step is the force calculation. This can be estimated using a simple recursive function. We start at the root of the tree, which contains the entire computational domain. $l$ represents the side length of the current cell and $r$ is the distance from the cell's centre of mass to a particle, $p$. If $l/r < \theta$, where $\theta$ is an accuracy parameter approximately equal to 1, we treat the cell as a source of long-range forces and use its centre of mass. Otherwise, we recursively visit the child cells in the tree.

Throughout this thesis, we make use of two different tree codes - GADGET-4[1] (Springel et al., 2021) and a modified version of `treecode`, written by Barnes and with improvements by Dubinski (Dubinski, 1996).

## 2.2.2   Timesteps

To accurately integrate the equations of motion, the choice of timestep size is crucial. In general, a smaller timestep will give a more accurate result, but this must

---

[1]GADGET-4 also has an optional particle-mesh scheme for use with long-range gravitational forces.

be balanced against the increased computational costs. In cosmological simulations, not only is there a huge variation in spatial scales, there is also a large variation in time scales. The timescales at the centre of a halo are far smaller than those in the low density regions. To account for this disparity and to improve efficiency, modern $N$-body codes use adaptive timesteps, where the size of the timestep can vary for individual particles and at different times (Power et al., 2003).

By default, GADGET-4 uses a variable timestep in order to improve efficiency, which we have used throughout this thesis unless otherwise stated. It is also possible to force equal timesteps by setting the `FORCE_EQUAL_TIMESTEPS` configuration option at compilation time. This adopts a global timestep for all particles, by forcing all particles to take the minimum timestep desired by any of the particles. However, this timestep may still vary in size from one timestep to another. By setting the minimum timestep and the maximum timestep to be equal in the parameter file and by enabling the `NO_STOP_BELOW_MINTIMESTEP` configuration option, we can force all particles to take the same size timestep for all timesteps.

### 2.2.3 Softening

Softening is used in $N$-body simulations to prevent the forces and deflection angles from getting too large when two particles get very close to each other. These strong deflections are unphysical and a source of error, as the particles used in $N$-body simulations are more massive than cold dark matter candidates. Softening is done by modifying the potential due to each particle with the addition of $\epsilon$, the softening parameter. The simplest example of this is Plummer softening, where each body is replaced by a Plummer sphere of scale radius $\epsilon$:

$$\Phi = -\frac{1}{\sqrt{r^2 + \epsilon^2}} \tag{2.1}$$

GADGET-4 uses a cubic spline kernel of the form
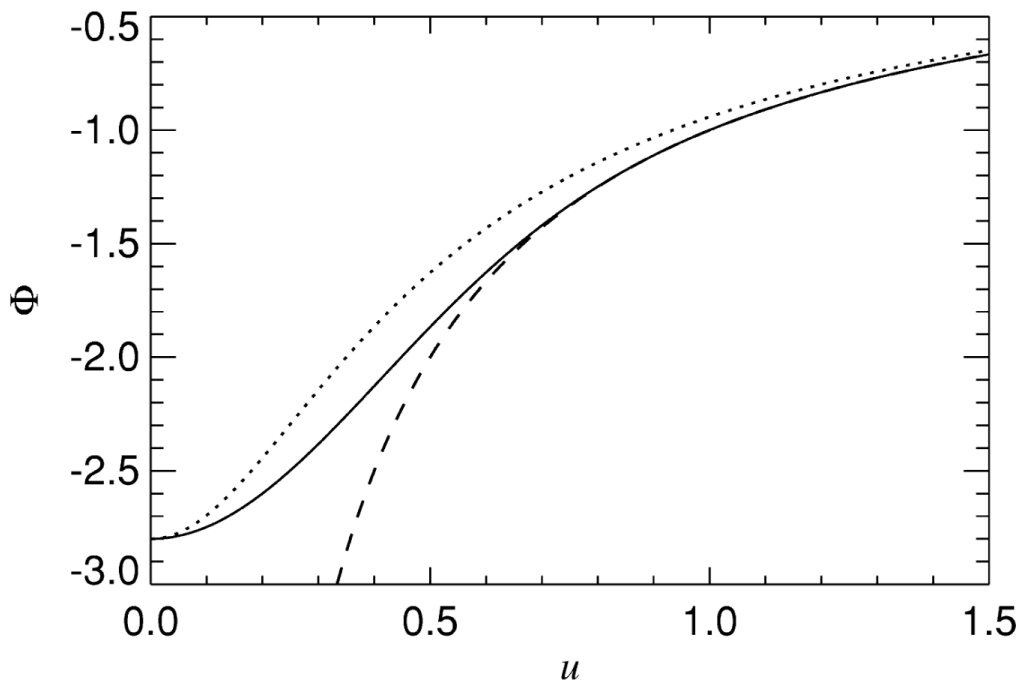
Figure 2.1: Comparison of spline-softened (solid line) and Plummer-softened (dotted line) potential of a point mass with the Newtonian potential (dashed line). *Figure 13 from Springel et al. (2001b).*

$$
W_2(u) = \begin{cases} \frac{16}{3}u^2 - \frac{48}{5}u^4 + \frac{32}{5}u^5 - \frac{14}{5}, & 0 \leq u \leq \frac{1}{2}, \\[2mm] \frac{1}{15}u + \frac{32}{3}u^2 - 16u^3 + \frac{48}{5}u^4 - \frac{32}{15}u^5 - \frac{16}{5}, & \frac{1}{2} \leq u < 1, \\[2mm] -\frac{1}{u}, & u \geq 1. \end{cases} \tag{2.2}
$$

Figure 2.1 shows a comparison of the spline-softened and the Plummer-softened potentials of a point mass with the Newtonian potential. For a given softening length in GADGET, $h$, the Plummer equivalent softening is defined as $\epsilon = h/2.8$.

Softening is also linked to the size of the timestep, $\Delta t$, in GADGET according to the equation

$$
\Delta t = \min\left[\Delta t_{max}, \left(\frac{2\eta\epsilon}{|\mathbf{a}|}\right)^{\frac{1}{2}}\right] \tag{2.3}
$$

where $|\mathbf{a}|$ is the particle acceleration, $\eta$ is an accuracy parameter and $\Delta t_{max}$ is the maximum allowed timestep (Springel, 2005).

### 2.2.4 Force Accuracy

GADGET offers two options of opening criterion when setting the force accuracy - the standard Barnes-Hut option with an opening angle of $\theta$ and a relative criterion, where a Barnes-Hut opening criterion is used for the initial timestep and a dynamical updating criterion is used subsequently (Springel et al., 2021). In this case, an approximation of the expected force error is compared to the magnitude of the total force,

$$\frac{M}{r^2}\left(\frac{l}{r}\right)^p < \alpha|\mathbf{a}| \tag{2.4}$$

where $|\mathbf{a}|$ is the magnitude of the acceleration of the particles in the previous timestep. $M$ is the mass of particles within a cell with side length, $l$, and $r$ is the distance of the target coordinate to the node's centre of mass. $\alpha$ controls the force accuracy. Nodes which are most important for the final force are assessed with greater force accuracy than less important contributions, meaning that the process is more economical than the traditional Barnes-Hut opening criterion.

## 2.3 Identifying Substructure

The task of identifying substructures in the results from $N$-body simulations is a non-trivial one. The Friends-Of-Friends (FOF) algorithm (Press and Davis, 1982) is commonly used to identify groups of particles which compose haloes in a cosmological simulation. It does so through a dimensionless parameter known as the linking length, $b$, which is a fraction of the mean inter-particle separation. A linking length of $b = 0.2$ is typically used in cosmological simulations. Any two particles which are within a distance of $b$ from each other are *friends*. A particle is also linked to all particles which are linked to its friends, i.e. its *friends-of-friends*. Groups with fewer than 20 particles are discarded.

FOF groups are composed of particles which are physically close to each other, but this does not necessarily mean that they are gravitationally bound. To find gravitationally bound substructures a halo-finder algorithm must be used. There are various algorithms used to identify bound groups of particles within these FOF groups. In this thesis, we use SUBFIND Springel et al. (2001a). SUBFIND begins by looking for bound structures in FOF groups. It estimates the local density of all the dark matter particles in the group and searches for overdense regions. It groups together nearby particles until it finds a saddle point and for each group it iteratively removes unbound particles. If more than 20 particles remain after this stage, the group becomes a subhalo. Each particle can belong to at most one subhalo. However, subhaloes can be found within other subhaloes.

Merger trees can be used to track subhaloes through a simulation (Lacey and Cole, 1993). The merger trees which we are concerned with are constructed using Dhaloes. They are intended as input to the GALFORM semi-analytic model of galaxy formation (Cole et al., 2000). The algorithm used to produce these is described in Jiang et al. (2014). In order to produce them, one starts with FOF haloes which have been decomposed into subhaloes using SUBFIND. The subhaloes are then tracked between snapshots and grouped into Dhaloes. One advantage of Dhaloes over FOF haloes (particularly when it comes to semi-analytic modelling) is that Dhaloes maintain consistent membership over time. Once a subhalo becomes part of a Dhalo, it stays a member for the duration of the simulation. This is true even of satellite subhaloes which have passed through the main halo and are so distant that they are no longer linked to the corresponding FOF group. Some FOF haloes are also split into multiple Dhaloes when substructures are well separated and they are only linked into a single FOF halo by bridges of low-density particles. The merger tree can then be traversed and used to track the orbits and masses of subhaloes within the simulation.

## 2.4 NFW Fitting

Navarro et al. (1996, 1997), hereafter NFW, found that the density profile of a dark matter halo may be approximated by a simple formula with just two free parameters:

$$\frac{\rho(r)}{\rho_{crit}} = \frac{\delta_c}{\left(\frac{r}{r_s}\right)\left(1 + \frac{r}{r_s}\right)^2},$$ (2.5)

where $\rho_{crit} = 3H_0^2/8\pi G$ is the critical density and $\delta_c$ and $r_s$ are a characteristic density and scale radius. This formula holds true for all haloes which are formed hierarchically and are close to virial equilibrium, regardless of their mass or the cosmological model used and has been confirmed by many others (see, for example, Cole and Lacey (1996); Kravtsov et al. (1997); Aylett-Bullock et al. (2021)).

In order to generate initial conditions which match haloes from the Aquarius simulations, we must first fit the haloes that we wish to replicate with NFW profiles. To do this, we use the method outlined by Neto et al. (2007). We calculate a differential density profile for the candidate halo by binning the halo mass in 32 equally spaced bins between the virial radius, $r_{vir}$, and $\log_{10}(r/r_{vir}) = -2.5$. For haloes with a mass $< 10^8$ $M_\odot$, we reduce the number of mass bins to 22.

The two free parameters, $\delta_c$ and $r_s$, are adjusted to minimise the root mean square deviation, $\sigma_{fit}$, between the binned $\log(\rho(r))$ and the NFW profile,

$$\sigma_{fit}^2 = \frac{1}{N_{bins} - 1} \sum_{i=1}^{N_{bins}} \left[\log_{10} \rho_i - \log_{10} \rho_{\text{NFW}}(\delta_c; r_s)\right]^2.$$ (2.6)

The initial conditions code which we are using to generate equilibrium haloes requires values for the virial mass, $M_{vir}$, which is the mass contained within the virial radius, and the halo concentration, $c_{vir} = r_{vir}/r_s$. So once we have the values for $\delta_c$ and $r_s$, we can calculate these. The concentration and the characteristic density are related by,

$$\delta_c = \frac{200}{3} \frac{c^3}{\ln(1+c) - c/(1+c)}. \tag{2.7}$$

For defining virial quantities in this thesis, we set $\Delta = 200$, apart from in Chapter 3 where $\Delta = 97$.

We also tested the fitting procedure outlined in Ludlow et al. (2014). This is similar to Neto et al. (2007), but the NFW profile is fit between 0.05 $r_{200}$ and 0.6 $r_{200}$. We found that in our case there was very little difference between the two. Ludlow et al. (2014) also restrict their analysis to haloes that contain at least 5000 particles within their virial radius. However, as a number of the haloes in which we are interested have fewer than 5,000 particles, we extended the fitting procedure to include all haloes with at least 1,000 particles, i.e. all haloes with a mass greater than $1.37 \times 10^7$ M$_\odot$ for the Aq-A-2 halo used in Chapter 4.

We tested our fitting procedure on a number of subhaloes from the Aquarius simulation, by looking at how the concentration calculated from fitting changes as the number of particles in the halo decreases. A sample of the results are shown in Figure 2.2. On the left, with a density profile in blue, is a subhalo which had 416,625 particles at $z = 1.11$. On the right, with a density profile in red, is a subhalo which had 145,412 particles at $z = 1.11$. We fit an NFW profile to their density profiles and then calculate the concentration. We then randomly reduce the number of particles in the halo by 10%, eventually stopping at 1% of the original number of particles. For each reduction in particles, we re-calculate the density profile, fit an NFW and re-calculate the concentration.

We can see that the concentration, shown beside each density profile, stays relatively constant for both examples, even when we get to just 1% of the number of particles. We therefore took 1,000 particles as our limit.
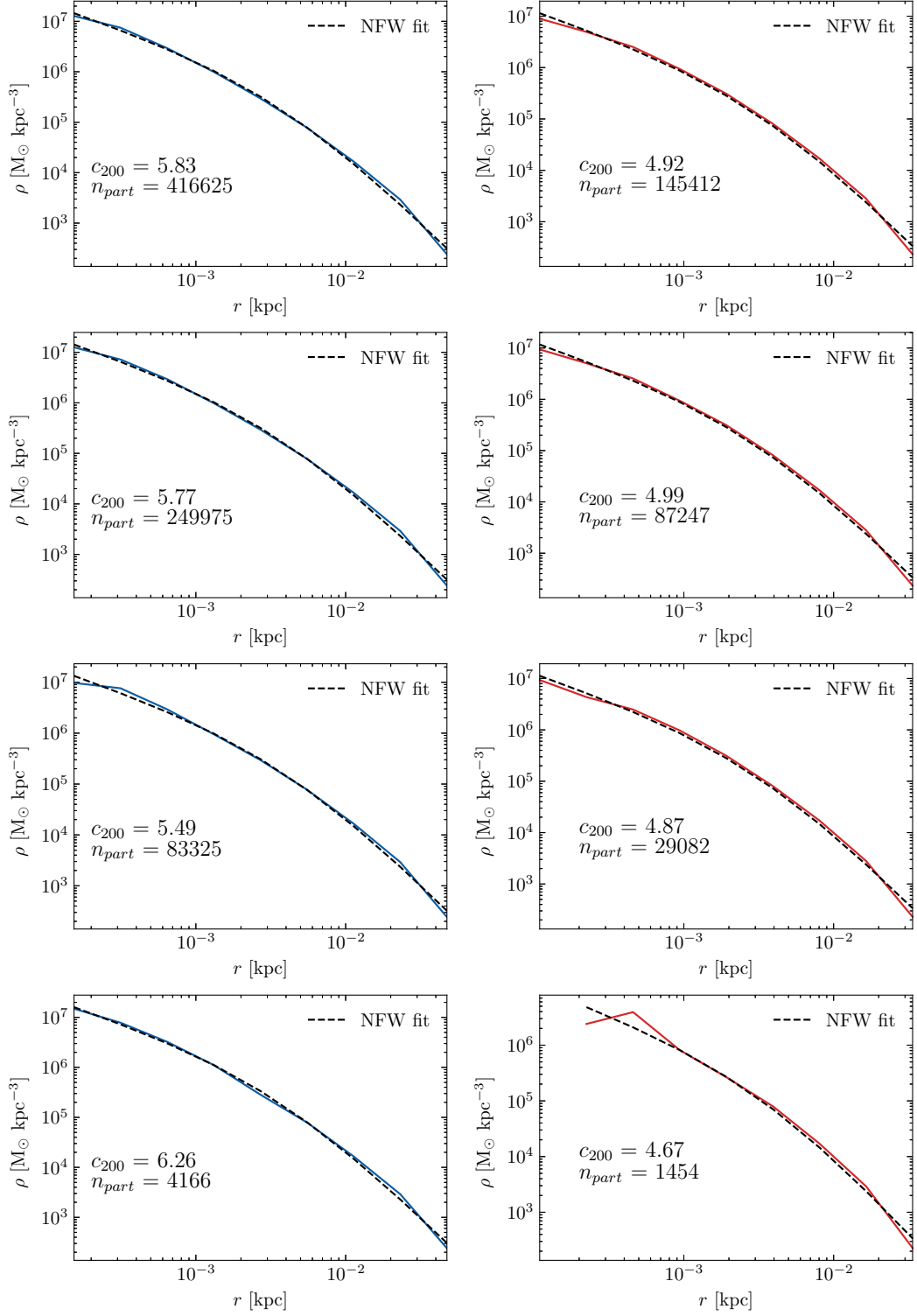
Figure 2.2: The effect that randomly reducing the number of particles in a subhalo has on its calculated concentration for two example subhaloes. The differential density is shown in blue (red) for the subhalo on the left (right). The fitted NFW profile is shown with a dashed line.

## 2.5 Initial Conditions for Modelling Infalling Haloes

To generate equilibrium $N$-body realisations, we use code introduced in Kazantzidis et al. (2004). In this method, the phase-space distribution function (DF) is sampled in order to generate the subhalo. It is assumed that the DF is only a function of the energy so that the velocity distribution is isotropic (Binney and Tremaine, 2008).

In an NFW profile the cumulative mass diverges as $r \to \infty$, so the profile must be truncated at some point. Kazantzidis et al. (2004) propose using an exponential cut-off for $r > r_{vir}$, as suggested by Springel and White (1999). This cut-off takes the form of

$$\rho(r) = \frac{\rho_s}{c^\gamma (1+c^\alpha)^{(\beta-\gamma)/\alpha}} \left(\frac{r}{r_{vir}}\right)^\epsilon \exp\left(-\frac{r-r_{vir}}{r_{decay}}\right) \quad (r > r_{vir}), \qquad (2.8)$$

where $(\alpha, \beta, \gamma) = (1, 3, 1)$ for an NFW profile. To ensure a smooth transition between both sections of the profile, $\epsilon$ is set to

$$\epsilon = \frac{-\gamma - \beta c^\alpha}{1 + c^\alpha} + \frac{r_{vir}}{r_{decay}}. \qquad (2.9)$$

For all simulations, we have set this $r_{decay}$ factor to ten times the virial radius, unless otherwise specified.

## 2.6 HEX

The Halo Expansion (HEX) code was written by Ben Lowing (Lowing et al., 2011). The HEX code approximates the potential of a simulated halo using a basis function expansion. It makes it possible to integrate the orbits of subhaloes in a time-varying halo potential at a much lower computational cost than the original simulation. This means it is possible to re-create very expensive halo simulations and replay them with different parameters. It is ideal for examining orbits and the tidal stripping of subhaloes.

The basis function expansion used in HEX is based on the self-consistent field (SCF) method first devised by Ostriker and Mark (1968), where it was used to find the equilibrium structure of rapidly rotating stars. The SCF method then used to model the potential of simple galaxies by Clutton-Brock (1972, 1973). The method was further developed by Hernquist and Ostriker (1992) and HEX is based on their formulation, where Poisson's equation is solved by expanding the density and potential in a set of basis functions. Differentiating the potential series gives the acceleration and this can then be used to self-consistently evolve the particles.

In order to avoid the high cost (both in time and money) of a full cosmological $N$-body simulation, approximations are often used. One of the most common is to use a static analytical potential to represent the main halo and to perform a simulation of the small-scale component. This makes it easy to vary the parameters of the dark matter halo - something that is not possible in a full $N$-body simulation - but it is also not very realistic. HEX provides an alternative.

We use the HEX code to accurately represent the halo potential from the Aquarius simulation (Springel et al., 2008). In order to investigate the small-scale structure and substructure of haloes, we need a high mass resolution. The Aq-A-2 simulation has a mass resolution of $\sim 10^4$ M$_\odot$. This is too low to analyse the evolution of the smallest subhaloes. We use a potential expansion including terms up to order $n_{max} = l_{max} = 30$, with a fixed scalelength of 33 kpc and sum over all particles within 340 kpc of the halo centre, to approximate the Aq-A-2 halo. A set of coefficients is generated for each snapshot, approximately every 155 Myrs. Figure 2.3 shows the orbits and radial distances of four subhaloes from the Aquarius simulation over 8 Gyr, integrated in the HEX-approximated Aquarius potential. The original orbits are shown in blue, while the resimulated HEX orbits are shown in red. HEX reproduces the orbits well. The orbit of the subhalo in the top left is reproduced particularly well, with both pericentres matching closely.

The equations used to describe this method are lengthy. The derivation is given in Lowing et al. (2011), with a correction in Kelly (2022).
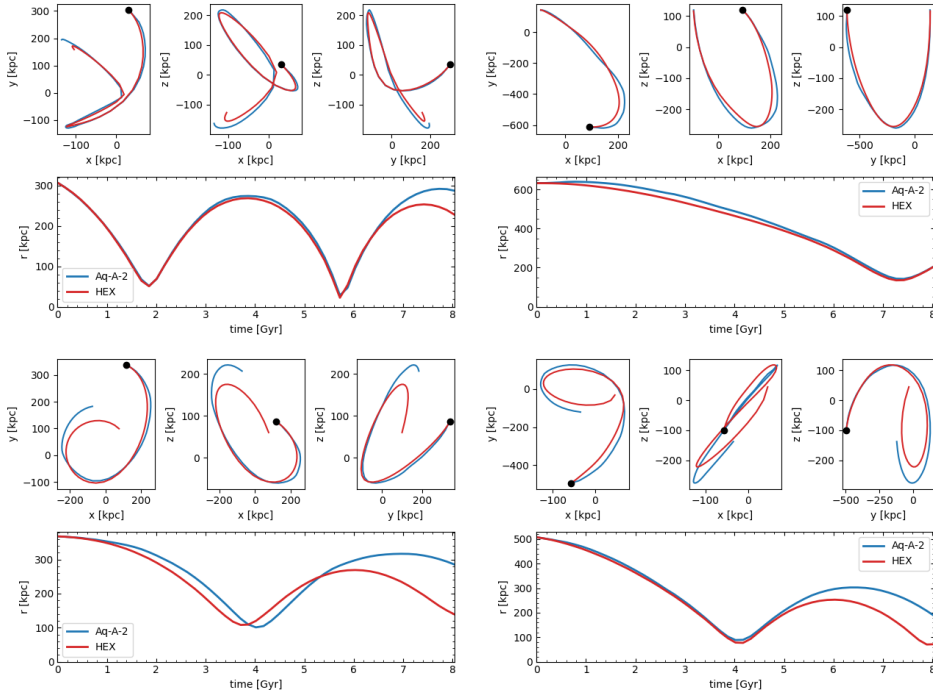
Figure 2.3: The orbits and radial profiles of four different subhaloes taken from Aquarius-A-2. The original is shown in blue, while the HEX resimulation is shown in red.

Following Lowing et al. (2011), Sanders et al. (2020) also used basis function expansions to reproduce the evolution of a Milky Way-like halo. The authors of this paper looked at the orbits of Milky Way subhaloes and the plane of satellites. Sanders et al. (2020) introduced an inertial correction to the basis function expansion method. This is to account for the fact that the halo centre is a non-inertial reference frame. The co-moving coordinate of the halo centre can be given as

$$\mathbf{x}(t) = \mathbf{r}(t)/a(t) \tag{2.10}$$

where $a$ is the cosmological scale factor and $\mathbf{r}$ is the position of the halo centre. The peculiar velocity of the halo centre can be calculated as

$$\mathbf{u} = \dot{\mathbf{r}} - H(t)\mathbf{r} = \dot{\mathbf{x}}a(t) \tag{2.11}$$

where $H(t) = \dot{a}(t)/a(t)$ is the Hubble parameter. The acceleration of the reference frame is thus given by $\dot{\mathbf{u}}$. The force on the particle is then

$$\mathbf{F}(x, t) = -\nabla \Phi(\mathbf{x}, t) - \dot{\mathbf{u}}(t) \tag{2.12}$$

where $\Phi(\mathbf{x}, t)$ is the halo potential and can be reconstructed through a basis function expansion. Without this correction, the agreement between the original orbits and the reconstructed HEX orbits is much less accurate for orbits far from the centre of the halo. This correction is not present in the original work by Lowing et al. (2011), as this focused on subhaloes near the centre, but has since been implemented in HEX by Kelly (2022). The HEX code used to produce all simulations in this thesis contains this inertial correction.

Now, having discussed the methods used in this thesis, we will look at reproducing the work of van den Bosch and Ogiya (2018) in the next chapter.

# Substructure Convergence Studies

## 3.1 Introduction

In this chapter, we take a look at the work of van den Bosch and Ogiya (2018) (hereafter vdBO18 in this chapter). The authors of this work use a large suite of idealised numerical simulations to examine the tidal evolution and disruption of subhaloes. In these simulations, the subhalo is represented by an $N$-body system which is integrated in a static, analytical, external potential representing the main halo. Both the host halo and the subhalo are assumed to be spherical and to have an NFW density profile at the start of the simulation. The authors look at several parameters which can affect the accuracy of numerical simulations, such as force softening, force accuracy and the number of particles which are used to simulate the subhalo.

The host halo is represented by a static, spherical potential with an NFW density profile. The authors define the virial radius, $r_{vir}$, as the radius within which the average density is 97 times the critical density. The virial velocity, $V_{vir} = \sqrt{GM_{vir}/r_{vir}}$, is defined as the circular velocity at the virial radius. The concentration is defined as $c = r_{vir}/r_s$. The host halo is 1000 times more massive than the subhalo.

The subhalo is represented by an $N$-body system. The initial conditions (ICs)

are generated by sampling particles from the distribution function (DF) using the acceptance-rejection method (Press et al., 1992). The subhalo also has an initial NFW density profile. As mentioned in Chapter 2, in an NFW profile the cumulative mass diverges as $r \to \infty$, so the profile must be truncated at some point. Kazantzidis et al. (2004) propose using an exponential cut-off for $r > r_{vir}$, as suggested by Springel and White (1999). vdBO18 argue that as their subhaloes are immediately thrown into an external tidal field, there is little point in having a subhalo which is in perfect equilibrium at the start of the simulation. Therefore, the subhalo is truncated sharply at a radius $r_{max}$, where $r_{max} = r_{vir}$.

The majority of the simulations in vdBO18 are run using a modified version of **treecode**, written by Barnes and with improvements by Dubinski. The code uses a Barnes and Hut (1986) octree and a simple second-order leapfrog integration scheme. Forces between objects are softened using Plummer softening. A second $N$-body code was used for simulations where the number of particles, $N_p$, $> 10^6$. The authors dub this code **OTOO+** (Nakasato et al., 2012; Ogiya et al., 2013). This code was designed to run on GPU clusters and uses a second-order Runge-Kutta integrator. The majority of the simulations which we shall discuss use **treecode**. Only the fiducial runs have $N_p > 10^6$.

Throughout this paper, the authors are concerned with the bound fraction of the mass of subhaloes over time. The bound fraction is defined as

$$f_{bound}(t) \equiv \frac{m_s(t)}{m_{s,0}} = \frac{N_{bound}}{N_p}, \tag{3.1}$$

where $m_s(t)$ is the bound mass of the subhalo at time $t$, $m_{s,0}$ is the initial mass of the subhalo, $N_{bound}$ is the number of bound particles in the subhalo and $N_p$ is the total number of particles in the simulation. As discussed in Chapter 2, identifying substructure in an $N$-body simulation is a non-trivial task. In their paper, the authors use an iterative process to identify the bound fraction. A particle, $i$, is considered bound to the subhalo if its binding energy

$$E_i \equiv \frac{1}{2}m_i v_{\text{int},i}^2 - \sum_{j \neq i} \frac{Gm_i m_j Q_j}{\left(|\mathbf{r}_j - \mathbf{r}_i|^2 + \epsilon^2\right)^{\frac{1}{2}}} < 0. \tag{3.2}$$

$Q_j$ is equal to one if the particle is bound and zero otherwise. At the start of the simulation, it is assumed that all particles are bound. Subsequently, $Q_i$ is assumed to be the same as in previous outputs. $E_i$ is calculated for each particle using Equation 3.2, $Q_i$ is updated and $f_{bound}$ is calculated. The centre-of-mass position and velocity of the subhalo are calculated, taking the average position and velocity of the most bound particles. The velocity of each particle is updated using this centre-of-mass velocity. These steps are repeated until the changes in the new centre-of-mass position and velocity are less than $10^{-4} r_{vir}$ and $10^{-4} V_{vir}$ respectively.

The authors test a number of parameters which can affect the accuracy of numerical simulations, such as the size of the timestep, force accuracy, force softening and the number of particles used in the simulation. Figure 3.1 shows how the bound fraction of a subhalo is affected by a change in the size of the timestep, $\Delta t$, (left) and the opening angle, $\theta$, (right). For the timestep, the results are converged so long as $\Delta t \leq 0.4$. Here, $\Delta t = 1$ (in model units) corresponds to 63.4 Myr. For the force accuracy, the results are converged for $\theta < 2.0$. The authors thus take $\Delta t = 0.02$ and $\theta = 0.7$ as conservative values to use for the rest of their simulations.

The key plot from the paper is shown in Figure 3.2. This shows the bound fraction as a function of time for a number of simulations where the orbiting subhalo is modelled with different numbers of particles, $N_p$, and different softening lengths, $\epsilon$. The softening lengths are in units of the initial scale radius of the subhalo. The simulations follow a subhalo, represented by an $N$-body system, placed on a circular orbit in a static potential. The orbit is at a distance of one tenth of the virial radius of the main halo. The black line represents the converged results, which is based on a simulation with $N_p = 10^7$ and $\epsilon = 0.003$. Within each box, the blue lines show the results from ten different simulations with differing random
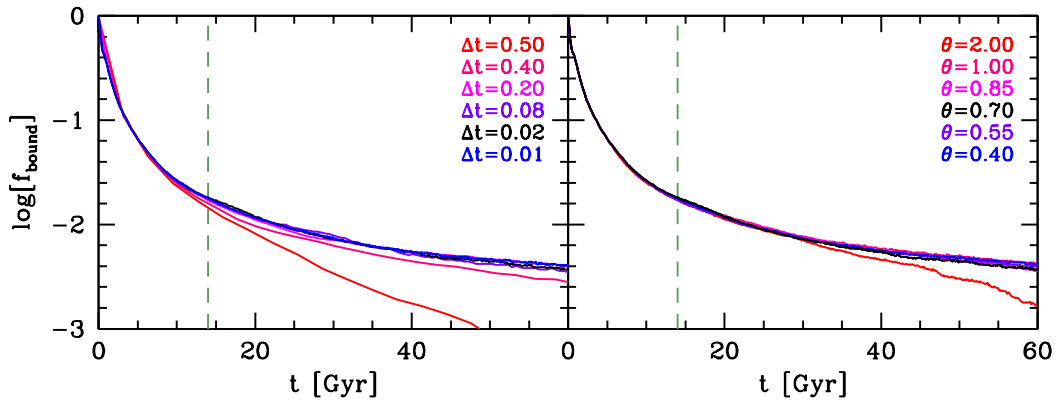
Figure 3.1: The effect that changing the timestep (left) and the opening angle (right) has on the bound fraction of a subhalo over time. The subhalo ($N_p = 10^5$) is placed on an orbit at a distance from the main centre of 20% of $r_{vir}$ of the main halo. The softening length is set to $\epsilon = 0.05$ in units of $r_s$. *Figure 5 from vdBO18.*

realisations of the initial conditions and the red line represents the average of these lines. The vertical dashed line shows the Hubble time. The yellow band indicates the typical values of $N_p$ and $\epsilon$ used in cosmological simulations.

We can see that along this yellow band the results either do not converge, or they are converging to an incorrect value. The authors therefore claim that most state-of-the-art cosmological simulations do not have the necessary force softening or a sufficient number of particles to properly resolve the dynamical evolution of subhaloes. They conclude that most disruption of substructure in $N$-body simulations is numerical in origin, rather than physical, and that cosmological simulations suffer from significant overmerging, mainly due to inadequate softening.

However, as pointed out in the paper itself there are a number of limitations to this work. The primary one is the highly idealised and artificial setup. The host halo is represented by a static, spherical potential, when haloes are dynamic and vary with time. The subhaloes are placed on circular orbits, whereas realistic cosmological orbits are usually elliptical. The subhaloes are also placed on an orbit that is very close to the centre of the main halo for the entirety of the simulation.

Figure 3.3 shows similar results, but here the subhalo is placed at a distance from the main centre of twenty percent of the virial radius. The requirements for con-
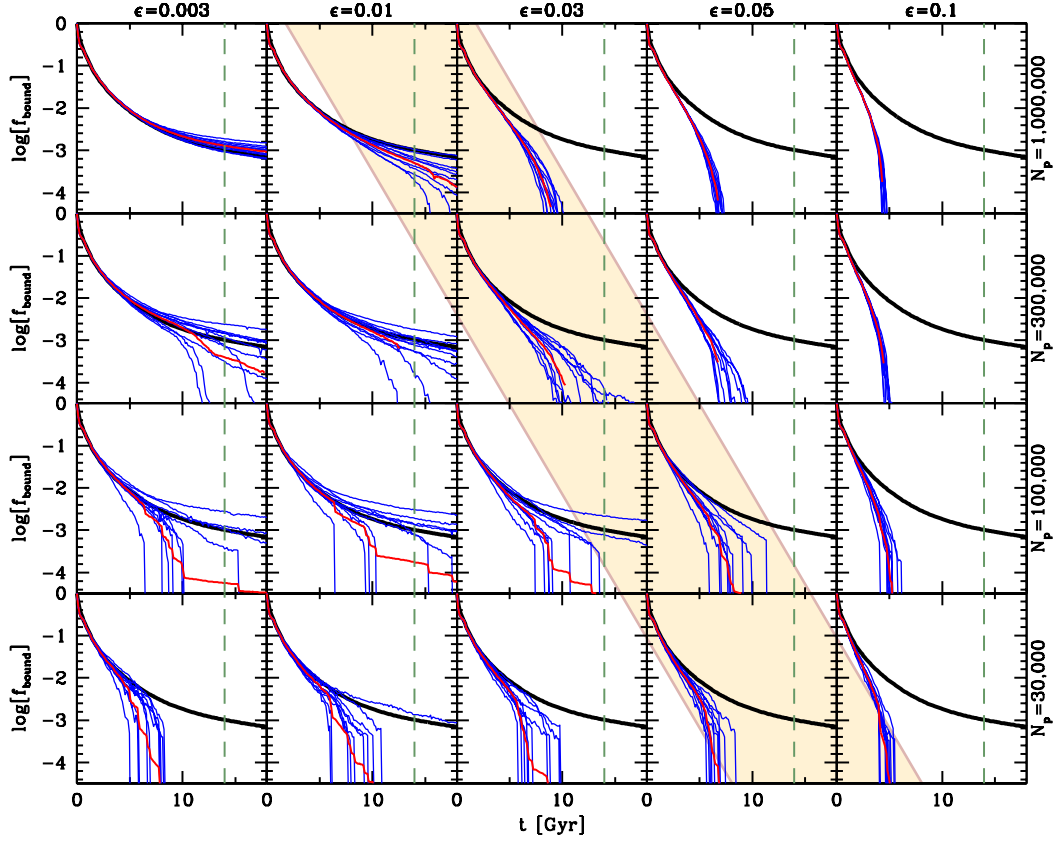
Figure 3.2: The bound fraction as a function of time for simulations with a varying number of particles, $N_p$, (shown along the right-hand side) and softening lengths, $\epsilon$ in units of $r_s$, (shown along the top of the plot). All simulations follow the evolution of a subhalo placed on a circular orbit at a distance from the main centre of 10% of $r_{vir}$ of the main halo. The subhalo has $c = 10$, while the main halo has $c = 5$ and is 1000 times more massive than the subhalo. Within each box, the blue lines show the results from ten different simulations with differing random realisations of the initial conditions and the red line represents the average of these lines. The black line represents the converged results, which is based on a simulation with $N_p = 10^7$ and $\epsilon = 0.003$. The vertical dashed line shows the Hubble time. The yellow band indicates the typical values of $N_p$ and $\epsilon$ used in cosmological simulations. *Figure 10 from vdBO18.*
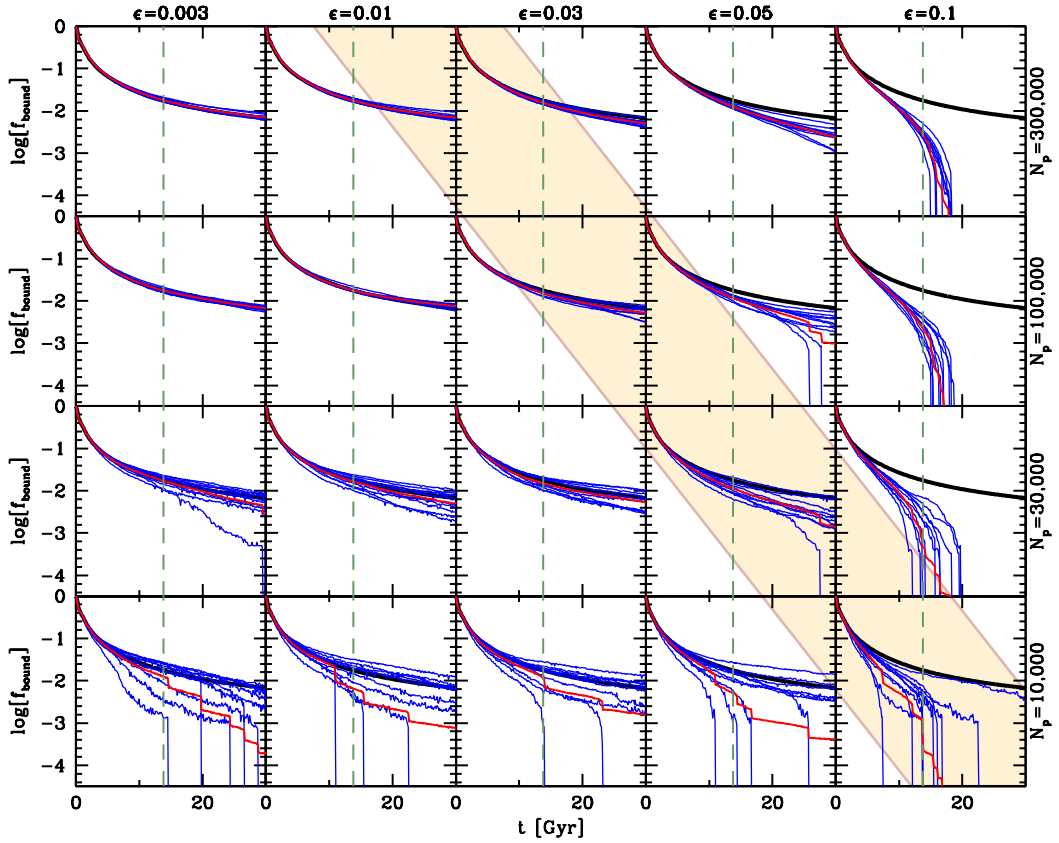
Figure 3.3: The same as Figure 3.2 but this time the subhalo is placed on an orbit at a distance from the main centre of 20% of $r_{vir}$ of the main halo. *Figure 11 from vdBO18.*

vergence are less strict than in Figure 3.2 and along the yellow band the results are converged when $N_p \gtrsim 10^5$. However, the effect of discreteness noise, which arises when a finite number of particles is used to simulate a continuous distribution, is still apparent, especially for small $N_p$.

The vdBO18 paper has been quite influential, cited by over 200 papers at the time of writing. In this chapter, we first check that these results are reproducible before looking at more realistic conditions. In Section 3.2 we introduce the setup used to reproduce the work of vdBO18 and we discuss the difficulties of identifying substructure in artificial conditions. We then test some of the numerical parameters used in $N$-body simulations and reproduce the key plots from vdBO18 in Section 3.3. In Section 3.4 we perform a similar analysis using the original code from

vdBO18, before concluding in Section 3.5.

## 3.2  Reproducing van den Bosch and Ogiya (2018)

We first set out to make sure it is possible to closely reproduce the results presented in vdBO18 independently of the codes that they used. Our host halo is represented by a static potential with an NFW density profile. Rather than using an analytical potential to represent the host, we used HEX (see Section 2.6) to generate the static potential. We did this as we wanted a single setup that could be used both for these simulations and to resimulate orbits from the Aquarius simulation later on. We model the host using code introduced by Kazantzidis et al. (2004). The mass of the halo was set to $10^{12}$ M$_\odot$, the concentration, $c = 5$ and the number of particles, $N_p = 10^6$. We generate coefficients using HEX for this halo and then run our $N$-body code with this static set of HEX coefficients to simulate a static potential. We checked that the resulting potential was a very accurate fit to an analytic NFW.

We also model the subhalo using the same initial conditions code. Following the approach taken by vdBO18, we introduce a sharp cut-off at the virial radius of our subhalo in order to prevent an infinite mass. All simulations follow the evolution of a subhalo with $c = 10$. The mass of each subhalo was set to $10^9$ M$_\odot$, $10^3$ times less massive than the main halo. The number of particles per subhalo varied from $3 \times 10^4$ - $1 \times 10^6$ (m$_p = 3.33 \times 10^5$ - $10^3$ M$_\odot$). The subhalo is placed on a circular orbit at a distance from the centre of the main halo of 10% of the virial radius of the main halo.

The integration of the subhalo evolution was carried out using the GADGET-4 code (Springel et al., 2021). The simulations are run as non-cosmological and the code was set up to be as close as possible to the one used in the original paper. The timestep was fixed to $\Delta t = 0.02$, where $\Delta t = 1$ was equal to 63.4 Myr. The type of opening criterion was set to a geometric opening criterion (see Section 2.2.4), which

is primarily governed by the opening angle $\theta$, where $\theta = 0.7$. As discussed in the previous chapter, GADGET-4 does not use Plummer softening, but rather a cubic-spline (W2) kernel. However, the GADGET-4 parameter file takes the equivalent Plummer softening lengths and these values were matched to those of the original. There are, however, slight differences in the shape of the Plummer softened and spline softened potentials as shown in Figure 2.1.

## 3.2.1 Identifying Substructure

In order to calculate $f_{bound}$ we used the halo-finder SUBFIND. Due to the artificial setup of the simulations - the subhalo is placed on a circular orbit, very close to the centre of the main halo - we find that SUBFIND sometimes had trouble keeping track of the bound remnant of the subhalo. Figure 3.4 shows the subhalo after three orbits (7.42 Gyr). The bound remnant as found by SUBFIND is shown in red and the particles which have been stripped from the subhalo are shown in black. The orbit which the subhalo is placed on, at a distance from the main centre of 10% of the virial radius of the main halo, is shown in green. The subhalo is surrounded by and continuously moving through its own tidal debris. This makes it difficult for SUBFIND to correctly identify the bound remnant.

When the linking length of the FOF algorithm was set to its default value, SUBFIND tended to miss the subhalo, sometimes for several snapshots in a row. We ran a number of simulations with varying FOF linking lengths to determine the best linking length to use for these experiments. The number of particles in the subhalo is $10^5$, $\theta = 0.7$ and $\Delta t = 0.02$. The subhalo was placed on a circular orbit at a distance from the centre of the main halo of 20% of the virial radius of the main halo and the simulation was run for 60 Gyr.

The `FOF_LINKLENGTH` value in GADGET is generally set to 0.2 times the mean interparticle separation. The code estimates the mean particle spacing from the dark matter density and the mean particle mass of all the particles selected with

7.42 Gyr
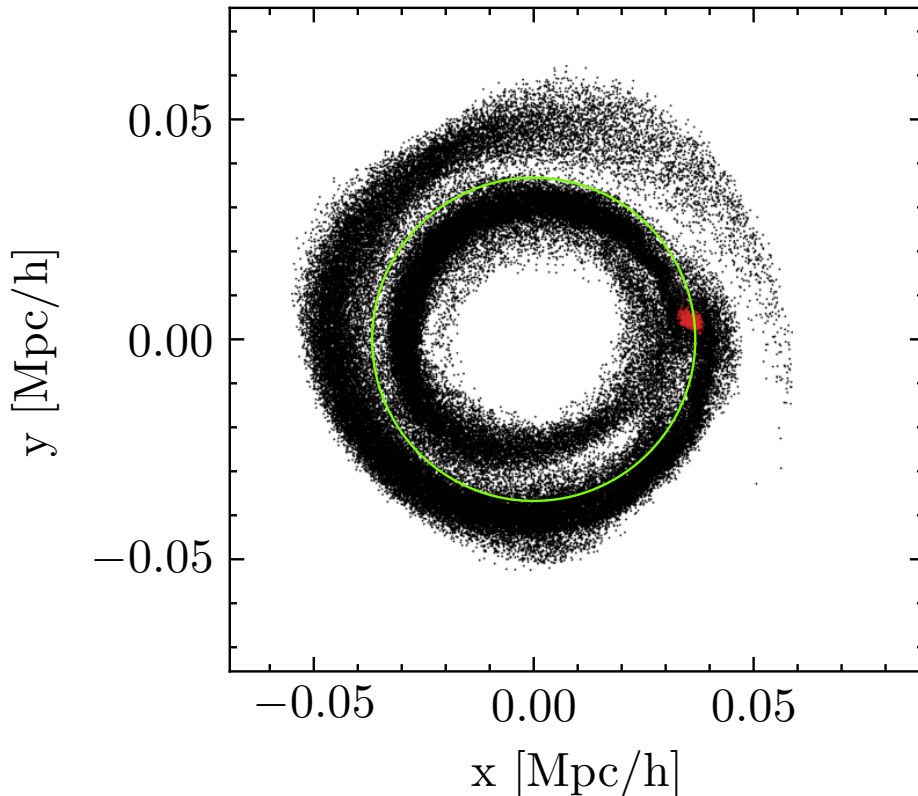


Figure 3.4: The subhalo after three orbits. The bound remnant is shown in red and the stripped particles are shown in black. The subhalo's orbit (at 10% $r_{vir}$) is shown in green.

the `FOF_PRIMARY_LINK_TYPES` mask. We tested values from 0.5 to 0.02. Figure 3.5 shows the FOF groups as calculated by SUBFIND for linking length values of 0.2, 0.14 and 0.08 after three orbits or 7.42 Gyr. When the linking length is set to 0.2, the largest FOF group (shown in light grey) contains almost all the particles in the simulation. This is apparent when we compare the leftmost panel of Figure 3.5 to Figure 3.4. The bound remnant is shown in red on the centre right of each panel.

We looked at the mass of the bound remnant of the subhalo as found by SUBFIND for different linking lengths. Figure 3.6 shows the mass of the bound fraction as a function of time. Different linking lengths are represented by different colours. The mass seems relatively insensitive to the linking length, unless the linking length is
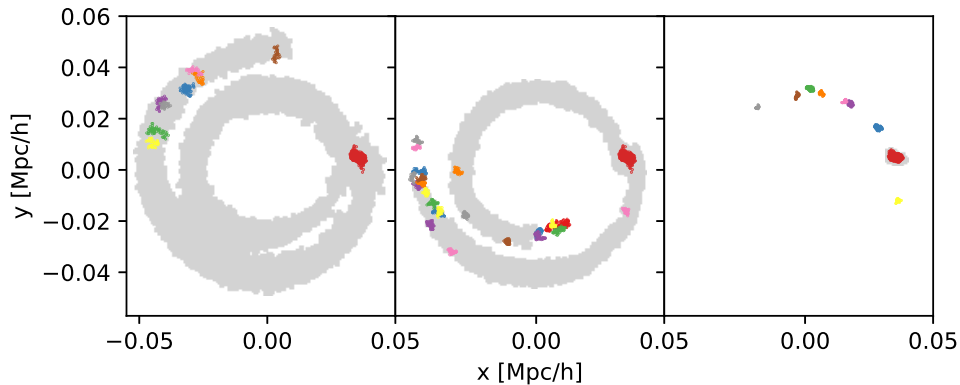
Figure 3.5: FOF groups as found by SUBFIND after three orbits, when the linking length is set to 0.2, 0.14 and 0.08 (left to right) times the mean interparticle separation. Different FOF groups are represented by different colours. The bound fraction as identified by SUBFIND is shown in red.

very small. Gaps are visible in the lines representing the larger linking lengths tested 0.50 - 0.17. This is where SUBFIND lost track of the subhalo. We find that a linking length of 0.08 was the largest value where SUBFIND was able to successfully identify the bound remnant without missing a snapshot. There is an initial difference in the mass of the subhalo when using 0.08 - the mass of the subhalo is about half that found when using the default value of 0.2. However, the two subhaloes' masses converge as the simulation progresses. Therefore, we decided that 0.08 was a good value for the linking length and it was used for the following simulations.

We also looked at the bound fraction as calculated by SUBFIND for different values of linking length. The results are shown in Figure 3.7. Here the bound mass is
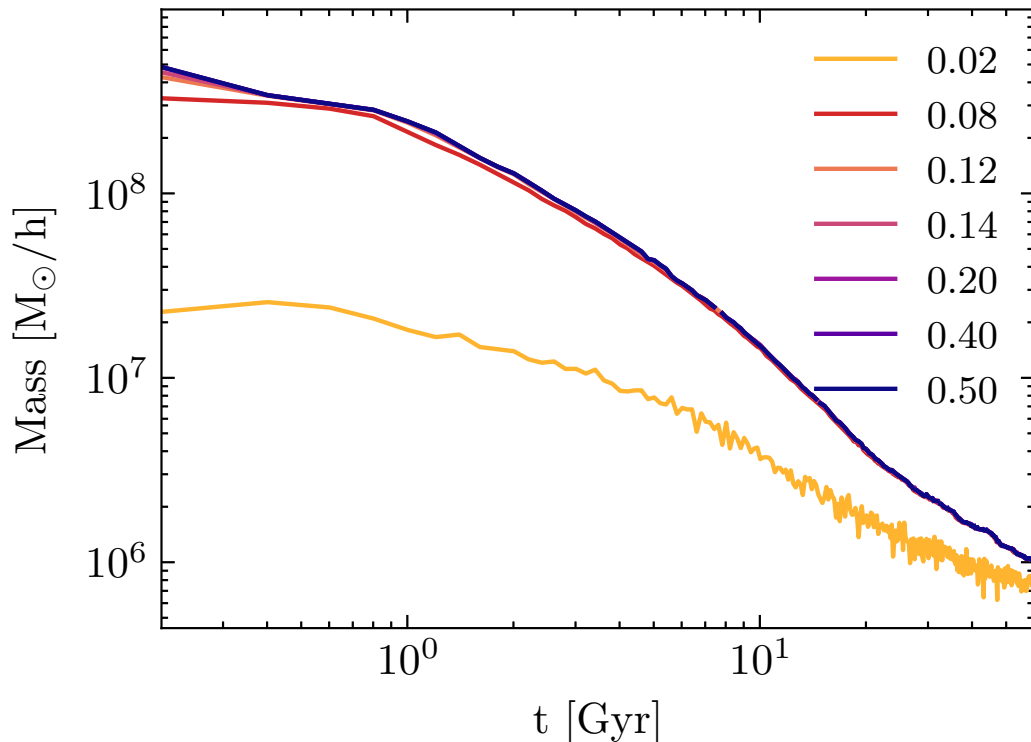
Figure 3.6: The effect that changing the FOF linking length has on the mass of a subhalo. The different colours represent different linking lengths.

normalised by the initial mass, as seen in Figure 3.6. Once again, small gaps are seen in the lines representing the larger linking lengths, where SUBFIND loses track of the subhalo.

## 3.3 Testing Parameters

### 3.3.1 Timesteps

We looked at the effects of changing the timestep, $\Delta t$, and the opening angle, $\theta$, on the bound fraction, as in Figure 5 of vdBO18 (Figure 3.1). We first performed a number of simulations with different fixed $\Delta t$. The subhalo was placed on a circular orbit at a distance from the centre of the main halo of 20% of the virial radius of the main halo. For all these simulations the number of particles in the
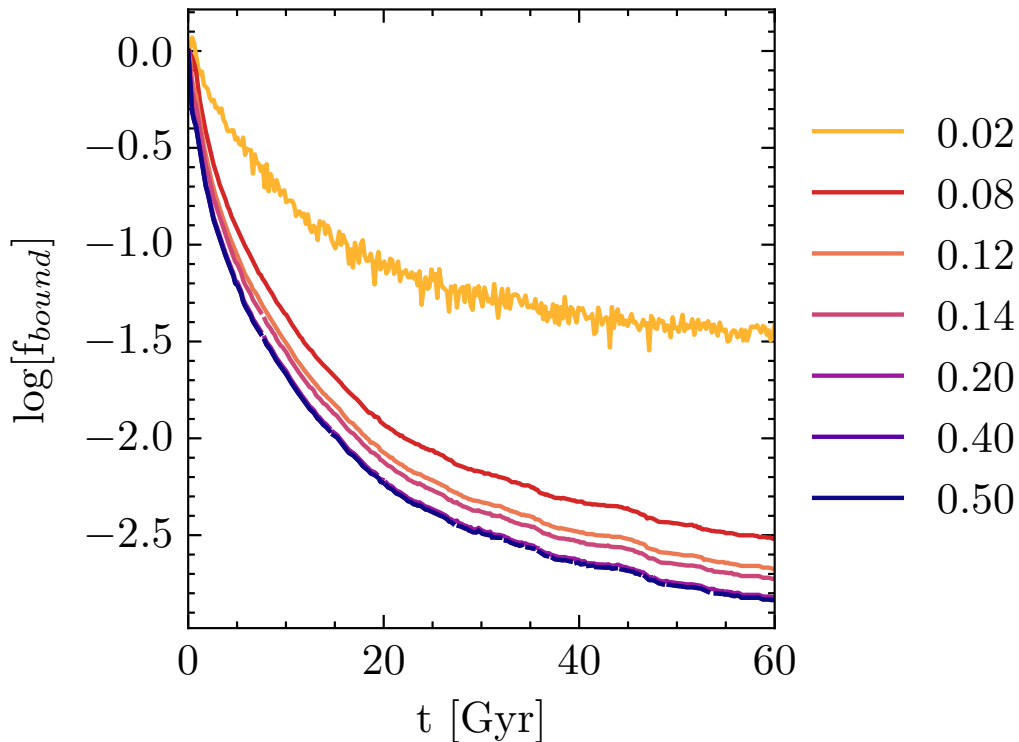
Figure 3.7: The effect that changing the FOF linking length has on the bound fraction of a subhalo. The different linking lengths are represented by different colours.

subhalo were set to $N_p = 10^5$, the softening length was set to $\epsilon = 0.05$ times the initial scale radius of the subhalo and the opening angle was set to $\theta = 0.7$. The simulations were then run for 60 Gyrs. We looked at six different fixed timesteps and then calculated the bound fraction of the subhalo for each using SUBFIND. Figure 3.8 shows the results. Similar to vdBO18, we find that the results converge quite well for $\Delta t \leq 0.4$. In keeping with vdBO18, we used $\Delta t = 0.02$ for all further simulations.

### 3.3.2 Force Accuracy

We next looked at the effects of changing the opening angle, $\theta$, on the bound fraction. Again, the subhalo was placed on a circular orbit at a distance from the centre of the main halo of 20% of the virial radius of the main halo. For
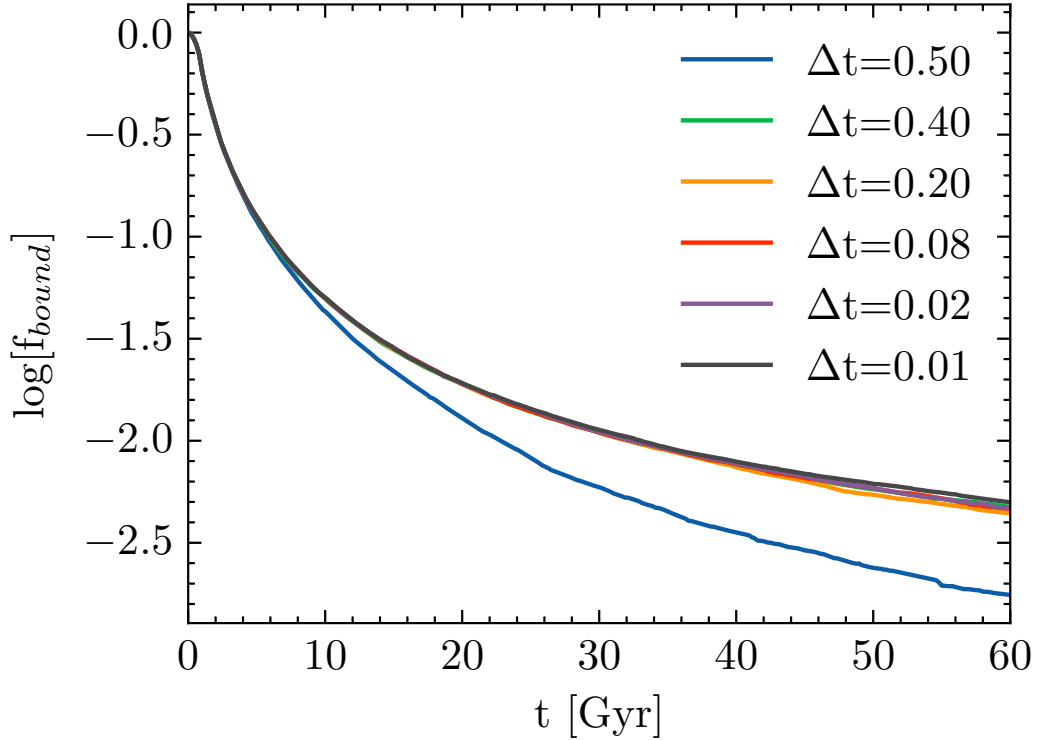
Figure 3.8: The effect that a changing timestep, $\Delta t$, has on the bound fraction of a subhalo. The subhalo is placed on a circular orbit at a distance from the centre of the main halo of 20% of the virial radius of the main halo. All simulations use $N_p = 10^5$ and $\epsilon = 0.05$.

all simulations the number of particles in the subhalo was set to $N_p = 10^5$, the softening length was set to $\epsilon = 0.05$ times the scale radius of the subhalo and the timestep was set to $\Delta t = 0.02$. The simulations were then run for 60 Gyrs, with six different opening angles and the bound fraction of the subhalo was calculated using SUBFIND. Figure 3.9 shows the results. We find that the results converge well and that there was no discernible difference in the bound fraction for different opening angles until after approximately 10 Gyrs. In keeping with vdBO18, we used $\theta = 0.7$ for all further simulations.
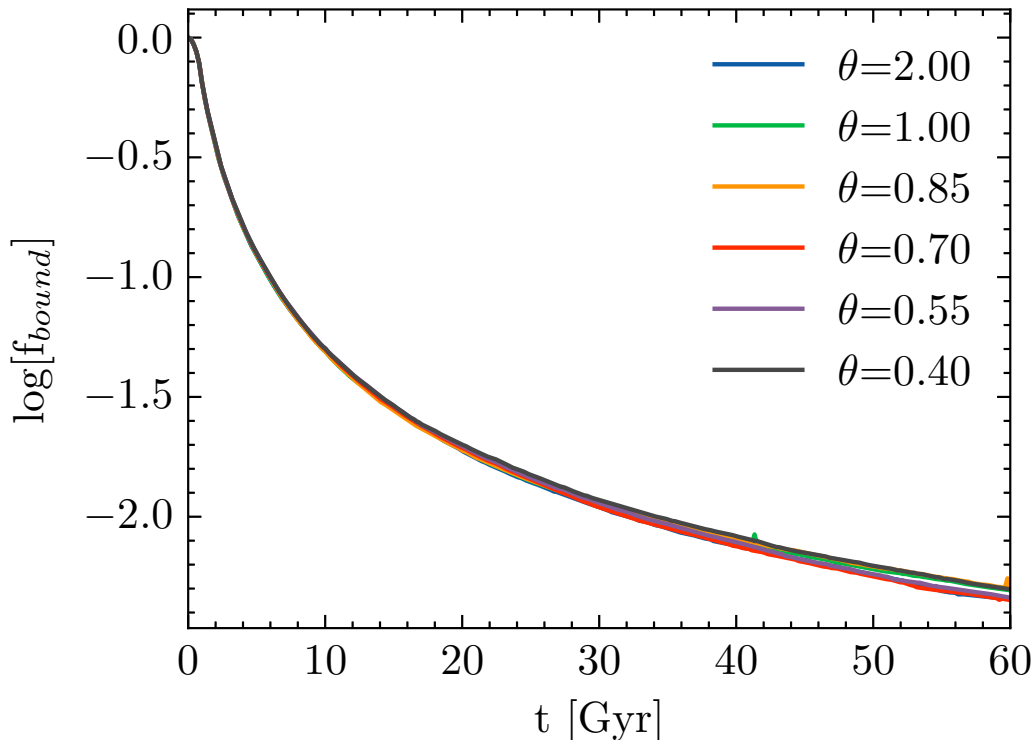
Figure 3.9:   The effect that a changing opening angle, $\theta$, has on the bound fraction of a subhalo. The subhalo is placed on a circular orbit at a distance from the centre of the main halo of 20% of the virial radius of the main halo. All simulations use $N_p = 10^5$ and $\epsilon = 0.05$.

### 3.3.3   Force Softening and Number of Particles

We next attempted to reproduce the key plots from vdBO18 - Figure 10 and Figure 11 (Figures 3.2 and 3.3). These depict a detailed resolution study, where both the softening length, $\epsilon$, and the number of particles in the subhalo, $N_p$, were varied, in an attempt to discover which values led to convergence. Five different softening lengths were used, ranging from $0.1 - 0.003$ times the initial subhalo scale radius, and four different mass resolutions were used, from $N_p = 30,000$ to $N_p = 1,000,000$, giving a total of twenty different simulations. For each combination of $\epsilon$ and $N_p$, ten different random seeds were used to generate the subhaloes. In total, over 200 simulations were run for each plot.

The subhalo was first placed on a circular orbit at a distance from the centre of
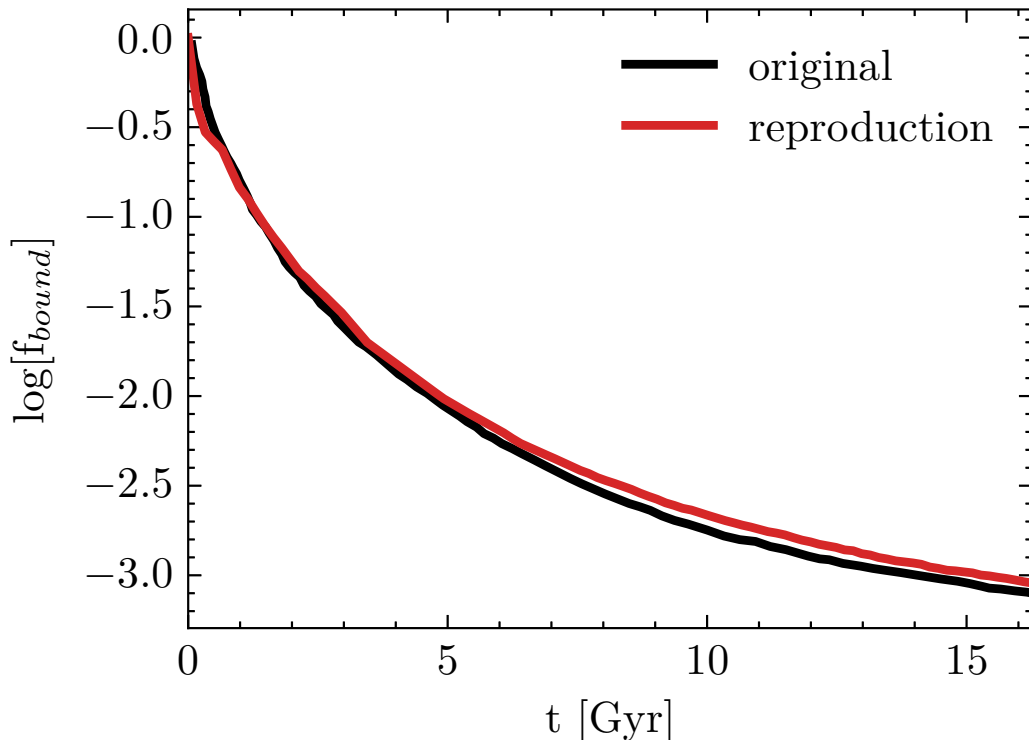
Figure 3.10: The bound fraction over time for a fiducial simulation with $N_p = 10^7$ and $\epsilon = 0.003$. The subhalo is placed on an orbit at a distance of 10% $r_{vir}$ from the main centre. The original run from vdBO18 is shown in black and our reproduction is shown in red.

the main halo of 10% of the virial radius of the main halo. The simulation was run for 17 Gyrs and the remaining bound fraction was then calculated using SUBFIND.

We first looked at reproducing the converged result from Figure 10 of vdBO18. This is a subhalo with $N_p = 10^7$ and $\epsilon = 0.003$, and it is represented by the black lines in Figure 3.2. We generated the initial conditions for this subhalo, integrated its orbit using GADGET-4 and used SUBFIND to calculate the bound fraction over time. The result is shown in Figure 3.10. The values for the original line are extracted directly from Figure 3.2. The two runs match closely. This is both surprising and encouraging. We have used two different methods, with completely different sets of codes, and found that our results match each other well. We would expect smaller systems to also agree.
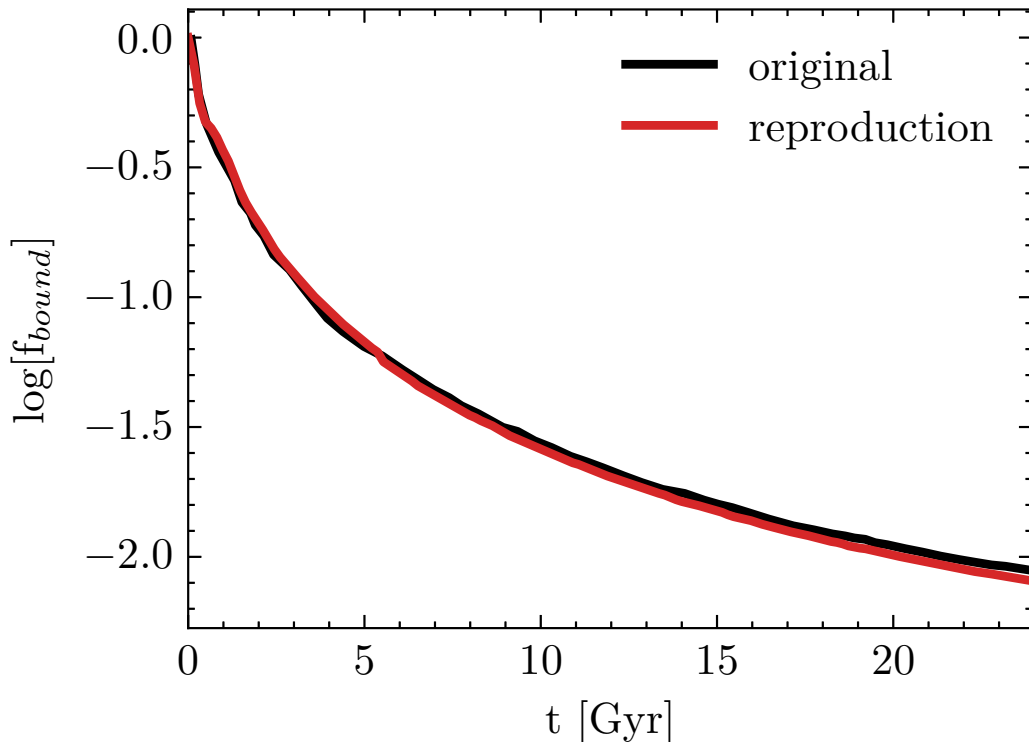
Figure 3.11:   The same as Figure 3.10 but this time the subhalo is placed on an orbit at a distance of 20% $r_{vir}$ from the main centre.

Figure 3.11 shows the same result as 3.10, but this time the subhalo is placed on an orbit at a distance of 20% $r_{vir}$ from the main centre. Once again, we compared our converged result to that found by vdBO18 and the two runs also match closely.

The full results are shown in Figure 3.12. Each blue line represents a different randomly generated subhalo and the red line represents the mean of these ten runs. The vertical green dashed line shows the Hubble time.

Our results differ slightly from vdBO18. We find that our subhaloes are disintegrating less and at later times. For example, in the middle box on the top row, all subhaloes survive for the duration of the simulation and for longer than the Hubble time. In the equivalent box in Figure 3.2, all subhaloes are completely disrupted within 10 Gyrs. In the box directly below this one, the majority of subhaloes survive up to the Hubble time, whereas in the original, almost all subhaloes are gone by this point. As these two boxes are along the yellow band in Figure 3.2,
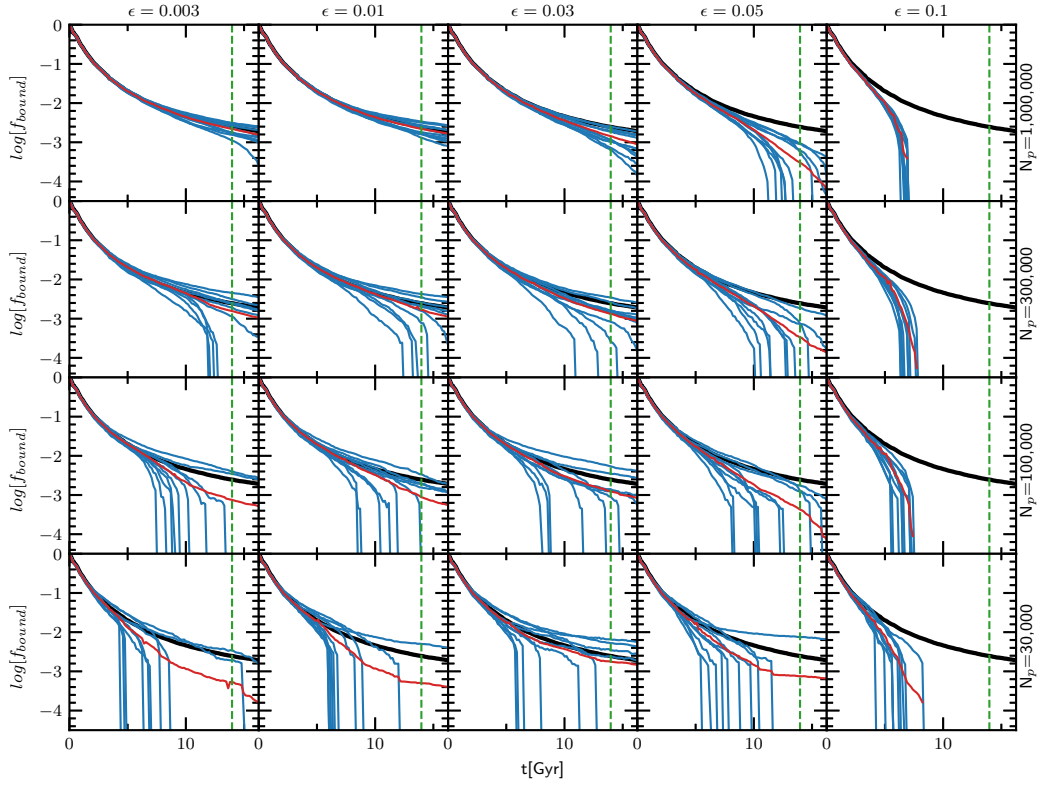
Figure 3.12: The bound fraction vs time for a number of simulations with different numbers of particles, $N_p$, and softening lengths, $\epsilon$. The black line represents the fiducial subhalo. The blue lines represent simulations of subhaloes created using different random seeds. The red line is the mean. The green dashed line shows the Hubble time. The subhaloes are placed on a circular orbit a distance of 10% of the virial radius of the main halo.

showing the typical values used in cosmological simulations, one might question the assertion that such simulations do not have the necessary force or mass resolution to properly resolve subhaloes.

While the results differ quantitatively, they do agree, in general, qualitatively. For smaller $N_p$ and for larger $\epsilon$ the simulations fail to converge and we see a lot of discreteness noise, particularly along the bottom two rows. However, these differences are surprising, considering the care that was taken to ensure that our setup matched the original. This prompted further investigation.

Figure 3.13 shows the same results as Figure 3.12, but this time for a subhalo on a circular orbit a distance from the main centre of 20% of the virial radius of the
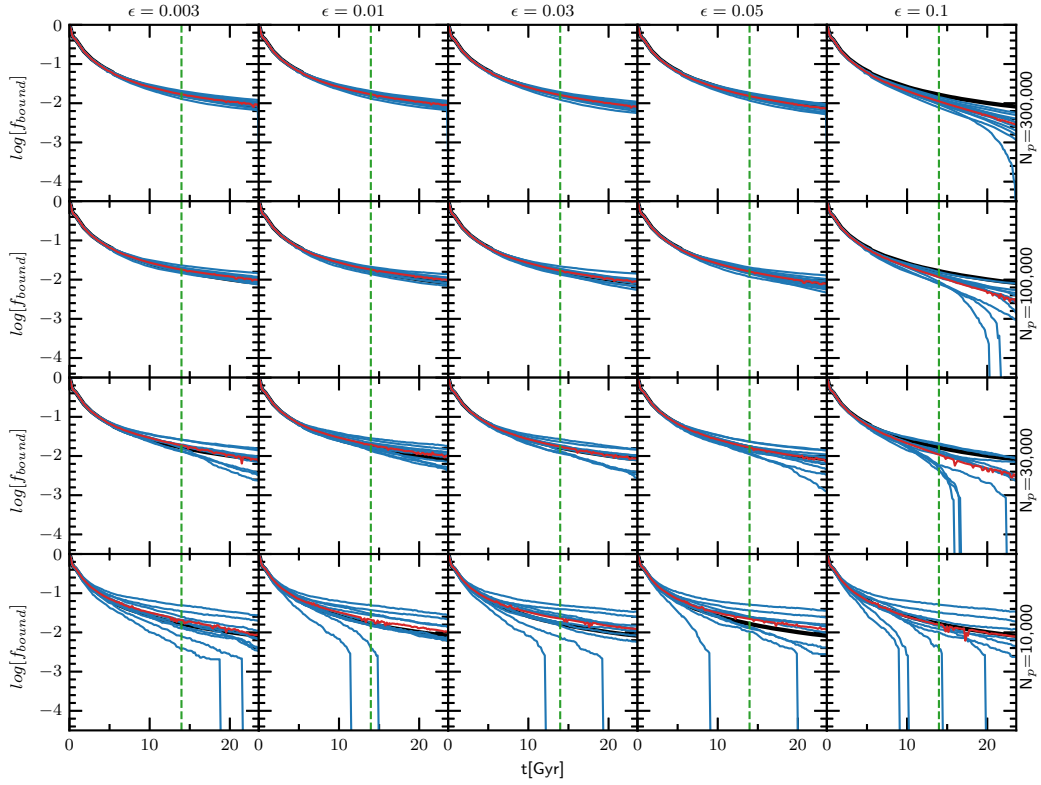
Figure 3.13: The same as Figure 3.12, but this time the subhaloes are placed on a circular orbit a distance of 20% of the virial radius of the main halo.

main halo. This simulation was run for over 20 Gyrs and a smaller number of particles was used. The greater distance from the centre means that the subhaloes are subjected to smaller tidal forces, and thus require a smaller number of particles to be properly resolved. Again, our results differ slightly from vdBO18. The subhaloes are disintegrating less often and at later times. Along the bottom row especially, we see some discreteness noise again.

## 3.4 Testing the original code

The results agree qualitatively, but there are differences. Our results use independent codes as the codes used in vdBO18 are not in the public domain. To make further progress in trying to understand the difference we contacted Frank van den Bosch. We then used the original code kindly supplied by Frank van den Bosch

to attempt to reproduce the exact results. All parameters were set to match those used in the original work. We even attempted to use the same random numbers to generate our initial conditions.

## 3.4.1   Comparison of Analysis Codes

In order to determine what was causing the discrepancies between our results and those of vdBO18, we ran some comparison tests between the halo-finder algorithm used in the original paper and SUBFIND. We ran four simulations for four different objects using the `treecode`, converted the simulation output to GADGET-4 output format and then ran SUBFIND over these results. The results are shown in Figure 3.14. The bound fraction as found by the original analysis code is shown in blue, while the results from SUBFIND are indicated by the red dashed line. For the most part, the results are in good agreement. There is a slight offset between the two lines, where the red SUBFIND line is consistently above the blue line, but this could be reduced by tuning the FOF linking length.

The original analysis code displays some erratic fluctuations once the bound fraction becomes very small, i.e. when the subhalo has disrupted. To handle this, if the bound fraction is found to be 1.5 times higher than in the previous snapshot, we assume that the subhalo has disappeared completely. The value of 1.5 was reached after extensive testing by Frank van den Bosch, who found that the subhalo is completely disrupted at this point and the code starts to produce junk values. This seems to match the results as given by SUBFIND - the erratic behaviour in the blue line begins once SUBFIND has lost track of the subhalo. These results mean that we can be confident that the discrepancies between our results and those of vdBO18 are not due to the post-simulation analysis.
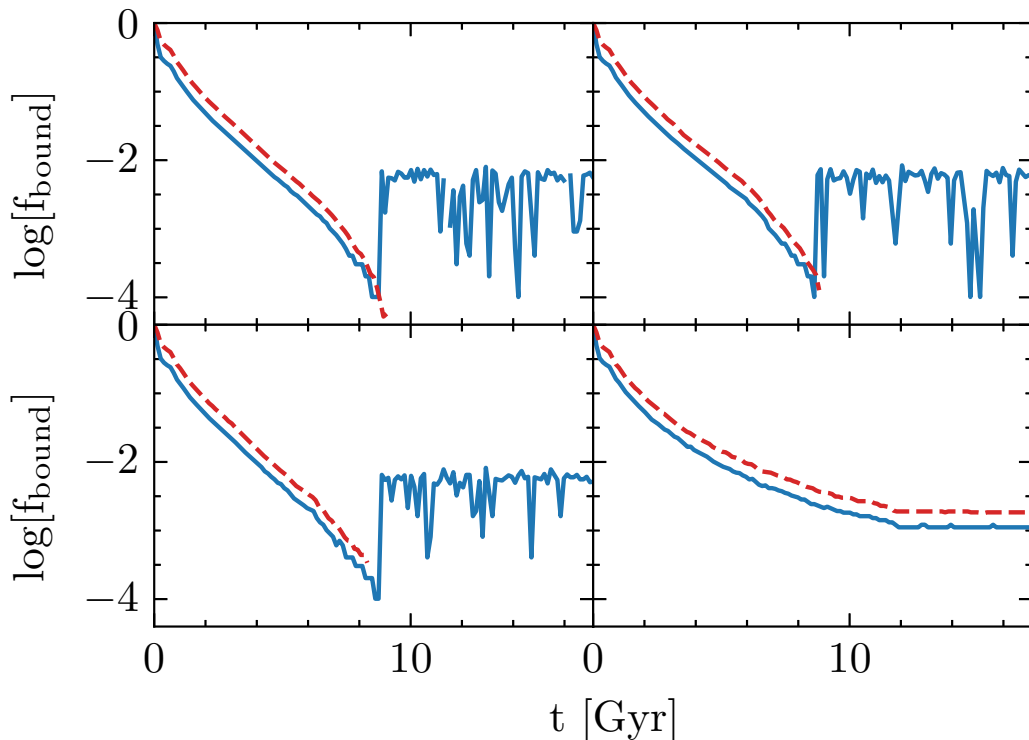
Figure 3.14: A comparison between SUBFIND and analysis code provided by Frank van den Bosch for four different objects. The bound fraction as calculated by SUBFIND is shown as the red-dashed line, while the bound fraction calculated by the van den Bosch analysis code is shown in blue.

## 3.4.2 Reproduction of Key Figure

Figure 3.15 shows the results of a series of simulations with varying softening lengths and numbers of particles using the original code from vdBO18. The initial conditions code was used to generate each subhalo and the $N$-body `treecode` was used to integrate the orbits. The subhalo is once again placed on a circular orbit, at a distance from the main centre of 10% of the virial radius of the main halo. As above, each blue line shows the bound fraction over time for a subhalo generated using a different random seed. The red line represents the mean of the ten runs in each box and the vertical green line shows the Hubble time. The black line shows the converged result with $N_p = 10^7$ and $\epsilon = 0.003$. This was also run using the original `treecode`.

Although every measure was taken to ensure that the setup was as similar to vdBO18 as possible and all the same codes were used, there are still differences between Figure 3.15 and Figure 3.2 (Figure 10 from vdBO18). In general, the subhaloes here last for longer before being fully disrupted. For example in the third box on the top row, half of the subhaloes survive up until the Hubble time, whereas in the equivalent box in vdBO18 all subhaloes are destroyed within 10 Gyrs. In the bottom row, several subhaloes survive until the end of the simulation, losing less mass than in even the fiducial simulation, whereas others are completely disrupted within 5 Gyr. While this is a good demonstration of the effects of discreteness noise, it is also another example of our subhaloes being slightly more robust than those from the original paper.

We were unable to track down exactly why these discrepancies occur. The codes used were the same and all parameters were set to the values used in the original simulations. The same random numbers were used to generate the initial conditions. Correspondence with Frank van den Bosch did not yield any insights as to what could be causing the differences.

Comparing the results of our simulations using GADGET-4 (Figure 3.12) and those using the vdBO18 code (Figure 3.15), we can see that the GADGET-4 subhaloes are much more resilient. In the first three columns of the top row of Figure 3.12 all subhaloes survive for the duration of the simulation and in the fourth column, a third of the subhaloes do not get disrupted. In contrast, in Figure 3.15, while all subhaloes survive in the first two columns, only a third survive in the third column and none survive in the fourth column. In the fifth column, with the largest softening length, the results are similar for both.

## 3.5 Conclusions

Work by vdBO18 has shown that subhaloes undergo artificial disruption due to numerical effects. In Sections 3.2 and 3.3 we attempt to reproduce the results of
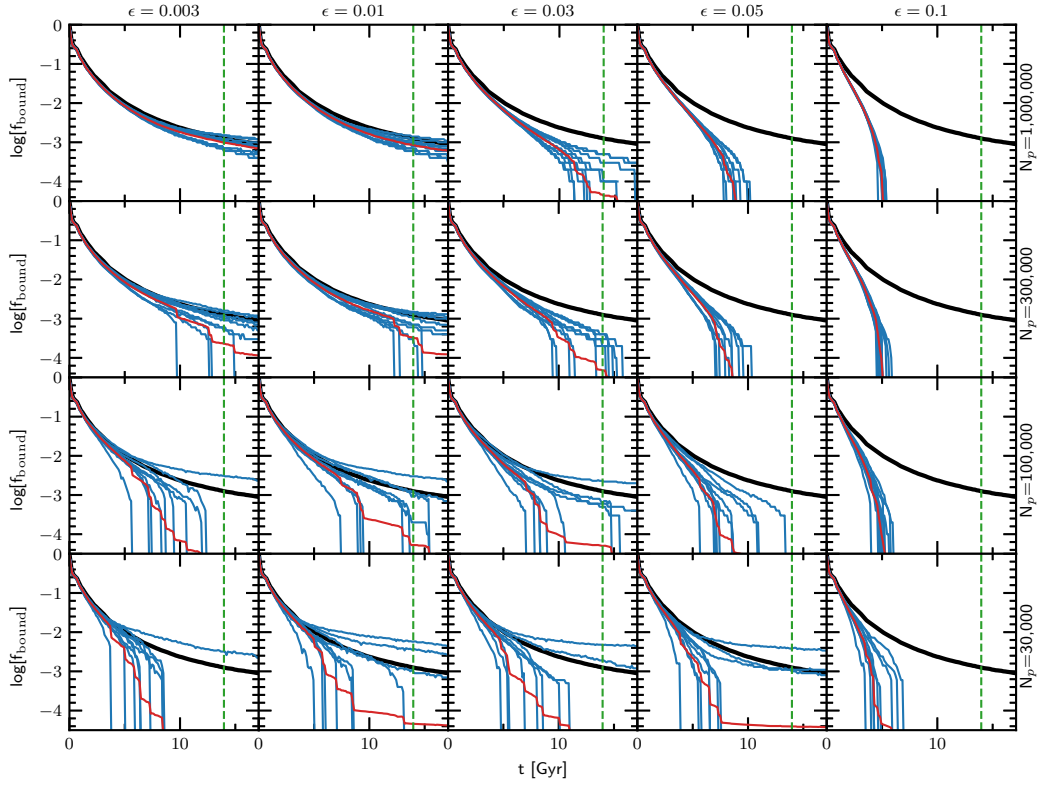
Figure 3.15: The same as Figure 3.12, but run using code provided by Frank van den Bosch. The bound fraction vs time for a number of simulations with different numbers of particles, $N_p$, and softening lengths, $\epsilon$. The black line represents the fiducial subhalo. The blue lines represent simulations of subhaloes created using different random seeds. The red line is the mean. The green dashed line shows the Hubble time. The subhaloes are placed on a circular orbit a distance of 10% of the virial radius of the main halo.

vdBO18 using independent software, and in Section 3.4 we use the same codes as are used in vdBO18. Every attempt was made to reproduce the methods used by the authors of the original paper - we even used the same random numbers to generate our initial conditions. In spite of this, as well as with help from Frank van den Bosch, it has not proven possible to reproduce these results exactly. However, our results do agree qualitatively. The subhaloes do get disrupted, particularly those with low $N_p$ and high $\epsilon$. The effects of discreteness noise are also evident in simulations with low mass resolution. Our subhaloes have proven to be slightly more robust, but they do end up disintegrating, especially at longer softening lengths and fewer numbers of particles. Comparing the results of our simulations using

GADGET-4 (Figure 3.12) and those using the vdBO18 code (Figure 3.15), we can see that the GADGET-4 subhaloes are much more resilient. Our findings, therefore, do disagree with those of vdBO18, but qualitatively the results are nonetheless similar.

However, the conditions under which these simulations are run are idealistic and extremely artificial. The subhaloes are placed on circular orbits, which are very close to the centre of the main halo for the entire duration of the simulation, and the main halo is represented by a static, external potential. In the next chapter, we will investigate the disruption of subhaloes in more realistic conditions using the HEX code.

# Milky Way-Type Halo Substructure

## 4.1 Introduction

In this chapter, we examine the substructure of a Milky Way-like halo from the Aquarius simulation and its tidal evolution. We follow a similar analysis to the previous chapter, but here we use more realistic conditions. van den Bosch and Ogiya (2018) followed the tidal evolution of subhaloes which were placed on circular orbits, in a static analytical potential. In order to use more realistic orbits, we take objects directly from the Aquarius simulation and resimulate their evolution using HEX. By selecting these objects we get a sample of the distribution of subhaloes and their properties. By using HEX, we also have a time-dependent potential to represent the main halo, rather than a static, analytic potential. HEX also accurately captures the shape of the halo, whereas we used a spherically symmetric potential in the previous chapter.

We use the Aquarius merger trees (see Section 2.3), provided by John Helly (Jiang et al., 2014), to track these subhaloes through the simulation. The Aquarius merger trees cover the high-resolution region of the Aquarius simulations. The Aquarius A-2 halo was selected for our initial analysis for a couple of reasons. Firstly, to

| $m_p$ | $\epsilon$ | $N_{hr}$ | $N_{lr}$ | $M_{200}$ | $r_{200}$ |
|---|---|---|---|---|---|
| [M$_\odot$] | [pc] | | | [M$_\odot$] | [kpc] |
| $1.370 \times 10^4$ | 65.8 | 531,570,000 | 75,296,170 | $1.842 \times 10^{12}$ | 245.88 |

Table 4.1: Properties of the Aq-A-2 simulation. $m_p$ is the particle mass, $\epsilon$ is the Plummer equivalent softening length, $N_{hr}$ is the number of high-resolution particles and $N_{lr}$ is the number of low-resolution particles. $M_{200}$ is the virial mass of the halo, while $r_{200}$ is the corresponding virial radius. *Excerpt from Table 1 in Springel et al. (2008).*

| $V_{max}$ | $r_{max}$ | $\delta_V$ | $c^*_{NFW}$ | $z_{form}$ |
|---|---|---|---|---|
| | [km s$^{-1}$] | [kpc] | | |
| 208.49 | 28.14 | $2.060 \times 10^4$ | 16.19 | 1.93 |

Table 4.2: Basic structural properties of the main halo in Aq-A-2. $V_{max}$ is the maximum circular velocity, $r_{max}$ is the radius at which this occurs, $\delta_V$ is the characteristic density contrast based on the peak of the circular velocity curve and $c^*_{NFW}$ is the same value converted to an equivalent NFW concentration. $z_{form}$ is the redshift of the halo formation, here defined as the point at which $M_{200}$ of the main halo progenitor exceeds half its final value. *Excerpt from Table 2 in Springel et al. (2008).*

produce accurate results, HEX requires a high rate of temporal sampling. Aq-A-2 has a high time resolution output compared to the other Aquarius simulations, with snapshots every $\sim 155$ Myr. This is essential for HEX to produce an accurate potential. Secondly, the Aq-A-2 simulation is run at a very high resolution, with 531,570,000 high-resolution particles and a mass per high-resolution particle of $1.37 \times 10^4$ M$_\odot$. However, the Aq-A-2 main halo also has a high concentration for its mass (16.19), which is significantly higher than the median concentration (9.76) of the other level 2 Aquarius haloes. For this reason, we would like to model the other relaxed Aquarius haloes at a later date. Some basic parameters of the Aq-A-2 simulation are shown in Table 4.1, while some structural properties of the main halo are shown in Table 4.2.

In this chapter, we first discuss substructure in the Aquarius A-2 simulation in Section 4.2. We will discuss two different types of subhalo and the distribution of their infall times and mass distribution. In Section 4.3 we will discuss the numerical setup and the pipeline used for these simulations. We will then talk about the use of HEX in Section 4.3.2, before discussing subhalo selection in Section 4.4.1. Finally,

we discuss our results and conclusions in Section 4.5.

## 4.2 Substructure in Aq-A-2

Once a relatively lower mass dark matter halo is accreted by another halo, it becomes a subhalo of its host. In our analysis, we take the time of accretion, or infall, to be the time that a subhalo crosses $r_{200}$ of the main halo. We classify this substructure depending on whether a bound object is identified by SUBFIND at $z = 0$. If an object still exists at the end of the simulation at $z = 0$, it is called a type 1 subhalo (Font et al., 2011). This means that it contains 20 or more particles according to SUBFIND. If a subhalo loses enough mass to fall below the 20 particle limit according to SUBFIND, it is called a type 2 subhalo. Once an object becomes a type 2 subhalo, we can no longer track its orbit through the merger trees, but we can still follow its most bound particle. The position of this most bound particle at the end of the simulation is taken as the final position of the subhalo. There are approximately ten times as many type 2 subhaloes in the Aq-A-2 halo as there are type 1 subhaloes. However, the majority of these are very small and close to the 20 particle limit at their time of infall into the main halo.

We first look at the distribution of type 1 and type 2 subhaloes. Figure 4.1 shows all type 1 subhaloes which are within $r_{200}$ of the main halo in Aq-A-2. The red circle indicates $r_{200}$ of the main halo. We can see that the concentration of subhaloes is denser in the middle of the halo, but that there is not much clustering of substructure. It is also evident from the distribution of substructure that the halo is not spherical in shape. Figure 4.2 shows all type 2 subhaloes which are within $r_{200}$ at $z = 0$. Many of the type 2s are associated with larger substructures and they are more centrally located than the type 1s.

The subhaloes that exist within the main halo today were accreted over time. Figure 4.3 shows the distribution of infall times into the main halo for all subhaloes, type 1s and type 2s, in Aq-A-2. The red line indicates the infall times of all
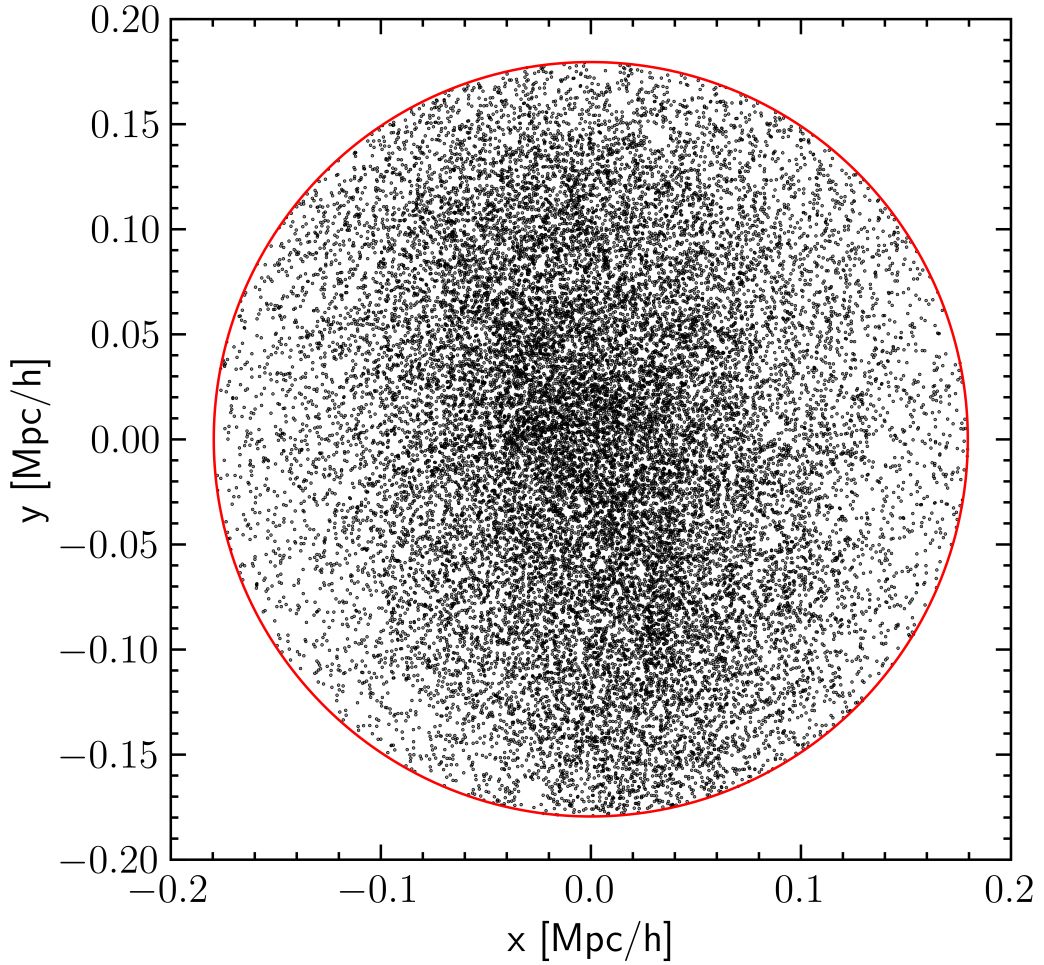
Figure 4.1: All type 1 subhaloes which are within $r_{200}$ (the red circle) at $z = 0$.

subhaloes, while the blue bars show the infall times of all subhaloes that end up within $r_{200}$ at $z = 0$. We are primarily concerned with subhaloes that infall after $z = 1$, as this is as far back as our HEX coefficients have been tested. We are also, for the moment, focusing on objects which lie within $r_{200}$ at $z = 0$. Therefore, we are looking at the blue bars from $z = 1$ onwards. The peak mass distribution of all subhaloes within $r_{200}$ at $z = 0$ is shown in Figure 4.4. The peak mass is the highest mass which a particular subhalo reaches throughout the duration of the simulation.

Figure 4.5 shows the distribution of infall times for just the type 1 subhaloes. There are two peaks in the distribution at lookback times of approximately 5 Gyr and 1.5
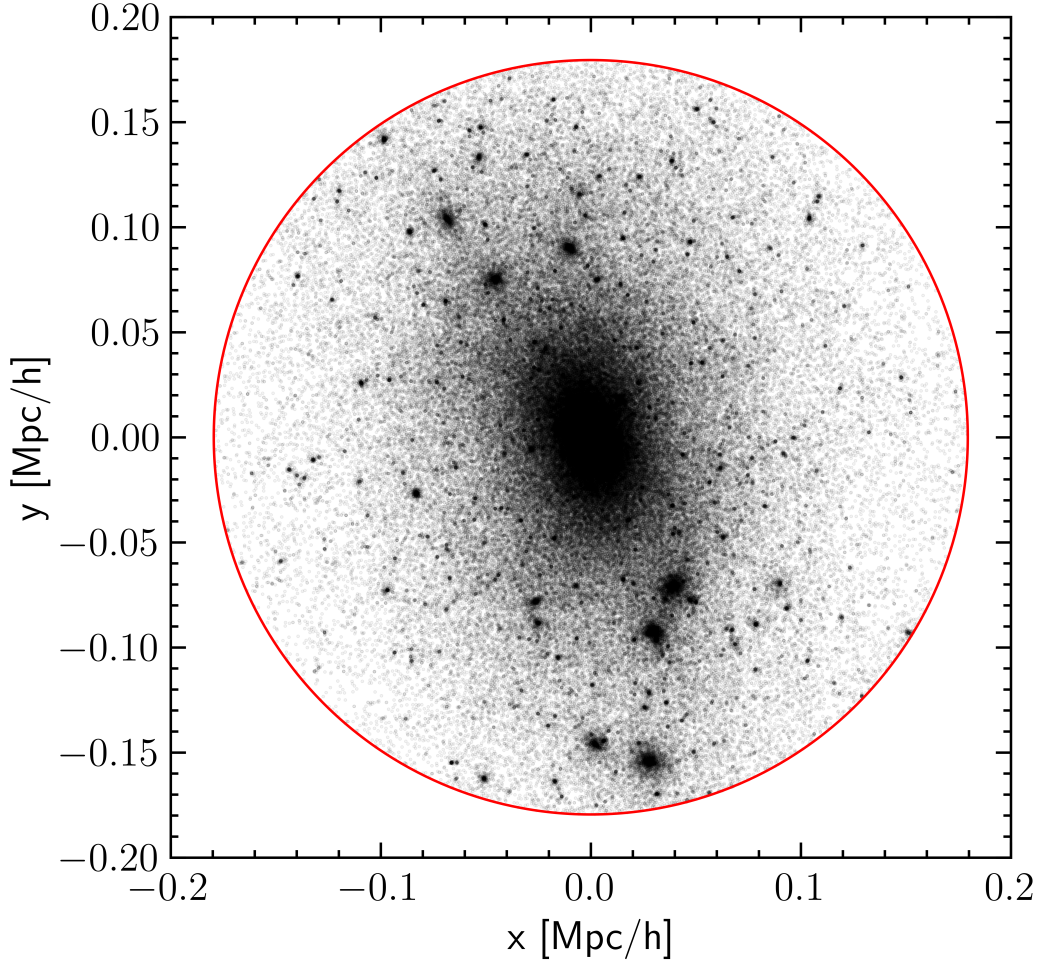
Figure 4.2: All type 2 subhaloes which are within $r_{200}$ (the red circle) at $z = 0$.

Gyr, with a dip in the middle. The dip can be explained by the fact that subhaloes which infall at this time (2.5 - 4 Gyr lookback time) tend to be at a stage in their orbital period where they are at their first apocentre, outside of $r_{200}$ at $z = 0$. The red line at this point demonstrates that there are many subhaloes which infall at this time.

The peak mass distribution of the type 1 subhaloes within $r_{200}$ at $z = 0$ is shown in Figure 4.6. This extends from $\sim 10^5 - 10^{10}$ M$_\odot$ and peaks at a mass of $10^6$ M$_\odot$, which in the Aq-A-2 simulation is approximately 200 particles. There is a sharp drop-off at about $2 \times 10^5$ M$_\odot$, which is 20 particles and is the point when a type 1 subhalo becomes a type 2. As our fitting procedure (see Section 2.4) requires at
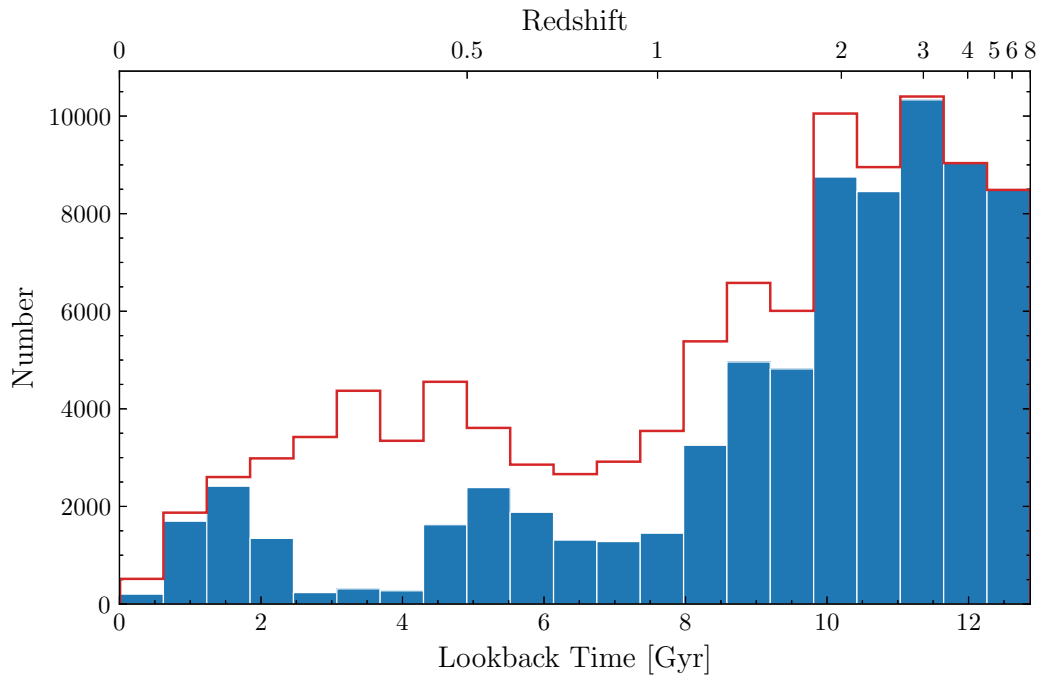
Figure 4.3: The distribution of infall times for all subhaloes in Aq-A-2. The infall times of subhaloes which are within $r_{200}$ at $z = 0$ are shown in blue, while the red indicates the infall times of all subhaloes.
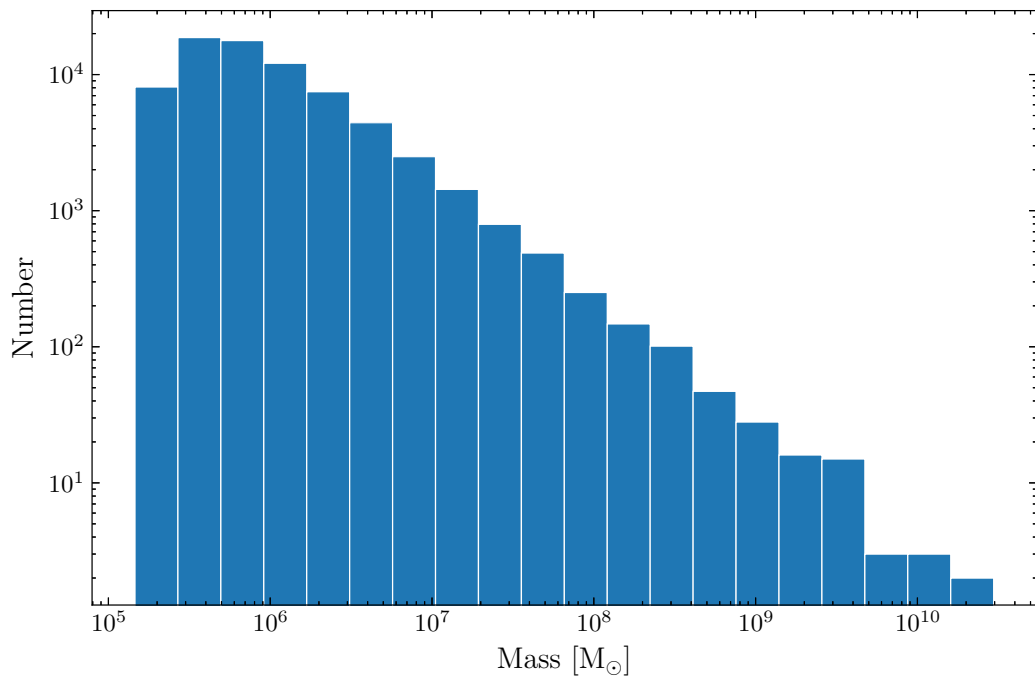


Figure 4.4: The peak mass distribution of all subhaloes within $r_{200}$ at $z = 0$ in the Aq-A-2 simulation.
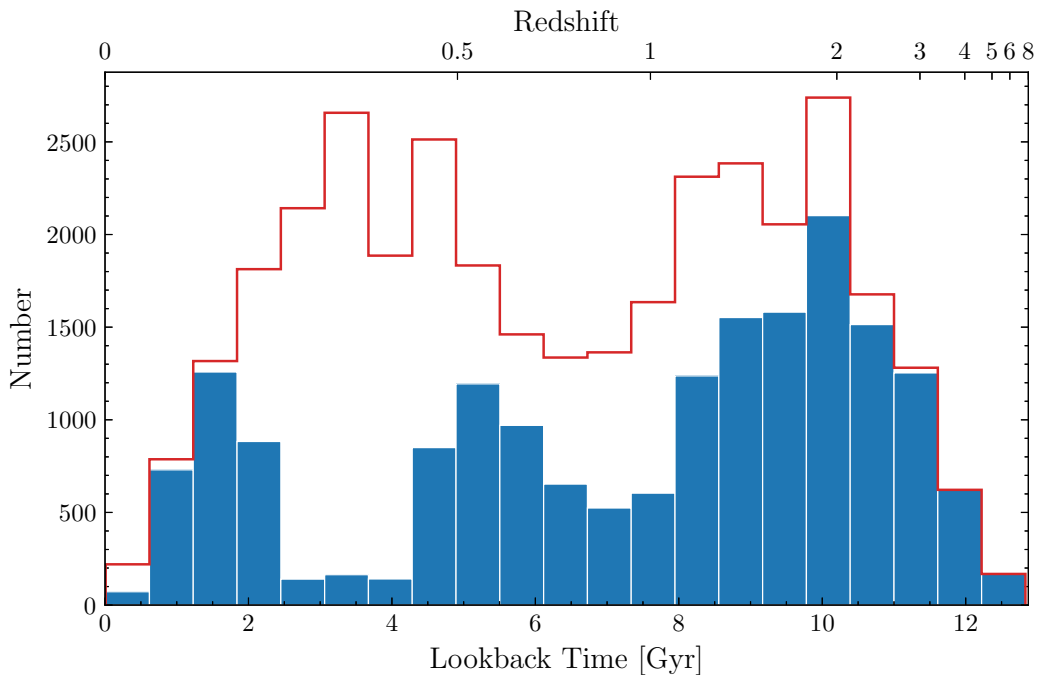
Figure 4.5: The distribution of infall times of type 1 subhaloes. The infall times of subhaloes which are within $r_{200}$ at $z = 0$ are shown in blue, while the red indicates the infall times of all type 1 subhaloes.

least 1,000 particles to get an accurate fit, we concern ourselves with the higher mass subhaloes for the moment. Specifically, we are first looking at subhaloes with an infall mass of at least $10^8$ M$_\odot$.

Figure 4.7 shows the distribution of infall times for all type 2s in Aq-A-2. Here, the vast majority of subhaloes infall before $z = 2$, after which point the numbers drop sharply. This makes sense, as the longer an object orbits close to the main centre and the greater number of pericentric passages it has, the more mass it loses and the more likely it is to become disrupted. With each pericentric passage, an object will lose a certain amount of mass. Once again, we see a dip in the distribution between $\sim 2.5 - 4$ Gyr lookback time.

The peak mass distribution of the type 2 subhaloes within $r_{200}$ at $z = 0$ is shown in Figure 4.8. Similar to the type 1 peak mass distribution, this extends from $\sim 10^5 - 10^{10}$ M$_\odot$. However, it peaks at a slightly lower mass, closer to the cut-off point for type 1 subhaloes. Thus, the majority of the objects which become type
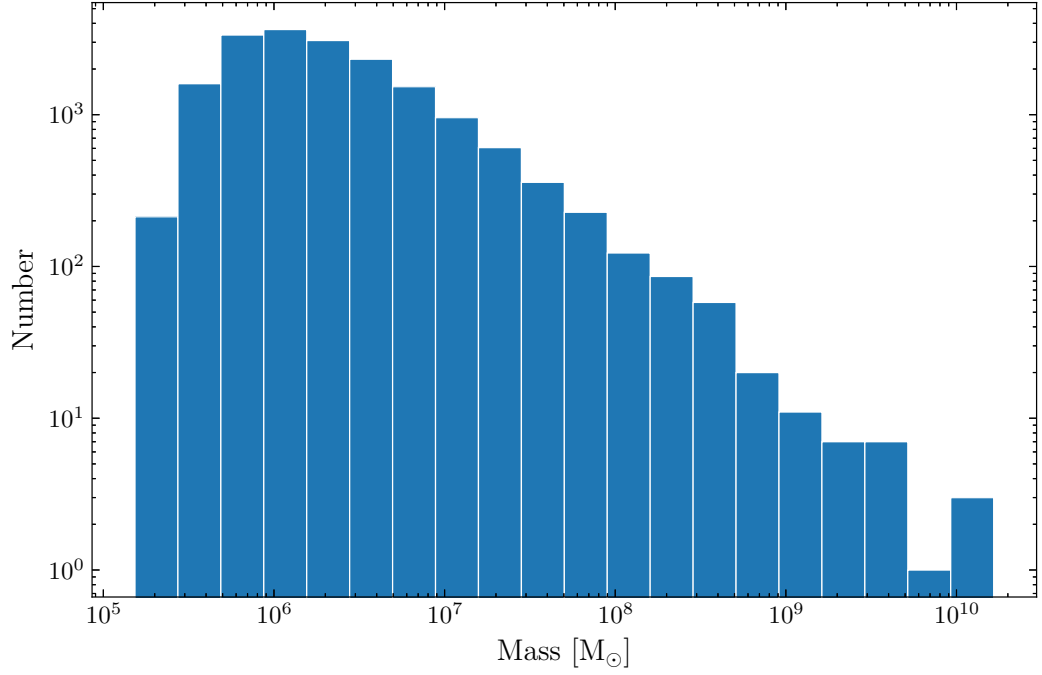
Figure 4.6: The peak mass distribution of type 1 subhaloes within $r_{200}$ at $z = 0$ in the Aq-A-2 simulation.
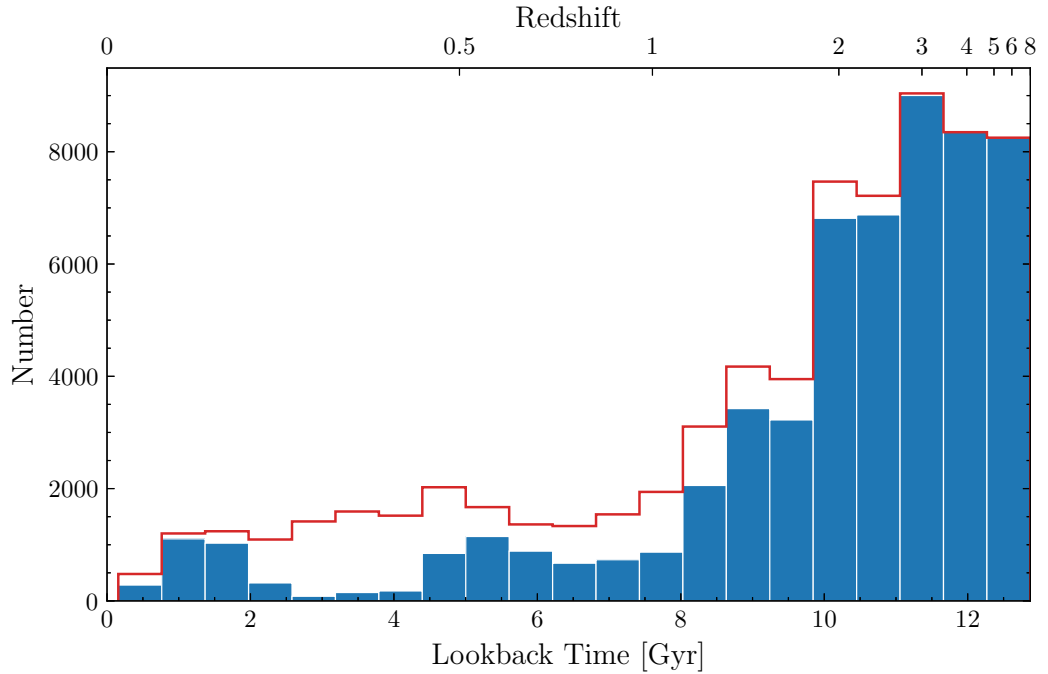


Figure 4.7: The distribution of infall times of type 2 subhaloes. The infall times of subhaloes which are within $r_{200}$ at $z = 0$ are shown in blue, while the red indicates the infall times of all type 2 subhaloes.
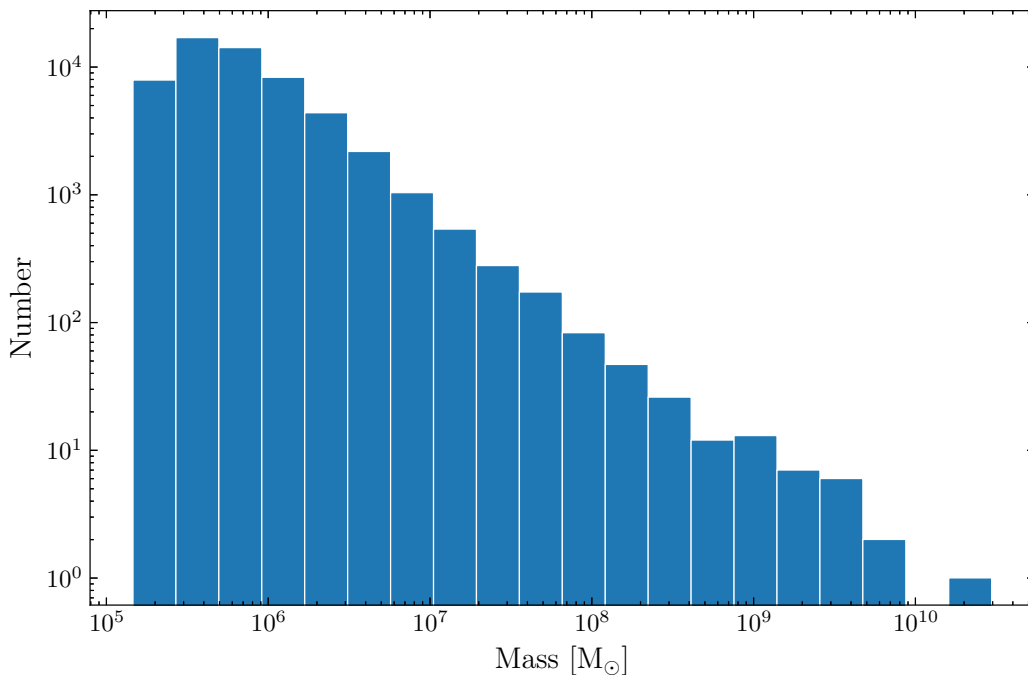
Figure 4.8: The peak mass distribution of type 2 subhaloes within $r_{200}$ at $z = 0$ in the Aq-A-2 simulation.

2 subhaloes are very small, composed of not many more than 20 particles for the duration of their existence.

## 4.3 Numerical setup

To examine the tidal evolution of substructure in the Aq-A-2 simulation, we want to run thousands of simulations, each containing hundreds or thousands of particles. In order to simplify and speed this process up, we construct a pipeline, which is described in the following section.

### 4.3.1 Pipeline

We use the pipeline shown in Figure 4.9. The pipeline is composed of multiple codes in different languages, comprising Python, Fortran and C++. Each stage has a well-defined input and output.

We use a pipeline for the following reasons:

- The pipeline follows the *Batch Sequential* software design pattern, which has the advantages of increased simplicity and ability to modify compared to one big codebase (Avgeriou and Zdun, 2005). In addition, stages can be reused for different purposes if necessary.

- We can easily use different languages for different parts of our problem. Different programming languages are suited to different tasks. As well as this, some parts of our pipeline use third-party programs such as GADGET-4 which we did not want to modify. A pipeline lets us incorporate third-party codes without requiring modification.

- The pipeline offers flexibility in that a single stage can be run or multiple stages can be chained together.

- Usage of a pipeline is especially important on HPC clusters where we have a fixed amount of time in which a job must be completed. Using a pipeline allows us to design our stages so that they can be completed within this time. Each stage was designed so that it would fit within the COSMA7[1] three day batch job time limit. However, if the entire pipeline was run in a single go, it would exceed the time limit.

- Each stage can be tested and verified independently.

The stages of the pipeline are illustrated in Figure 4.9. First, we track objects to their time of infall using the Aquarius merger trees described in Section 2.3. Here, we take time of infall to be when the subhalo crosses $r_{200}$ of the main halo for the first time. Then we select a sample of subhaloes to analyse by taking cuts based on their time of infall and their mass at time of infall. We find the particles belonging to each of these subhaloes at some point before their time of infall and before the subhalo becomes disrupted. The density profiles of these subhaloes

---

[1]COSMA7 is a HPC system that is part of the Durham Memory Intensive DiRAC facility.
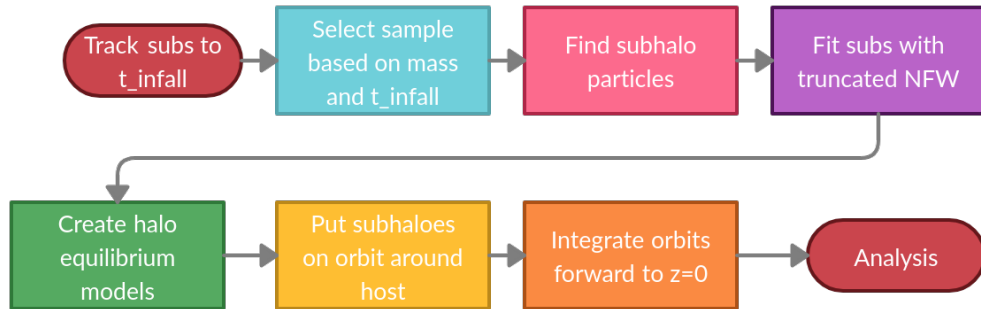
Figure 4.9: The pipeline used to resimulate Aq-A-2 orbits using HEX.

are then calculated and fitted with a truncated NFW profile, using the procedure outlined in Section 2.4. We can then create equilibrium models of these subhaloes using the code introduced by Kazantzidis et al. (2004) and described in Section 2.5. Using this procedure means that we can create lower and higher resolution mass subhaloes in order to examine the effect that the number of particles has on the evolution of substructure. We next convert units and formats and re-centre the halo at the origin for use with HEX. These equilibrium model haloes are then put on an orbit around the host matching the orbit of the original subhalo from Aquarius. Using GADGET-4 with HEX, we can then integrate the orbits forward to $z = 0$, experimenting with different values for the softening length, $\epsilon$. We use GADGET-4's adaptive, variable timestep, rather than using a fixed timestep as in Chapter 3. We also use the relative cell-opening criterion, described in Section 2.2.4. We do this to ensure that our conditions are closer to realistic cosmological simulations. We do, however, use the same softening length calculations that were used in Chapter 3. The results of these resimulations are then analysed.

The pipeline was originally designed using the logically simplest implementation. However, several stages failed to complete within the three day COSMA7 batch job time limit. Therefore, we spent time optimising many of the stages. The optimisations performed on the various stages of the pipeline are as follows:

- Track subhaloes to time of infall: We originally started with a logically simple,

but slow solution, which we optimised to heavily reduce reads. After doing performance tests, we found that almost all the time was spent in disk access, so we optimised this by reducing the number of reads from disk performed. Before optimisation, for each snapshot file we checked, we read in all particle information: particle IDs, particle positions and particle velocities. Reads were optimised by only reading the particle IDs from the file originally and doing a pre-check, if the IDs were found, then the particle positions and velocities corresponding to these IDs were extracted from the data. This resulted in significantly fewer disk reads. We also used Python multiprocessing to reduce the total time to completion. As the problem was easily parallelisable, Python multiprocessing was ideal to utilise all the cores available on the COSMA7 nodes. NumPy was used to vectorize the operations for a further speed-up.

- Select sample: NumPy was used to vectorize the calculations, which resulted in a significant speedup considering the large amounts of data used.

- Find particle data: Similar to the first stage above, we first started with a logically simple, but slow solution, which spent large amounts of time reading data. This was optimised by reading in the data once, instead of multiple times. This resulted in a sufficiently large speed-up.

- Fit subhaloes with NFW profile: The first solution that we came up with was adequate for our needs.

- Create halo equilibrium models: This is third-party code, written in Fortran, which was already sufficiently fast.

- Put subhaloes on orbit around host: The simplest version was already sufficient for our needs. No further optimisation was required.

- Integrate orbits to $z = 0$: GADGET-4 is already heavily optimised for HPC systems.

| # | Pipeline stage | Input (format) | Output (format) |
|---|---|---|---|
| 1 | Track subhaloes to time of infall | Aquarius merger trees (HDF5) | Subhalo infall time information (JSON) |
| 2 | Select sample | Stage 1 output (JSON) | Serialised Python dictionaries and lists of subhalo information (custom ASCII) |
| 3 | Find particle data | Stage 2 output (custom ASCII), Aquarius snapshots (Binary) | Particle data (HDF5) |
| 4 | Fit subhaloes with NFW | Stage 3 output (HDF5) | Halo fitting information (custom ASCII) |
| 5 | Create halo equilibrium models | Stage 4 output (custom ASCII) and **make_halo** parameter file (ASCII) | Equilibrium halo positions and velocities (CSV) |
| 6 | Put subhaloes on orbit | Stage 5 output (CSV) | GADGET initial conditions format (HDF5) |
| 7 | Integrate subhalo orbits to $z = 0$ | Stage 6 output (HDF5) | GADGET snapshots and subfind output (HDF5) |
| 8 | Analysis | Stage 7 output (HDF5) | Various |

Table 4.3: A summary of the inputs and outputs of each stage of the pipeline.

The inputs and outputs at each stage of the pipeline are summarised in Table 4.3.

## 4.3.2 HEX

In order to resimulate the orbits of the Aquarius subhaloes using different $N_p$ and $\epsilon$, we used HEX (see Section 2.6). HEX had previously been implemented in GADGET-2. For this thesis, we ported the code to work with GADGET-4, which is written in a more modern C++ style. This necessitated some changes to the code. Some work also had to be done to make the code work with GADGET-4's new integer coordinate system.

We experimented with using different numbers of coefficients to represent the main halo potential. Higher order terms resolve smaller-scale structure, however there is also a trade-off to be made with the time taken to run the simulations. Lowing et al. (2011) found that in Aq-A-2 a force inaccuracy of less than 1% could be achieved using $n_{max} = l_{max} = 20$ and was sufficient for most purposes. Following

this, Kelly (2022) demonstrated that there was little improvement by exceeding $n = 22$ and $l = 10$. We looked at $n_{max} = l_{max} = 20$, $n_{max} = l_{max} = 30$ and $n_{max} = l_{max} = 45$. Figure 4.10 shows some of the results. The eight examples shown are a representative set of different orbital parameters. The blue lines represent the original orbit from Aq-A-2. The orange, red and purple lines are HEX resimulations, where $n_{max} = l_{max} = 20$, $n_{max} = l_{max} = 30$, $n_{max} = l_{max} = 45$, respectively. For the most part, there is very little difference for the majority of subhaloes tested between the three different potentials. We therefore chose to use $n_{max} = l_{max} = 30$ for the remainder of the simulations in this chapter, as this seemed to be a good balance between accuracy and simulation time. In most cases, this has been sufficient. However, close encounters between objects are not always properly resolved.

## 4.4 Results

### 4.4.1 Type 1 Subhaloes

We first examined the larger type 1 subhaloes from the Aq-A-2 simulation. We looked at all subhaloes which infall after $z = 1$ with infall masses $> 10^8$ M$_\odot$ and which lie within $r_{200}$ at $z = 0$. There are 45 such subhaloes and Figure 4.11 shows the distribution of their infall times. One of these subhaloes was excluded from our sample as it gained quite a lot of mass from $z = 1$ to $z = 0$ and is therefore not representative. Four subhaloes were also excluded as we could not get a good fit to an NFW profile pre-infall. This left us with 40 subhaloes in our sample, which we resimulate from $z = 1.11$ to $z = 0$.

The orbits and radial profiles of seven of these 40 subhaloes are shown in the top half of Figures 4.12 - 4.18. For each of these subhaloes, the orbit from the highest resolution HEX resimulation is shown, with $N_p = 100,000$ and $\epsilon = 0.003$. These examples show some of the more interesting orbits, where the subhalo came close
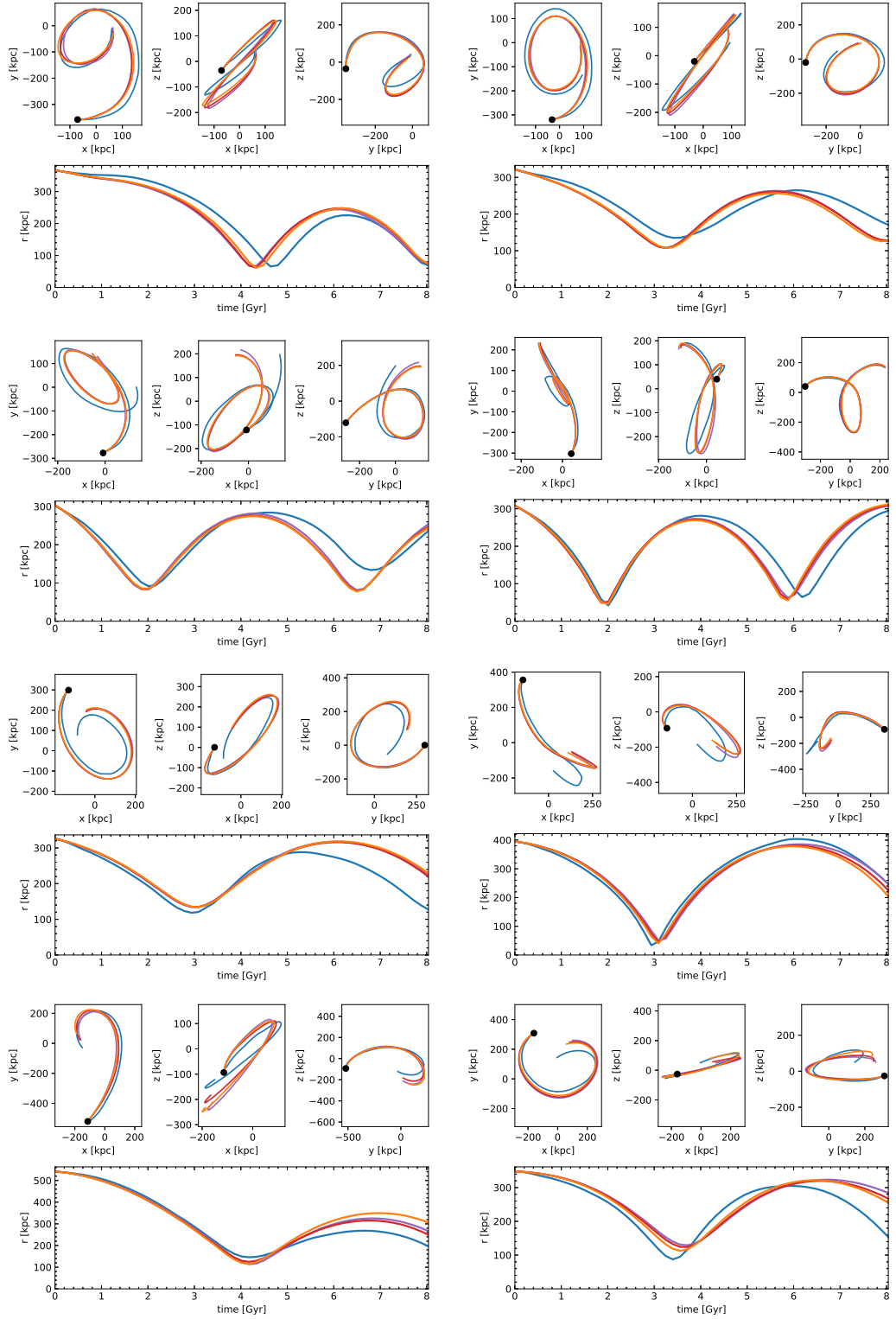
Figure 4.10: The orbits and radial distances for a selection of eight subhaloes using a different number of coefficients in HEX. The blue lines represent the original orbit from Aq-A-2. The orange, red and purple lines are HEX resimulations, where $n_{max} = l_{max} = 20$, $n_{max} = l_{max} = 30$, $n_{max} = l_{max} = 45$, respectively. The filled, black circle in the orbit plots shows the start of the simulation.
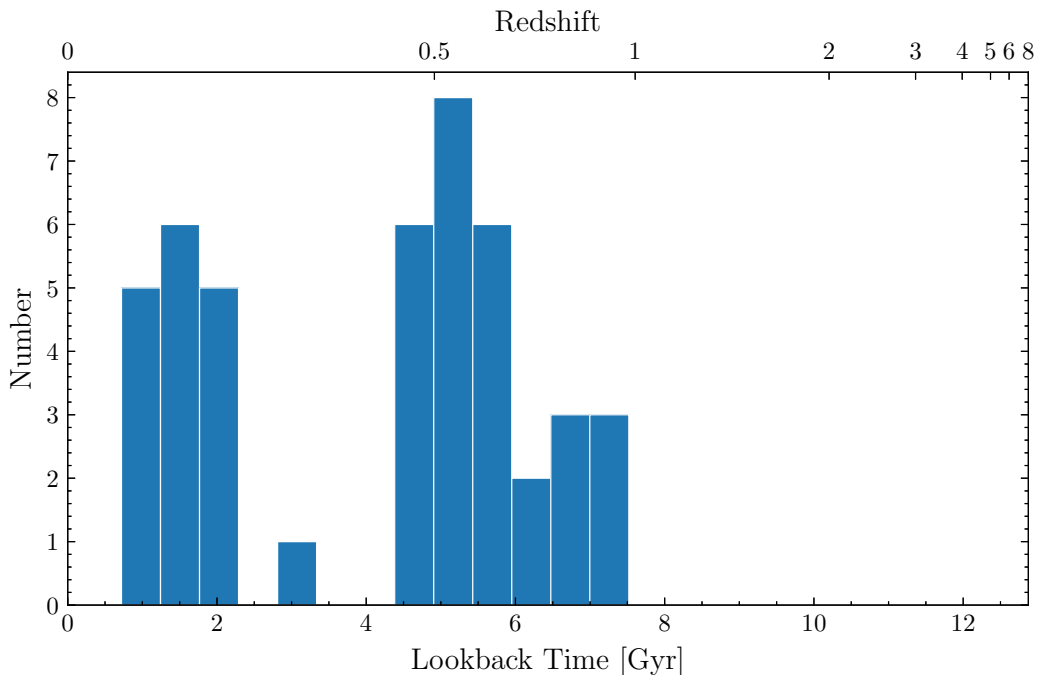
Figure 4.11: Distribution of infall times for type 1 subhaloes with mass $> 10^8$ M$_\odot$, which infall after $z = 1$.

to the centre of the main halo and lost a significant amount of mass. Most of the orbits are reproduced reasonably well by HEX. The orbit depicted in Figure 4.18, in particular, is very close to the original Aquarius orbit. HEX seems to have a slight problem capturing the apocentres, but does a better job reproducing the pericentres. Sometimes an object will encounter another object, which leads to a wiggle in its orbit as seen in the blue line in Figure 4.12. HEX does not capture these orbits very well, but increasing the number of coefficients used to represent the potential may lead to improvements.

For these 40 subhaloes, we conducted a resolution study, similar to that in Chapter 3. A selection of the results, from the six subhaloes are shown in the bottom half of Figures 4.12 - 4.18. Six different softening lengths were used, ranging from $0.3 - 0.003$ times the initial subhalo scale radius, and seven different mass resolutions were used, from $N_p = 100$ to $N_p = 1,000,000$, giving a total of 42 different simulations. For each combination of $\epsilon$ and $N_p$, five different random seeds were

used to generate the subhaloes. In total, over 200 simulations were run for each plot.

The subhaloes all infall into the main halo after $z = 1$. We generate the initial conditions and run the simulations from $z = 1.11$, in order to capture the subhaloes before they became disrupted due to the gravitational forces close to the centre of the main halo. Some subhaloes gained mass between $z = 1.11$ and their time of infall in the original Aquarius simulation. This can be seen in Figure 4.14, where the black line rises above 0 just before it crosses $r_{200}$ of the main halo. Such an increase in mass cannot really occur in our setup, as the subhalo is isolated and there is no chance for encounters with other objects. Despite this, the blue lines of Figure 4.14 closely match the black line of the original Aquarius simulation, especially in the higher-resolution runs.

As these subhaloes all have masses above $10^8$ M$_\odot$ and the mass resolution of the Aquarius A-2 simulation is $1.37 \times 10^4$ M$_\odot$, this means that each of the 40 subhaloes have at minimum 7,000 particles, with some having significantly more. Therefore, it is not too surprising that for most of these subhaloes, the mass lost by the subhaloes in the HEX simulations when $N_p$ is high closely matches the mass lost in the original simulations. For the most part, the bound fraction remaining at the end of the simulation is very similar to that found in Aquarius, so long as $N_p \geq 1,000$. For $N_p \leq 300$, subhaloes start to become completely disrupted before the end of the simulation, often straight after their first pericentre, and the effects of discreteness noise become more and more obvious. Even the subhalo represented in Figure 4.12, which does not come particularly close to the centre of the main halo and only loses about 40% of its mass by $z = 0$, is almost always disrupted when $N_p = 100$.

The results of all of these simulations are summarised in Figure 4.19. This shows the results of 8,400 simulations. Each of the 40 subhaloes is represented by a coloured dot in each of the boxes. This dot is the mean final value of $f_{bound}$ from HEX (for a given $N_p$ and $\epsilon$) vs the final value of $f_{bound}$ from the original Aquarius
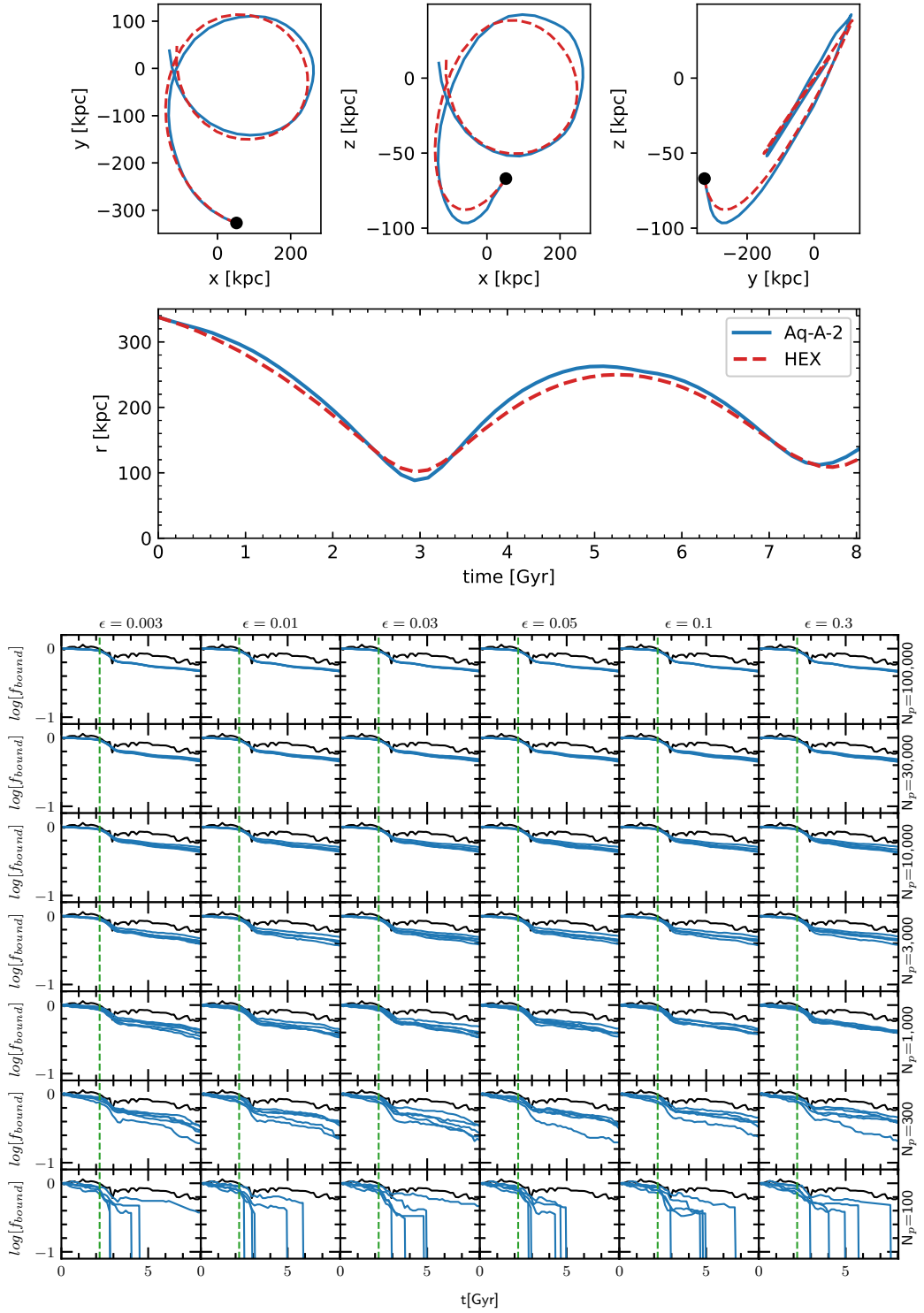
Figure 4.12: Above: The orbit and radial distance of a subhalo. The blue line represents the original orbit from Aq-A-2, while the red dashed line shows the resimulated HEX orbit. The mass of the subhalo prior to accretion was $3.65 \times 10^8$ M$_\odot$. Its first pericentre is at 88.37 kpc and first apocentre is at 262.79 kpc. Below: The bound fraction vs time for simulations with different $N_p$ and $\epsilon$. The black line represents the original bound fraction from the Aq-A-2 simulation. The blue lines represent different random realisations of the same subhalo. The green dashed line shows the time of infall.
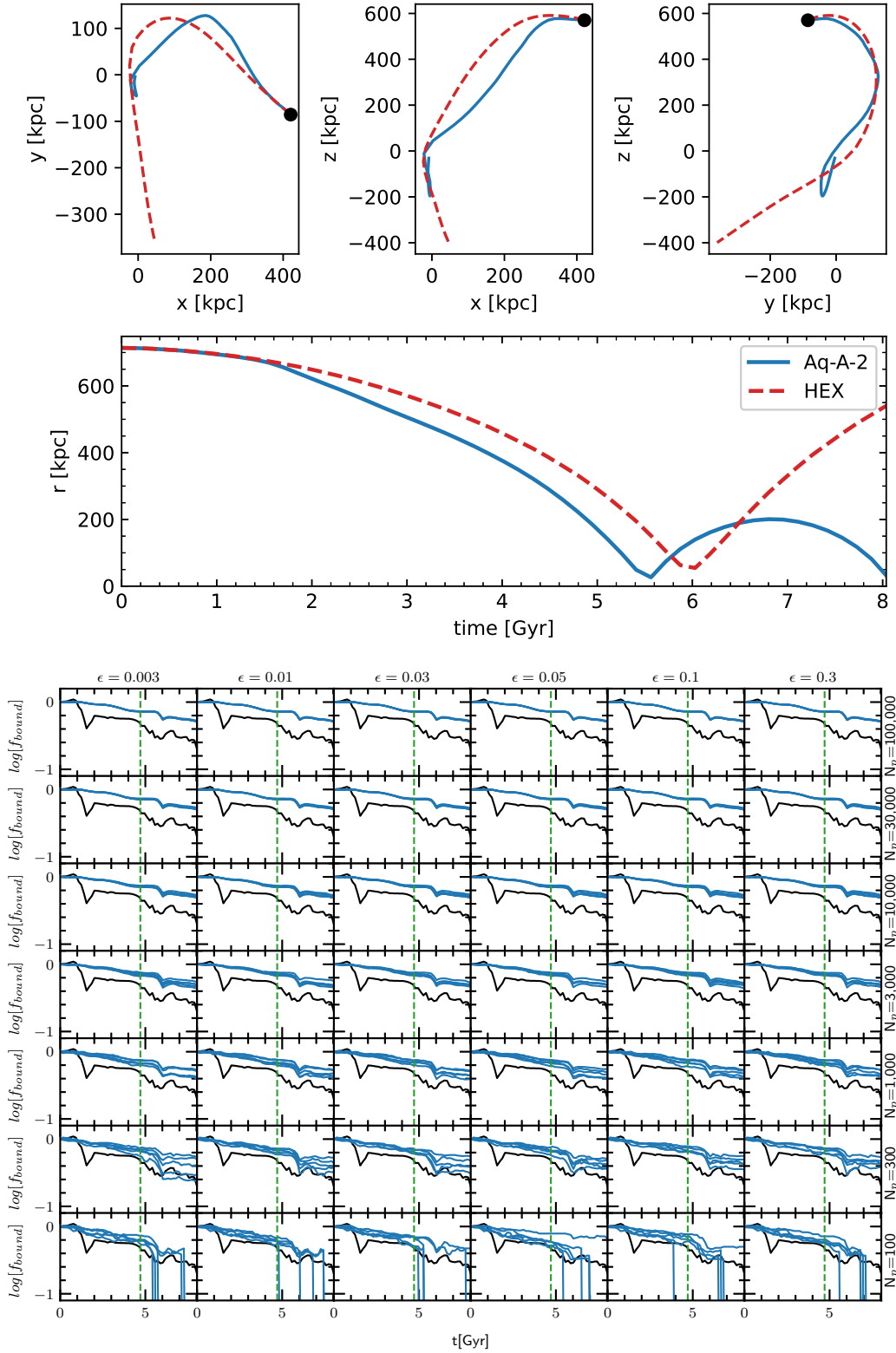
Figure 4.13: The same as Figure 4.12, but for a different subhalo. The mass of the subhalo prior to accretion was $7.67 \times 10^8$ $M_\odot$. Its first pericentre is at 26.63 kpc and first apocentre is at 200.84 kpc.
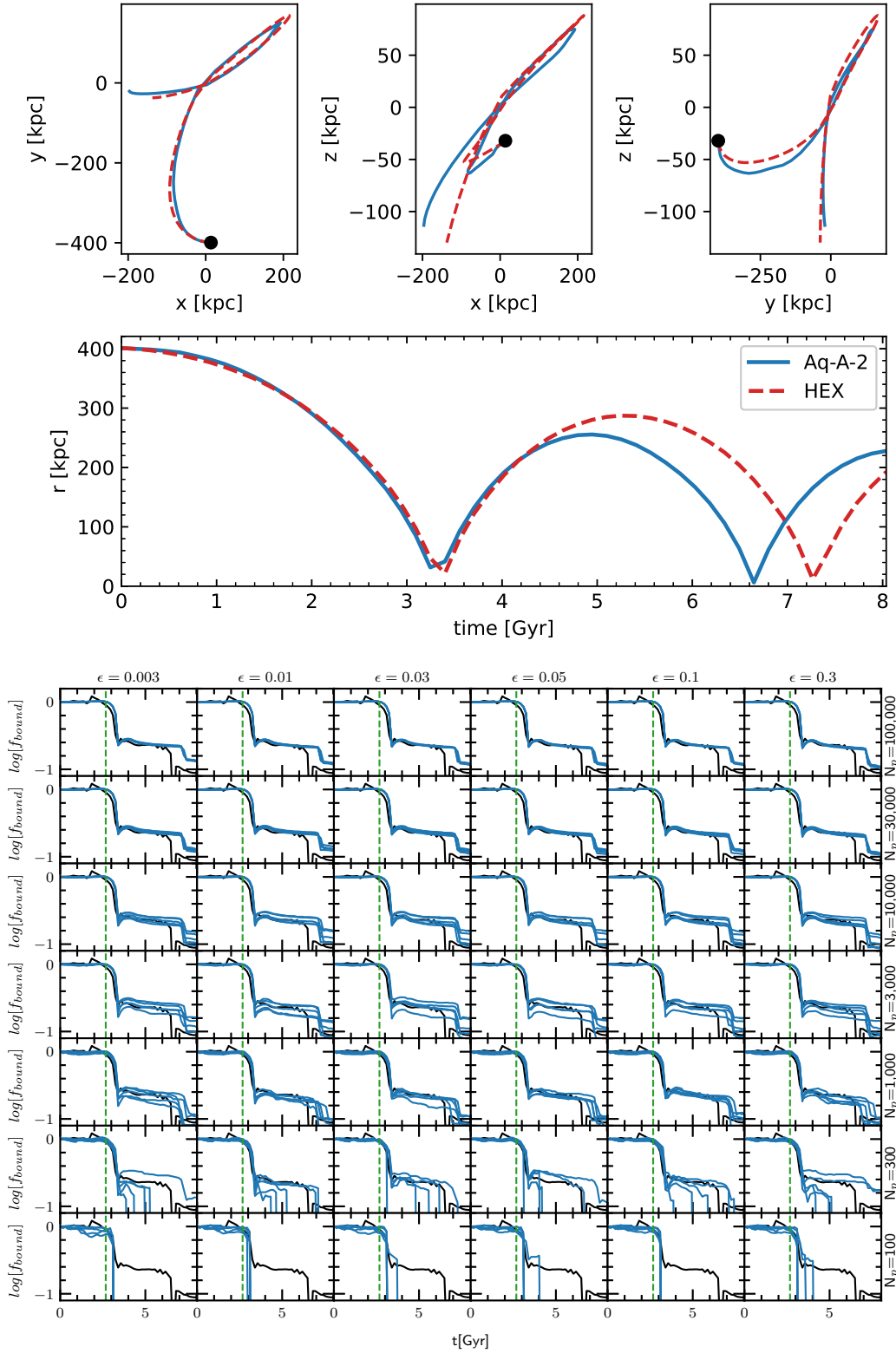
Figure 4.14: The same as Figure 4.12, but for a different subhalo. The mass of the subhalo prior to accretion was $1.61 \times 10^9$ M$_\odot$. Its first pericentre is at 31.73 kpc and first apocentre is at 255.58 kpc.
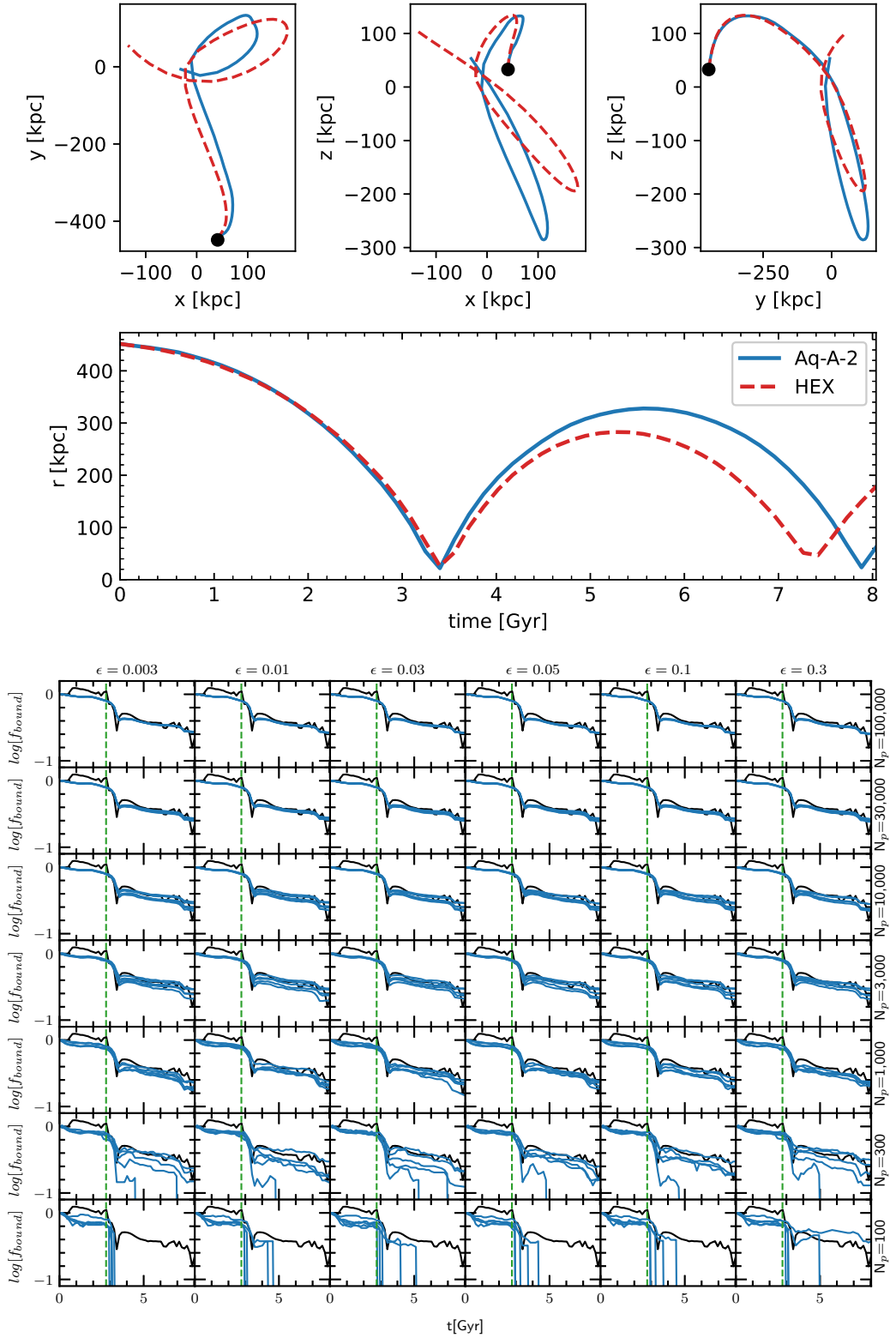
Figure 4.15: The same as Figure 4.12, but for a different subhalo. The mass of the subhalo prior to accretion was $2.35 \times 10^8$ M$_\odot$. Its first pericentre is at 21.95 kpc and first apocentre is at 327.97 kpc.
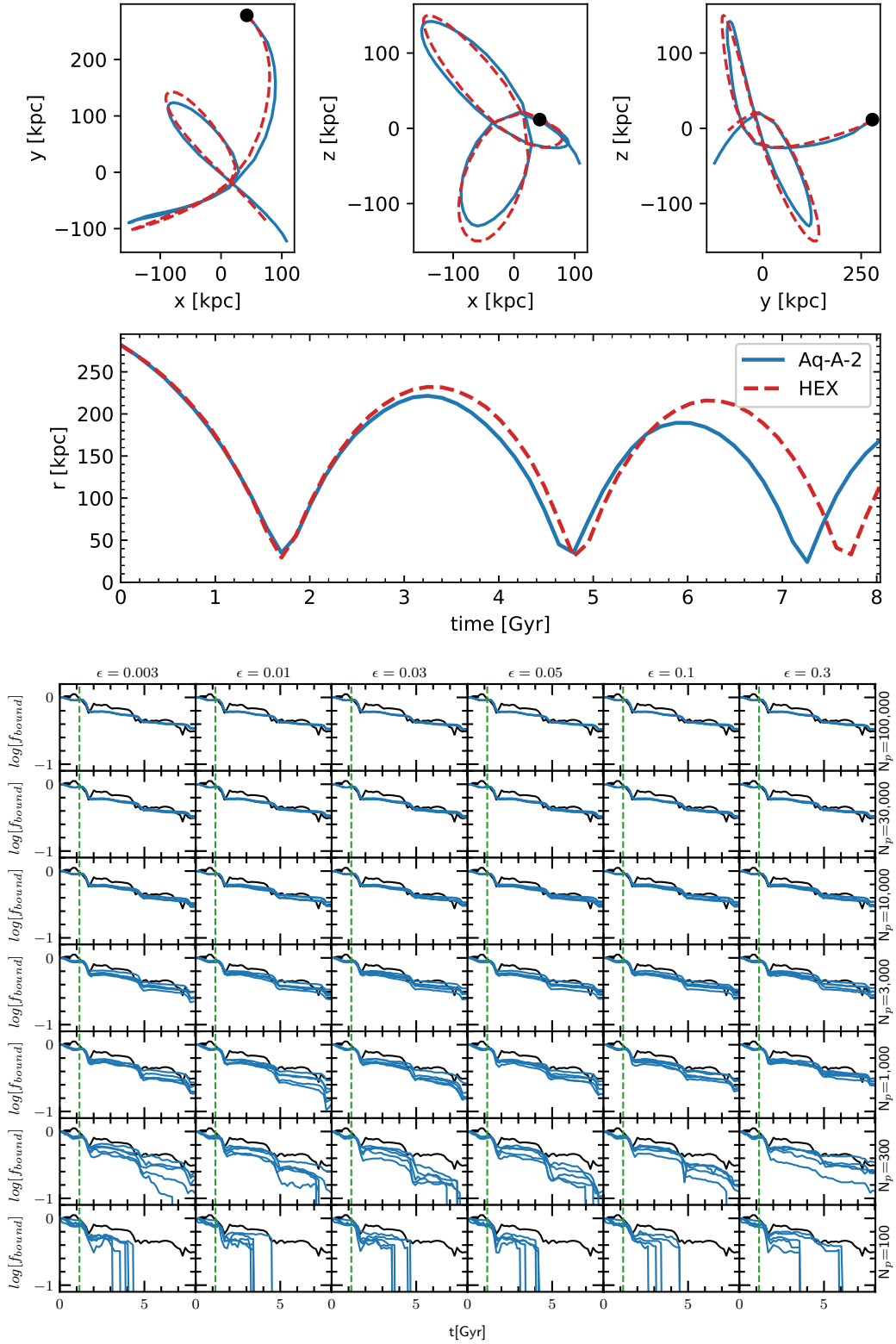
Figure 4.16: The same as Figure 4.12, but for a different subhalo. The mass of the subhalo prior to accretion was $1.29 \times 10^9$ M$_\odot$. Its first pericentre is at 34.91 kpc and first apocentre is at 221.54 kpc.
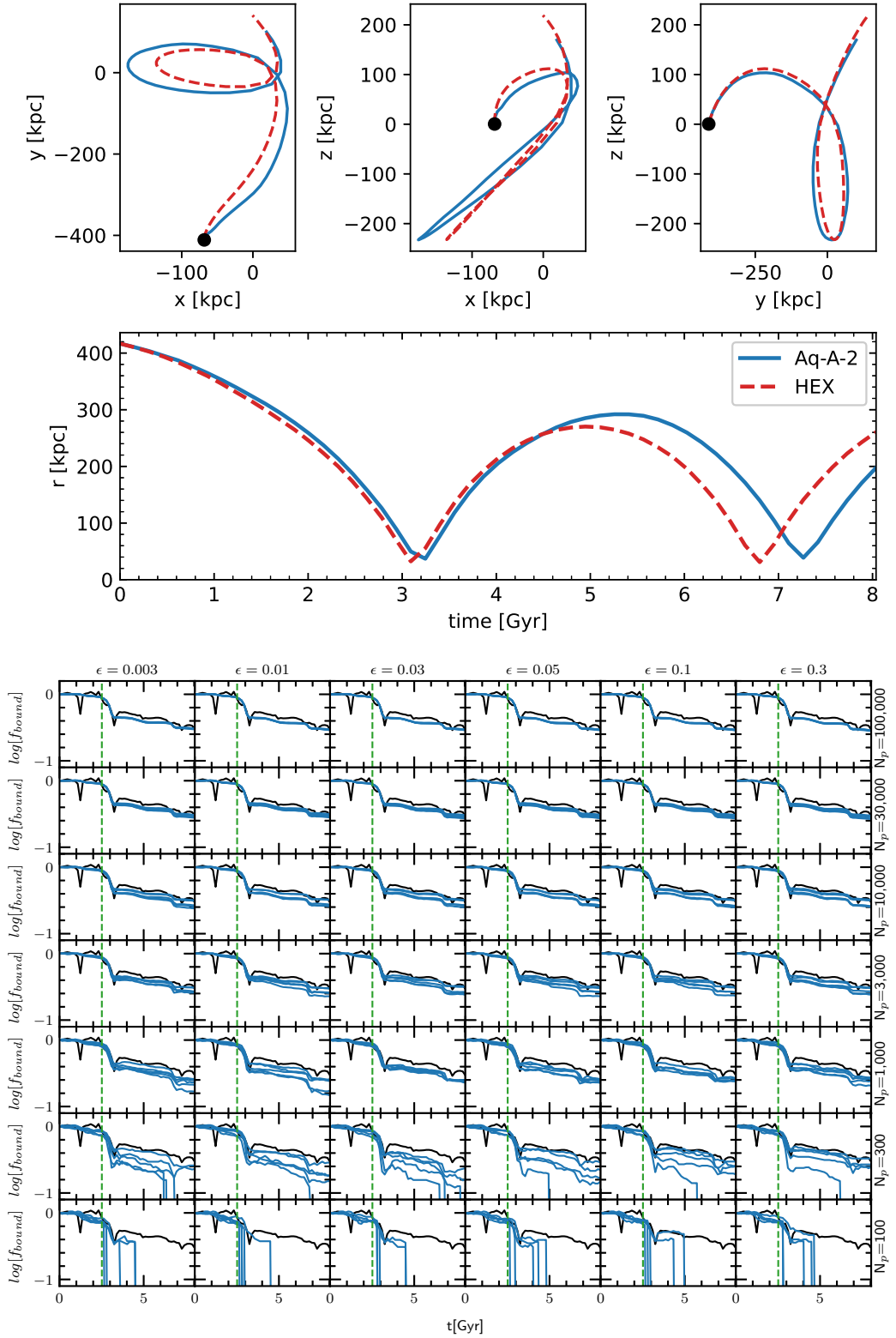
Figure 4.17: The same as Figure 4.12, but for a different subhalo. The mass of the subhalo prior to accretion was $6.66 \times 10^8$ M$_\odot$. Its first pericentre is at 37.05 kpc and first apocentre is at 291.97 kpc.
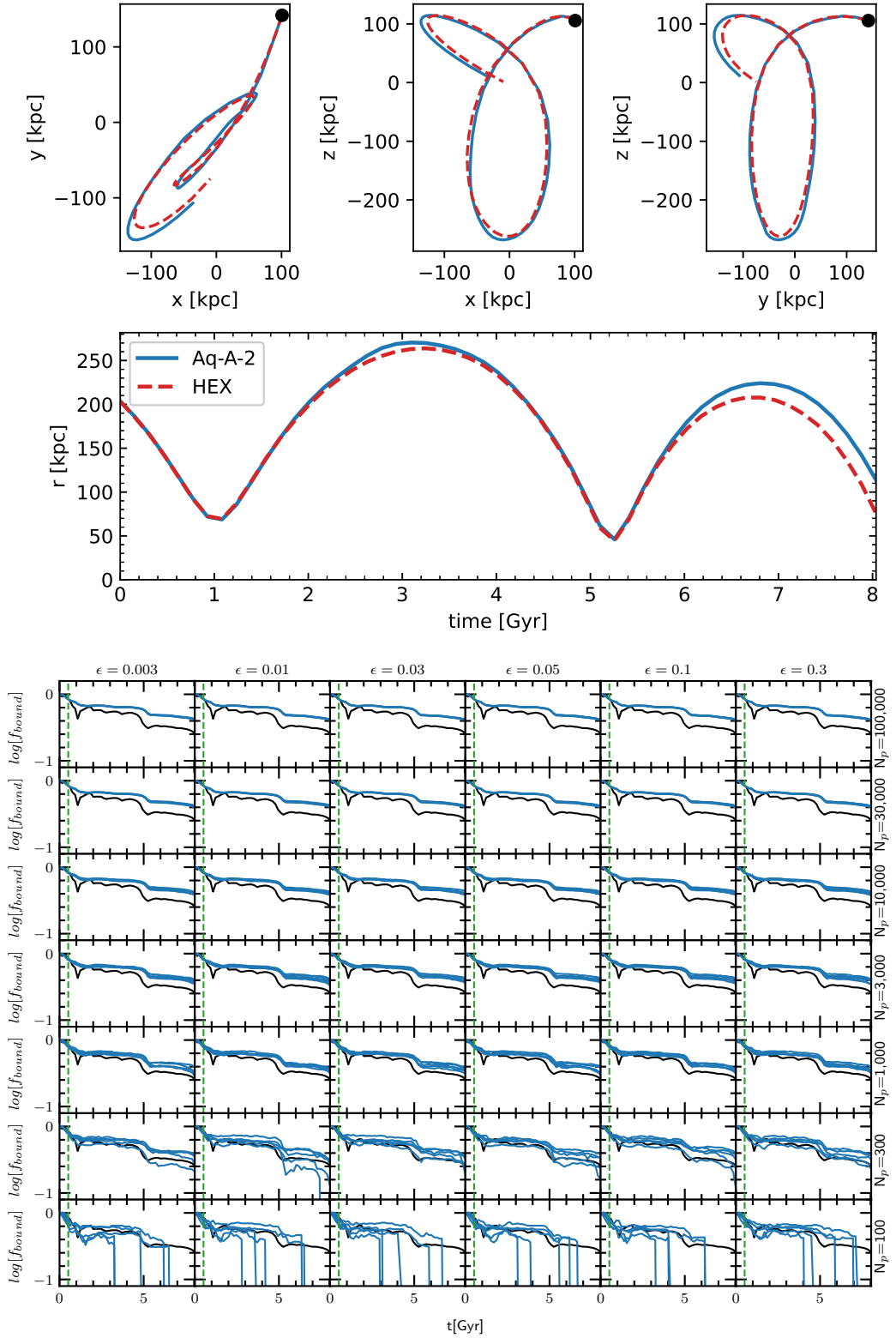
Figure 4.18: The same as Figure 4.12, but for a different subhalo. The mass of the subhalo prior to accretion was $7.64 \times 10^9$ M$_\odot$. Its first pericentre is at 68.71 kpc and first apocentre is at 270.51 kpc.
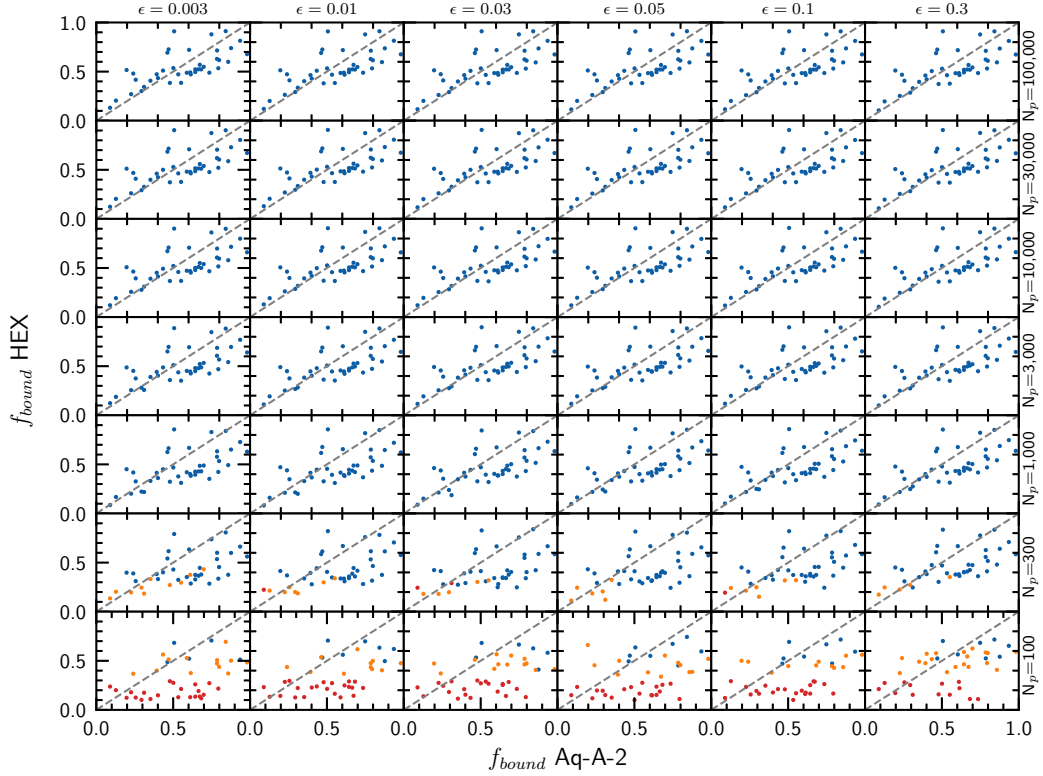
Figure 4.19:   The ratio between the final subhalo bound fraction from the HEX resimulations, with different $N_p$ and $\epsilon$, and the final bound fraction from the original Aq-A-2 simulation for 40 different subhaloes. The value for HEX $f_{bound}$ is the mean of five different simulations with different random realisations. Blue dots indicate that the subhalo survived until $z = 0$ for all five HEX simulations. Orange dots represent subhaloes that survived for at least one of the five simulations, while red dots indicate that the subhalo was disrupted in all five simulations. The position of the red dots is therefore not indicative of the final mass of the subhalo - they are placed along the $x$-axis according to their bound fraction in Aq-A-2, but their position along the $y$-axis is random.

simulation. If in all five simulations with a different random seed, the subhalo at a particular $N_p$ and $\epsilon$ survives until $z = 0$, the dot is blue in colour. On the other hand, if the subhalo is disrupted in all five simulations, it is represented by a red dot. An orange dot indicates that the subhalo survived until $z = 0$ for at least one of the five simulations. The red dots are placed on the $x$-axis according to their $f_{bound}$ value in the Aquarius simulation, though their position on the $y$-axis is random. For the orange dots, the mean of the bound fraction of all the surviving subhaloes is taken.
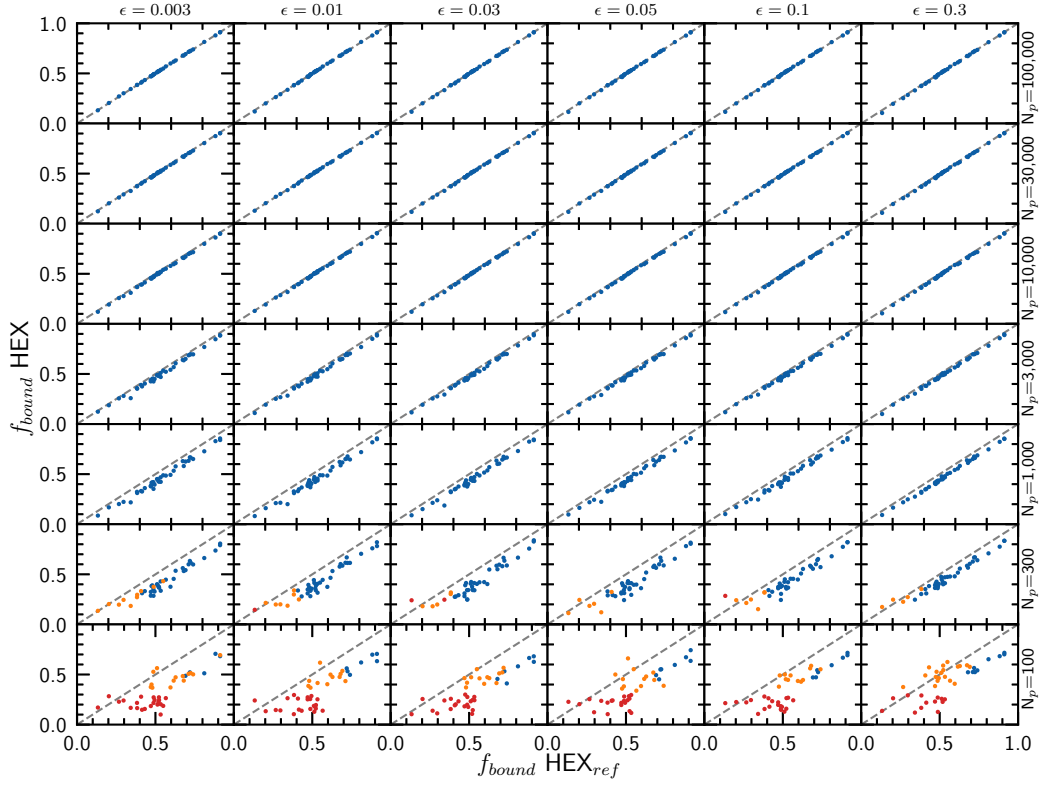
Figure 4.20: The ratio between the final subhalo bound fraction from the HEX resimulations, with different $N_p$ and $\epsilon$, and the final bound fraction from the highest resolution HEX resimulation, $\text{HEX}_{ref}$, (i.e. the top-left box) for 40 different subhaloes. The value for HEX $f_{bound}$ is the mean of five different simulations with different random realisations. Blue dots indicate that the subhalo survived until $z = 0$ for all five HEX simulations. Orange dots represent subhaloes that survived for at least one of the five simulations, while red dots indicate that the subhalo was disrupted in all five simulations. The position of the red dots is therefore not indicative of the final mass of the subhalo - they are placed along the $x$-axis according to their bound fraction in $\text{HEX}_{ref}$, but their position along the $y$-axis is random.

While the HEX method is very good, it is still not perfect at reproducing the orbits of the original subhaloes. As we have seen, this results in small differences between the original simulation and the reconstructed orbits. Much of the scatter in this plot, especially at the higher resolutions, seems to be due to these small differences. We have therefore generated a similar plot, Figure 4.20, where the highest resolution HEX simulation is used as the reference, rather than the Aq-A-2 simulation. As in our six samples above, we see that for simulations with $N_p \geq 1,000$, the dots are all blue and are close to the identity line with some scatter when $N_p = 1,000$. When $N_p = 300$, some subhaloes start to disrupt before the end of the simulation and the scatter of the blue dots also increases. A significant number of subhaloes disrupt entirely when $N_p = 100$ and only a few blue dots remain in the final row.

Perhaps surprisingly, while there is a strong dependence on $N_p$ for the survival of the subhaloes, there seems to be little dependence on softening length, $\epsilon$. In fact, the box with the smallest softening length contains more subhaloes that were completely disrupted than the one with the longest length. This may be due to artificial scattering caused by a too small softening length. This is interesting as inadequate force softening was what van den Bosch and Ogiya (2018) mainly attributed to overmerging in modern cosmological simulations. While the effects of different softening lengths were evident in our results from Chapter 3, in these simulations with more realistic conditions, i.e. realistic orbits and a time-varying potential, changing the softening length used, even by as much as a factor of 100, seems to make little difference to the results.

## 4.4.2 Type 2 Subhaloes

While we did not have time to perform a proper analysis of type 2 subhaloes, we did run a full set of simulations for a couple of samples. Although there are many more type 2 subhaloes at $z = 0$ of Aq-A-2, the majority of these are very close to the particle limit of 20 at their time of infall. Subhaloes which start out massive but later become type 2s are much rarer. We are therefore more limited in our

sample selection and could not find any type 2s which have large peak masses ($>$ $10^7$ M$_\odot$) and infall after $z = 1$. The subhaloes shown here were randomly selected from a sample of type 2 subhaloes with infall times before $z = 1$, but which still had masses of at least $10^7$ M$_\odot$ at $z = 1$. The results are shown in Figure 4.21.

The subhalo depicted in Figure 4.21 has $\sim 1,802$ particles at $z = 1.11$ when we begin our HEX resimulations. It first infalls to the main halo at $z = 2.16$. We can see on the radial distance plot the point at which the subhalo becomes disrupted and the blue line stops. This point is depicted as a vertical, dashed red line on the grid plot below. We can see that for higher $N_p$, although the subhalo loses over 90% of its mass, a bound remnant does remain by the end of the simulation. When $N_p \leq 1,000$, the subhalo is destroyed before $z = 0$. The results are similar for the subhalo depicted in Figure 4.22, which has 1,394 particles at $z = 1.11$ and first infalls into the main halo at $z = 1.97$.

Both of these examples have a much shorter orbital period compared to our type 1 examples. They therefore have more pericentric passages and come very close to the centre of the main halo, which contributes to the greater tidal disruption that they experience compared to the type 1 subhaloes. Further work is required, however, to say anything about the entire population of type 2 subhaloes. There are approximately 30 type 2 subhaloes with infall times before $z = 1$, but which still had masses of at least $10^7$ M$_\odot$ at $z = 1$. We will run simulations of the remaining subhaloes in this sample to see if they too survive to $z = 0$ once the mass resolution is sufficiently high. This could help give us an insight into the distribution of dark matter substructure within the Milky Way and help us to put constraints on the nature of dark matter.

## 4.5 Conclusions

In this chapter, we discussed substructure in the Aquarius A-2 simulation. We have examined the tidal evolution of subhaloes using realistic conditions. We have
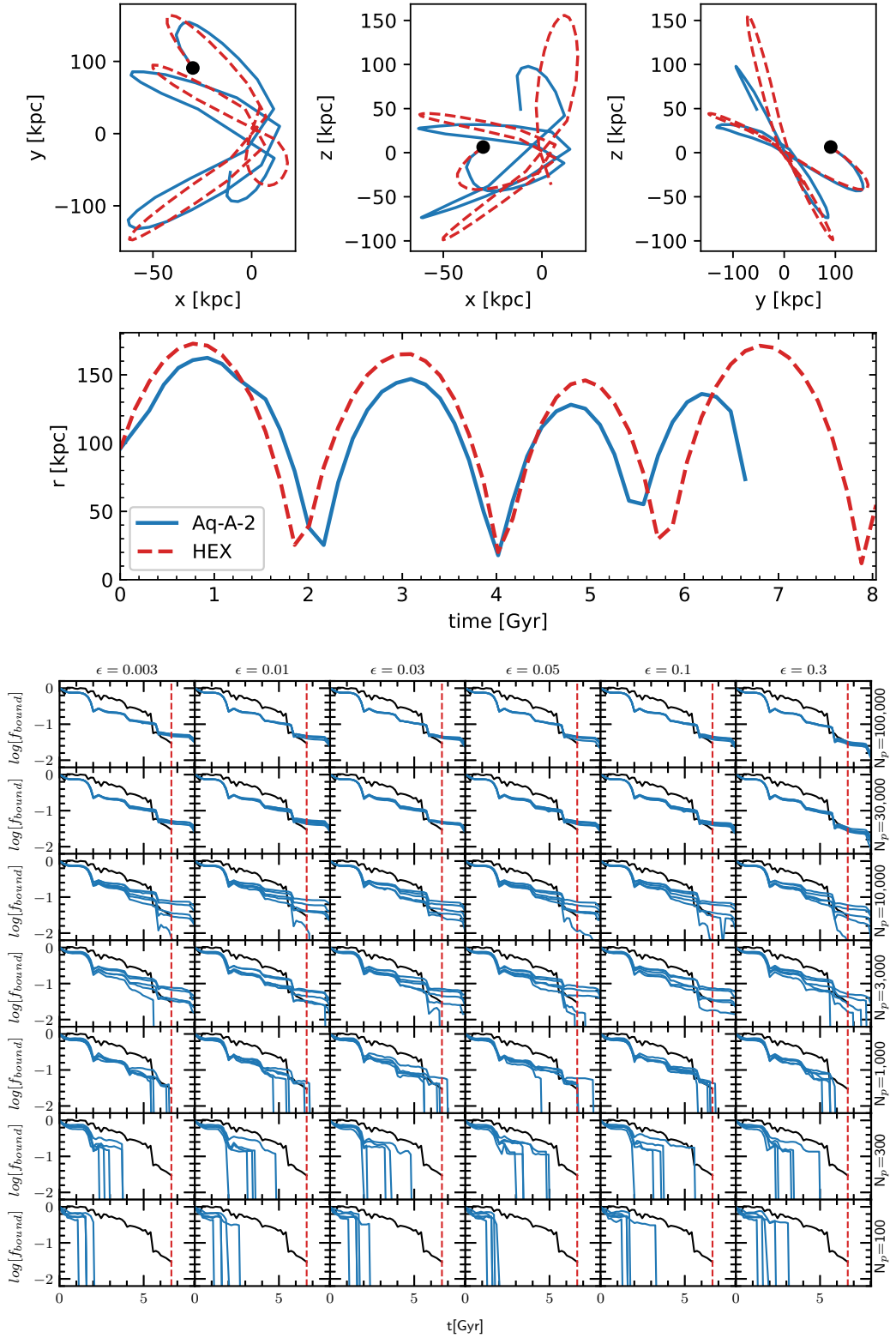
Figure 4.21: The same as Figure 4.12, but for a type 2 subhalo. The red dashed vertical line in the grid plot represents the time that the subhalo becomes a type 2.
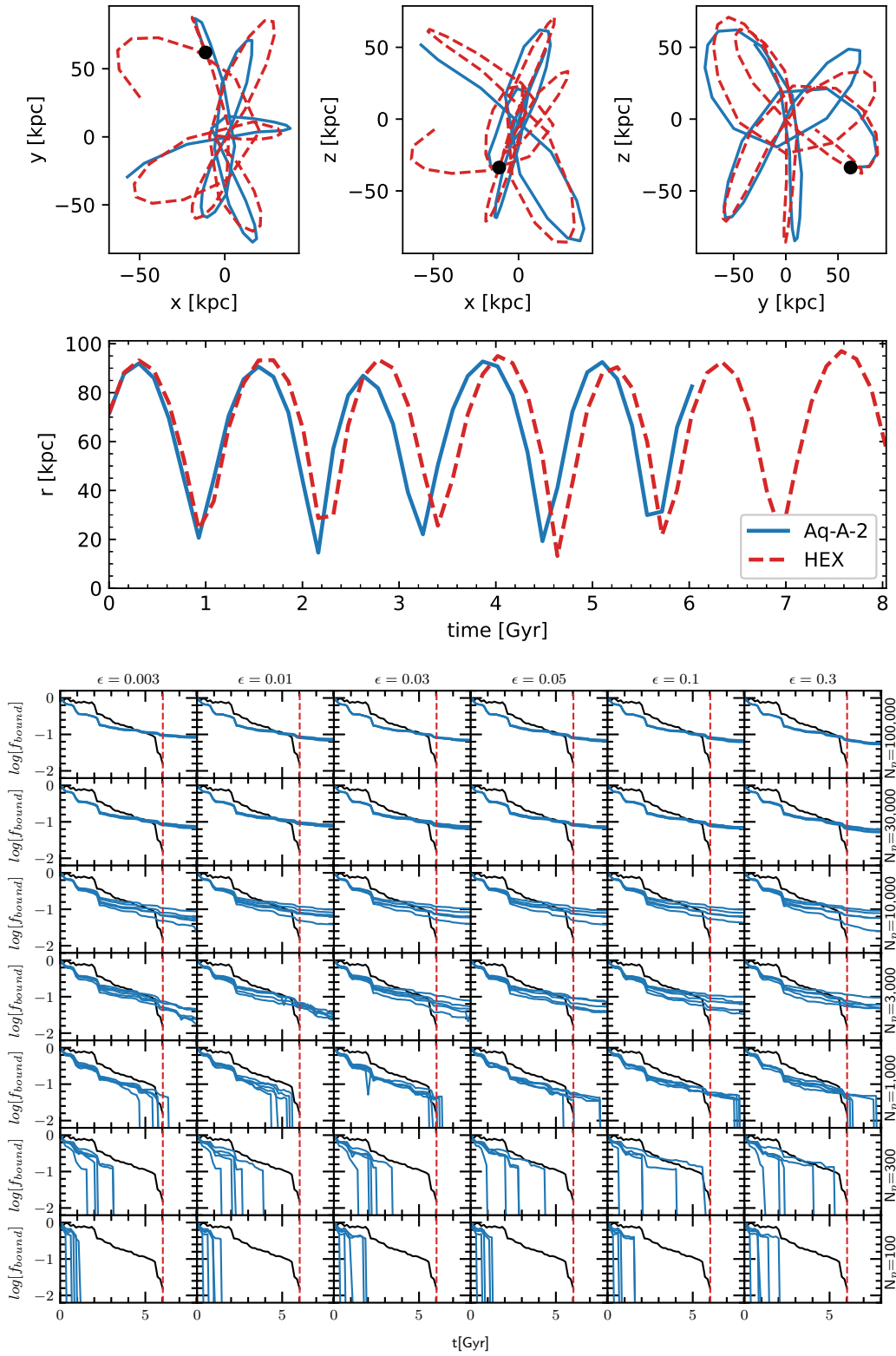
Figure 4.22: The same as Figure 4.21, but for another type 2 subhalo.

used the HEX code to model the potential of our main halo and we have selected a sample of subhaloes and orbits from the Aquarius A-2 simulation to perform a convergence study on. We have constructed a pipeline to sample the Aquarius merger trees, extract subhalo information and run resimulations using GADGET-4 and HEX, in order to examine the effect that softening length and number of particles has on the tidal evolution of subhaloes.

We have found that for type 1 subhaloes with infall masses $> 10^8$ M$_\odot$, HEX replicates the original orbits well. While we do see very slight divergence from masses once $N_p < 30,000$, in general the bound fraction is close for different random realisations of the initial conditions. Subhaloes only start to fully disrupt once $N_p \leq 300$ and the effects of discreteness noise also start to become apparent. Even at this stage, for all of our 40 examples at least one of the five random realisations survived until the end of the simulation. It is only once we get to $N_p = 100$ that the majority of subhaloes are destroyed. We have shown that over 8 Gyr, subhaloes which are placed on realistic orbits in a realistic potential are quite resilient.

As well as the number of particles, we have also examined the effect that softening length has on the tidal evolution of subhaloes. We have looked at softening length over two orders of magnitude, from $0.003 - 0.3$ times the initial scale radius of the subhalo. We find that softening seems to have little effect on the survival of subhaloes which are placed on these realistic orbits, contrary to the findings of van den Bosch and Ogiya (2018).

We have also examined some examples of type 2 subhaloes, i.e. subhaloes which did not survive to the end of the original simulation. Our samples have more pericentric passages than do our type 1s and tend to spend longer closer to the centre of the main halo and are thus subject to stronger tidal forces. We find that when $N_p$ is increased to a sufficient value, these subhaloes do survive for the duration of the simulation. However, type 2 subhaloes which have large masses before infall are quite rare and so are not representative of the entire population. We therefore conclude that the resolution recommendations suggested by van den Bosch and

Ogiya (2018) may be excessive for realistic simulations, such as Aquarius, where most objects do not spend the majority of their time at or close to pericentre.

## 4.5.1 Further Work

In this chapter, we found that large type 1 subhaloes modelled using HEX closely follow their tidal evolution in the Aquarius simulation. We also looked at a couple of examples of type 2 subhaloes. However, in order to produce better statistics, we need to examine a representative sample of the entire population of subhaloes in Aq-A-2. In particular, we want to examine type 2 subhaloes which are located in the central regions of the halo and are subject to stronger tidal forces.

As the Aquarius A halo is not completely typical and has a very high concentration for its mass, we plan to follow a similar analysis and model substructure from the other relaxed Aquarius haloes, Aq-B - Aq-E. Ideally, we would model all substructure in Aquarius and examine the properties of this substructure as a function of radius to determine whether a higher resolution would change anything.

Currently, our HEX coefficients only extend back to $z = 1.11$, meaning that we can only resimulate objects from this point onwards. Ideally, we would like to extend these coefficients back in time as far as possible. (agu) use $z = 2$ as the time of accretion for the majority of their subhaloes and we would like to compare these results to those of HEX for similar time of infall.

Finally, it would be interesting to investigate the effect of baryons on the tidal evolution of dark matter substructure using HEX. Kelly (2022) used the basis function expansion technique on the full magnetohydrodynamic simulation of the Au16 halo from the AURIGA Simulation (Grand et al., 2017). They found it necessary to increase the order of expansion compared to that used to model the dark matter only halo due to an increased central density. However, as this still fails to capture the thin disc accurately, they propose a technique to model the disc, by modelling the disc analytically and removing it from the basis function expansion. A good

candidate halo to test this on might be the halo described in Grand et al. (2021). This is a very highly-resolved resimulation of one of the Milky Way-sized haloes from the AURIGA Simulation, with a baryonic mass resolution of 800 $M_\odot$ and a dark matter particle resolution of $6 \times 10^3$ $M_\odot$, close to that of Aquarius.

# CDT Projects

## 5.1 Introduction

As my PhD was part of the Centre for Doctoral Training in Data Intensive Science, I had the opportunity to undertake a three-month placement in industry with Jumping Rivers[1], an analytics company based in Newcastle-upon-Tyne.

During my time with Jumping Rivers I was involved with the Diffify[2] project, which provides information on additions and updates to CRAN (the repository for R packages)[3]. At the time of my placement, I was the sole contributor to this project. All of the code shown below was written by myself.

I also spent time during my PhD working on the JUNE project. This was a multidisciplinary project to develop a model for the simulation of epidemics. Work was contributed by a large group of fellow CDT students and academics, as well as clinicians from the NHS. The project paper is included in Appendix A. A brief description of my contribution is included in Section 5.3.

---

[1] https://www.jumpingrivers.com/
[2] previously called Cowberries
[3] https://diffify.com/

## 5.2 Diffify

### 5.2.1 Introduction

R is a programming language used primarily for statistical computing and graphics[4]. The Comprehensive R Archive Network (CRAN) is a network of web servers which stores up-to-date versions of and documentation for R packages[5]. Most R packages can be found here. CRANberries (Eddelbuettel, 2007) is a website which provides an RSS feed of all updates to CRAN [6]. It provides a list of all new packages with metadata. It also provides a list of updates to current packages. There are approximately 20,000 R packages on CRAN and dozens of these packages are updated each day. CRANberries is indiscriminate; it provides information on all of these packages, as well as perhaps another few dozen new packages each day. Much of this information is often unnecessary and can be overwhelming, as one probably only cares about changes to the packages that one uses. In addition to the information overload, the diff that CRANberries provides for updated packages is in the form of a git diff, which is not human-readable.

The aim of this project, Diffify, is to provide a more tailored experience. Users can request information on specific packages or groups of packages. In addition, Diffify provides a more human-readable diff, giving the kind of information that users are most concerned about, i.e. changes to package dependencies, etc. from one version to another.

### 5.2.2 Method

The starting point and the inspiration for this project was the CRANberries website. CRANberries uses the R function `available.packages()`, which returns all

---

[4]`https://www.r-project.org/`
[5]`https://cran.r-project.org/`
[6]`http://dirk.eddelbuettel.com/cranberries/`

available R packages along with some metadata, such as version numbers. This information is then recorded in an SQL database and updated every two hours. The differences in packages between points in time is computed and stored in the SQL database which is used to calculate the RSS feed. Information on new, updated and removed packages is then added to the RSS feed (Eddelbuettel, 2007).

Jumping Rivers specified a number of requirements for this project.

Requirements included that:

- the code should be packaged as a Python package

- the code should expose a representational state transfer (REST) application programming interface (API)

- a database should be used for data storage and retrieval

- tests should be implemented

- documentation should be written

Throughout the course of the project I gave regular updates and incorporated feedback into the project, through a process of iterative development (Larman and Basili, 2003).

### 5.2.2.1 Python package

The technologies that were used in this project include:

- Poetry: a tool to handle dependency management and packaging in Python

- SQLAlchemy: a Python SQL toolkit and Object Relational Mapper (ORM) that enables one to switch easily between different types of SQL database

- Flask: a micro web framework written in Python - it is lightweight and allows for the quick set up of web applications

The Python package was called `cran-diff`. Poetry[7] was used to build the package and handle dependencies. The `cran-diff` package contains scripts and modules that can check CRAN for new packages or updates to existing packages, extract relevant metadata from new and updated packages and store this in a database, compare two versions of a package and provide a human-readable diff and provide a REST API to access this information.

A relational SQL database, set up using SQLAlchemy, was used to host the R package metadata. The database schema, as specified through SQLAlchemy's ORM (Object-Relational Mapping), is shown below.

```python
class Packages(Base):
    __tablename__ = 'packages'
    id = Column(Integer, primary_key=True,
        nullable=False)
    name = Column(String(250), nullable=False)
    title = Column(String(250), nullable=False)
    version = Column(String(250), nullable=False
        )
    date = Column(DateTime, nullable=False)
    description = Column(String(500), nullable=
        False)
    maintainer = Column(String(250), nullable=
        False)
    url = Column(String(250), nullable=False)
    bugreport = Column(String(250), nullable=
        False)
```

---

[7] https://python-poetry.org/

```python
class Imports(Base):

    __tablename__ = 'imports'

    id = Column(Integer, primary_key=True,
        nullable=False)

    package_id = Column(Integer, ForeignKey('
        packages.id'), nullable=False)

    name = Column(String(250), nullable=False)

    version = Column(String(250), nullable=False
        )


class Suggests(Base):

    __tablename__ = 'suggests'

    id = Column(Integer, primary_key=True,
        nullable=False)

    package_id = Column(Integer, ForeignKey('
        packages.id'), nullable=False)

    name = Column(String(250), nullable=False)

    version = Column(String(250), nullable=False
        )


class Exports(Base):

    __tablename__ = 'exports'

    id = Column(Integer, primary_key=True,
        nullable=False)

    package_id = Column(Integer, ForeignKey('
        packages.id'), nullable=False)

    name = Column(String(250), nullable=False)
```

The `id` in the `Packages` table is used as a foreign key in the `Suggests`, `Imports` and `Exports` tables. This allows for the linking of these tables together and efficient querying of relevant data.

A REST API was also created using Flask. This provided a high-level API to query for information from the database. All data returned from the API is in JavaScript Object Notation (JSON) format, which is widely used for REST APIs. A shell script, "run.sh", was created to automate the process of starting up the Flask web server. This started a web server listening on IP address 127.0.0.1 and port 5000. The app queries the database, using parameters supplied via the API, and returns a HTTP 400 response code if the package or version is not found, in keeping with the modern convention of using HTTP error codes for communicating errors.

### 5.2.3 Problems

A number of problems were encountered throughout the course of this project.

A preferred database was not specified in the original requirements list. As no preference was given, the Python package `pymysql` was originally used to generate the database due to ease of use. This is tied to the MySQL database server software and prevents other SQL databases from being used. Later on it was decided that the database should be set up using SQLAlchemy, a Python SQL toolkit and Object Relational Mapper. This made it easy to switch between using different types of SQL database, but meant that the code had to undergo a significant re-write.

Originally, an SQLite database was used for testing the package. However, SQLite does not enforce the length of a VARCHAR[8]. Despite putting limits on all of the columns in the database, some ended up being significantly longer. This was only realised and rectified once the database had been switched to a PostgreSQL database for use in production, which strictly enforces the length of VARCHAR fields.

---

[8]See `https://www.sqlite.org/faq.html#q9`

The metadata of the vast majority of R packages (especially more recent packages) is encoded in UTF-8, with a small number encoded in latin1. If package metadata contains non-latin characters, the encoding must be specified in the package `DESCRIPTION` file or the package will not be accepted by CRAN. However, some older packages do not meet this criterion and problems were encountered when adding their metadata to the database. The decision was made to only include packages which had been updated in the past two years in the database. This removed most non-conforming packages. A check was also added to the code to check for latin1 encoding and to convert from this encoding to UTF-8 if necessary.

## 5.2.4 Results

The Diffify API has the following endpoints:

`GET /names`: Returns list of all package names in the database.

Example:

```
$ curl -X GET http://127.0.0.1:5000/names

{
    "package_names": [
    "A3",
    "aaSEA",
    "ABACUS",
    "abbyyR",
    "abc",
    "abc.data",
    "ABC.RAP",
    "abcADM",
    ...]
}
```

`GET /versions/<package_name>`: Returns a list of all versions of a given package in the database.

Example:

```
$ curl -X GET http://127.0.0.1:5000/versions/
    dplyr

{
    "versions":[
    "0.8.5",
    "0.8.4",
    "0.8.3",
    "0.8.2",
    "0.8.1",
    "0.8.0.1",
    "0.8.0",
    "0.7.8",
    "0.7.7",
    "0.7.6",
    "0.7.5"
    ]
}
```

`GET /imports/<package_name>/<version>`: Returns a JSON object of all imports and their versions for a given version of a given package.

Example:

```
$ curl -X GET http://127.0.0.1:5000/imports/
    dplyr/0.8.5
```

```
{
    "imports":[
        {"R6": "", "Rcpp": "1.0.1", "assertthat"
            : "0.2.0", "ellipsis": "",...}
    ]
}
```

GET /suggests/<package_name>/<version>: Returns a dictionary of all suggests and their versions for a given version of a given package. Suggests are other packages upon which <package_name> does not depend, but which would be nice to have.

Example:

```
$ curl -X GET http://127.0.0.1:5000/suggests/
    dplyr/0.8.5


{
    "suggests":[
        {"DBI": "", "Lahman": "", "MASS": "", "
            RMySQL": "",...}
    ]
}
```

GET /exports/<package_name>/<version>: Returns a list of all exports for a given version of a given package. Exports are used to make accessible parts of an R module to users of that module.

Example:

```
$ curl -X GET http://127.0.0.1:5000/exports/
    dplyr/0.8.5


{
    "exports":[
        [".data", "add_count", "add_count_", "
            add_row", "add_rownames", "add_tally"
            ,...]
    ]
}
```

GET /diff/<package_name>/<version1>/<version2>: Compares two versions of a package and returns a structure containing the differences between imports, suggests and exports, with separate keys for added, changed, and removed items.

Example:

```
$ curl -X GET http://127.0.0.1:5000/diff/dplyr
    /0.8.5/0.7.5


{
    "exports":{
        "added":["group_by_drop_default", "
            group_cols", "group_nest", "
            validate_grouped_df", "as_label", "
            distinct_if", "new_grouped_df",...],
        "removed":["exprs"]
    },
    "imports":{
        ...
```

```
        }

        ...

    }
```

Figure 5.1 shows an example of the `/names` endpoint used in a web browser.

Figure 5.2 shows an example of the `/diff` endpoint used in a browser.

Error handling is included in the API. Figure 5.3 shows an example of a non-existent package version request.

Populating the database from scratch takes approximately 5.5 hours. However, this is only necessary the first time that the database is set up, as later updates only include new additions or changes to CRAN. The API responds in milliseconds.

Documentation was included and written in the form of a Jumping Rivers user manual, an internal system for creating web-based documentation in a consistent manner. This documentation was used to facilitate the handover of this project.
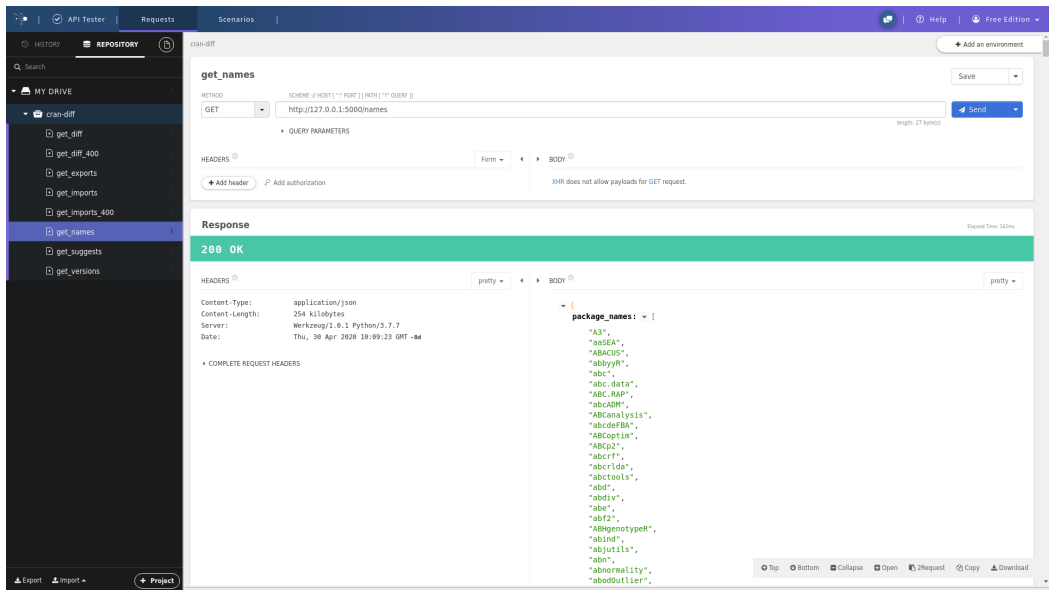
Figure 5.1: Example of an API request for all package names included in the database.
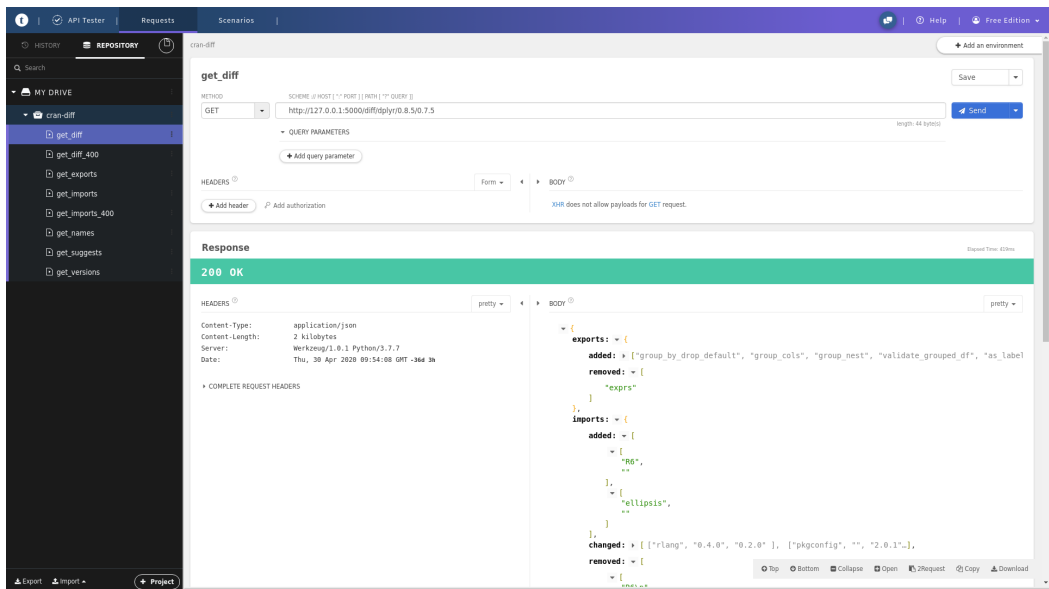


Figure 5.2: Example of an API request for the diff between two versions of the dplyr package.
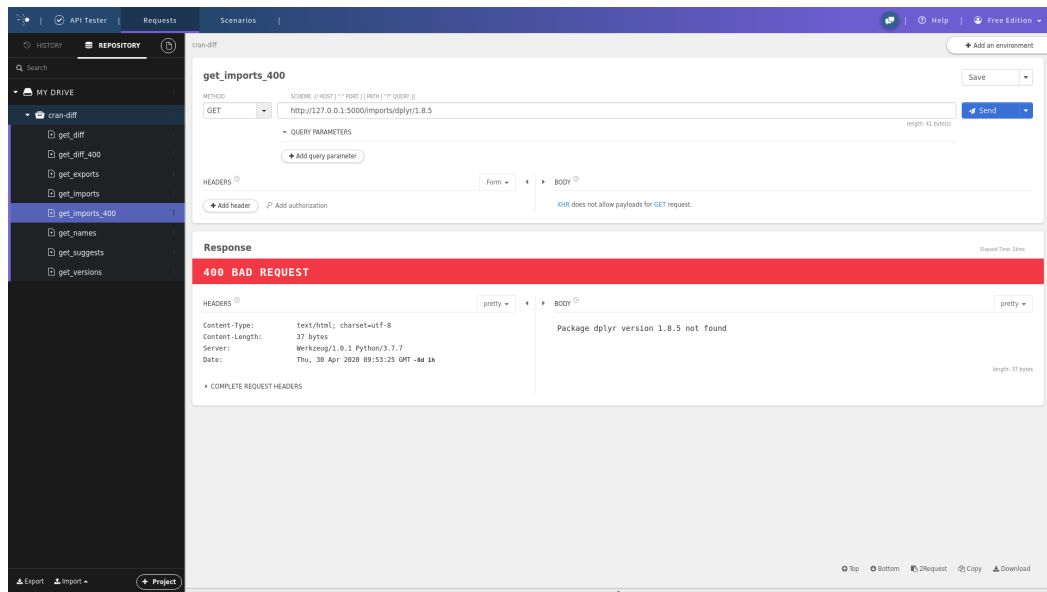
Figure 5.3: Example of an API request for the imports of a version of dplyr which is not in the database.

## 5.3 The JUNE Project

Following the outbreak of the SARS-CoV-2 virus, a multidisciplinary team of researchers from Durham University worked to develop JUNE, an open-source model for the simulation of epidemics. JUNE is an agent-based model (ABM), where each agent represents an individual member of the population. The agents are spatially distributed according to population density and with the demographics taken from census data. As the spread of COVID-19 is highly variable over both time and location (Bertozzi et al., 2020) and as it affects different parts of the population in different ways (Williamson et al., 2020), it requires a tool with a high level of precision and ABMs are ideal for this.

The published paper, containing a full description of the model and calibration is reproduced in Appendix A.

My contribution to this project was in the area of data analysis and visualisation. I analysed survey data and generated visualisations of the output to help improve the accuracy of the model, as well as creating plots for the publication. I looked at

time use survey data to work out where the population spends its time and created contact matrix plots of the simulation to compare to previous studies, such as the BBC Pandemic Experiment (Klepac et al., 2020). I wrote the code to generate the contact matrix plots, shown in Figure A.9 and reproduced below.
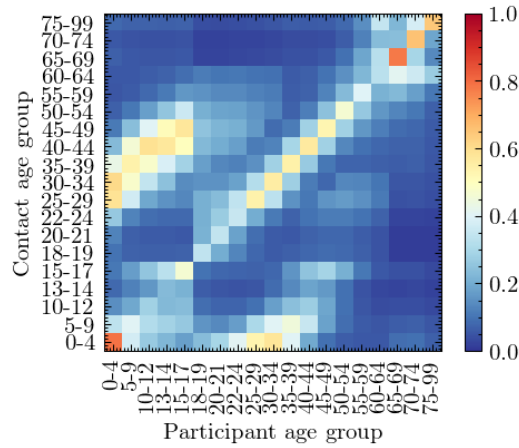
(a) Household
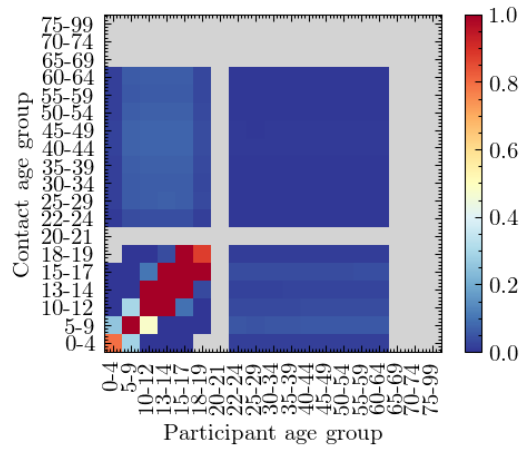


(b) School
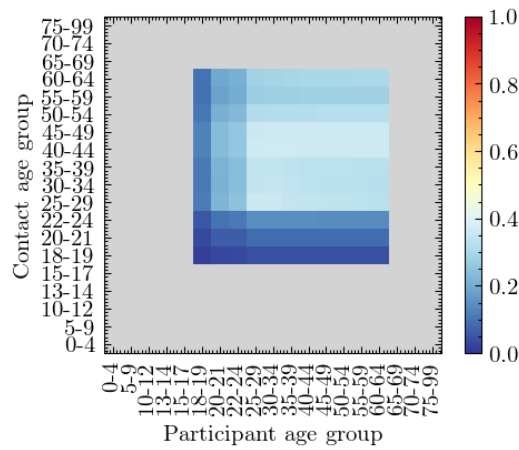


(c) Company

Figure 5.4: Social contact matrices for England derived from JUNE, before any mitigation strategies are implemented. Colour bars show (average) number of contacts in social settings between age groups, with all colour scales truncated at one to show differences between settings, while still clearly showing the structure in the matrices.

# Conclusions

## 6.1 Summary

Substructure is estimated by modern $N$-body simulations to comprise 5-10% of the mass of a dark matter halo (Gao et al., 2004). Dark matter substructure is important, not just for understanding galaxy formation, but also for modelling galaxy clustering and constraining the nature of dark matter. It was not until the late 1990s that cosmological simulations were able to resolve dark matter substructure in Milky Way-sized haloes (Moore et al., 1998). However, recent work suggests that current cosmological simulations, such as Aquarius (Springel et al., 2008), may still suffer from overmerging and that the majority of disruption is artificial (Penarrubia et al., 2010; van den Bosch et al., 2018; van den Bosch and Ogiya, 2018). In this thesis, we investigated the tidal evolution of dark matter subhaloes.

In Chapter 3, we attempted to reproduce the work of van den Bosch and Ogiya (2018) (vdBO18), first by using Gadget-4 with subfind, then by using code supplied by Frank van den Bosch. After performing a series of convergence tests, vdBO18 determined that modern cosmological simulations, such as Aquarius, still suffer from overmerging and that the majority of substructure disruption is not physical in nature, but is instead due to numerical factors. Specifically, they conclude that this disruption is caused by insufficient force softening and discreteness

noise caused by inadequate mass resolution.

We attempted to reproduce some of the key figures from vdBO18. Despite every attempt being made to reproduce the methods of the authors, even using the original codes, it has not proven possible to reproduce these results exactly. However, our results do agree qualitatively. The subhaloes do get disrupted, particularly those with low number of particles, $N_p$, and high softening length, $\epsilon$. The effects of discreteness noise are also evident in simulations with low mass resolution. Our subhaloes have proven to be slightly more robust, but they do end up disintegrating, especially at longer softening lengths and fewer number of particles. Our findings, therefore, do disagree with those of vdBO18, but qualitatively the results are nonetheless similar.

However, the conditions under which these simulations are run are idealistic and extremely artificial. The subhaloes are placed on circular orbits, which are very close to the centre of the main halo for the entire duration of the simulation, and the main halo is represented by a spherical, static, external potential. In Chapter 4, we investigated the tidal evolution of dark matter substructure using more realistic conditions. We selected a sample of subhaloes from the Aquarius A-2 simulation (Springel et al., 2008) and resimulated their orbits using the halo expansion (HEX) code. The HEX code approximates the potential of a simulated halo using a basis function expansion. It makes it possible to integrate the orbits of subhaloes in a time-varying halo potential at a much lower computational cost than the original simulation. This allowed us to easily re-create very expensive halo simulations and replay them with different parameters. We constructed a pipeline to sample the Aquarius merger trees, extract subhalo information and run resimulations with GADGET-4 and HEX, in order to examine the effect that softening length and number of particles has on the tidal evolution of subhaloes.

For type 1 subhaloes, we found that HEX replicates the original orbits well for subhalo mass $> 10^8$ M$_\odot$. When $N_p < 30,000$ we do start to see slight divergence but in general the bound fraction is close for different random realisations of the

initial conditions. However, once $N_p \leq 300$ subhaloes start to fully disrupt and the effects of discreteness noise also start to become more apparent. Despite this, for all of our 40 examples at least one of the five random realisations survived until the end of the simulation. It is only once we get to $N_p = 100$ that the majority of subhaloes are destroyed. We have shown that over 8 Gyr, subhaloes which are placed on realistic orbits in a realistic potential are quite resilient.

As well as the number of particles, we have also examined the effect that softening length has on the tidal evolution of subhaloes. We have looked at softening length over two orders of magnitude, from $0.003 - 0.3$ times the initial scale radius of the subhalo. We found that softening seems to have little effect on the survival of subhaloes which are placed on these realistic orbits.

We have also examined some examples of type 2 subhaloes. We found that when $N_p$ is increased to a sufficient value, these subhaloes do survive for the duration of the simulation, as shown in van den Bosch and Ogiya (2018). However, type 2 subhaloes that reach large peak masses are quite rare and so are not representative of the entire population.

In Chapter 5, we discussed Diffify, a tool which provides a comparison between different versions of any R package stored on CRAN. This was developed as a Python web application, which exposed an HTTP REST API. This used an SQL database which was populated using the openly-available package metadata from CRAN. We followed modern software development practices and included documentation and tests.

In Chapter A, we introduced June, an open-source model for the simulation of epidemics. June was primarily designed to simulate the COVID-19 outbreak in England, however its modular design allows it to theoretically fit any epidemic or geographical area. It is an individual-based model, which allows for fine-grained simulation of behaviour, including work patterns and leisure patterns. The simulation results compared favourably with official epidemiological data.

## 6.2   Further Work

In Chapter 4, we used the HEX technique to model both type 1 and type 2 subhaloes from the Aquarius simulation. However, we only looked at a small sample of subhaloes. In order to produce better statistics, we would ideally like to model all substructure from the Aquarius simulations. Type 2 subhaloes which are centrally located and are subject to the strongest tidal forces are of particular interest. Once we model all substructure, we could examine the subhaloes' properties as a function of radius to determine how accurate the results from Aquarius are. Does a higher resolution give different results? Is there a radius at which numerical effects are most important?

We have mentioned in Chapter 4 that the Aquarius A halo is not completely typical and has a very high concentration for its mass. We plan to follow a similar analysis and model substructure from the other relaxed Aquarius haloes, Aq-B - Aq-E.

At the moment, we have only tested our HEX coefficients to $z = 1.11$ and are thus limited to resimulating objects from this point onwards. Ideally, we would like to extend these coefficients back in time as far as possible. (agu) use $z = 2$ as the time of accretion for the majority of their subhaloes and we would like to compare these results to those of HEX for similar time of infall.

Finally, it would be interesting to investigate the effect of baryons on the tidal evolution of dark matter substructure using HEX. Following the work of Kelly (2022) on the Au16 halo from the AURIGA Simulation (Grand et al., 2017), we could test the HEX technique on the halo described in Grand et al. (2021).

# JUNE: open-source individual-based epidemiology simulation

Joseph Aylett-Bullock[*,1,2], Carolina Cuesta-Lazaro[*,1,3], Arnau Quera-Bofarull[*,1,3], Miguel Icaza-Lizaola[**,1,3], Aidan Sedgewick[**,1,4], Henry Truong[**,1,2], Aoife Curran[1,3], Edward Elliott[1,3], Tristan Caulfield[7], Kevin Fong[8,9], Ian Vernon[1,6], Julian Williams[5], Richard Bower[1,3], Frank Krauss[1,2]

[1]Institute for Data Science, Durham University, Durham DH1 3LE, UK
[2]Institute for Particle Physics Phenomenology, Durham University, Durham DH1 3LE, UK
[3]Institute for Computational Cosmology, Durham University, Durham DH1 3LE, UK
[4]Centre for Extragalactic Astronomy, Durham University, Durham DH1 3LE, UK
[5]Institute of Hazard, Risk & Resilience, Durham University, Durham DH1 3LE, UK
[6]Department of Mathematical Sciences, Durham University, Durham DH1 3LE, UK
[7]Department of Computer Science, University College of London, London WC1E 6BT, UK
[8]Department of Science, Technology, Engineering and Public Policy, University College London, London WC1E 6BT, UK
[9]Department of Anaesthesia, University College London Hospital, London NW1 2BU, UK
* Equal contribution
** Equal contribution

**Abstract**

We introduce JUNE, an open-source framework for the detailed simulation of epidemics on the basis of social interactions in a virtual population constructed from geographically granular census data, reflecting age, sex, ethnicity, and socio-economic indicators. Interactions between individuals are modelled in groups of various sizes and properties, such as households, schools and workplaces, and other social activities using social mixing matrices. JUNE provides a suite of flexible parameterisations that describe infectious diseases, how they are transmitted and affect contaminated individuals. In this paper we apply JUNE to the specific case of modelling the spread of COVID-19 in England. We discuss the quality of initial model outputs which reproduce reported hospital admission and mortality statistics at national and regional levels as well as by age strata.

## A.1 Introduction

The spread of SARS-CoV-2 in populations with largely no immunological resistance, and the associated COVID-19 disease, have caused considerable disruption to health care systems and a large number of fatalities around the globe. The assessment of policy options to mitigate the impact of this and other epidemics on the health of individuals, and the efficiency of healthcare systems, relies on a detailed understanding of the spread of the disease, and requires both short-term operational forecasts and longer-term strategic resource planning.

There are various modelling approaches which aim to provide insights into the spread of an epidemic. They range from analytic models, formulated through differential or difference equations, which reduce numerous aspects of the society–virus–disease interaction onto a small set of parameters, to purely data-driven parametrizations, often based on machine learning, which inherently rely on a probability density that has been fitted to the current and past state of the system in an often untraceable way. As another class of approaches, agent-based

models (ABMs) are "*particularly useful when it is necessary to model the disease system in a spatially-explicit fashion or when host behavior is complex[.]*" Russell et al. (2017) p2:5.[1] Being the traditional tool of choice to analyse behavioural patterns in society, they find ample use in understanding and modelling the observed spread of infections and in leveraging this for intermediate and long-term forecasting (Auchincloss et al., 2012; El-Sayed et al., 2012; Rockett et al., 2020). Such models also provide the flexibility to experiment with different policies and practices, founded in realistic changes to the model structure, such as the inclusion of new treatments, changes in social behaviour, and restrictions on movement.

To simulate pandemics, specific realisations of ABMs, individual-based models (IBMs), have been developed in the past two decades, for example Ferguson et al. (2006); Chao et al. (2010). In these models, the agents represent individuals constituting a population, usually distributed spatially according to the population density and with the demographics - age and sex - taken from census data.[2] Within the existing taxonomy of agent-based models in epidemiology, see for instance Hunter et al. (2017); Abar et al. (2017), these models often use a disease-specific modelling framework. Interactions between individuals in predefined social settings, systematically studied for the first time in Mossong et al. (2008), provide the background for disease spread, formulated in probabilistic language and dependent on the properties of the individuals and the social setting. The sociology of the population and the transmission dynamics are constrained separately using external datasets and available literature, and connected in the description of the spread of the disease. Calibration of such models to observed disease outcomes, such as hospital admission and mortality rates, is therefore reduced to the specific interface between the disease and the varying physiology across the broad population. Policy interac-

---

[1]Indeed, many models also feature some optimising behaviour of individuals as artificial intelligence-type actors against randomly drawn welfare functions, see for example Brandon et al. (2018).

[2]We will use the term "sex" in regard to chromosomal differentiation throughout this paper rather than gender. At the time of writing a full classification of the impact of chromosomal sex versus gender identification on the epidemiology of COVID-19 is unavailable. Within our modelling framework, nested and non nested identifiers can be constructed to map sex and gender should more granular statistical data be available.

tions and mitigation strategies can be flexibly encoded in detail as modifications of the social setting, and allow precise analysis of their efficacy that is not readily available in other approaches.

Evidence from disease data such as COVID–19 fatality statistics suggests that case and infection fatality rates are correlated, amongst other factors, to the age and socio-economics status of the population exposed to the etiological agent (Williamson et al., 2020). This necessitates the construction of a model with exceptional social and geographic granularity to exploit highly local heterogeneities in the demographic structure. In this publication we introduce a new individual-based model, JUNE,[3] a generalisable modular framework for simulating the spread of infectious diseases with a fine-grained geographic and demographic resolution and a strong focus on the detailed simulation of policy interventions. JUNE reaches a geographic resolution of societal factors similar to models that focus on single-site infection models, such as Zoellner et al. (2019), where space, location and distance are carefully modelled. In addition, similar to approaches such as the STHAM model Lund et al. (2020), the individuals in JUNE follow detailed spatio-temporal activity profiles that are informed by available data including time surveys, geographic and movement data. In contrast to such models that are usually constrained to a few tens of thousands of agents, JUNE simulates, simultaneously, the full population of a country in its spatio-temporal setting, and how a disease spreads through its population mediated by contacts between individuals. JUNE allows for flexible and precise parameterisations of policies that affect groups of individuals selected according to any of their characteristics. This allows modelling of policies to mitigate the further spread of a disease, realised as changes and restrictions on movement, to which we add the effectiveness of changes in social behaviour such as social distancing. The major cost for this level of detail in the model is in computational load; indeed, models such as JUNE would likely not have been possible prior to 2010

---

[3]A full open source code base and implementation examples are linked here:
GitHub: https://github.com/IDAS-Durham/JUNE
and PyPI: https://pypi.org/project/june/.
version used for this paper is v1.0

without using a prohibitive amount of computing power, see for instance Mei et al. (2018).

As a first application of JUNE, we model the spread of COVID–19 in England. In this context, JUNE uses census, household composition, and workplace data to ensure that each of the 53 million people in England are assigned a specific, identifiable location at any point in time. Their activities, health, age and other demographic attributes are then modelled at a fine-grained geographical level, which helps to ensure that local heterogeneity in population and movement characteristics are well recovered. This societal structure, generated by the model, is validated against a series of datasets (among others this includes: surveys of household size and composition, location and size of businesses, size and type of schools by region). The calibration to observed data from the actual spread of SARS-CoV-2 is then limited to how the virus is transmitted in the community through person-to-person 'contacts' (in the sense of sufficient proximity and timing to transmit). This component of the infection is calibrated to the spatio-temporal development of hospitalisations and casualties during the COVID–19 outbreak in England, starting in early March 2020. Preliminary observations demonstrate that a detailed large-scale model of this type has important implications for intermediate- to long-term modelling of the SARS-CoV-2 spread in the UK and elsewhere.

The remainder of this paper is as follows. A.2 provides an overview of the structure of the JUNE framework. In A.3, we detail the construction of a virtual population including a variety of demographic attributes. For the example case of England, we demonstrate that the constructed population reproduces the distributions of age, gender, ethnicity, socioeconomic indices, and the composition of the households they live in, all with a granularity of a few hundred people. The static properties of the population also include the assignment of students to schools and Universities and of employment in companies dis-aggregated by 21 industry sectors. In A.4 we discuss the dynamics of the population model. We demonstrate how JUNE correctly reproduces the average time-profile of daily activities of individuals in England. We

also describe in detail how we reconstruct movement and daily commute patterns based on publicly available data. Social interactions in various settings are modelled through parameters informed by social mixing matrices derived from surveys such as `PolyMod` (Mossong et al., 2008) and the BBC Pandemic project (Klepac et al., 2020). In contrast to other models, JUNE also incorporates interactions in various social venues such as pubs, restaurants, cinemas and shopping, outside the more structured settings of households, work places and schools. A.5 introduces the generalisable disease model with specific applications to COVID-19 — its transmission properties and the impact it has on infected individuals. We employ a probabilistic model for the former, while for the latter we incorporate data from the UK and other countries to characterise the journey of infected people through the healthcare system. In A.6 we describe how JUNE models the impact of various policy interventions and other mitigation strategies. In A.7 we show some first indicative results of JUNE highlighting its potential for future, more detailed studies. A.8 introduces our approach to fitting the model using Bayesian Emulation. We summarize our work in A.9, and conclude the paper with discussion of future work and improvements to the model.

## A.2 The Structure of the JUNE Modelling Framework

The JUNE framework is built on four interconnected layers: `population`, `interactions`, `disease` and `policy`, the layers and their interfaces are illustrated in Figure A.1. In the context of this publication, we focus on the application of JUNE to England's population, the spread of the COVID-19 disease, and policies that have been enacted by the UK Government in 2020. Clearly, a different population with different behavioural patterns will not only affect the distribution of individuals according to their personal characteristics, but it will also necessitate the adaptation of, e.g., social venues to these patterns and corresponding changes to the `population` and `interactions` layers. Similarly, modifications to the `disease`

layer will allow application of the JUNE framework for a different disease or, possibly, even a range of competing diseases. This flexibility and adaptability is even more pronounced in the `policy` layer where the introduction of new policies in reaction to an epidemic depends on behavioural patterns or societal norms.

The `population` layer encodes the individuals in the model and constructs static social environments such as the households they live in, the schools and universities they study in, and the workplaces where they work. The construction of the virtual population is informed by demographic data such as age, sex, and ethnicity distributions, the geographic location of their residence, and its composition. Depending on their age, individuals will attend school or university, work, or be retired.

The `interaction` layer models the social interactions of individuals, based on data about the frequency and intensity of contacts with other people in social settings. In addition to daily patterns of regular interactions with fixed groups of individuals such as household members, students and teachers in schools, and work colleagues, the `interaction` layer also models more randomised interactions. These include daily commute patterns to and from work, and more dynamic activities such as visits to restaurants, pubs, cinemas, and visits to other households.

The `disease` layer, which sits on top of the `population` and `interaction` layers, models the characteristics of disease transmission and the effects it has on those infected. In terms of disease transmission, the model incorporates the varying susceptibility of individuals, how likely individuals are to become infected when they mix with others in various locations, and how infectious they are over the course of their infection. In terms of disease progression, the model captures how likely individuals are to experience symptoms with varying severity, to be hospitalised, to be admitted to intensive care, or to die, as well as the timings associated with these events.

In response to the spread of a disease through its population, a government might

introduce policy measures designed to control and reduce the impact of the disease. In the case of COVID-19 in England and many other countries, policies have included social distancing measures, the closure of schools, shops, restaurants, and other leisure venues, and restrictions on movement. In JUNE, these are modelled in the `policy` layer. The high level of detail present in the `population` and `interaction` layers allows policies to be modelled at a corresponding granularity. This enables JUNE to describe the impact of policies that can be applied to specific geographical regions, to specific venues or sectors, or to individuals with specific characteristics. Examples include, but are not restricted to, the closure of targeted different types of (or even singular) venues, the inclusion or exclusion of specific age groups when going to school, shielding of the older parts of the population, modification to inter-household visits, and self-isolation measures for infected individuals and their contacts, including variations of compliance with these measures.

## A.3   Population and its Static Properties

JUNE creates a detailed virtual population at the individual level through its `population` layer, utilising a cross-section of demographic and geographic information. Since JUNE relies on multiple datasets, and is built to dynamically adapt to varying types of input, the approaches described in this section are generalisable to other settings with similar or complementary data availability. Given that different settings, e.g. countries, may have different methods and types of data collection, many of the input parameters described here are optional, allowing JUNE to be more easily adapted to differences in reporting.

### A.3.1   Geography and demography

To facilitate generalisability across multiple settings, JUNE models the geographic distribution of a population using a hierarchy of three layers – regions, super areas
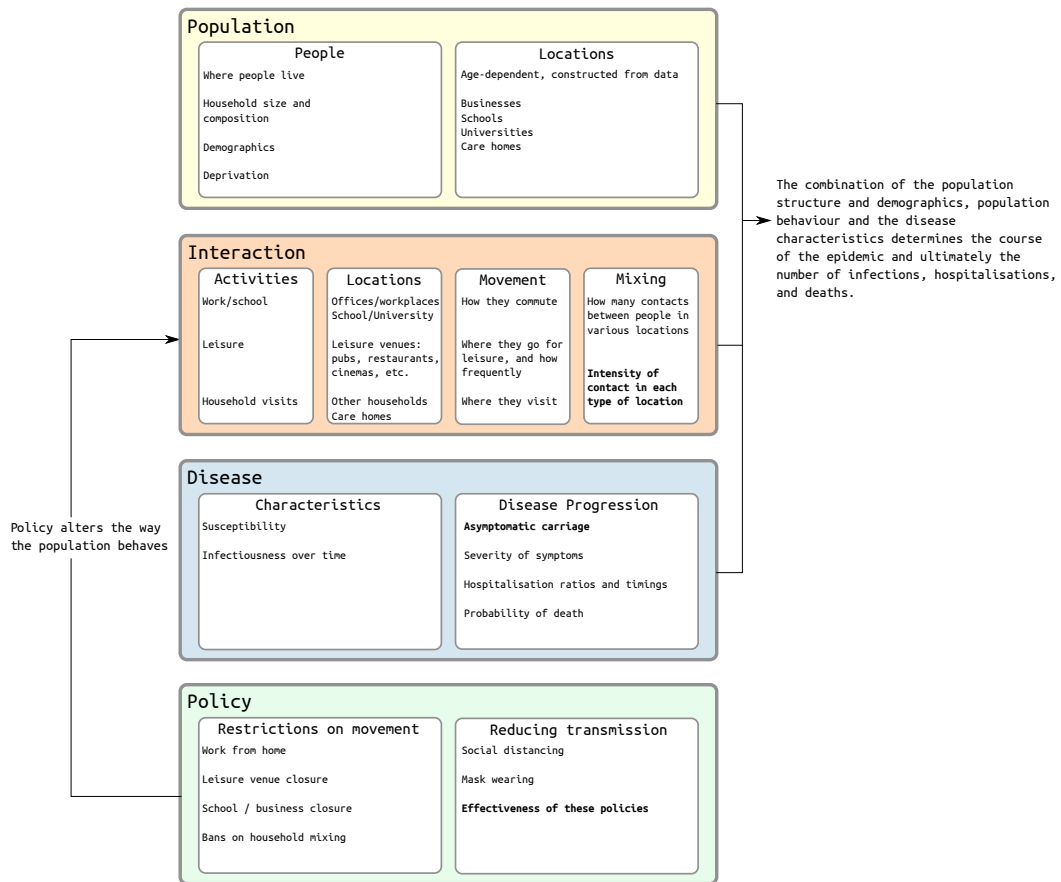
Figure A.1: Overview of the structure of JUNE. Fitted parameters are shown in bold.
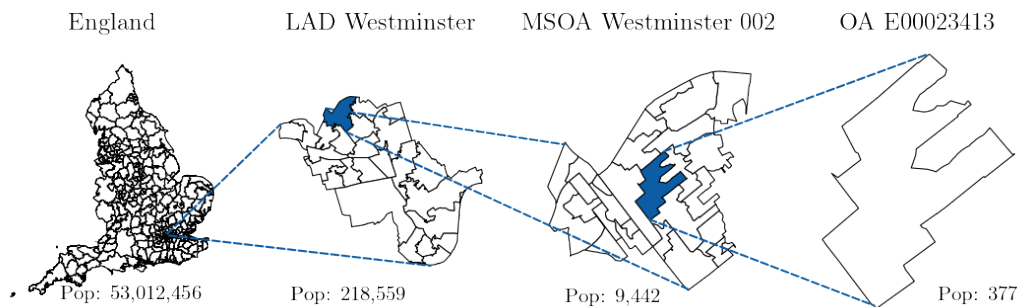


Figure A.2: Graphical representation of how the census data for England are structured, from the level of local authority districts (LAD), down to the level of output areas (OA), with middle layer super output area (MSOA) in between .

and areas. Layering these geographies allows the use of data at different levels of aggregation and enables simple statistical projections of data between these levels.

For the case of England, the construction of the virtual population in JUNE is largely

based on data from the latest UK census, which was carried out in 2011. This data is accessible through Nomis, an open-access database provided by the Office for National Statistics (ONS), and each dataset varies in its degree of aggregation. The three hierarchical geographical layers represented in Figure A.2 are:

1. regions – London, East Midlands, West Midlands, the North West, the North East, etc.;

2. super areas – approximately 7,200 middle layer super output areas (MSOAs);

3. areas – approximately 180,000 output areas (OAs).
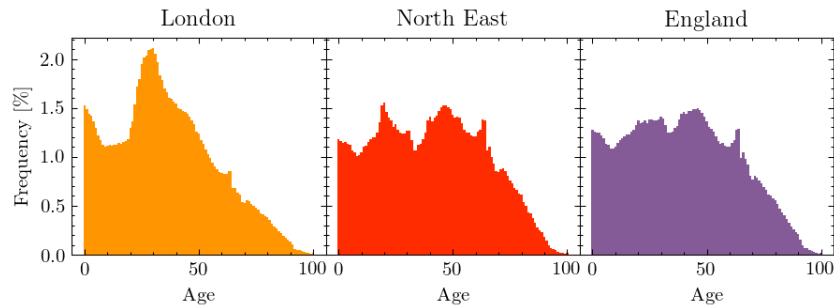


Figure A.3: Age profiles in different regions of England, taken from the ONS database and implemented in June

The individuals in June's virtual population are constructed according to age and sex dis-aggregated information, the minimal information required by June. In the case of England, the ONS census data provides this information at the OA (area) level (Office for National Statistics, 2011a,k) such that June naturally captures the population density at the most fine-grained level. In Figure A.3 we show age distributions in different regions. We use data derived from the ONS to additionally assign one of five broad ethnic categories to individuals based on their age, sex and location of residence (Office for National Statistics, 2011g) and follow a similar procedure for the socio-economic index, which we divide into centiles, according to the ranked English Index of Multiple Deprivation (IMD) (Ministry of Housing, Communities & Local Government, 2019).

## A.3.2 Household construction

The virtual population within June is placed into households of varying types. Depending on the structure of the available public records, households in June can be allocated with an arbitrary degree of granularity, taking into account multiple demographic attributes.

For the UK the ONS census datasets provide a detailed record of both household type and composition in England at the OA (area) level. That is, for each OA there is a set of summary statistics across a number of criteria, choices can then be made in regard to aggregating those frequency measurements at different resolutions. In terms of data categories for households the OA (area) level provides the following occupancy type counts: single, couple, family, student, communal, and other (Office for National Statistics, 2011j), and further specifies them by the number of old adults, aged over 65, adults, dependent adults (such as students), and children, providing around 20 distinct classes contingent on the underlying census information. Given the data structure, it is impossible to recover the exact composition for each household type. For example, the number of non-dependent children (people over the age of 18 living with their parents), the number of multi-generational families, and the exact distribution of adult groups sharing a household are not specified in these data-sets. However, these features can be statistically extrapolated using a mix of further secondary data and validated against various aggregate survey information at the regional and national level.[4] Households are populated iteratively giving preference to those household types with the most precise available data. The exact procedure for the UK is documented in Appendix B.1.1.

Similarly, to households, care homes are classified by type, positioned and populated using ONS data Office for National Statistics (2011d) at the OA (area) level. The ONS collects information on the age distribution and sex of residents of

---

[4]A forthcoming publication will discuss the use of secondary data to further constrain the uncertainty in the household construction and the subsequent impact on simulating the spread of COVID–19.

communal establishments at the MSOA (super area) level (Office for National Statistics, 2011c). By combining these datasets, we infer the age and sex distribution of the care home population.

Other communal establishments specified in the census, including student accommodations and prisons can be flexibly added with sufficient datasets. Within the presented version of June we do not model explicitly the age and sex distribution of these other communal establishments, however, since the age and sex distribution of the OA (area) level's population will be biased towards these communal residents, their resident characteristics are deemed to be realistic. Cross checking of case studies suggests that the communal allocation does capture the age and sex very accurately.

### A.3.3 Construction of virtual schools and universities

Schools and universities are two locations where a resident population will visit and interact. Every location can have universal and specific attributes flexibly initiated within the modelling frame work depending on the detail of available information. From public records June locates and enrolls schools according to their precise geo-coordinates and the publicly reported age ranges and numbers in attendance at each school. Students are sent to one of the $n$ nearest schools to their place of residence, according to which schools cater for their age. We form year groups which include all students of the same age. The formation of year groups, and classes within them, allows June to control mixing within and between children of different ages within the school environment.

To model schools in England, we use data provided by the UK Register of Learning Providers (UKRLP)[5] to determine the location of schools and their age brackets. Based on the current enrolment requirements for the UK, we assume that children between the ages 0-19 can attend school, with mandatory attendance between

---

[5] https://www.ukrlp.co.uk/

5-18. Since 19 year-olds can attend school, university, work or none of these, the institution they attend is determined by the number of vacancies in schools accepting students of that age group. We send children to one of the $n = 10$ nearest schools where classes sizes are limited to 40. One way in which we validate our assumptions is by comparing average travel distance to schools of different types. In JUNE, we find 1.7 km and 5.0 km for primary and secondary school students, compared with 2.6 km and 5.5 km respectively from the 2014 national travel survey (Department for Transport, 2014). Teachers are allocated to a school by randomised sampling from the available population — i.e. people over the age of 21 (to allow them to have attended university), who live in the same MSOA (super area) as the school, and who have been assigned the "Education" as their work sector (see next section for more details on work sectors). The number of teachers assigned to a particular school, and therefore the number of classes, is determined by sampling the ratio of students to teachers from a Poisson distribution with mean equal to the UK national average, separately for primary (mean of 21) and secondary schools (mean of 16), or a random choice of the two for mixed schools (Department for Education, 2011). JUNE's recovered student-teacher ratios are 22.0 and 17.8 for primary and secondary schools respectively.
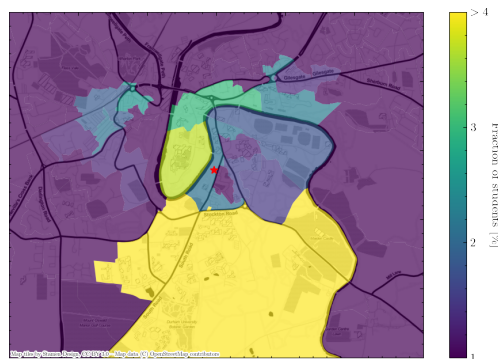


Figure A.4: A geographical visualisation of the location of student residences in Durham in JUNE, with the university location represented as a red star in the middle. Output areas are colour-coded according to the fraction of students they host. Note that the large southern area is where most of the university accommodation blocks are located.

Similarly universities are located according to their address as recorded in the
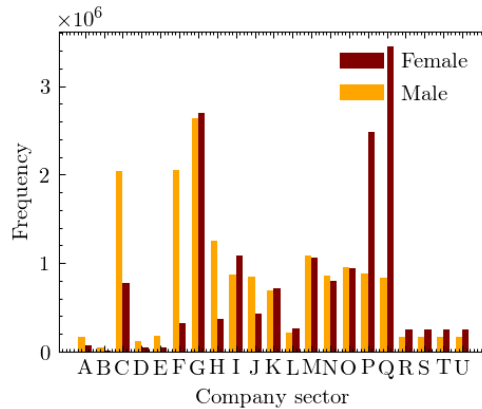
Figure A.5: Number of workers by sex and company sector (denoted by SIC code identifiers, see Table A.1) in June.

UKRLP. Students are enrolled in a university using the UKRLP enrollment data. The enrolled students are assigned from a subset of the local population to the university, reflecting the fact that the ONS census uses the term-time address of students.

Students are sampled from adults between the ages 18-25 with a preference given to those previously assigned to living in student or communal households in a given radius around the university. The concentrations of students expected by June in a particular area can be matched to secondary data on student living within any given city in the UK. Figure A.4 shows an example of a university city, Durham, in which we highlight the modelled regions inhabited by students. To date, we have not explicitly constructed the employees at Universities and their interactions with the student body.

## A.3.4   Construction of workplaces

Workplaces are constructed for the subset of the population in employment according to public records. We divide employment structures into three categories: work in companies with employees; work outside fixed company structures; work in hospitals and schools. The number of employees in each MSOA (super area) is data driven from the workforce information in that specific MSOA. To distribute

the workforce over workplaces, JUNE first initialises companies based on data containing their locations, sizes and the sectors in which they operate. In a next step, individuals who are eligible to work (i.e. between the ages of 18-65) are assigned an industry sector based on the geographic distribution of where the workforce live by sector. This results in origin-destination matrices which are used to match workers to their workplace and to optimize the distribution of individual company to reproduce sector-dependent distributions.

In England, the ONS database contains information on companies and workforce structured by industry type. Industries and companies are categorised according to 21 sectors following the Standard Industrial Classification (SIC) code convention (Office for National Statistics, 2007) (see Table A.1) and information about company numbers per sector, and company sizes is available at the MSOA (super area) level (Office for National Statistics, 2011f). Similarly, the ONS data also contain the size and sex distribution of the workforce by sector at the MSOA level, as well as the location of their employment (Office for National Statistics, 2011e,h). This enables the construction of an origin-destination matrix and allows us to distribute the workforce accordingly. More details on this specific procedure for initialising companies in JUNE and matching working individuals to these companies can be found in Appendix B.1.3.

The resulting distribution of our procedure assigning individuals an industry sector can be seen in Figure A.5. JUNE captures many of the sex-dependent features of the job market such as females dominating the healthcare profession and males the manufacturing sector. Recovering these sector-level sex imbalances can be crucial to reproducing and predicting potential sex imbalances in disease spread.

JUNE locates employment using data specifying the physical position of, for instance, company buildings. This, however, does not capture other modes of employment. We model people working from home through the specification of single-person companies in the same location as their place of residence. It should be noted that we do not currently explicitly model those workers who may not work in formal

| SIC code identifier | Description |
| --- | --- |
| A | Agriculture, forestry and fishing |
| B | Mining and quarrying |
| C | Manufacturing |
| D | Electricity, gas, steam and air conditioning supply |
| E | Water supply; sewerage, waste management and remediation activities |
| F | Construction |
| G | Wholesale and retail trade; repair of motor vehicles and motorcycles |
| H | Transportation and storage |
| I | Accommodation and food service activities |
| J | Information and communication |
| K | Financial and insurance activities |
| L | Real estate activities |
| M | Professional, scientific and technical activities |
| N | Administrative and support service activities |
| O | Public administration and defence; compulsory social security |
| P | Education |
| Q | Human health and social work activities |
| R | Arts, entertainment and recreation |
| S | Other service activities |
| T | Activities of households as employers; undifferentiated goods-and services-producing activities of households for own use |
| U | Activities of extraterritorial organisations and bodies |

Table A.1: Standard Industrial Classification (SIC) code identifiers for the 21 work place sectors modelled in JUNE and used by the ONS to categorise companies (Office for National Statistics, 2007).

company buildings but also do not work from home, such as contractors who may interact with a household of the people they are visiting for building improvements or maintenance work.

Hospitals play a dual role in JUNE, both as an essential part of patient's possible medical journey and as workplaces. We will discuss the role of hospitals for the former case in Section A.5. For both purposes, hospitals are initialised like many other locations in JUNE, based on available data regarding their location and capacity. Hospitals can be modelled individually or as clusters; in the latter case we represent the full cluster by one hospital. For our simulation of COVID–19 in

England, we define the relevant National Health Service (NHS) trusts as those that reported disease-related casualties – this amounts to a total of 129 trusts – and we cluster them into single hospitals.[6] The clustering of hospitals is in fact a better representation of the situation in England. The aggregation of data by NHS trust allows for a more detailed comparison of the number and geographical spread of hospital admissions with available data. We assign medical workers to hospitals based on the same origin-destination matrix at the MSOA (super area) level as derived above, by choosing from those who work in the healthcare sector ("Q"), with the additional constraint of assuming a fixed ratio of 10 hospital beds per medic – nurse or doctor. Teachers are chosen from the population in a similar matter by using the origin-destination matrix and choosing from those in the education sector ("P").

## A.4   Simulating Social Interactions

The `interaction` layer maps the spatial movement, location and intensity of social interactions, of the virtual population. To ensure a close match to real world dynamics summary information on the virtual population is calibrated to equivalent observed summary data. Comparable IBMs, such as Ferguson et al. (2006) and Chao et al. (2010), simulate social interactions in either static environments, such as households, schools, or work places, in a similar manner to that described in the previous section, or in an less specific way determined by gravity models. In contrast, JUNE allows for the specification of additional social settings, and directly connects them to geographical locations, such as shops or restaurants. We can also model transport routes of different types between specified geographical start and end-points. This granularity is further increased through the addition of social mixing matrices which parametrise differences in frequency and intensity of contacts between individuals in various settings (Mossong et al., 2008; Klepac et al., 2018).

---

[6]Some NHS trusts share resources and exchange patients across regions in an ad-hoc manner, however, this is not modelled explicitly.

## A.4.1 A virtual individual's day

Calendar days, decomposed into time-steps of varying length given in units of hours, are the background for our simulation of the social interactions of our virtual population.

Time in JUNE occurs in discrete time-steps of varying length measured in hours. Every time step in JUNE is tagged to a calendar day. The use of calendar time allows JUNE to distinguish between week-day and weekend activity profiles, which is relevant for time spent at work or in school. Each day can have a number of fixed, static, activities, such as 8 hours of work at the workplace or 10 hours at home overnight, supplemented with other activities, denoted as "other", that are distributed dynamically. Time-steps apply to all individuals, and are chosen to best approximate an "average" individual's day. The default time-steps are described more explicitly in Appendix B.2. During each time-step during which an "other" activity is allowed, each person who is not otherwise occupied, for example they are working or ill and in hospital, is assigned a set of probabilities for undertaking other activities in the model. These probabilities are part of a flexible social interaction model and depend on the age and sex of the person.[7] Given $N$ possible activities with associated probabilities per hour given by $\lambda_1, \ldots, \lambda_N$, for a person with characteristic properties $\{p\}$, the overall probability $\mathcal{P}$ of being involved with any activity in a given time interval $\Delta t$ is modelled through a Poisson process,

$$\mathcal{P} = 1 - \exp\left(-\sum_{i=1}^{N} \lambda_i(\{p\})\Delta t\right). \tag{A.1}$$

If the individual is selected to participate in one of these activities, the chosen activity, $i$, is then selected according to its probability

$$\mathcal{P}_i = \frac{\lambda_i(\{p\})}{\sum_{j=1}^{N} \lambda_j(\{p\})}. \tag{A.2}$$

---

[7]These probabilities can be generalised to depend on any attributes of the individual given reliable data.

The person is then moved to the relevant location corresponding to this activity. If no activity is selected, the individual will stay at home.

A summary of how much time is spent each week on various activities as a function of age is reported in Figure A.6(a). In Figure A.6(b) we show a comparison of the amount of time spent at home, work, grocery shopping, eating at restaurants/pubs, and commuting between JUNE and the UK Time Use Survey, 2014-2015 (Gershuny and Sullivan, 2015). Care home and cinema visits are not accounted for in the time survey.
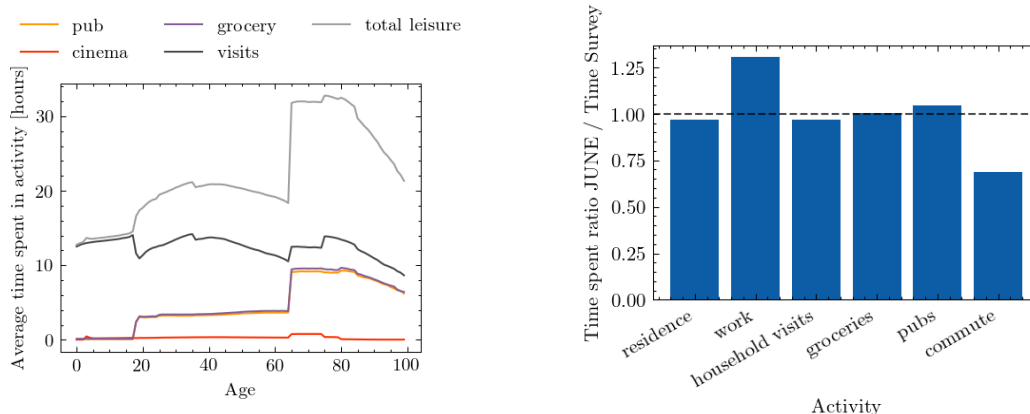


(a) Time spent in leisure by age in JUNE.

(b) Comparison of the fraction of time spent in different activities in JUNE and the time survey.

Figure A.6: Leisure activities in JUNE

## A.4.2 Localised activities

Within JUNE social activities outside the static home, work and school settings can be specified and given their own specific interaction model. Indeed, collections of sub-models can be specified, with substitutable activity choices. For instance, for the English population, these activity models are informed and parameterized by time surveys available from the ONS (Office for National Statistics, 2011h), which identify a variety of activity types including time spent at home, work, or in school. In addition, we have identified five additional settings which we assume are similarly relevant for the spread of the disease and have a similar level of social mixing: visits

to pubs or restaurants ("pubs"), cinema visits ("cinemas"), shopping ("groceries"), visiting friends or relatives in their homes ("household visits"), or visiting family members in care homes ("care home visits").

For England, we have located 120,000 pubs and restaurants according to their geo-coordinates, as well as 32,000 stores and 650 cinemas, with data from Open-StreetMap (OpenStreetMap contributors, 2017). Each time a person is assigned to any of "pubs", "groceries", or "cinemas", we pick a random venue from the $n$ venues closest to their place of residence, or the closest venue if the distance to any of them is greater than 5km. We have chosen $n = 7$ for pubs, $n = 15$ for shopping stores, and $n = 5$ for cinemas. Note that there are no permanent "workers" in these venues who return to a single venue daily; only "attendees" who choose their venue at random. Further locations such as gyms and places of worship can be easily added to the activity model, and, of course, it can easily be adjusted to other societies.

In addition, we model interactions in naively constructed social networks, by linking each household to a list of up to $\mathcal{N}$ other households in the same super area. One of the households in this list is selected, if "household visits" is chosen as activity during a time step. Residents will stay at home to receive the incoming visitor, who in turn may also bring their whole household with them according to a probability described by an external parameter. Comparison with national surveys suggests that setting the number of linked households $\mathcal{N} = 3$ provides realistic movement profiles. While care home residents in JUNE cannot visit other people, each resident is connected to a household of them in the local super output area from whom they can receive visitors. JUNE also models the interactions that result from elderly people needing help in their daily activities. Each person older than 65 years old has a probability, increasing by age, of needing some kind of assistance in their daily activities. We therefore assign a member of the local super area to be the carer of an elderly person, following the data available in Age UK (2019). Every weekday, the carer spends their leisure time visiting the household of the person
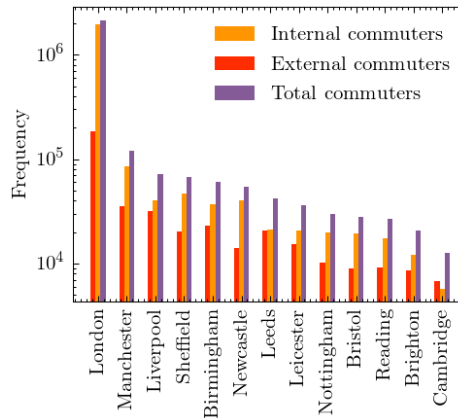
Figure A.7: Number of internal and external commuters by city as modelled in JUNE.

needing domestic care.

## A.4.3   Modelling mobility: commuting patterns

Mobility is modelled in JUNE through a number of transport types that collect and move the virtual population within a pre-specified region or connecting regions within a simulated country. JUNE permits an arbitrary number of transport networks of different types with different interactions (e.g. bus networks, train networks and road networks). For any given movement of the population JUNE ensures that each individual is singularly accounted for with an equivalent end or return location. Travellers move between nodes on these transport networks, and may share their means of transportation and potentially interact in a time consistent manner.

To model commuting and rail travel in England we utilize data provided by the UK Department for Transport (UK Department for Transport, 2011). Large metropolitan areas are selected as the major transit node for the network. Commuting induces social mixing between many people who may not normally come into contact and reflects the importance of transport as a mechanism for promoting the geographical spread of infection supplementing the spread from individuals moving to a new location and infecting other individuals at that location.

To fill our origin–destination matrix we use information contained in the ONS data-
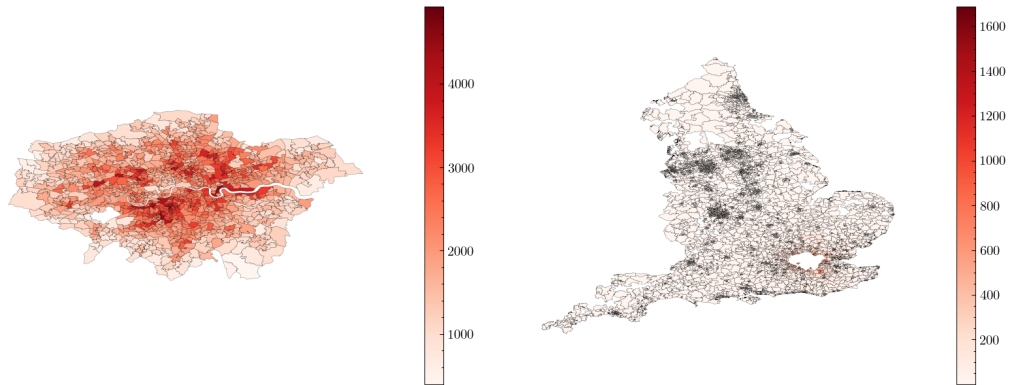
(a) Number of internal commuters in London.

(b) Number of external commuters in London.

Figure A.8: Commuting maps for London as derived from JUNE. Any visible super area (MSOA) which is not completely white has at least one commuter from that location.

base concerning the mode of commuting of individuals at the area (OA) level (Office for National Statistics, 2011b), to distribute commuting modes probabilistically. We define two modes of public transport, "external" which defines those commuting in and out of metropolitan areas, and "internal" which defines those commuting within these areas. Metropolitan areas are defined using data obtained from the ONS (Office for National Statistics, 2015). For the sake of computational efficiency we model only the travel patterns of those working inside metropolitan areas, who in fact represent the overwhelming majority of public transport commuters. This includes commuters who live and work in the city, as well as those who are entering the metropolitan area from outside. The number of internal and external commuters by city in England is given in Figure A.7. The cities included are geographically spread across England thereby accounting for major commuting patterns in most regions modelled. In total, we explicitly model commuting into 13 out of a possible 109 cities in England, which accounts for 60% of all metropolitan commuters and 46% of all those using public transport to commute to work. Figure A.8 shows maps of the residences of internal and external commuters in two cities in our model, where the inner section in white denotes the respective metropolitan areas. Specifically, from Figure A.8(b) we can see that, given the large commute

radius of cities like London (we observe a similarly large radius for Birmingham and several other cities), commuting can be a key driver for the inter-regional spread of infectious-diseases.

Travelling within a metropolitan area, i.e. the internal commuting mode, is modelled as a self-connected loop – practically speaking this means that internal commuters may in principle interact, irrespective of the actual movement inside the city. For external commuting, the travel into and out from the metropolitan area, we identify shared routes for commuters living in neighbouring areas and super areas. The number of possible routes into each city, and therefore the number of ways to divide regions around the cities, is informed by the approximate number of rail network lines into each city – currently this is set to eight in London and four for each of the other 12 cities (National Rail, 2015).

We randomly partition people sharing the same commuting route into subgroups, "carriages", which define the environment in which social interactions take place. The commuting time-step is run twice a day and in each run the travellers are randomly distributed into carriages. The number of people per carriage is determined by city-dependent data obtained from the UK Department for Transport (UK Department for Transport, 2011). More details on the specific algorithm for modelling commuting in JUNE can be found in Appendix B.1.4.

## A.4.4  Social interaction frequencies and intensities

Social contact matrices (Klepac et al., 2020; Mossong et al., 2008) provide information about the age-dependent frequency and intensity of in-person contact in different social settings, an important ingredient to many epidemiological simulations. They measure the average daily number of conversational and physical contacts between individuals of different ages. This means that they are normalised to the size of the population in the respective age bins, but do not account for whether they can take part in such contacts. To use them within JUNE we therefore have

to account for the fact that social settings define the group of people coming into contact with each other. To exemplify this, consider the case of contacts between adults and students in schools. While the social contact matrices in the literature normalise the number of contacts of a 30-year old with children of a certain age to the number of 30-year old adults in the population, in JUNE only a subset of 30-year old adults work as teachers and can therefore interact with the children. In the construction of matrices specific for JUNE, we therefore combine the results from Klepac et al. (2020) and Mossong et al. (2008) with simple assumptions about possible participants in contacts.

Averaging over age ranges in different settings, we arrive at simplified social mixing matrices, $\chi_{si}^{\mathscr{L}}$, which will be comparable to the inputs from literature upon combination with the model results for the composition of social environments. Below we list our simplified social mixing matrices inferred from literature, with $\mathscr{L} \in \{(H), (S), (W)\}$ (home, school, work place), as well as the relative proportions, $\phi_{si}^{\mathscr{L}}$, of physical contacts. The latter are relevant, since in line with standard approaches, closer physical contact in JUNE is proportional to a higher propensity for transmission for the etiological agent.

For the households social mixing matrices, we define four categories, young children ($K$), young dependent adults of age 18 or more ($Y$) that still live with their parents, adults ($A$), and older adults ($O$) of age 65 and over. We use:

$$\chi_{ij}^{(H)} = \begin{pmatrix} 1.2 & 1.69 & 1.69 & 1.69 \\ 1.27 & 1.34 & 1.47 & 1.50 \\ 1.27 & 1.30 & 1.34 & 1.34 \\ 1.27 & 1.50 & 1.34 & 2.00 \end{pmatrix} \text{ and } \phi_{ij}^{(H)} = \begin{pmatrix} 0.79 & 0.70 & 0.70 & 0.70 \\ 0.70 & 0.34 & 0.40 & 0.40 \\ 0.70 & 0.40 & 0.62 & 0.40 \\ 0.70 & 0.40 & 0.40 & 0.56 \end{pmatrix}$$

(A.3)

For household visits, we make the simplifying assumption that the same matrices also describe the contacts between visitors and residents. For visits to care homes, we believe that visitors come into contact only with residents and care home workers, and not with other visitors. We therefore hypothesize 6 conversational contacts

with residents and 1.5 with care home workers.

Social contacts in schools identify teachers (T) and students (S), the latter are organised in year groups and further divided into classes of up to 40 students. In our age-averaging, we implicitly assume that the number and character of teacher-student contacts is independent of the age of the students. Student-student contacts are assumed to be most frequent within a class or year group, and fall off steeply with the age difference. This behaviour is captured by fitting a matrix with values for the age-diagonal elements and a fall-off per year age-difference by a factor of 3. Therefore we have

$$\chi_{ij\in\{T,S\}}^{(S)} = \begin{pmatrix} 4.8 & 0.75 \\ 15 & \chi_{SS}^{(S)} \end{pmatrix} \quad \text{and} \quad \phi_{ij\in\{T,S\}}^{(S)} = \begin{pmatrix} 0.05 & 0.08 \\ 0.1 & \phi_{SS}^{(S)} \end{pmatrix}, \tag{A.4}$$

with the student-student matrices taking the following form

$$\chi_{SS}^{(S)} = \begin{pmatrix} 2.5 & 0.75 & 0.25 & \dots \\ 0.75 & 2.5 & 0.75 & \dots \\ 0.25 & 0.75 & 2.5 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \text{and} \quad \phi_{SS}^{(S)} = 0.15 \;\; \forall i,j \in \{S\}. \tag{A.5}$$

For the contacts at work we do not take into account of any age-dependence and, in the absence of data, do not model any sector-dependent variation of their number of intensity, thus

$$\chi_{si}^{(W)} = 4.8 \quad \text{and} \quad \phi_{si}^{(W)} = 0.07, \tag{A.6}$$

In Appendix B.3 we detail the algorithms used to construct the social mixing matrices used in June including the matrices for other locations not listed here.

These social mixing matrices in June are defined for a setting-specific characteristic time $t_{\text{char}}$, so the total number of contacts in a time interval $\Delta t$ in a given setting is then modified by a factor $\Delta t / t_{\text{char}}$.

To validate these simplified matrices, we include them within June where they are combined with the composition of the specific social settings. In Figure A.9

(a) Household



(b) School



(c) Company
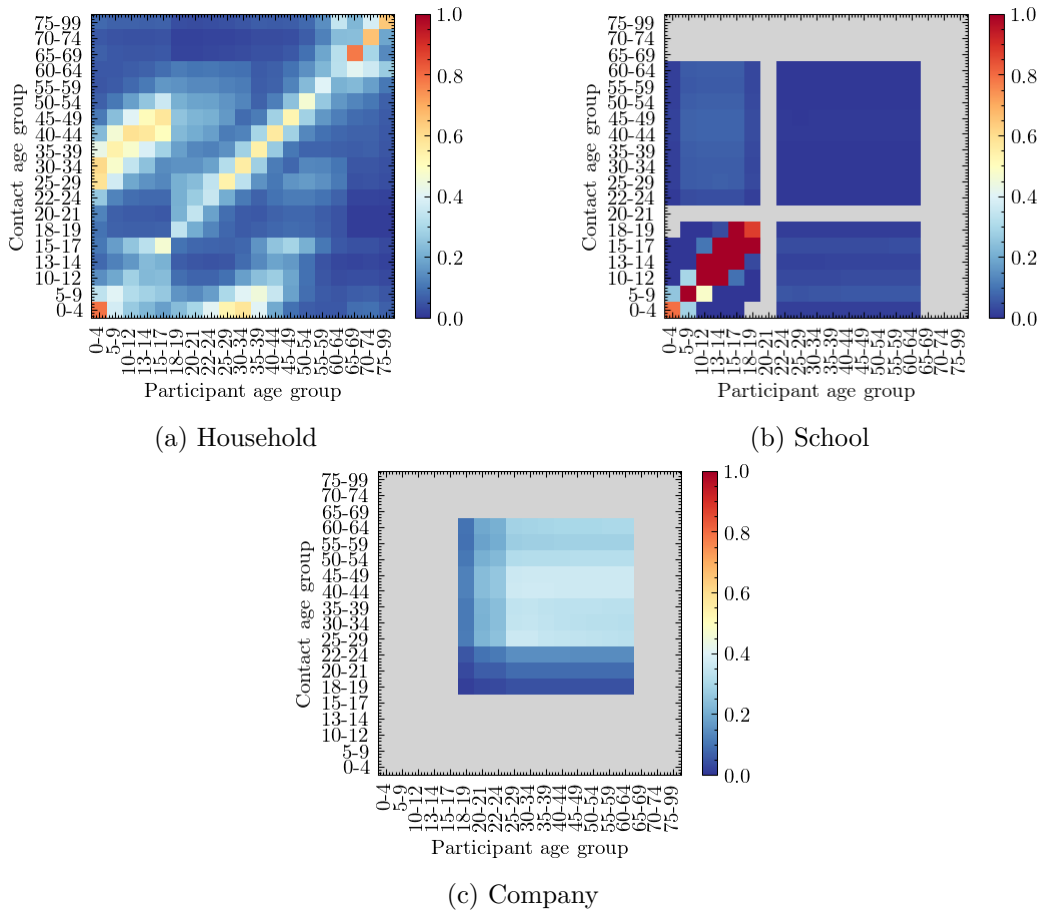
Figure A.9: Social contact matrices for England derived from JUNE, before any mitigation strategies are implemented. Colour bars show (average) number of contacts in social settings between age groups, with all colour scales truncated at one to show differences between settings, while still clearly showing the structure in the matrices.

we show the resulting contact matrices as "measured" from the JUNE simulation. The effect of the combination with the composition is most pronounced in the household matrices which exhibit textures that can be directly traced back to the age intervals of children, dependent children/young adults, adults and older adults that JUNE inherits from the ONS data. These matrices naturally recover much of the structure present in those recorded in Klepac et al. (2020) and Mossong et al. (2008). Further details on the methodology for extracting these matrices from JUNE can be found in Appendix B.3.5.

# A.5 Infection Modelling: Spreading and Health Impact

The transmission of infection through social interactions described in the `interaction` layer, and the progression of the disease and its impact on the individual, are both modelled in the `disease` layer. Although we focus on the case of COVID–19 here, this layer is designed to be generalisable and can contain more than one circulating etiological agent and or types of agent.

Throughout this Section and the rest of the paper, we will use two definitions of COVID-19 "cases". The first is when we refer to cases in the model itself — here, a case of COVID-19 is an infected agent which may be symptomatic or asymptomatic. The second is when referring to cases in reality — here, a case someone who has tested positive for COVID-19. Since the latter is subject to testing coverage, capacity and efficacy, we do not use these for fitting or validation purposes.

## A.5.1 Infection Transmission

JUNE models the transmission of an infection from infecting individual, $i$, to susceptible individual, $s$, in a probabilistic way. The probability of infection in a social setting within a group of people, $g$, at a location, $L$, depends on a number of factors:

- the number, $N_i$, of infectious people $i \in g$ present;

- the infectiousness of the infectors, $i$, at time $t$, $I_i(t)$;

- the susceptibility, $\psi_s$, of the potential infectee, $s$;

- the exposure time interval, $[t, t + \Delta t]$, during which the group, $g$, is at the same location;

- the number of possible contacts, $\chi_{si}^{(L)}$, and the proportion of physical contacts, $\phi_{si}^{(L)}$, in location $L$, both taken from Eqs. (A.3)-(A.6) in Section A.4.4;

- and the overall intensity, $\beta^{(L,g)}$, of group contacts in location $L$.

Most of these ingredients depend on the time, $t$, of the contact. For example, the number of contacts, $\chi_{si}^{(L)}$, and the proportion of physical contacts, $\phi_{si}^{(L)}$, and the overall contact intensity, $\beta^{(L,g)}$, will change with the implementation of social distancing policies. To simplify notation, we introduce a combined contact intensity for a group $g$ with size $N_g$ at location $L$,

$$\beta_{si}^{(L,g)}(t) = \beta^{(L,g)} \cdot \frac{\chi_{si}^{(L)}(t)}{N_g} \left\{ 1 + \phi_{si}^{(L)}(t) \left[ \alpha(t) - 1 \right] \right\} , \qquad (A.7)$$

where the ratio $\chi/N_g$ provides a simple parametrization of the probability of $s$ being in contact with another individual in the group, and $\alpha(t) > 0$ describes the relative impact of close physical contacts. Both the factor $\alpha(t)$, which we assume to be the same for all locations, and the location- and group-specific contact intensities, $\beta^{(L,g)}$, are taken from fits to data.[8]

In the construction of an infection probability for a susceptible individual, $s$, we make a number of assumptions. First of all, we model the probability of being infected as a Poisson process. In keeping with the probabilistic process, the argument of the Poisonnian is given by a sum over individual pairs of infectious individuals with the susceptible person, implying a simple superposition of individual infectiousness. The underlying individual transmission probabilities are written as the product of the susceptibility of the susceptible individual, the infectiousness of the infected person, and the contact intensity, all integrated over the time interval in which the interaction occurs. The integration over time ensures that the transmission probability increases with the time of exposure. We therefore arrive at the

---

[8] In reality, the $\beta$ parameters are fitted in a location specific way, irrespective of the group — i.e. a location of type $L$ in containing one group of people, $g_1$, and another location of the same type, but with a different group of people, $g_2$ (e.g. two pubs in different places) will have the same $\beta$.

transmission probability, i.e. a probability for $s$ to be infected as:

$$\bar{\mathcal{P}}_s(t,\, t+\Delta t) = 1 - \exp\left[-\psi_s \sum_{i\in g} \int_t^{t+\Delta t} \beta_{si}^{(L,g)}(t')\mathcal{I}_i(t')\mathrm{d}t'\right].\qquad(\mathrm{A.8})$$

Note that in the actual implementation, we approximate the integral over time with a simple product,

$$\int_t^{t+\Delta t} \beta_{si}^{(L,g)}(t')\mathcal{I}_i(t')\mathrm{d}t' \;\longrightarrow\; \beta_{si}^{(L,g)}(t)\mathcal{I}_i(t)\Delta t\,.\qquad(\mathrm{A.9})$$

This leaves us to fix the last two ingredients in A.8, the individual susceptibility, $\psi_s$, and the infectiousness, $\mathcal{I}_i(t)$. Contemporary peer reviewed academic research on susceptibility to infection by the etiological agent with or without the onset of disease symptoms is sparse and inconsistent. Following some evidence, for example in Dong et al. (2020) and Lee et al. (2020), on transmission and susceptibility of children (using the UN classification), we fix $\psi_s = 0.5$ for children under the age of 12, and $\psi_s = 1$ for everybody else. The infectiousness of individuals, $\mathcal{I}_i$, changes with time, and it is not directly measurable. To model its behaviour we use the temporal dependence of viral shedding as a proxy for infectiousness. Studies in the context of COVID–19 have shown that viral shedding peaks at or slightly before the onset of symptoms, and then begins to decrease (He et al., 2020). In June we use a globally defined temporal dependence of infectiousness, $f_I(t)$, and multiply it with a peak value, $\mathcal{I}_{i,max}$, which depends on the infected individual,

$$\mathcal{I}_i(t') = \mathcal{I}_{i,max} \cdot f_I(t').\qquad(\mathrm{A.10})$$

We choose the maximal infectiousness according to a log-normal distribution parameterised by its median $\exp(\mu) = 1$ and shape $\sigma = 0.25$. The long right tail of the log-normal distribution allows for small numbers of highly infectious individuals more likely to precipitate superspreading events (SSEV). We also capture the conjectured reduced infectiousness of individuals with no or only mild symptoms. Following a similar parameterisation to that in Hinch et al. (2020), we multiply the maximal infectiousness of asymptomatic individuals by 0.5. In Figure A.10, we

show an example of the time evolving profile for an infected individual in JUNE, comparing the resulting infectiousness for different symptoms. For the time-dependent profile, we use the gamma distribution as fitted in He et al. (2020),

$$f_I(\tau = t' - t_0 - t_{inc}, a) = \frac{\tau^{a-1}e^{-\tau}}{\Gamma(a)}, \qquad (A.11)$$

where $t_0$ is the time of infection, $t_{inc}$ is the incubation period, sampled from a normal distribution centered at two days prior a possible onset of symptoms and with a width of half a day, $a$ is the shape parameter of the gamma distribution, and $\Gamma(a)$ is the the gamma function.
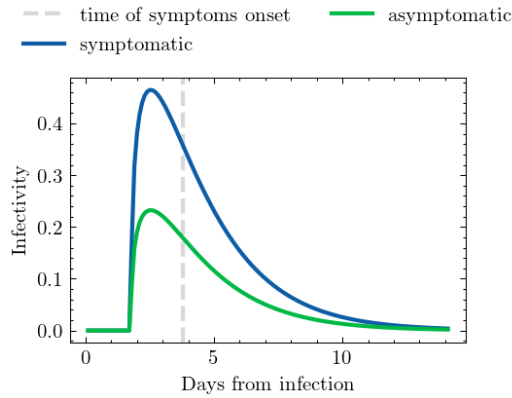


Figure A.10: Time-dependent infectiousness profile, $f_I(t')$, shown for the same realisation of the infection but where the infected person is symptomatic or asymptomatic.

## A.5.2  Infection Progression

When an individual is infected, they will experience different impacts on their health. Figure A.11 presents the paths available in JUNE for the progression of the infection that aim to capture different symptom severities, outcomes, and their operational impact on the healthcare system, i.e. whether patients are hospitalised or admitted to intensive care or treatment units (ICU/ITU). Once an individual is infected, JUNE selects their specific complete path according to these probabilities. These paths are codified as a sequence of possible different stages of the disease ("infected", "asymptomatic", "mild", "severe", "hospitalised", "ICU/ITU", "dead",
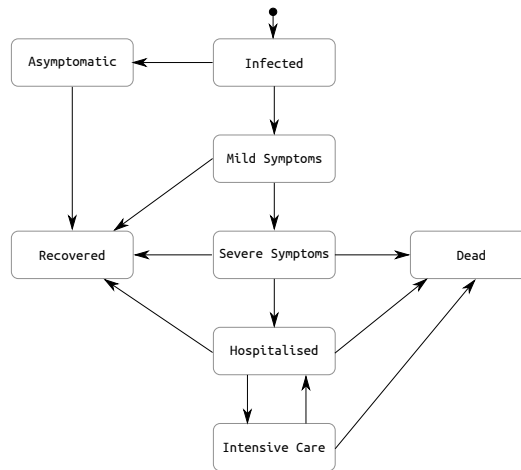
Figure A.11: Pathways for the infection progression and possible outcomes. Note that in our model a patient can only go to the intensive care once, and that a patient that returns from the intensive care to the hospital will survive.

"recovered") in addition to characteristic time intervals for each stage. The latter are chosen randomly according to probability functions informed by available data. The paths terminate with the individuals either dead and taken out of the simulation, or recovered, in which case their susceptibility is set to 0, making them immune to re-infection.[9]

JUNE distinguishes the following different routes for the progression of the infection with rates depending on the characteristics of the infected individual (currently age, sex), summarily denoted by $p$:

1. asymptomatic individuals, rate $R_{\text{I}\rightarrow\text{A}}(p)$, continue their life normally;

2. individuals with mild symptoms, rate $R_{\text{I}\rightarrow\text{M}}(p)$, usually continue their lives as normal, except if certain policies are activated;

3. individuals with severe but not lethal symptoms, rate $R_{\text{I}\rightarrow\text{S}}(p)$, stay at home until recovery;

4. individuals with severe symptoms who will eventually die in their residences, with rate $R_{I\rightarrow\text{DR}}(p)$;

---

[9] JUNE could also model reinfection of individuals but to date there are no data constraining this in the context of COVID–19.

5. individuals who are admitted to hospital but will recover, with rate $R_{\mathrm{I}\to\mathrm{H}}(p)$;

6. individuals who are ultimately admitted to ICU/ITU before recovering, with rate $R_{\mathrm{I}\to\mathrm{ICU}}(p)$

7. individuals who are admitted to hospital and will die there, with rate $R_{\mathrm{I}\to\mathrm{DH}}(p)$.

8. individuals who are admitted to ICU/ITU and die there, with rate $R_{\mathrm{I}\to\mathrm{DICU}}(p)$.

The determination of probabilities for the different paths is based on COVID–19 data that are not entirely sufficient to develop a complete and detailed picture. As a consequence, we supplement them with assumptions by inferring some properties through cross-relating datasets. In the following, we will outline our procedure which is largely predicated by our choice of the example at hand - the spread of COVID–19 in England. We will use a notation where $N_X(p)$ denotes the number of cases satisfying criterion $X$ for people with characteristic properties $p$.

The construction of reasonable progression paths, and their probabilistic distribution, relies critically on the knowledge of how many people have been infected, as well as the dependence on attributes such as age and sex. COVID–19 tests between February and May 2020 in the UK were mostly administered to people presenting symptoms or people that have been in close contact with confirmed cases in hospital, thereby biasing the results. We therefore need to infer the number of infections from other controlled studies, such as antibody tests. In Ward et al. (2020) the seroprevalence, $r_{\mathrm{sp}}(p)$, of COVID–19 in the adult population in England was determined through a sample of more than 100,000 adults, showing a reduction in seroprevalence with increasing age. Because the seroprevalence is an estimate of all people that were infected up to the time of the test and – most importantly – survived, we need to correct for those who died of the disease until this point. This turns out to be an important correction, especially in older age bins due to the non-negligible probability of elderly who died. We therefore add the age– and sex–dependent number of deaths, $N_{\mathrm{D}}(p)$, reported by the ONS (Office for National

| Quantity | Source |
|---|---|
| Population by age, sex and residence type | Office for National Statistics (d), Office for National Statistics (b) |
| Seroprevalence in GP by age | Ward et al. (2020) |
| Seroprevalence in CH by age | Department of Health & Social Care (2020) |
| Deaths by place of occurrence and residence type | Office for National Statistics (c) |
| Deaths profile by age and sex | Office for National Statistics (c) |
| Deaths in CH profile by age and sex | Office for National Statistics (2020a) |
| Hospital deaths profile by age, sex | Scientific Advisory Group for Emergencies (2020) |
| Hospital deaths in CH profile by age, sex | Public Health England (PHE) (2020) |
| ICU/ITU deaths profile by age, sex | Public Health England (PHE) (2020) |
| Total hospital admissions | NHS |
| Hospital admissions profile by age, sex | Scientific Advisory Group for Emergencies (2020) |
| ICU/ITU admissions profile by age, sex | Public Health England (PHE) (2020) |
| Hospital admissions in CH profile by age, sex | Public Health England (PHE) (2020) |

Table A.2: Datasets used in the derivation of mortality and hospitalisation rates. GP stands for people living in a household, and CH stands for people living in care homes. If not specified, datasets involve people from both populations. All data is taken until the 13th of July 2020, consistently with the seroprevalence study Ward et al. (2020).

Statistics, c), to the corresponding numbers inferred from the seroprevalence to arrive at the total number of cases, $N_{\text{tot}}(p)$:

$$N_{\text{tot}}(p) = r_{\text{sp}}(p)N(p) + [1 - r_{\text{sp}}(p)]N_{\text{D}}(p)\,, \qquad (A.12)$$

where $N(p)$ is the total population number in England with characteristics $p$. We note that there were two population groups excluded from the serology survey: people under the age of 18, and care home residents. For the former, we assume that their seroprevalence by age is identical to the population group aged 18, while for the latter, we set a flat seroprevalence by age at 11% value as reported in the Vivaldi report of the UK Department of Health and Social Care (Department of Health & Social Care, 2020) in the beginning of July 2020.

Health outcomes given a simulated infection are captured in $R_{\text{I}\rightarrow\text{X}}$, where X is
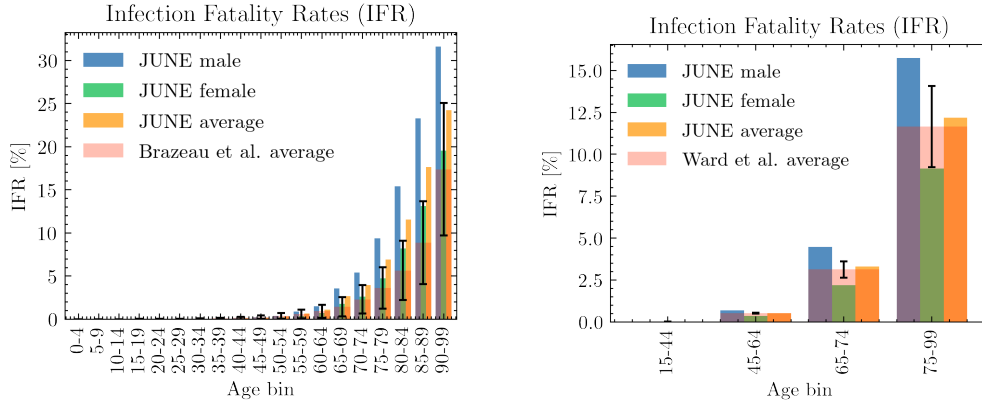
(a) IFR comparison of JUNE with NF Brazeau (2020)

(b) IFR comparison of JUNE with Ward et al. (2020)

Figure A.12: IFR comparison of JUNE with various estimates of community transmission. Error bars show 95% CI on the IFRs as estimated from data.

one of the 8 trajectories listed in Figure A.11. The asymptomatic rate, $R_{I \to A}$, and the mild case rate, $R_{I \to M}$, are taken from a calibration done in Hinch et al. (2020) from Pollán et al. (2020) and Riccardo et al. (2020). To calculate the different hospitalisation and fatality rates, we have used a series of datasets listed in Table A.2, all of them containing data until 13th July, 2020, to be consistent with the considered seroprevalence values. In order to avoid possible irregularities in our results derived from the use of different data sources, we normalise all our death data to the ONS reported numbers of total deaths (51,443), hospital deaths (32,164), and residence deaths (19,279),[10] and then use more granular data to distribute deaths by age and sex for each place of death occurrence (Public Health England (PHE), 2020; Scientific Advisory Group for Emergencies, 2020). Likewise, the total number of hospital admissions is taken from NHS, and distributed by age, sex, and residence type also using Public Health England (PHE) (2020) and Scientific Advisory Group for Emergencies (2020). The number of deaths in care homes reported in Office for National Statistics (a) is only reported by age until late June, so we assume that the distribution does not change until the 13th July, 2020. We also ensure that we correctly account for differences in reporting

---

[10]https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/
deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales

times. As a first step, we calculate the overall infection fatality rate (IFR) for the general population outside care homes (GP),

$$R_{\text{I}\to\text{D}}^{\text{GP}}(p) = \frac{N_{\text{D}}(p) - N_{\text{D}}^{\text{ch}}(p)}{N_{\text{tot}}(p) - N_{\text{tot}}^{\text{ch}}(p)}, \tag{A.13}$$

which can be directly compared to the results from the REACT2 study Ward et al. (2020) (right panel of Figure A.12), and the Imperical College London COVID-19 report 34 NF Brazeau (2020) (left panel of Figure A.12). The remaining rates just follow from the same methodology,

$$R_{\text{I}\to\text{X}}^{\text{GP}}(p) = \frac{N_{\text{X}}(p) - N_{\text{X}}^{\text{ch}}(p)}{N_{\text{tot}}(p) - N_{\text{tot}}^{\text{ch}}(p)}, \tag{A.14}$$

$$R_{\text{I}\to\text{X}}^{\text{CH}}(p) = \frac{N_{\text{X}}^{\text{ch}}(p)}{N_{\text{tot}}^{\text{ch}}(p)}, \tag{A.15}$$

$$\tag{A.16}$$

where X refers to one of deaths or hospital admission in the normal hospital ward or in the ICU/ITU. The rate of non hospital deaths is computed by subtracting the hospital death rates from the overall IFRs. Finally, the probability of having severe symptoms but recovering at home is given by:

$$R_{\text{I}\to\text{S}}(p) = 1 - \sum_{i \neq S} R_{\text{I}\to\text{X}_i}(p). \tag{A.17}$$

The results of computing the individual infection outcome rates by age, sex, and residence type are shown in Figure A.13. The most important visible difference is the disparity on the fatality rates between care home residents and the general population. This could be the reflection of various reasons, including, for example, a generally poorer health condition of the care home population, or differences in admission policies to hospitals. Consistent with the ONS data Office for National Statistics (a), most of the care home deaths occur within the care home residence itself, while the probability of being admitted to the hospital decreases with age. Likewise, both for the general population and the care home population, people aged 55-70 years old are the group most likely to be admitted in the ICU/ITU.
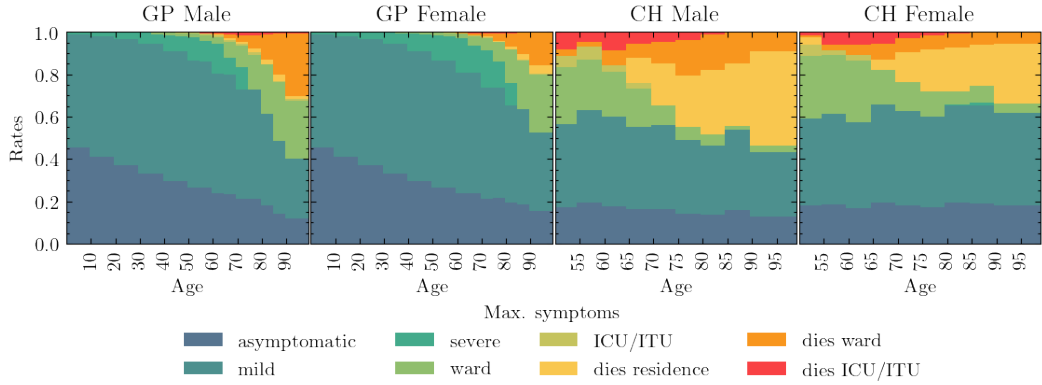
Figure A.13: Rates of different infection outcomes for males and females living in households and care homes. For care home residents, we only show the rates for people aged over 50, as the younger ones are assumed to follow the general population rates.

| Trajectory | Stages | |
|---|---|---|
| asymptomatic | $I[\beta_I]$ $A[C_{14}]$ | R |
| mild | $I[\beta_I]$ $M[C_{20}]$ | R |
| severe | $I[\beta_I]$ $M[C_{20}]$ $S[C_{20}]$ | R |
| death at home | $I[\beta_I]$ $M[LN_M]$ $S[C_3]$ | D |
| ward | $I[\beta_I]$ $M[LN_M]$ $H[\beta_H]$ $M[C_8]$ | R |
| death in ward | $I[\beta_I]$ $M[LN_M]$ $H[\beta_D]$ | D |
| ICU/ITU | $I[\beta_I]$ $M[LN_M]$ $H[LN_{ICU}]$ $ICU[e_{ICU}]H[e_H]$ $M[C_3]$ | R |
| death in ICU/ITU | $I[\beta_I]$ $M[LN_M]$ $H[LN_{ICU}]$ $ICU[e_D]$ | D |

Table A.3: List of different trajectories through disease progression, with stages and, in brackets, the distribution from which corresponding timings are drawn. For their definition see B.3. The available stages are Infected, Asymptomatic, Mild and Severe symptoms, admitted to a regular Hospital or an ICU/ITU ward, and, finally, as outcomes, Recovered or Dead.

Females are less likely in general to develop a severe infection of COVID-19, with fatality rates roughly equivalent to those of a male 5 years younger.

Once an infection outcome has been determined, the infected individual follows a symptoms trajectory composed of different stages. The time spent at each stage is sampled from different distributions derived from different data sources. In Table A.3, we list the different stages per trajectory by infection outcome, and the details on the various timings are listed in B.4. In the left panel of Figure A.14, we show the probability density functions for the incubation time, and the time to die or recover in hospital.
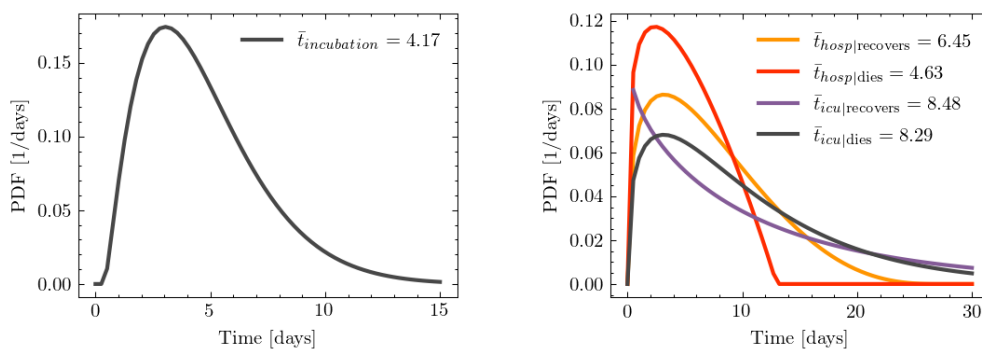
(a) Time taken for an infected individual to develop symptom.

(b) Time spent in hospital by patients given their infection.

Figure A.14: Probability density functions for symptom and progression timing.

## A.5.3 Seeding Infections

In the absence of sufficiently detailed knowledge of how epidemics arrive in a country, we seed infections using secondary information such as the number and regional distribution of observed cases. In the example of the simulating the spread of COVID–19 in England, we use the number of COVID–19-related deaths recorded in hospitals to estimate initial infection numbers and their regional distribution. Accounting for the time delay between infection and possible death, and for the probability of admitted patients to die, we have

$$N_{\text{tot}}(t, x) = \frac{1}{\bar{R}_{\text{H} \to \text{D}}(x)} \, N_{\text{H} \to \text{D}}(t + \Delta t_{\text{D}}, x) \,, \tag{A.18}$$

where $N_{\text{tot}}(t, x)$ is the estimated number of cases in a region, $x$, on day, $t$, $N_{\text{H} \to \text{D}}(t, x)$ is the number of observed deaths in the region at date $t$, and $\bar{R}_{\text{H} \to \text{D}}(x)$ is the rate for people dying in hospital in the region $x$, where the average is over the characteristics $p$ is given by:

$$\bar{R}_{\text{H} \to \text{D}}(x) = \frac{1}{N_{\text{H} \to \text{D}}(t + \Delta t_{\text{D}}, x)} \sum_{i \in N_{\text{H} \to \text{D}}(t + \Delta t_{\text{D}}, x)} R_{\text{H} \to \text{D}}(p_i) \,. \tag{A.19}$$

The relatively large statistical fluctuations in the initial phase of an epidemic, and possibly differing time profiles across regions, translate into the need for a region-specific seeding. This difference is highlighted by contrasting the seeding for London, where we introduce initial infections over two days only ($28^{th}$-$29^{th}$

February 2020) with the North East of England and Yorkshire, where we seeded infections for a week, $28^{th}$ February - $5^{th}$ March 2020. We introduce the estimated number of daily cases in each of the regions until the following criterion is met,

$$N_{\text{tot}}\left(t < T(x), x\right) > 0.1 N_{\text{tot}}(t_{\max}, x) \tag{A.20}$$

where $T(x)$ is the number of days over which we seed new infections in region $x$, and $N_{\text{tot}}(t_{\max}, x)$ is the maximum number of cases that region $x$ would reach in any given day, estimated from the maximum number of daily deaths in hospital. It is important to define the seeding for the infection based on the maximum number of cases each region will have, since the different regions are experiencing different stages of the epidemic at any given time.

## A.6 Mitigation Policies and Strategies

Policies and interventions, often enacted by governing bodies, are introduced in an attempt to mitigate and control the spread of infectious diseases. In general, such policies are highly dependent on the type of infection and social norms in the affected population, and may include guidelines on how to change individual patterns of behaviour or the closure of certain venues where transmission is estimated to be highly likely. The modular nature of JUNE allows policies to be dynamically activated and deactivated at different points in time to allow for changes in policy decisions. Due to JUNE's granularity, these policies can be implemented at a highly localised level: by type and place of social interactions, by geographical region, by industry sector or venue type. JUNE can also model the population's compliance with the measures, again with high granularity. In this section we present a variety of policies which can be implemented in JUNE and exemplify their application through those measures that have been enacted by the UK Government to mitigate the spread of SARS-CoV-2.
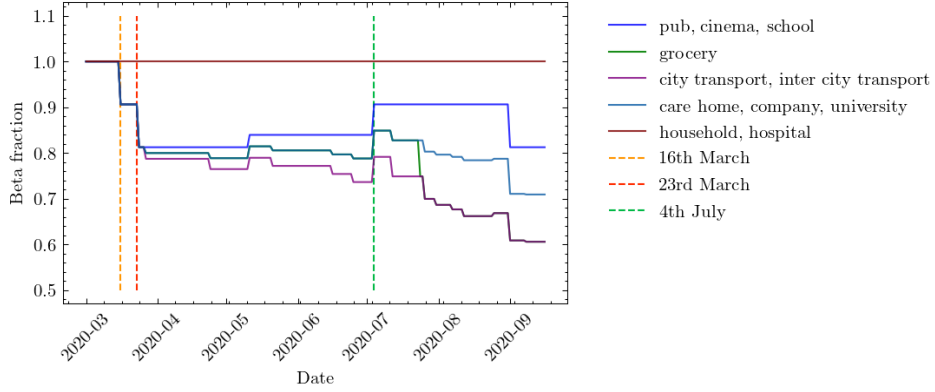
Figure A.15: Example scenario of different intensity parameters, $\beta^{(L,g)}$, over time normalised to unity (see Equation A.7). The parameters change due to the effects of compliance with social distancing and mask wearing advice and regulations.

## A.6.1 Behavioral Changes

There are a variety of changes in behavioral patterns that are designed to reduce the probability of viral transmission, ranging from simple social distancing, increased hygiene and mask wearing, to quarantining of infected individuals or those who have been in sufficiently close contact with them, and the shielding of vulnerable parts of the population. We model the impact of the former set of measures, social distancing, increased hygiene and mask wearing, through multiplicative reductions in the location-specific contact-intensity parameters, $\beta^{(L,g)}$, see Figure A.15 for an example. The impact of compliance with social distancing and other, similar measures can be recorded both nationally and sometimes even in specific locations. This allows us to calculate the reduction in the corresponding intensity parameters as follows:

$$\beta^{(L,g)} = M^{(L,g)}\beta^{(L,g)} \tag{A.21}$$

$$= \left[1 - C^{(N)} \cdot C^{(L)} \cdot (1 - E)\right]\beta^{(L,g)}, \tag{A.22}$$

where $M^{(L,g)}$ is the location and group specific modification factor, $C^{(N)}$ is the national compliance (i.e. percentage of the population following guidelines), $C^{(L)}$ is the compliance in a given location or social setting $L$, and $E$ denotes the efficiency of the measure. Quarantining is simulated by keeping the individuals in

question in their homes for a certain amount of time, and allowing them to interact with members of their household in an otherwise unchanged household setting only. In JUNE, we have the ability to apply different policies to those with mild and severe symptoms, and to quarantine household members of symptomatic individuals. Similarly, JUNE also allows the definition of vulnerable individuals – typically by characteristics such as age – and of a prescription of how shielding policies are enacted relative to this group.

We will now turn to discuss our choices for specific measures. There have been a variety of studies on the effectiveness of social distancing with respect to COVID–19 and other infectious diseases. A comprehensive systematic review and meta-analysis (Chu et al., 2020) suggested that the relative risk of infection decreases by approximately a factor of 2 per meter distance. In practice, however, the efficiency of social distancing is highly dependent on external factors, in terms of both physical and social environment. We therefore use this literature as a benchmark, assuming on average 1 meter social distancing, $E = 0.5$, and fit the effects of social distancing to data where possible (see A.6.3).

We simulate mask wearing according to Equation A.22, i.e. by multiplicatively reducing the $\beta$ parameters in different locations. There is a significant body of literature on the effectiveness of mask wearing, including differences based on the material of the mask and the locations in which they are worn (Fischer et al., 2020; Chu et al., 2020; Howard et al., 2020), as well as changes in efficiency due to re-using or washing them (Suen et al., 2020; Toomey et al., 2020). In general, we focus on the wearing of masks by non-healthcare workers in settings outside the home and estimate mask effectiveness, $E$, to be 50% (Liang et al., 2020), irrespective of the specific of the location. However, after adjustments for compliance the actual, intensity parameter reduction may be much lower than this, which leads us to believe that this represents a conservative estimate.

In JUNE, quarantining of infected people with mild or severe symptoms is relatively straightforward: afflicted individuals do not leave their household for a pre-defined

period of time – usually 7 to 14 days – but do not change interaction patterns with the residents in their household. In some versions of quarantine policies, household members must also stay at home and isolate themselves. This is modelled in June along the same lines, only the possibly time-dependent compliance of the population with quarantine measures. Clearly, infected individuals with severe symptoms will always stay at home, until they are either recovered, moved to hospital, or died. It should be noted that quarantine sits on top of other, less individual-driven policy interventions included in June which may restrict movement, such as the closure of companies and leisure venues.

Given the additional danger infectious diseases may pose to the more vulnerable and elderly populations, various policies, usually referred to as "shielding" can be introduced with an aim to protect these individuals. In June, shielding is realised similar to quarantine: vulnerable individuals – usually defined by their age or other characteristics – stay at home and do not interact with others outside their household. Apart from the definition of relevant characteristics, this only leaves a compliance probability to be introduced, which reduces the participation in any other social settings other than households.

## A.6.2   Closure of Venues

Mitigation strategies that aim at reducing infection transmission through changes in individual behaviour may have to be further supplemented through partial or complete closure of certain parts of public life such as companies, transport, schools and universities.

Starting with the closure of companies, June can realise this important measure in a sector–specific way. June allows the definition of "key" and "furloughed" workers, again in a sector–specific way. While the former represent those parts of the work force that continue with their essential work as usual, the latter never goes to work and instead is given the chance to take part in available leisure activities or stay

at home during regular working hours. For the rest of the work force, JUNE allows the definition of flexible work patterns by assigning daily probabilities for workers go to their companies.

School and university closure is handled similarly to the closure of companies in JUNE. However, in the case of schools we are able to close individual year groups as well as entire schools, and we can identify the children of key workers and have them continue going to school. Since the return to in-person schooling may also be voluntary at certain points, or children may only go to school on certain days of the week, we also can apply a compliance factor at the year group level which is used to probabilistically determine which children attend school on any particular day.

In addition to the partial or complete closure of companies in some industry sectors and of schools or universities, government policies may also close or limit the number or people attending leisure venues, such as restaurant and pubs, cinemas, or similar. In JUNE we are able to fully or partially close different types of leisure venues either nationally or at a more local level, down to super areas. Partial closure is enacted through a change in the probabilities that people attend different venues, which is both sex and age dis-aggregated. Modifications to other leisure activities, such as household visits, are also simple to realise in JUNE, by directly modifying the daily probabilities for such activities to take place.

## A.6.3   Policies in the UK

The population, interaction and disease layers of JUNE will have time inhomogenous states and parameterisations dependent on the public policy response and the response of individuals in changing behaviour as an infection spreads. For ex-post analysis, policies can be imposed on the simulation using a set of policy levers with varying effectiveness. For ex-ante prediction, scenarios of responses to different policy combinations can be realtime of pseudo out-of-sample forecasted.

To simulate, ex-post, the spread of COVID–19 in England we impose a set of policies restricting movement and attempting to reduce transmission. Table A.4 lists the operational policy interventions enacted by the UK government from the beginning of March 2020 to October 2020 in an effort to reduce the spread of SARS-CoV-2.

In order to estimate the effects of social distancing on the epidemiological development of COVID–19, we implement multiple staggered social distancing steps during the first wave of the pandemic between 16th March - 4th July, 2020 and then again going into September 2020 as schools and universities begin to fully reopen. We fit the national compliance, $C^{(N)}$, with social distancing between 24th March - 11th May, 2020 in the range 20-100% when fitting the rest the parameters (see A.7). This is taken to be the harshest social distancing step against and others are determined relative to this fit. The location specific compliance, $C^{(L)}$, is set to be 100% in all locations during fitting to avoid parameter degeneracy and then altered manually thereafter. No social distancing is assumed between household members. We derived the compliance with mask wearing from a YouGov survey YouGov (2020b), and we further stratify the results by social environment or locations. Specifically, we assume complete (100%) compliance with mask wearing during commuting, 50% in care homes and no compliance in pubs, schools or in the household. Compliance with mask wearing in grocery stores is assumed to be at 50% before 24th July, 2020, after which we assume complete compliance given the change in government regulations. Since we already assume low intensity parameters in hospitals due to the significant amount of personal protective equipment (PPE) being worn in these scenarios, we do not apply any additional mask wearing in these settings.

On 16th March, 2020, the UK Government encouraged people with COVID-19 symptoms to quarantine in their household for 7 days and all those in their household to quarantine for 14 days from symptom onset. We assume that compliance with this measure varies with time as people become more aware of the dangers

| Date (dd/mm/yy) | Policy | Implemented |
|---|---|---|
| 04/03/2020 | Encourage increased hand-washing | |
| 12/03/2020 | Case isolation at home | * |
| 16/03/2020 | Voluntary household quarantine | * |
| 16/03/2020 | Stop all non-essential travel | ** |
| 16/03/2020 | Stop all non-essential contact | ** |
| 16/03/2020 | Voluntary working from home | * |
| 16/03/2020 | Voluntary avoidance of leisure venues | * |
| 16/03/2020 | Encourage social distancing of entire population | * |
| 16/03/2020 | Shielding of over 70s | * |
| 20/03/2020 | Closure of schools and universities | * |
| 21/03/2020 | Closure of leisure venues | * |
| 21/03/2020 | Stopping of mass gatherings | ** |
| 23/03/2020 | 'Stay at home' messaging | ** |
| 11/05/2020 | Multiple trips outside are allowed in England only | |
| 13/05/2020 | Encouraged to go back to work if they can while distancing | * |
| 01/06/2020 | Meeting in groups of up to 6 outside allowed | ** |
| 01/06/2020 | Shielding of over 70s relaxed | * |
| 01/06/2020 | School reopening for Early Year and Year 6 students | * |
| 13/06/2020 | 'Support bubbles' allowed | |
| 15/06/2020 | School reopening for Year 10 and 12 students for face-to-face support | * |
| 04/07/2020 | Leisure venues allowed to reopen | * |
| 04/07/2020 | Household-to-household visits permitted along with overnight stays | * |
| 24/07/2020 | Mask wearing compulsory in grocery stores | * |
| 01/08/2020 | Shielding is paused | * |
| 01/08/2020 | 'Eat Out to Help Out' scheme introduced | * |
| 31/08/2020 | 'Eat Out to Help Out' scheme ends | * |
| 01/09/2020 | Schools and Universities allowed to reopen | * |
| 01/09/2020 | 'Rule of 6' introduced | |
| 14/10/2020 | Tiered local lockdown system introduced | * |

Table A.4: List of policies introduced in England by the UK Government at different points in time. * indicates policies directly implemented in the model, ** indicates policies which are indirectly implemented — i.e. other policies effectively implement this one by default.

of COVID-19. Between the 16th March - 23rd March, 2020 (i.e. the week leading up to the nationwide 'lockdown') we fit compliance with the quarantine policy of those symptomatic to be between 5-45%, and the probability that the rest of the

household of a symptomatic individual complies is set to the same fitted value. After 'lockdown' comes into effect, the government tightened these rules to only leave the house for essential trips and one form of exercise per day. To account for this, we increase the symptomatic and household compliance with quarantine to be double their fitted value. In addition, the UK government strongly suggested that people over the age of 70 were to shield, from 16th March, 2020. As in the case of quarantine, we assume people become more compliant with this policy over time and that the initial compliance with the shielding policy for this age bracket increased from 20% in the first week to 70% afterwards. Indeed, one of the reasons the compliance was set to only 70% even after lockdown is due to the fact that people in this age bracket already have a reduced mobility and interaction potential. A 70% compliance therefore still allows them a small chance to interact with others, e.g. in grocery stores, and any higher compliance figures would mean a complete and unrealistic decoupling of this critical population from any social interactions. The shielding policy initially runs until 1st August, 2020 and after which the UK government paused the policy.

To model the partial or complete closure of industry sectors, it is important to understand the descriptions of key workers provided by the UK Government (Office for National Statistics, 2020c), and match these up with the relevant 5-digit SIC codes (Office for National Statistics, 2007). This ultimately allows us to deduce the proportion key workers in each sector and assign the corresponding key worker attribute probabilistically according to these proportions. In our simulation we encode findings from the ONS (Office for National Statistics, 2020c), reporting that 33% of the total workforce were key workers in 2019 with 14% able to work from home. We therefore set the proportion of key workers, i.e. those who go to work each day, at 19% of the workforce. We use the same logic to also decide which workers are furloughed in June by identifying the 5-digit SIC codes of the relevant affected industries and proportionally assigning the relevant percentage of a given sector to be furloughed. We derive the relevant SIC codes from the

Institute for Fiscal Studies in the UK (Institute for Fiscal Studies, 2020), and we dynamically correct for any over or underestimation of furloughed workers by defining the proportion of the workforce who should be furloughed at any particular time, derived from Government reports (HM Revenue & Customs, 2020). A similar dynamic correction is also applied to the key work force. To model the more random work patterns of the remaining part of the work force, we derive a probability that a random worker goes into the company for work from a YouGov survey (YouGov, 2020a). We note that in many surveys, including this and others undertaken (e.g. by the ONS (Office for National Statistics, 2020b)), the methodology does not explicitly state if key or furloughed workers were included. We believe, however, that our use of these surveys presents at least a conservative estimate of work attendance.

From 20th March, 2020, all schools and universities in England were asked to close, with the exception that children of key workers could still attend school. To account for the partial school reopening of Early Years (nursery and reception age children) and Year 6 students on 1st June 2020, we open up these year groups in June with an attendance compliance based on data derived from the Department for Education (DfE) (UK Department for Education, 2020). While the government also asked schools to offer face-to-face support for Year 10 and 12 students from 15th June, 2020, we do not include this as the sessions were generally limited and had an attendance rate around the 10% level (UK Department for Education, 2020). Figure A.16 shows the good agreement in the number of children attending school as derived from June compared with DfE data. The slight deviation from data after 15th June, 2020, can be explained by not fully capturing the partial return of Year 10 and 12 students. The good agreement between June and the DfE data before 1st June 2020, is of particular note since this option was available only for children where all parents in the household were classified as key workers. This serves as an implicit partial validation of our method of selecting which individuals are key workers, as well as the household and company sector distribution
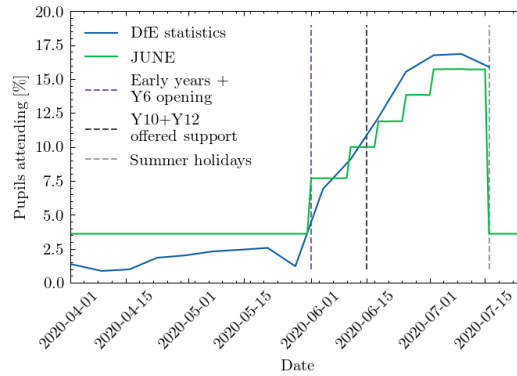
Figure A.16: School attendance in JUNE compared to data collected by the UK's Department for Education (UK Department for Education, 2020)

algorithms. From 1st September, 2020, we reopen schools fully in JUNE, while accounting for a closure for the national school holidays. While the timings of this week-long holiday varies across the country, we assume all schools share the same holiday period 26th October - 30th October, 2020. Similarly, universities are opened from 1st September, 2020, but with more restrictive social distancing measures in place. Given the modelling of where university students live, their inter-mixing is naturally captured in the household component of JUNE (see A.3.2).
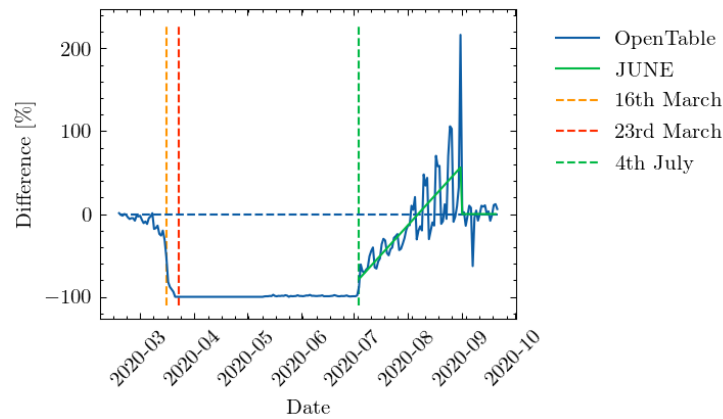


Figure A.17: Year-on-year restaurant attendance from OpenTable (OpenTable, 2020) including a fit to the simulated reopening change in probabilities used to derive the probability that people attend restaurants in JUNE.

On 16th March, 2020, the UK Government encouraged people to avoid going to leisure venues such as bars and restaurants, although this rule was not imposed through the closure of such venues. However, on the 21st March, 2020, this closure

took place. We model these policies first by reducing the probability that people leave the house from 16$^{th}$ March, 2020 followed by the closure of all relevant leisure venues included in the simulation – cinemas, pubs, and restaurants – from 21$^{st}$ March, 2020. Visits to care homes are also halted from this time. Since many of these venues were permitted to reopen from the 4$^{th}$ July, 2020, we assumed all venues reopen at this point. Additionally, data collected by OpenTable suggests that restaurant attendance after that date saw a significant increase likely encouraged by the UK Government's 'Eat Out To Help Out' scheme which we capture in JUNE (see Figure A.17) (OpenTable, 2020; Government of the United Kingdom, 2020). For the simulation of other leisure activities, and in particular household-to-household visits, we assume a drop in compliance and a consequently increasing number of such visits. In line with data collected by the ONS (Office for National Statistics, 2020d), we model this be gradually increasing the probability of visiting another household from mid-May until 4$^{th}$ July, 2020, when overnight visits were permitted.

## A.7  Discussion of model outputs

In this section we finally highlight the ability of JUNE to capture intricate social dynamics through a number of model outputs. It is worth noting that the realisations of JUNE presented in the following were run at parameter settings sampled from the "non-implausible" region of the global parameter space, as defined in Section A.8 and Appendix B.5. See B.4 for the ranges of the global parameter space. A more complete uncertainty analysis and parameter exploration will be performed in Bullock et al. (2021).

In Figure A.18 we exhibit results for the number of daily deaths in hospital for regions of England and England itself. In addition, in Figure A.19 we show the same realisations for daily deaths in England stratified by age. The agreement with data is satisfying and while there are minor discrepancies for certain outputs, we
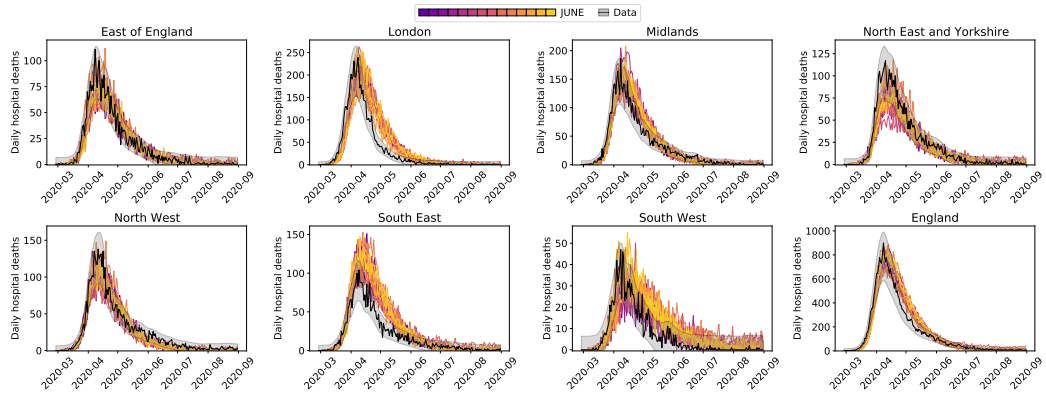
Figure A.18: Daily hospital deaths for each region in England, and England itself, for 14 realisations of JUNE as described in this section. Each realisation is illustrated as a separate colour for visibility. Observed data in black with 3 standard deviation error bands. Data from CPNS (CPN).
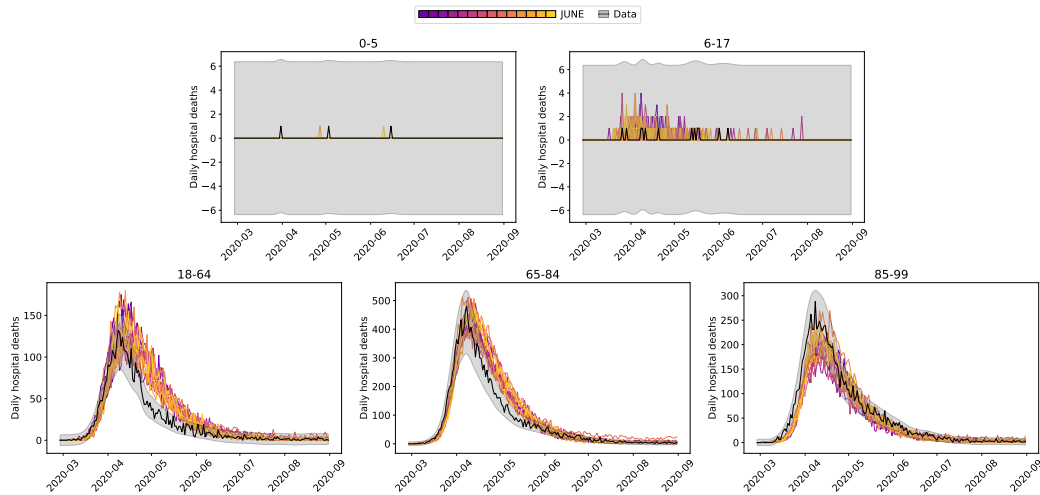


Figure A.19: Daily hospital deaths in England stratified by age, for the same realisations as in Figure A.18. Observed data in black with 3 standard deviation error bands. Data from CPNS (CPN).
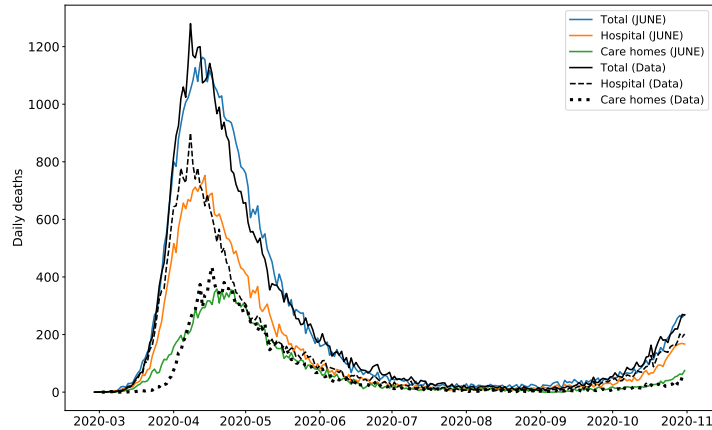
Figure A.20: Deaths in England illustrated as different lines for total deaths, hospital deaths, and deaths within care homes. Note that the total curve is the sum of hospital deaths, and residence deaths (care homes as plotted, and usual households which are not plotted). Data from ONS (Office for National Statistics, c).

would like to stress that all of these outputs are simultaneously fit by JUNE without any region-specific parameters.

Along with deaths in hospitals, there have been a non-negligible number of fatalities in care homes in England during this pandemic. JUNE successfully models both deaths in hospitals, and deaths within care homes as illustrated in Figure A.20 where there is good agreement with data even into the second wave of the pandemic.

We would like to emphasise that the outputs shown here are illustrative of the capabilities of JUNE to capture the social dynamics of a heterogeneous population giving rise to large differences in disease spread to different age strata and regions.

All interactions resulting in infections are stored in full detail in the model's output, enabling further ex-post analysis of the sociological nature of disease spread and outcomes for all individuals modelled in the simulation. A simple example of such an analysis is shown in Figure A.21 where locations of infections are compared for one of the realisations shown in Figure A.18. Remaining realisations manifest a similar hierarchy of infection locations demonstrating JUNE's physical consistency across parameter space. A further, more involved, example of this type of detailed ex-post analysis can be found in IDAS Covid group (2021).
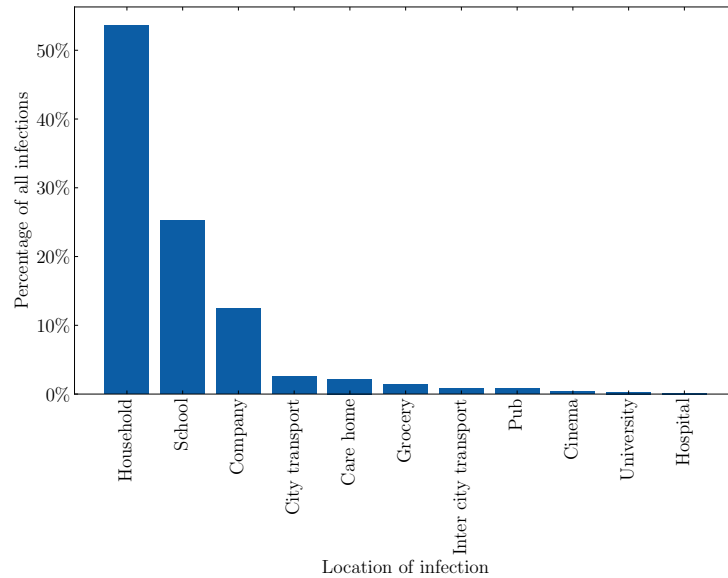
Figure A.21: Locations where infections take place in one realisation of June from Figure A.18. This is a simple illustrative example of the type of analysis that you can carry out using the detailed outputs of June.

## A.8 Fitting via Bayesian Emulation

We now discuss efficient calibration strategies which form a critical part of our ability to extract core insights from June. Fitting a complex model such as June to observed data presents a challenging task. This is mainly due to a) the detailed nature of June and the inevitable computational expense of performing model evaluations, b) the large number of input parameters that we may wish to explore, c) the stochastic nature of the output of June and d) the various uncertainties present in the comparison between model and data. A typical full England run of June like those shown in Figure A.18 would take approximately 10 hours to complete on 64 cores (Intel Xeon Skylake) and 128GB of memory. The combination of computational expense and high dimensional input parameter space precludes the use of many parameter exploration methods that rely upon large numbers of model evaluations (including many standard optimisers, sampling approaches such as MCMC, etc.). The stochastic output, which implies we will be exploring a much more complex surface, requires methods developed to deal with stochastic functions. Even

more challenging is that the substantial uncertainties present imply that we may not even want to optimise for a single "best fit to data" as it may have limited statistical relevance, but instead search for the set of all input parameter values that give acceptable matches between model output and observed data, thereby fully capturing the induced parametric uncertainty.

We hence employ the Bayes linear emulation and history matching methodology (Craig et al., 1997; Vernon et al., 2010a; Andrianakis et al., 2015), a widely applied uncertainty quantification approach designed to facilitate the exploration of large parameter spaces for expensive to evaluate models of deterministic or stochastic form. This approach centres around the concept of an *emulator*: a statistical construct that mimics the slow to evaluate scientific model in question, providing predictions of the model outputs with associated uncertainty, at as yet unevaluated input parameter settings. In contrast to the model, the emulator is extremely fast to evaluate: for example, in the case of JUNE, the emulator exhibited a speed increase of nine orders of magnitude. The emulator provides insight into the model's structure and, thanks to its speed, can be used to perform the global parameter search far more efficiently than approaches that attempt to use the comparatively slow scientific model itself. Here we give a brief overview of emulation and history matching, but for more details see Appendix B.5. See also Andrianakis et al. (2015, 2017b,a); McCreesh et al. (2017) for further examples of its application within epidemiology, McKinley et al. (2018) for a comparison to Approximate Bayesian Computation in an epidemiological setting, and Vernon et al. (2018) for a tutorial introduction in the context of systems biology. For an extensive treatment see Vernon et al. (2010a) along with the discussion in Vernon et al. (2010b). See also O'Hagan (2006) for a general introduction to emulation.

Initially, we identify a large set of input parameters to search over, primarily composed of interaction intensity parameters at the group level, along with associated broad ranges, as given in Table B.4. We then identify a set of particular model outputs to match to corresponding observed data. Here we focus on hospital deaths

(CPNS CPN) and total deaths (ONS) at well-spaced time points throughout the period of the first wave of the epidemic. We then construct Bayes linear emulators for each of the model outputs at each of the chosen time points. The emulators are trained using a set of JUNE runs, initially designed using a 18-dimensional Latin hypercube, and seek to mimic the behaviour of each of the JUNE outputs as a function over the 18-dimensional parameter space. The emulators provide, at each unevaluated input location, an expectation for the possible JUNE model output value and a position dependent variance representing the emulator's uncertainty about this estimate. Close to known runs the emulator's uncertainty will be low, however it will increase appropriately as we move to less well explored regions of the parameter space (Vernon et al., 2018). Note that we deliberately choose to emulate the direct physical outputs of the model as this has multiple benefits for emulator construction, in contrast to emulating a combined metric such as the likelihood (for discussion of this point see Vernon et al. (2010a,b, 2018)).

Due to the emulators' speed, they are ideal for global parameter exploration. This is performed by constructing an implausibility measure, that gives the distance between the emulator's expected JUNE model output and the observed data we are trying to match, standardised by all the major uncertainties present: observational errors, emulator uncertainty and structural model discrepancy, the latter being a direct acknowledgement that the model is an imperfect representation of reality (see Appendix B.5 for details). The implausibility measures are used to rule out large regions of the input parameter space that will not provide acceptable matches, and the analysis then proceeds in iterations: a second batch of JUNE runs is performed over the remaining region of parameter space, new emulators constructed (which are only defined over this region), new implausibility measures formed and more parameter space removed. This process is referred to as iterative History Matching (Andrianakis et al., 2015; Vernon et al., 2010a). See for example Andrianakis et al. (2017a) where it was successfully applied to a stochastic disease model with 96 input parameters.

For the June model, we constructed emulators for hospital deaths and total deaths at eight time points over the period March to June, for England and for each of the seven regions, and for the age bins (defined by the SITREP dataset) $0-5, 6-17, 18-64, 65-84, 85^+$. The emulators were trained in three iterations formed from 125 June evaluations each. The emulators were then evaluated at 500,000 locations across the 18-dimensional input space, taking 10 minutes on a single processor. The results of the global parameter search are given in the optical depth plots (Vernon et al., 2010a) of Figure A.22, which shows the location of the "non-implausible" region of interest in various 2-dimensional projections of the 18-dimensional parameter space for all combinations of 12 of the most interesting input parameters (the remaining six inputs were only loosely constrained, jointly with other parameters, if at all). The June runs discussed in the preceding section were sampled from this region. Note the various joint constraints on the input parameter space imposed by the matching process, for example the strong reciprocal relationship that is required between $\beta_{school}$ and $\beta_{household}$. Similar but more complex trade-offs are identified between several other parameters e.g. $\beta_{company}$ and $\beta_{houshold}$; $\beta_{grocery}$ and $\beta_{citytransport}$, and between $\beta_{household}$ and $M_{quarantine\ household\ compliance}$. Most parameters were not individually identifiable, however, $\beta_{company}$ and $\beta_{carevisits}$ were reasonably well constrained. For more details of this approach, of emulator diagnostics, and further output plots see Appendix B.5.

We can see that the Bayes linear emulation and history matching methodology facilitates the efficient exploration, development and calibration of the highly complex June model using a modest number of runs, a process which would be extremely challenging to perform directly. While here we have performed a provisional exploration of the parameter space as part of the model development, for a full uncertainty analysis of the June model, including the emulator driven generation of full probabilistic forecasts incorporating all major sources of uncertainty, see Bullock et al. (2021).
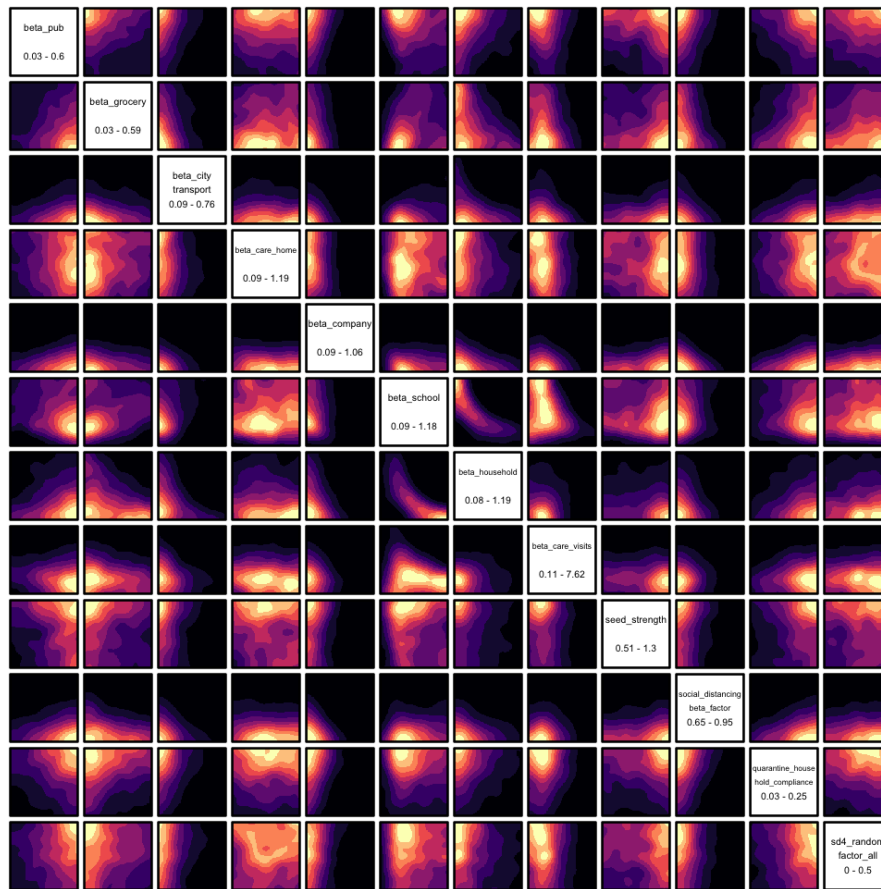
Figure A.22: 2-dimensional projections of the 18-dimensional input space, for the 12 most interesting input parameters, coloured by the optical depth of the non-implausible region, which gives the depth or thickness of the non-implausible region conditioned on the two given inputs (Vernon et al., 2018). The ranges for each parameter are given below the parameter name in the diagonal panels. These plots are formed from 500,000 emulator evaluations over the input space. The emulators were trained on three iterations of 125 June model evaluations.

## A.9 Summary

In this paper we introduced the new June model to simulate the spread of epidemics through a population. June is an individual-based model (IBM) enabling a highly granular geographical and sociological resolution. The frequent and persisting perception that IBMs such as June are heavily parameterised and therefore lack predictive power is misleading. As noted in Marathe and Ramakrishnan (2013), many of the properties and building blocks of these types of model are not globally

fitted to observed cases or fatalities, as is the case for deterministic and stochastic models built from differential equations. Instead, the JUNE framework separates the uncertainty arising from unknown disease dynamics from uncertainties in the population structure, where the latter is informed by demographic statistics and other available data.

The model is formulated and encoded in four distinct layers, `population`, `interaction`, `disease`, and `policy`. Its modular structure allows not only the flexible and seamless addition of many details and novel features, but it also lends itself to application to other populations with different sociological setups. As a first example we discuss its application to the case of the spread of COVID–19 in England, with convincing results underlining the quality of the model and its ability to understand the spread of an epidemic in great detail and with high geographical and sociological resolution.

Studies where JUNE is applied to different settings are forthcoming (Aylett-Bullock et al., 2021). One of the strengths of the model is its ability to capture differences in geographical and sociological structure with unprecedented resolution, facilitated through the hierarchical structure in which the `population` is organised. JUNE also allows a flexible yet detailed modelling of daily activities of the virtual population, by combining the geographical position of buildings and other structures with the social interactions taking place. In contrast to other models this enables a very granular understanding of work patterns, leisure activities, etc.. In forthcoming publications we will exploit this high level of detail to try and answer pertinent questions relating to social imbalances in the impact of COVID–19.

## Acknowledgements

---

[11]https://ddis.physics.dur.ac.uk/

This paper made use of Python (Van Rossum and Drake, 2009) and the following Python libraries: Matplotlib (Hunter, 2007), Numpy (Harris et al., 2020), Pandas (pandas development team, 2020; Wes McKinney, 2010), Scipy (Virtanen et al., 2020), SciencePlots (Garrett, 2020).

# JUNE Appendix

# B.1 Algorithms

## B.1.1 Constructing Credible Households

The ONS divides households into the following broad categories: single, couple, family, student, communal, and other (Office for National Statistics, 2011j). We populate the households in this ordering, giving preference to those types for which we have the most precise and unambiguous data.
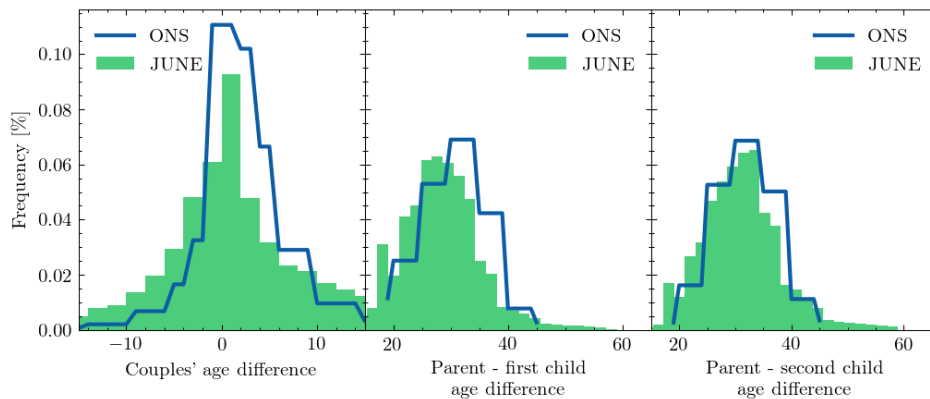


Figure B.1: Distributions of age differences between partners (left), between parents and their first (middle) and second child (right): outputs of June compared with the input data from the ONS database.

We define and construct households types as follows:

1. Single: These are households with a single person living in it. The census data differentiate single households occupied by an adult or an older adult ($\geq 65$ years old), and we fill the households accordingly.

2. Couple: These are households occupied by a couple without children. Again, the census differentiates between household with adults or older adults living in them. We preferentially fill these households with two people of different sex, with an age difference sampled from the corresponding UK distribution of age differences at the time of marriage (Office for National Statistics, 2017b) (see also the left panel Figure B.1).

3. Family: These households are defined by the number of adults (singles or couples) and the number of children. A difficulty here is that the census data does not stratify beyond "two or more" children. To compensate for this, we introduce a distribution to select the number of children in these households. To fill a family household, we allocate a female adult first. If there are no female adults available (because they have already been allocated somewhere else), we chose a male adult. In case of families with two adults, we match the person with a partner, preferentially with different sex, and an age difference sampled from the same dataset we use for couples. The census data provides us with the number of dependent children for each OA (area), and we add a suitable number of children according to the age difference between the mother and the *n*-th child as given by ONS data collected on birth characteristics (Office for National Statistics, 2017a) (see also the middle and right panels of Figure B.1).

4. Students: From the census data, we know how many student households there are and how many students live in a given OA (area). We uniformly distribute students among their households, assuming a constant ratio of the number of students per household. Students are selected from the population aged between 18 and 25 years old.

5. Communal: We use census data on the number of people in an OA (area) living in a communal establishment, as well as the number of such establishments, such as care homes (Office for National Statistics, 2011d). The communal establishments are filled last after the types described above their residents will be those who do not live in any of the other household types. As in the case of student households, we assume a constant ratio of the number of communal residents per establishment.

6. Other: This category encapsulates the uncertain household compositions given by ONS. These may include groups of adults living together, multi-

family or multi-generational families. In a similar manner to the communal households, these are filled last with those people that have not yet been allocated.



(a) Household sizes.

(b) Household composition by age.
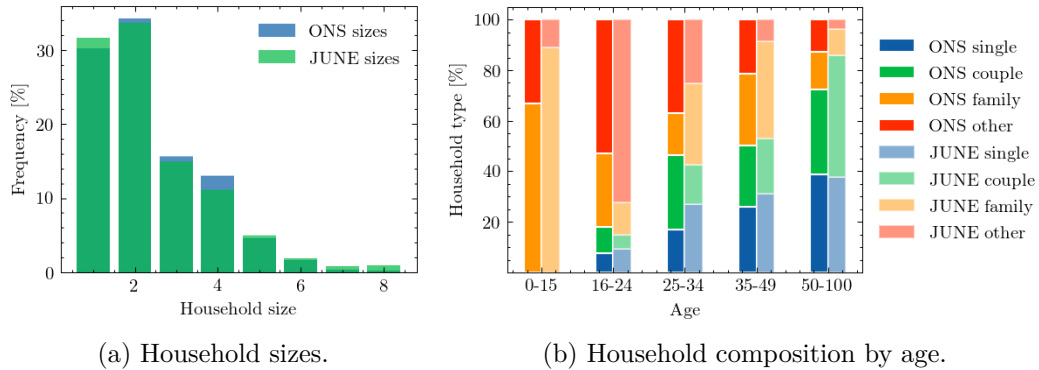
Figure B.2: Comparisons between outputs of JUNE and data from the ONS database for all England.

As a further test of our household populating algorithm against available data, we compare the JUNE household size distribution and age dependence of people living in different household types with that given by ONS (Office for National Statistics, 2011j). Figure B.2 demonstrates that the JUNE household composition algorithm clearly produces a household size distribution in good agreement with the census data. We also observe the impact of our assumptions on the composition of families and more complex household compositions ("other"). Given the unknown specifics of certain household composition types, we believe our overall household composition by age to be in reasonable agreement with the data.

## B.1.2    Schools

The procedure for assigning children and teachers to schools throughout England is specified in Section A.3.3.

Following our algorithm, we arrive at a distribution of school sizes displayed in Figure B.3, which we see to be in reasonable agreement with the data. Similarly, Figure B.4 shows the full distribution of class sizes in JUNE. In the case of COVID–
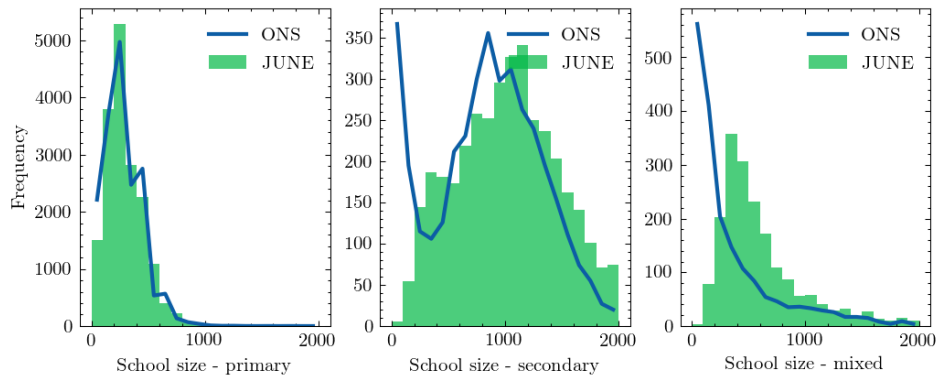
Figure B.3: Distribution of school sizes comparing the JUNE simulation with data
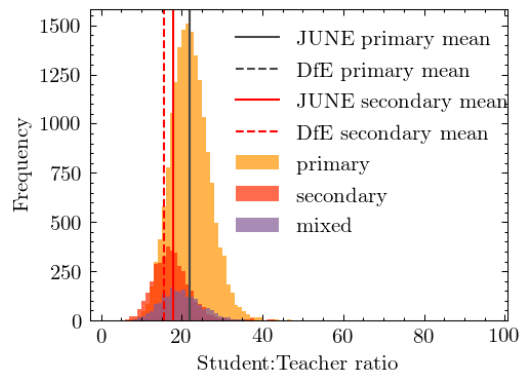


Figure B.4: Distribution of student to teacher ratios for primary schools, secondary schools, and mixed schools.

19, most countries have prioritised the return of children to school from younger age brackets. Therefore, recovering good agreement with data particularly in these age brackets is crucial.

## B.1.3 Work places

We use ONS data on industries and companies in England categorised according to 21 sectors following the Standard Industrial Classification (SIC) code convention (see Table A.1) (Office for National Statistics, 2007) as our framework for differentiating between different types of work.

Companies are initialised according to ONS data on company sizes and sectors at the MSOA (super area) level (Office for National Statistics, 2011f). We use data

on the geographical distribution of company sizes to fix the number of companies at the MSOA (super area) level and use the data on the distribution of sectors to probabilistically assign an industry sector to these companies at the same geographical level. Since the ONS provides information on company sizes by binned size ranges, we take the median size of each bin and assign this to each company. The largest bin is 1000+ employees which we assume to be 1500. It should be noted that companies are not assigned a sector based on their size, but purely on their geography. This does not mean there is no correlation between company size and their sector in JUNE, but that this would arise implicitly based on the geographical distributions, rather than explicitly from data input.

Individuals are assigned a sector attribute probabilistically, following the distributions of sectors dis-aggregated by sex at the MSOA (super area) level (Office for National Statistics, 2011e). We determine the MSOA (super area) in which they work according to the ONS commuting origin-destination matrix (or 'flow' data) (Office for National Statistics, 2011h) which provides information on the number of people by sex travelling from one MSOA (super area) to another for work. Finally, a matching is carried about between people who work in a certain MSOA (super area), and the companies available to them based on their respective sector attributes. In future work we plan to use additional demographic attributes to assign individuals their sectors and companies.

## B.1.4 Commuting

The commuting structure in JUNE is built upon the national transport network constructed from nodes representing cities, and edges representing possible transit routes. Commuters are defined as either 'internal', i.e. they live and work in a city, or 'external', i.e. they live outside a city's metropolitan boundary and commute into it (see Section A.3.4 for more details on how people are assigned locations of work). The metropolitan boundary of each city is defined using data collected by
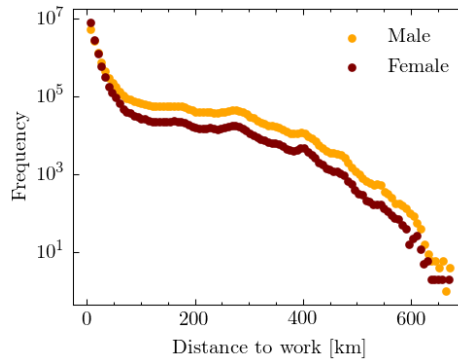
Figure B.5: Distance travelled to work by sex according to June. Here we see that men are more likely to travel further to work then women. This is in broad agreement with data presented in Klepac et al. (2018)

the ONS (Office for National Statistics, 2015) which specifies the MSOAs (super areas) that belong to the cities.

The following procedure is used to determine the groups within which people have the chance to mix during a commute.

1. For each city, we seed several additional nodes which act as 'gateway stations' outside the metropolitan area boundary. These serve as funnels into the city and determine the mixing of external commuters. In the case of London we seed eight stations which are placed evenly around the boundary of the metropolitan area. For all other stations we seed four evenly spaced stations North, South, East and West of the city boundaries. These figures are informed by the approximate and number of train lines entering each city, and the proportional differences between the number of London public transport links and those of other cities (National Rail, 2015).

2. We model the commuting of all people who travel by public transport into a city's metropolitan area. We assign all external commuters to the nearest gateway station to where they live. During each commuting time-step in the simulation, people travelling through the same gateway station are randomly split into 'carriages' containing people with whom they have the potential to

interact. Similarly, internal commuters are also split into carriages and able to interact with each other.

3. During a commute time-step, each carriage is assigned to be travelling at 'peak time' with an 80% probability.

4. The default number of people in an average carriage is fixed to 50 people. For each city this number is adjusted in proportion to data from the UK Department for Transport (DfT) data on overcrowding in trains (UK Department for Transport, 2011). This data also disaggregates at the level of peak or off-peak travel which is used to further adjust the filling of carriages.

5. The commuting time-step is run twice a day in order to simulate commuting in each direction.

We calculate the distance travelled to work by sex, in Figure B.5, and we see that men are more likely to travel further to work in our model than women. Our findings are in reasonably good agreement with the survey Klepac et al. (2018) and serve as an independent validation of our model.

## B.2 Time-steps

As mentioned in Section A.4.1, JUNE time-steps allow differentiation between weekdays and weekends, and have a number of allotted activities. The default time-steps are described in Tables B.1 and B.2.

When choosing the time-steps, we aimed to choose the lengths such that they are somewhat close to the characteristic time of interaction of activities allowed in that time-step, but also not choosing so many time-steps to overfragment the simulation. For instance, the weekday time-step with index 1 (09:00-17:00) is eight hours, and matches the primary activities of "school" and "work", even though the "leisure" activity (which is allowed for old adults who are not assigned a workplace) has a

| Index | Calendar time | Allowed activities |
|---|---|---|
| 0 | 08:00-09:00 | M, R, C |
| 1 | 09:00-17:00 | M, P, L, R |
| 2 | 17:00-18:00 | M, R, C |
| 3 | 18:00-21:00 | M, L, R |
| 4 | 21:00-08:00 | M, R |

Table B.1: Time-steps and allowed activities for a calendar weekday, where M, R, C, P, L are "medical facility", "residence", "commute", "primary activity" and "leisure", respectively.

| Index | Calendar time | Allowed activities |
|---|---|---|
| 0 | 08:00-12:00 | M, R, L |
| 1 | 12:00-16:00 | M, R, L |
| 2 | 16:00-20:00 | M, R, L |
| 3 | 20:00-08:00 | M, R |

Table B.2: Time-steps and allowed activities for a calendar weekend. Allowed activities are the same as in Table B.1

characteristic time of 3 hours. Breaking this in half would better match the leisure characteristic time (3 hours) for this time-step, but would mean that all individuals in the simulation would be reassigned an activity for the second of the two time steps. Even though the vast majority would be reassigned to their same, required "primary activity", causing needless computation.

## B.3    Contact matrices

We use the contact matrices from the BBC Pandemic survey (Klepac et al., 2020) and supplement them with the `PolyMod` matrices (Mossong et al., 2008) for interactions of children with other children in the age bracket of 0-12 years. When comparing the matrices that capture interactions in all settings given in the BBC study, an anomaly appears in the matrix describing physical contacts - the original `PolyMod` data approximately a factor of 3 higher than the BBC matrices in neighboring age bins. We account for that by a simple scaling of the physical contacts by 1/3 before using these data. To arrive at matrices including interactions at home, in school, or in other settings for the age brackets 5-12, missing in the BBC study,

we scale the `PolyMod` setting-inclusive results by a ratio of contacts in the respective setting for the age bin of 13-14 year old kids, while we assume that the interactions of 0-4 year old children are concentrated at home.

To extract mixing matrices that are suitable for our context-specific simulation, we have to correct for the fact that the reported matrices average over the corresponding age bins in the UK population. For example, contacts between teachers and school children are normalised to the full UK population in the respective age bin instead of the number of teachers in schools that actually participate in the interaction. This necessitates rescaling to the number of people in the social context to arrive at corrected social interaction matrices $\bar{\mathcal{M}}_{ij}^{(H,W,S,O)}$. This correction step will be detailed in the relevant subsections below.

## B.3.1 Social mixing at work

The matrices for the age-dependent interaction frequency at the work place show only a very mild correlation with age, typically favouring interactions of workers with a similar age by about a factor of 2. We will therefore not include age effects at the work place into the matrices used in June. To minimize effects due to early retirement, students etc. we average over the ages of 25-60 and we compare this to the average over the working age, 18-64, but correct for an employment rate of 75%. In so doing we arrive at the number of daily contacts for adults at work:

$$n_{AA}^{(W)} = \begin{cases} 4 & (0.35 \text{ physical}) \quad \text{for ages 25-60} \\ 4.8 & (0.35 \text{ physical}) \quad \text{for ages 18-64, corrected for employment rate} \end{cases}$$
$$\text{(B.1)}$$

In June we will use $n_{AA}^{(W)} = 4.8$, with a ratio of about 7% physical contacts. While it is obvious that different industrial sectors will in reality have very different numbers of daily contacts, with corresponding impact on their vulnerability towards infections, we have not made any attempt to account for such a sector-dependent modulation, apart from effects that naturally arise from different sizes of work forces in different companies.

## B.3.2 Social mixing in schools

We decompose school populations into year groups labelled with indices $i \in \{1, 2, \ldots, N\}$ for a school with $N$ year groups and denote teachers with $T$. Starting with the interaction of pupils in various year groups an apparent large asymmetry emerges between the summed number of interactions of pupils with adults in the school and of adults with pupils in the BBC data set. This, however, is easily explained by realising that the number of interactions in a given context is normalised to the fraction of the population in a given age bin, irrespective of whether they can participate in the interaction or not. This means that the number of interactions between teachers and pupils have to be renormalised to the ratio of teachers in the adult population – about 500,000 teachers out of 36,300,000 adults, with about 216,000 working in primary and 208,000 working in secondary schools.

Summing the number of interactions of children in the age range of 5-17 with adults in the range 25-65 in schools, and assuming the latter are all teachers yields an average of 0.75 pupil-teacher interactions ($0.06 = 8\%$ of them physical) per day with very little dependence on the children's age. Conversely, adults have about 0.2 ($0.02 = 10\%$ of them physical) interactions per day with children in schools, again, relatively independent of the age of the children. Normalising this to the number of teachers in the population, we arrive at about 15 teacher-pupil interactions per day, which fits very well to approximate teacher-pupil ratios of 1:20-1:25.[1] We therefore assume that the individual interaction frequency of one specific teacher-pupil pair is consistently described with 0.75/day. For interactions among adults in the school setting we include the interaction of parents with teachers and of parents among themselves, thereby blurring the picture. We therefore assume that teachers inherit the daily contact frequencies from the work place mixing above. Turning finally to the interactions amongst children, we see a very dominant correlation in age. In order to capture this, we assume that per year of age-difference the number of

---

[1]In fact, for primary schools, the average class size is about 21 pupils, while for secondary schools it is about 16 pupils (Department for Education, 2011).

interactions among children in school, $n_{KK}^{(S)}$, will be reduced by a factor $\xi$. By fitting to the combination of BBC and `PolyMod` studies we find, to good approximation, $xi = 0.3$ and $n_{KK}^{(S)} = 2.5$, with on average 15% of the interactions being physical.

As a consequence we obtain the following social interaction frequency matrix for individual pairings at schools:

$$
\bar{\mathcal{M}}_{ij}^{(S)} \approx
\begin{pmatrix}
4.8 & 0.75 & 0.75 & 0.75 & \dots \\
15 & 2.50 & 0.75 & 0.25 & \dots \\
15 & 0.75 & 2.50 & 0.75 & \dots \\
15 & 0.25 & 0.75 & 2.50 & \dots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix},
\tag{B.2}
$$

where the first row and the first column specify the interactions between teachers and students in different year groups, and the second and following row and columns are populated by interactions of the pupils with other pupils across year groups ordered by age.

### B.3.3   Social mixing at home

In our model we decompose the household population into four subgroups, namely children ($K$, ages 0-19), young adults ($Y$, 18-24), adults ($A$, 25-64), and older adults ($O$, 65+). We therefore arrive at a $4 \times 4$ matrix of corrected social interactions at home, $\bar{\mathcal{M}}_{ij}^{(H)}$, where the indices $i, j \in \{K, Y, A, O\}$. In the following we will detail how we arrive at the various matrix elements. When correcting for the impact of social environment, i.e. the household compositions, we will ignore household compositions which are listed as "other" in the ONS database, due to a lack of detailed information (see Section A.3.2 for more details). When using these data, we will use numbers in units of millions, $H_{OAYK}$ of households with a composition of $O$ older adults, $A$ adults, $Y$ independent children or young adults living at home, and $K$ children aged 0-19.

- $\bar{\mathcal{M}}_{OO}^{(H)}$: we ignore the case of care homes or other facilities with more than two residents. Then the average interaction frequency from the BBC data is given by $n_{OO}^{(H)} = 0.78$ (0.44 physical) and 0.62 at weekends.[2] With $H_{2000} = 2.131$ and $H_{1000} = 3.294$ .[3]

$$\bar{\mathcal{M}}_{OO}^{(H)} = 0.78 \cdot \frac{2H_{2000} + H_{1000}}{2H_{2000}} \approx 1.4\,. \tag{B.3}$$

- $\bar{\mathcal{M}}_{AA}^{(H)}$: the interaction frequency between adults aged 20-65 at home from the BBC data is given by $n_{AA}^{(H)} = 1.2$ (0.74 = 62% of them physical).

$$\bar{\mathcal{M}}_{AA}^{(H)} = 1.2 \cdot \frac{\sum\limits_{x,y} \left(2H_{02xy} + H_{01xy}\right)}{\sum\limits_{x,y} 2H_{02xy}} \approx 1.34\,, \tag{B.4}$$

where $\sum_{x,y} H_{02xy} = 8.751$ and $\sum_{x,y} H_{01xy} = 7.644$.

- $\bar{\mathcal{M}}_{YY}^{(H)}$: the interaction frequency between young adults age 18-26 at home from the BBC data is given by $n_{YY}^{(H)} = 1.3$ (0.4 = 34% of them physical). There is no obvious household correction that we can apply, but the number of contacts is relatively close to the value of $\bar{\mathcal{M}}_{AA}^{(H)} = 1.34$, so we will assume that young adults interact with each other with a frequency similar to that of adults:

$$\bar{\mathcal{M}}_{YY}^{(H)} = \bar{\mathcal{M}}_{AA}^{(H)}\,. \tag{B.5}$$

It is worth noting that the age range for young adults is relatively narrow, and that there will be edge effects that may effectively increase the interaction frequency.

- $\bar{\mathcal{M}}_{YA}^{(H)}$ and $\bar{\mathcal{M}}_{AY}^{(H)}$: we have $n_{YA}^{(H)} \approx 0.7$ with a relatively steep decline with the age of the young adults, which we attribute to the fact that with increasing age young adults move out of their parents' home. To obtain some better understanding of the situation, we look at the interaction of adults in the age

---

[2]One may speculate in how far this drop is a reflection of uncertainties in the data or a true "physical" effect, for example due to visitors, travel, or similar.

[3]Here and in the following, the numbers of different household configuations are taken from Office for National Statistics (2011i).

range 40-65 with young adults, aged 18-24. From this we arrive at an average of $n_{AY}^{(H)} = 0.17$ ($0.07 = 40\%$ of them physical).

To relate this to a corrected value we must make an assumption concerning the number of young adults in the three age bins that still live with their parents, which we take as 75%, 50%, and 40% for the three age bins. To correct the $AY$ number we assume that the majority of households with young adults living as non-dependent children with their parents is composed of households with one young adult adult. Therefore:

$$
\begin{aligned}
\bar{\mathcal{M}}_{YA}^{(H)} &= \frac{1}{3}\left[\frac{0.87}{0.75} + \frac{0.65}{0.5} + \frac{0.55}{0.4}\right] &\approx& \quad 1.3 \\
\bar{\mathcal{M}}_{AY}^{(H)} &= 0.17 \cdot \frac{\sum\limits_{xy}(2H_{02xy} + H_{01xy})}{\sum\limits_{y}(2H_{021y} + H_{011y})} &\approx& \quad 1.47\,,
\end{aligned}
\tag{B.6}
$$

where $\sum_y H_{021y} = 1.514$ and $\sum_y H_{011y} = 0.946$.

- $\bar{\mathcal{M}}_{KK}^{(H)}$: the average number of daily contacts at home between children age 0-17 is $n_{KK}^{(H)} = 0.47$ (79% of them physical). Assuming all children live as dependents with their parents, and demanding that households with "2 or more children" (ONS classification) have, on average, 2.3 children to account for the UK reproduction rate, we arrive at:

$$
\bar{\mathcal{M}}_{AA}^{(H)} = 0.87 \cdot \frac{\sum\limits_{x}(H_{02x1} + H_{01x1}) + 2.3(H_{02x2} + H_{0.1x2})}{\sum\limits_{x} 2.3(H_{02x2} + H_{0.1x2})} \approx 1.2\,.
\tag{B.7}
$$

- $\bar{\mathcal{M}}_{KA}^{(H)}$ and $\bar{\mathcal{M}}_{AK}^{(H)}$: to account for contacts of children with adults we will use sliding age windows in dependence on the age of the child, using that parents are usually between 20-40 years older than their children. We then arrive at $n_{KA}^{(H)} = 1.27$ (70% of them physical) and $n_{AK}^{(H)} = 0.67$, the former with an only mild dependence on the age of the child, while the latter shows clear edge effects for the first and last bins of the adult age distribution. These numbers

translate into:

$$\bar{\mathcal{M}}_{KA}^{(H)} = \hspace{8cm} 1.27 \quad \text{(B.8)}$$

$$\bar{\mathcal{M}}_{AK}^{(H)} = \quad 0.67 \cdot \frac{\sum_{x,y}(2H_{02xy} + H_{01xy})}{\sum_{[]x,y} 2(H_{02x1} + H_{02x2}) + (H_{01x1} + H_{01x2})]} \approx 1.69 \,.^4 \quad \text{(B.9)}$$

We will also assume that the interaction frequency and intensity of children and young adults living in the same household is determined by

$$\bar{\mathcal{M}}_{KY}^{(H)} = \bar{\mathcal{M}}_{KA}^{(H)} \quad \text{and} \quad \bar{\mathcal{M}}_{YK}^{(H)} = \bar{\mathcal{M}}_{AK}^{(H)} \quad \text{(B.10)}$$

- $\bar{\mathcal{M}}_{O,KYA}^{(H)}$ and $\bar{\mathcal{M}}_{KYA,O}^{(H)}$: we assume that interactions of children, young adults, and adults with older adults at home have three different realizations:

    1. as regular contacts in a multi-generational household, where we assume that older adults behave like adults in terms of interaction frequency and intensity;

    2. as regular contacts between children and their grand-parents who act as child-minders while the parents are at work;

    3. through regular or sporadic visits, where we again assume that interactions of older adults follow the pattern of adults.

As a result we obtain the following social mixing matrix

$$\bar{\mathcal{M}}_{ij}^{(H)} = \begin{pmatrix} 1.2 & 1.69 & 1.69 & 1.69 \\ 1.27 & 1.34 & 1.47 & 1.50 \\ 1.27 & 1.30 & 1.34 & 1.34 \\ 1.27 & 1.50 & 1.34 & 2.00 \end{pmatrix} = $$

|   | K | Y | A | O |
|---|---|---|---|---|
| K | 1.2 | 1.69 | 1.69 | 1.69 |
| Y | 1.27 | 1.34 | 1.47 | 1.50 |
| A | 1.27 | 1.30 | 1.34 | 1.34 |
| O | 1.27 | 1.50 | 1.34 | 2.00 |

$$\text{(B.11)}$$

where, for convenience, we have made the entries explicit.

---

[4]Note that the latter number becomes 2.5 if we only use the central bins of parent ages 35-50 and the corresponding number of contacts $n_{AK}^{(H)} = 0.96$, which however introduces a bias in favour of the more intense interactions with children in primary school age.

## B.3.4  Social mixing in other venues

Social venues ("pubs", "cinemas", and "groceries") in June are assumed to have only one subgroup, "attendees", meaning that the social mixing matrix for these interactions is a single-element; $\bar{\mathcal{M}}^{(P)} = 3$, $\bar{\mathcal{M}}^{(C)} = 3$, $\bar{\mathcal{M}}^{(G)} = 1.5$ for "pubs", "cinemas", and "groceries" respectively. These values were chosen heuristically according the estimated number of contacts in each location relative to the the number of contacts set elsewhere (as discussed above). Given that we do not consider different subgroups in these locations, making the matrix single-valued, these numbers only serve the purpose of intuitively introducing a hierarchy of contact intensities ($\beta$ parameters) into the model structure. Since the intensity parameters are fitted to data (see Section A.8), the form of Equation A.8 ensures that the choice these social mixing matrices values will not significantly affect the probabilities of transmission.

Hospitals have three subgroups: medical staff, ward patients, and ICU/ITU patients. The social mixing matrix for hospitals (where the superscript $M$ refers to "medical facility") is

$$\bar{\mathcal{M}}_{ij}^{(M)} = \begin{pmatrix} 5 & 10 & 10 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \text{ and } \phi_{ij}^{(M)} = \begin{pmatrix} 0.05 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \tag{B.12}$$

where $(i, j) \in \{S, W, I\}$ denoting the three subgroups, medical staff, ward patients, and ICU/ITU patients, respectively. The number of contacts between a medic and patients, 10, represents the average number of patients per medic. We assume that a patient is visited by a medic once per characteristic time, set to 8 hours for hospitals. The number of contacts between patients is irrelevant, as patients are by definition already infected, but is set to zero.

Social mixing in care homes considers three subgroups: workers, residents, and

visitors, with matrix

$$
\bar{\mathcal{M}}_{ij}^{(CH)} = \begin{pmatrix} 15 & 15 & 1 \\ 1.5 & 4 & 20 \\ 1.5 & 6 & 0 \end{pmatrix} \quad \text{and} \quad \phi_{ij}^{(CH)} = \begin{pmatrix} 0.05 & 1 & 0 \\ 1 & 0.4 & 1 \\ 0 & 0.5 & 0 \end{pmatrix}, \quad \text{(B.13)}
$$

with a characteristic time of 24 hours. The seemingly-high number of contacts between workers and visitors, and residents and visitors is to compensate for the characteristic time of 24 hours; if visitors were to be present in a care home for a full characteristic time, they would experience this many contacts, but visits are a day time activity which take only a few hours, resulting in fewer contacts.

Finally, universities are modelled as having six groups to represent professors and five distinct groups of students (for the moment based only on age 19-23), with diagonal elements $\bar{\mathcal{M}}_{i=j}^{(U)} = 2$ and off-diagonal elements $\bar{\mathcal{M}}_{i\neq j}^{(U)} = 0.75$, and all $\phi_{ij}^{(U)} = 0.25$.

### B.3.5 Deriving contact matrices from JUNE

We derive the contact matrices in Figure A.9 by simulating a week of pre-lockdown activity. For each person, in each subgroup $i$, in each venue, we choose the required $N_{ij}$ people (with replacement) for all (non-empty) subgroups $j$ in that venue (where $N_{ij}$ is from the relevant social mixing matrix). We populate "raw" contact matrices using these selected people. As these contacts are then uni-directional, we make the same corrections as in Klepac et al. (2020) to account for reciprocal contacts. We hope to produce contact matrices derived from constructing self-consistent (reciprocal) networks of contacts within groups in future work.

## B.4 Details on modelling health trajectories

For the times spent in different stages of disease progression we use a variety of functions, namely intervals of constant length, scaled and shifted $\beta$ functions, scaled

| Name | Function | Source |
|---|---|---|
| $C_T$ | constant with time $T$ | |
| $\beta_I$ | $\beta_{2.29,19.05,0.39,39.8}(t)$ | He et al. (2020) |
| $\mathrm{LN}_M$ | $\mathrm{LN}_{0.83,5.7}(t)$ | * |
| $\beta_H$ | $\beta_{1.35,3.68,0.05,27.1}(t)$ | Scientific Advisory Group for Emergencies |
| $\beta_D$ | $\beta_{1.21,1.97,0.08,12.9}(t)$ | Scientific Advisory Group for Emergencies |
| $\mathrm{LN}_{ICU}$ | $\mathrm{LN}_{1.41,0.9}(t)$ | ICNARC |
| $e_{ICU}$ | $e_{1.06,0.89,12}(t)$ | ICNARC |
| $e_D$ | $e_{1.23,1,9.69}(t)$ | ICNARC |

Table B.3: Characteristic functions and their parameters. *We constrained the time from symptom onset to hospitalisation through private communication with hospital physicians at early stages of the first COVID-19 wave of infections. We later checked this assumed profile against published data and found our values to be broadly consistent.

log-normal distributions, and exponential Weibull distributions, given by

$$
\begin{aligned}
C_{t_{end}}(t) &= \Theta(t_{end} - t) \\
\beta_{a,b,l,S}(t) &= \beta_{a,b}\left(\frac{t-l}{S}\right) \\
\mathrm{LN}_{s,S}(t) &= \mathrm{LN}_s\left(\frac{t}{S}\right) \\
e_{a,c,S}(t) &= e_{a,c}\left(\frac{t}{S}\right).
\end{aligned}
\tag{B.14}
$$

The trajectories and their building blocks to construct the corresponding time intervals infected individuals spend in various stages of the disease are listed in Table A.3.

## B.5 Calibration via Bayes Linear Emulation and History Matching

We now provide more details of the Bayes linear emulation and history matching process outlined in Section A.8. To set up the history matching problem, we identify a large set of 18 input parameters to the JUNE model to explore. This set is composed mainly of contact intensity parameters, but also contains such parameters governing social distancing effects, compliance and physical contact, with each parameter specified along with associated broad ranges, in Table B.4. We

| Input Parameter ($x_i$) | Type | Range |
|---|---|---|
| $\beta_{pub}$ | contact intensity | [0.02,0.6] |
| $\beta_{grocery}$ | . | [0.02,0.6] |
| $\beta_{cinema}$ | . | [0.02,0.6] |
| $\beta_{university}$ | . | [0.02,0.6] |
| $\beta_{city\ transport}$ | . | [0.08,0.77] |
| $\beta_{inter\ city\ transport}$ | . | [0.08,1.2] |
| $\beta_{hospital}$ | . | [0.08,1.2] |
| $\beta_{care\ home}$ | . | [0.08,1.2] |
| $\beta_{company}$ | . | [0.08,1.2] |
| $\beta_{school}$ | . | [0.08,1.2] |
| $\beta_{household}$ | . | [0.08,1.2] |
| $\beta_{care\ visits}$ | . | [0.1,8] |
| $\beta_{household\ visits}$ | . | [0.08,1.2] |
| $\alpha_{physical}$ | physical contact factor | [1.8,3] |
| $\alpha_{seed\ strength}$ | seeding | [0.5,1.3] |
| $M_{quarantine\ household\ compliance}$ | compliance | [0.034,0.26] |
| $M_{social\ distancing\ \beta\ factor}$ | social distancing | [0.65,0.95] |
| $M_{sd4\ random\ factor\ all}$ | social distancing | [0.004,0.5] |

Table B.4: The input parameters that form the 18-dimensional vector $x$ explored in the global parameter search, their type and their ranges that define the search region $\mathcal{X}_0$. The parameters $\alpha_{seed\ strength}$ and $M_{sd4\ random\ factor\ all}$ modulate the strength of the seeding process, and the social distancing policy active from the 7$^{th}$ of July respectively.

denote this set of parameters by the 18-dimensional vector $x$ and denote the initial search region defined by their combined ranges as $\mathcal{X}_0$. We identify a set of outputs to match to observed data, specifically the deaths and total deaths for England and for each of the seven regions, and for the age bins $0-5, 6-17, 18-64, 65-84, 85^+$, at the time points of the 20$^{th}$ March, 28$^{th}$ March, 5$^{th}$ April, 13$^{th}$ April, 21$^{st}$ April, 29$^{th}$ April, 12$^{th}$ May, 26$^{th}$ May, 8$^{th}$ June, 8$^{th}$ June and 23$^{rd}$ June, 2020. We represent the list of all these outputs as the vector $f$.

We note that the JUNE model can now be viewed as a function that maps the inputs $x$ to the vector of all outputs of interest $f(x)$. As we cannot evaluate the model $f(x)$ exhaustively over the full parameter space $\mathcal{X}_0$ due to computational expense, we mimic it using a fast to evaluate (but uncertain) Bayesian emulator. For an individual output $f_i(x)$, representing for example, the total deaths in England on

the 28th March, we construct an emulator of the form

$$f_i(x) = \sum_j b_{ij} g_{ij}(x_{A_i}) + u_i(x_{A_i}) + v_i(x) \qquad \text{(B.15)}$$

where the first term on the RHS is a regression term designed to capture global behaviour, composed of known deterministic functions $g_{ij}$ (with a common choice being low order polynomials) of the active variables $x_{A_i}$, which are a subset of the inputs that are found to be most influential for output $f_i(x)$, and of the $b_{ij}$ which are unknown regression coefficients. The middle term $u_i(x_{A_i})$ is a Gaussian process with various forms of correlation structure available, capable of mimicking large classes of functions, which has the flexibility to capture more local behaviour of $f_i(x)$, and $v_i(x)$ is an uncorrelated nugget that represents the effect of the remaining inactive input variables, and/or any stochasticity exhibited by the model.

We perform an initial space filling set of $n = 125$ runs $D = (f(x^{(1)}), f(x^{(2)}), \ldots,$ $f(x^{(n)}))$ with the $x^{(i)} \in \mathcal{X}_0$ chosen using a maximin Latin hypercube design. The emulators are updated by the runs $D$ using the Bayes linear update equations (Vernon et al., 2010a), and hence can give a prediction with corresponding uncertainty, of the unobserved $f(x)$ at a new, previously unevaluated input point $x$, in the form of the adjusted expectation $E_D(f_i(x))$ and the adjusted variance $Var_D(f_i(x))$ respectively. The emulators have to satisfy extensive diagnostics (Bastos and O'Hagan, 2008; Vernon et al., 2010a), an illustrative example of which is given in Figure B.6, left panel, which shows the emulator prediction $E_D(f_i(x))$ for $f_i(x)$ across several time points (the solid red line) and the prediction interval $E_D(f_i(x)) \pm 3\sqrt{Var_D(f_i(x))}$ (the red dashed lines) along with the held out run output $f(x)$ (the blue line) which the emulator has not previously seen, showing excellent agreement between emulator and model. The emulator evaluation takes a fraction of a second, and mimics the June model well.

By confronting the emulators with the observed data vector $z$ corresponding to the outputs in $f$, and incorporating major sources of uncertainty (e.g. observation error, structural model discrepancy, stochasticity), we can rule out large parts of
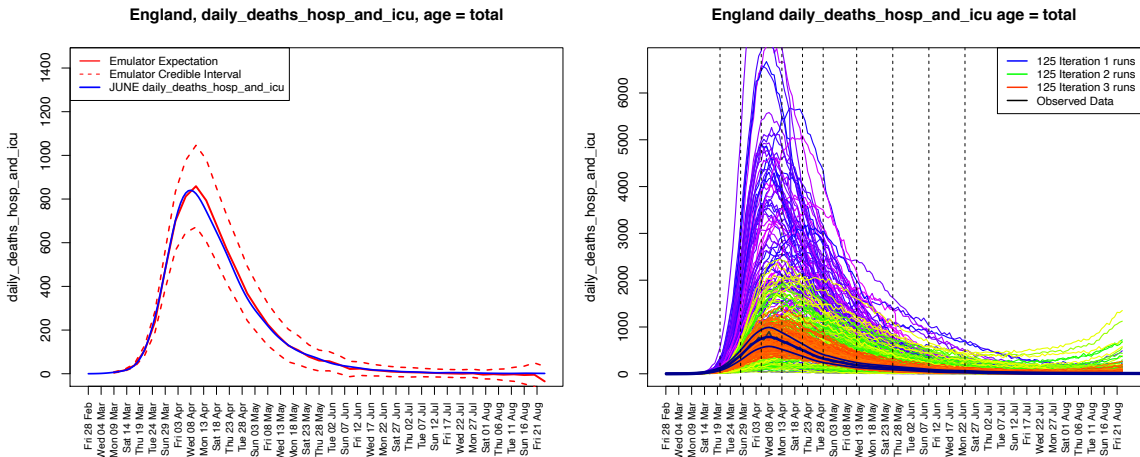
Figure B.6: Left Panel: an example diagnostic showing the emulator prediction $E_D(f_i(x))$ for $f_i(x)$ across several time points (the solid red line) and the prediction interval $E_D(f_i(x)) \pm 3\sqrt{Var_D(f_i(x))}$ (the red dashed lines) along with the held out smoothed run output $f(x)$ (the blue line). The emulator captures the behaviour of the JUNE model well. Right panel: daily hospital deaths for all of England, showing the progression of the runs from iterations 1, 2 and 3 used in the history matching process (in purple, green and red respectively). Observed data (smoothed and original) in black. Vertical dashed lines: emulated outputs.

the input parameter space $\mathcal{X}_0$ as implausible. We do this using an implausibility measure, for which the univariate version $I_i(x)$, is defined for each output as

$$I_i^2(x) \quad = \quad \frac{(E_{D_i}(f_i(x)) - z_i)^2}{Var_D(f_i(x)) + \sigma_{\epsilon_i}^2 + \sigma_{e_i}^2} \tag{B.16}$$

where $E_D(f_i(x))$ and $Var_D(f_i(x))$ are the emulator expectation and variance as before, $z_i$ is the observed data point corresponding to model output $f_i$, $\sigma_{e_i}^2$ is the variance of the observation error $e_i$ (a random quantity representing the imperfections of the measurement process), and $\sigma_{\epsilon_i}^2$ is the variance of the model discrepancy $\epsilon_i$ (an often neglected random quantity representing the imperfections of the model (Craig et al., 1997; Kennedy and O'Hagan, 2001; Vernon et al., 2010a)). If $I_i(x)$ is large, it is because the emulator expectation for $f_i(x)$ is very far from the observed data $z_i$, even given all the major sources of uncertainty, and therefore the input parameter $x$ is highly unlikely to yield model output similar to observed data were we to evaluate JUNE there, and hence $x$ could be discarded from further analysis. A typical cutoff maybe $I_i(x) < c$ where $c = 3$.

There are various ways to combine implausibility measures for each of the individual outputs, the simplest being to maximise: $I_M(x) = \max_i I_i(x)$, that is to take the maximum implausibility across all outputs of interest which is the measure chosen here, although we note that other more nuanced and/or robust versions are available, that capture more of the multivariate behaviour (Vernon et al., 2010a).

We now employ iterative history matching (Vernon et al., 2010a), a parameter search method that seeks to identify all parts of parameter space that would give rise to acceptable matches between model output and observational data. This proceeds at the $j$th iteration (or wave), by constructing emulators using the current set of runs, removing the implausible parts of the input space to define the new non-implausible region $\mathcal{X}_j = \{x \in \mathcal{X}_0 : I_M(x) < c\}$, designing and performing a new space filling set of runs across the reduced input space $\mathcal{X}_j$ and re-emulating, but now with a more accurate emulator defined only over the reduced region $\mathcal{X}_j$. For further discussion see Vernon et al. (2010a,b, 2018), but it suffices to note that the iterative nature of history matching is key, as it allows later iteration emulators to become far more accurate as they are only employed over far smaller parts of the input space, and are hence informed by a much higher density of runs.

The observed data for total deaths was obtained from the ONS, while the hospital deaths data is taken from CPNS - the Covid Patient Notification System (CPN). For each output corresponding to the element of $f$, the data was first smoothed slightly with a standard kernel smoother, to reduce the day-to-day stochasticity. The observation error and model discrepancy variances for each output were each decomposed into multiplicative and additive components to represent possible systematic biases, in addition to a scaled $\sqrt{n}$ component for the observation error only, to model the noisy count process. For example, we have the decompositions $\sigma_{\epsilon_i}^2 = \alpha_{mult,\epsilon_i}^2 z_i^2 + \gamma_{add,\epsilon_i}^2$, with $\alpha_{mult,\epsilon_i} = 0.06$ and $\gamma_{add,\epsilon_i}^2 = 3/2$, and $\sigma_{e_i}^2 = \alpha_{mult,e_i}^2 z_i^2 + \gamma_{add,e_i}^2 + (\delta_{corr,e_i}\sqrt{z_i})^2$, with $\alpha_{mult,e_i} = 0.06$ and $\gamma_{add,e_i}^2 = 3/2$ and $\delta_{corr,e_i} = 0.25$ governed by the mitigation of the smoothing process.

As described in Section A.8, we performed 3 waves of the history match with

125 runs each wave, finding that the emulators were of sufficient accuracy after the third wave. Figure B.6, right panel, shows the progression of the runs from iterations 1, 2 and 3 used in the history matching process (in purple, green and red respectively) for the daily hospital deaths in England output, with the data (original and smoothed) in black. We can see that the third iteration runs are vastly improved and surround the observed data. These allow accurate emulators to be constructed that can identify the region of input space of interest, which were used to construct Figure A.22, as discussed in Section A.8.

# Bibliography

Covid-19 Patient Notification System (CPNS) User Guide. `https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2020/09/CPNS-User-Guide-20200831.pdf`. Accessed: 14 December 2020.

S. Abar, G. K. Theodoropoulos, P. Lemarinier, and G. M. O'Hare. Agent Based Modelling and Simulation tools: A review of the state-of-art software. *Computer Science Review*, 24:13–33, 2017.

K. Abe, K. Hieda, K. Hiraide, S. Hirano, Y. Kishimoto, K. Kobayashi, S. Moriyama, K. Nakagawa, M. Nakahata, H. Nishiie, et al. Xmass detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 716:78–85, 2013.

Age UK. Briefing: Health and Care of Older People in England 2019. `https://www.ageuk.org.uk/globalassets/age-uk/documents/reports-and-publications/reports-and-briefings/health--wellbeing/age_uk_briefing_state_of_health_and_care_of_older_people_july2019.pdf`, 2019. Accessed: 14 December 2020.

N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini,

A. Banday, R. Barreiro, N. Bartolo, S. Basak, et al. Planck 2018 results-vi. cosmological parameters. *Astronomy & Astrophysics*, 641:A6, 2020.

A. Aguirre-Santaella, M. A. Sánchez-Conde, G. Ogiya, J. Stücker, and R. E. Angulo. Shedding light on low-mass subhalo survival and annihilation luminosity with numerical simulations. *Monthly Notices of the Royal Astronomical Society*, 518(1):93–110, 2023.

D. Akerib, C. Akerlof, D. Y. Akimov, A. Alquahtani, S. Alsum, T. Anderson, N. Angelides, H. Araújo, A. Arbuckle, J. Armstrong, et al. The lux-zeplin (lz) experiment. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 953:163047, 2020.

A. Albert, B. Anderson, K. Bechtol, A. Drlica-Wagner, M. Meyer, M. Sánchez-Conde, L. Strigari, M. Wood, T. Abbott, F. B. Abdalla, et al. Searching for dark matter annihilation in recently discovered milky way satellites with fermi-lat. *The Astrophysical Journal*, 834(2):110, 2017.

I. Andrianakis, I. Vernon, N. McCreesh, T. McKinley, J. Oakley, R. Nsubuga, M. Goldstein, and R. White. Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda. *PLoS Comput Biol.*, 11(1):e1003968, 2015.

I. Andrianakis, N. McCreesh, I. Vernon, T. McKinley, J. Oakley, R. Nsubuga, M. Goldstein, and R. White. Efficient history matching of a high dimensional individual based hiv transmission model. *SIAM/ASA Journal of Uncertainty Quantification*, 5(1):694–719, 2017a.

I. Andrianakis, I. Vernon, N. McCreesh, T. J. McKinley, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White. History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance

emulation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(4):717–740, 2017b.

A. H. Auchincloss, S. Y. Gebreab, C. Mair, and A. V. Diez Roux. A review of spatial methods in epidemiology, 2000–2010. *Annual review of public health*, 33: 107–122, 2012.

P. Avgeriou and U. Zdun. Architectural Patterns Revisited - A Pattern Language. In A. Longshaw and U. Zdun, editors, *EuroPLoP' 2005, Tenth European Conference on Pattern Languages of Programs, Irsee, Germany, July 6-10, 2005*, pages 431–470. UVK - Universitaetsverlag Konstanz, 2005.

J. Aylett-Bullock, C. Cuesta-Lazaro, A. Quera-Bofarull, A. Katta, K. H. Pham, B. Hoover, H. Strobelt, R. M. Jimenez, A. Sedgewick, E. S. Evers, D. Kennedy, S. Harlass, A. G. K. Maina, A. Hussien, and M. Luengo-Oroz. Operational response simulation tool for epidemics within refugee and IDP settlements. *medRxiv*, 2021. doi: 0.1101/2021.01.27.21250611. URL `https://www.medrxiv.org/content/early/2021/01/29/2021.01.27.21250611`.

J. Barnes and P. Hut. A hierarchical $\mathcal{O}(N \log N)$ force-calculation algorithm. *Nature*, 324(6096):446–449, 1986.

T. S. Bastos and A. O'Hagan. Diagnostics for Gaussian process emulators. *Tech*, 51:425–438, 2008.

C. L. Bennett, D. Larson, J. L. Weiland, N. Jarosik, G. Hinshaw, N. Odegard, K. Smith, R. Hill, B. Gold, M. Halpern, et al. Nine-year wilkinson microwave anisotropy probe (wmap) observations: final maps and results. *The Astrophysical Journal Supplement Series*, 208(2):20, 2013.

A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge. The challenges of modeling and forecasting the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117(29):16732–16738, 2020.

J. Binney and S. Tremaine. Galactic Dynamics. *Princeton University Press*, 2008.

N. Brandon, K. L. Dionisio, K. Isaacs, R. Tornero-Velez, D. Kapraun, R. W. Setzer, and P. S. Price. Simulating exposure-related behaviors using agent-based models embedded with needs-based artificial intelligence. *Journal of exposure science & environmental epidemiology*, pages 1–10, 2018.

J. Bullock, C. Cuesta-Lazaro, A. Quera-Bofarull, M. Icaza-Lizaola, A. Sedgewick, H. Truong, T. Caulfield, K. Fong, I. Vernon, J. Williams, R. Bower, and F. Krauss. JUNE: a Bayesian uncertainty analysis. *(in preparation)*, 2021.

D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini Jr. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Comput Biol*, 6(1):e1000656, 2010.

D. K. Chu, E. A. Akl, S. Duda, K. Solo, S. Yaacoub, and H. J. Schünemann. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *Lancet*, 395(10242):1973–1987, 2020.

M. Clutton-Brock. The gravitational field of flat galaxies. *Astrophysics and Space Science*, 16(1):101–119, 1972.

M. Clutton-Brock. The gravitational field of three-dimensional galaxies. *Astrophysics and Space Science*, 23(1):55–69, 1973.

S. Cole and C. Lacey. The structure of dark matter haloes in hierarchical clustering models. *Monthly Notices of the Royal Astronomical Society*, 281(2):716–736, 1996.

S. Cole, C. G. Lacey, C. M. Baugh, and C. S. Frenk. Hierarchical galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 319(1):168–204, 2000.

P. Coles and F. Lucchin. *Cosmology: The origin and evolution of cosmic structure.* John Wiley & Sons, 2003.

P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith. Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion). In C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi, and N. D. Singpurwalla, editors, *Case Studies in Bayesian Statistics*, volume 3, pages 36–93. SV, New York, 1997.

Department for Education. Class Size and education in England evidence report. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/183364/DFE-RR169.pdf`, 2011.

Department for Transport. National Travel Survey 2014: Travel to school. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/476635/travel-to-school.pdf`, 2014.

Department of Health & Social Care. Vivaldi 1: COVID-19 care homes study report. `https://www.gov.uk/government/publications/vivaldi-1-coronavirus-covid-19-care-homes-study-report/vivaldi-1-covid-19-care-homes-study-report#fn:1`, 2020.

J. Diemand, M. Kuhlen, and P. Madau. Dark matter substructure and gamma-ray annihilation in the milky way halo. *The Astrophysical Journal*, 657(1):262, 2007.

J. Diemand, M. Kuhlen, P. Madau, M. Zemp, B. Moore, D. Potter, and J. Stadel. Clumps and streams in the local dark matter distribution. *Nature*, 454(7205): 735–738, 2008.

Y. Dong, X. Mo, Y. Hu, X. Qi, F. Jiang, Z. Jiang, and S. Tong. Epidemiology of COVID-19 among children in China. *Pediatrics*, 145(6), 2020.

J. Dubinski. A parallel tree code. *New Astronomy*, 1(2):133–147, 1996.

D. Eddelbuettel. CRANberries, 2007. URL `http://dirk.eddelbuettel.com/cranberries/`.

A. M. El-Sayed, P. Scarborough, L. Seemann, and S. Galea. Social network analysis and agent-based modeling in social epidemiology. *Epidemiologic Perspectives & Innovations*, 9(1):1, 2012.

R. Errani and J. F. Navarro. The asymptotic tidal remnants of cold dark matter subhaloes. *Monthly Notices of the Royal Astronomical Society*, 505(1):18–32, 2021.

N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

E. P. Fischer, M. C. Fischer, D. Grass, I. Henrion, W. S. Warren, and E. Westman. Low-cost measurement of face mask efficacy for filtering expelled droplets during speech. *Science Advances*, 6(36):eabd3083, 2020.

A. S. Font, A. J. Benson, R. G. Bower, C. Frenk, A. Cooper, G. DeLucia, J. C. Helly, A. Helmi, Y.-S. Li, I. G. McCarthy, et al. The population of milky way satellites in the $\lambda$ cold dark matter cosmology. *Monthly Notices of the Royal Astronomical Society*, 417(2):1260–1279, 2011.

C. S. Frenk, S. D. White, M. Davis, and G. Efstathiou. The formation of dark halos in a universe dominated by cold dark matter. *The Astrophysical Journal*, 327:507–525, 1988.

L. Gao, S. D. White, A. Jenkins, F. Stoehr, and V. Springel. The subhalo populations of ΛCDM dark haloes. *Monthly Notices of the Royal Astronomical Society*, 355(3):819–834, 2004.

J. D. Garrett. SciencePlots (v1.0.6). Oct. 2020. doi: 10.5281/zenodo.4106650. URL http://doi.org/10.5281/zenodo.4106650.

J. Gershuny and O. Sullivan. United Kingdom Time Use Survey, 2014-2015. Technical Report SN: 8128, UK Data Service, 2015. URL http://doi.org/10.5255/UKDA-SN-8128-1.

Government of the United Kingdom. Get a discount with the Eat Out to Help Out Scheme. `https://www.gov.uk/guidance/get-a-discount-with-the-eat-out-to-help-out-scheme`, 2020.

R. J. Grand, F. A. Gómez, F. Marinacci, R. Pakmor, V. Springel, D. J. Campbell, C. S. Frenk, A. Jenkins, and S. D. White. The Auriga Project: the properties and formation mechanisms of disc galaxies across cosmic time. *Monthly Notices of the Royal Astronomical Society*, 467(1):179–207, 2017.

R. J. Grand, F. Marinacci, R. Pakmor, C. M. Simpson, A. J. Kelly, F. A. Gómez, A. Jenkins, V. Springel, C. S. Frenk, and S. D. White. Determining the full satellite population of a Milky Way-mass halo in a highly resolved cosmological hydrodynamic simulation. *Monthly Notices of the Royal Astronomical Society*, 507(4):4953–4967, 2021.

S. B. Green and F. C. van den Bosch. The tidal evolution of dark matter substructure–I. subhalo density profiles. *Monthly Notices of the Royal Astronomical Society*, 490(2):2091–2101, 2019.

S. B. Green, F. C. van den Bosch, and F. Jiang. The tidal evolution of dark matter substructure–II. The impact of artificial disruption on subhalo mass functions and radial profiles. *Monthly Notices of the Royal Astronomical Society*, 503(3):4075–4091, 2021.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R'ıo, M. Wiebe, P. Peterson, P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

X. He, E. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, and G. Leung. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26, 05 2020. doi: 10.1038/s41591-020-0869-5.

L. Hernquist and J. P. Ostriker. A self-consistent field method for galactic dynamics. *The Astrophysical Journal*, 386:375–397, 1992.

R. Hinch, W. J. M. Probert, A. Nurtay, M. Kendall, C. Wymatt, M. Hall, K. Lythgoe, A. Bulas Cruz, L. Zhao, A. Stewart, L. Ferritti, D. Montero, J. Warren, N. Mather, M. Abueg, N. Wu, A. Finkelstein, D. G. Bonsall, L. Abeler-Dorner, and C. Fraser. OpenABM-Covid19 - an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *medRxiv*, 2020. doi: 10.1101/2020.09.16.20195925. URL `https://www.medrxiv.org/content/early/2020/09/22/2020.09.16.20195925`.

HM Revenue & Customs. HMRC coronavirus (COVID-19) statistics. `https://www.gov.uk/government/collections/hmrc-coronavirus-covid-19-statistics#Coronavirus-Job-Retention-Scheme-Management-information`, 2020.

J. Howard, A. Huang, Z. Li, Z. Tufekci, V. Zdimal, H.-M. van der Westhuizen, A. von Delft, A. Price, L. Fridman, L.-H. Tang, et al. Face masks against COVID-19: an evidence review. 2020.

E. Hunter, B. Mac Namee, and J. D. Kelleher. A taxonomy for agent-based models in human infectious disease epidemiology. *Journal of Artificial Societies and Social Simulation*, 20(3), 2017.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

ICNARC. Icnarc report on covid-19 in critical care 10 july 2020. `https://www.icnarc.org/Our-Audit/Audits/Cmp/Reports`.

IDAS Covid group. Social differences in SARS-CoV-2 infection spread — Draft. *(in preparation)*, 2021.

Institute for Fiscal Studies. Sector shutdowns during the coronavirus crisis: which workers are most exposed? `https://www.ifs.org.uk/publications/14791`, 2020.

L. Jiang, J. C. Helly, S. Cole, and C. S. Frenk. *N*-body dark matter haloes with simple hierarchical histories. *Monthly Notices of the Royal Astronomical Society*, 440(3):2115–2135, 2014.

S. Kazantzidis, J. Magorrian, and B. Moore. Generating equilibrium dark matter halos: Inadequacies of the local Maxwellian approximation. *The Astrophysical Journal*, 601(1):37, 2004.

A. Kelly. *Exploring the properties of hot haloes around simulated galaxies.* PhD thesis, Durham University, 2022.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *JRSSB*, 63(3):425–464, 2001.

P. Klepac, S. Kissler, and J. Gog. Contagion! The BBC Four Pandemic – The model behind the documentary. *Epidemics*, 24:49 – 59, 2018. ISSN 1755-4365. doi: https://doi.org/10.1016/j.epidem.2018.03.003. URL `http://www.sciencedirect.com/science/article/pii/S1755436518300306`.

P. Klepac, A. J. Kucharski, A. J. Conlan, S. Kissler, M. Tang, H. Fry, and J. R. Gog. Contacts in context: large-scale setting-specific social mixing matrices from the BBC Pandemic project. *medRxiv*, 2020.

A. Klypin, S. Gottlöber, A. V. Kravtsov, and A. M. Khokhlov. Galaxies in *N*-body simulations: overcoming the overmerging problem. *The Astrophysical Journal*, 516(2):530, 1999.

A. A. Klypin, S. Trujillo-Gomez, and J. Primack. Dark matter halos in the standard cosmological model: Results from the bolshoi simulation. *The Astrophysical Journal*, 740(2):102, 2011.

A. Knebe, F. R. Pearce, H. Lux, Y. Ascasibar, P. Behroozi, J. Casado, C. C. Moran, J. Diemand, K. Dolag, R. Dominguez-Tenreiro, et al. Structure finding in cosmological simulations: the state of affairs. *Monthly Notices of the Royal Astronomical Society*, 435(2):1618–1658, 2013.

A. V. Kravtsov, A. A. Klypin, and A. M. Khokhlov. Adaptive refinement tree: a new high-resolution n-body code for cosmological simulations. *The Astrophysical Journal Supplement Series*, 111(1):73, 1997.

M. Kuhlen, J. Diemand, P. Madau, and M. Zemp. The via lactea incite simulation: galactic dark matter substructure at high resolution. In *Journal of Physics: Conference Series*, volume 125, page 012008. IOP Publishing, 2008.

C. Lacey and S. Cole. Merger rates in hierarchical models of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 262(3):627–649, 1993.

C. Larman and V. Basili. Iterative and incremental developments. a brief history. *Computer*, 36(6):47–56, June 2003. ISSN 0018-9162. doi: 10.1109/ MC.2003.1204375. URL http://ieeexplore.ieee.org/document/1204375/.

P.-I. Lee, Y.-L. Hu, P.-Y. Chen, Y.-C. Huang, and P.-R. Hsueh. Are children less susceptible to COVID-19? *Journal of Microbiology, Immunology, and Infection*, 2020.

M. Liang, L. Gao, C. Cheng, Q. Zhou, J. P. Uy, K. Heiner, and C. Sun. Efficacy of face mask in preventing respiratory virus transmission: A systematic review and meta-analysis. *Travel medicine and infectious disease*, 36:101751–101751, 2020.

B. Lowing, A. Jenkins, V. Eke, and C. Frenk. A halo expansion technique for approximating simulated dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 416(4):2697–2711, 2011.

A. D. Ludlow, J. F. Navarro, R. E. Angulo, M. Boylan-Kolchin, V. Springel, C. Frenk, and S. D. White. The mass–concentration–redshift relation of cold dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 441(1): 378–388, 2014.

A. M. Lund, R. Gouripeddi, and J. C. Facelli. STHAM: an agent based model for simulating human exposure across high resolution spatiotemporal domains. *Journal of Exposure Science & Environmental Epidemiology*, 30(3):459–468, 2020.

P. Madau, J. Diemand, and M. Kuhlen. Dark matter subhalos and the dwarf satellites of the milky way. *The Astrophysical Journal*, 679(2):1260, 2008.

M. V. Marathe and N. Ramakrishnan. Recent advances in computational epidemiology. *IEEE intelligent systems*, 28(4):96–101, 2013.

N. McCreesh, I. Andrianakis, R. N. Nsubuga, M. Strong, I. Vernon, T. J. McKinley, J. E. Oakley, M. Goldstein, R. Hayes, and R. G. White. Universal test, treat, and keep: improving ART retention is key in cost-effective HIV control in Uganda. *BMC Infectious Diseases*, 17(1):322, May 2017.

T. McKinley, I. Vernon, I. Andrianakis, N. McCreesh, J. Oakley, R. Nsubuga, M. Goldstein, and R. White. Approximate Bayesian Computation and simulation-based inference for complex stochastic epidemic models. *Statistical Science*, 33(1):4–18, June 2018.

S. Mei, H. Guan, and Q. Wang. An Overview on the Convergence of High Performance Computing and Big Data Processing. In *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 1046–1051. IEEE, 2018.

Ministry of Housing, Communities & Local Government. English indices of deprivation 2019. `https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019`, 2019.

B. Moore, N. Katz, and G. Lake. On the destruction and overmerging of dark halos in dissipationless n-body simulations. *The Astrophysical Journal*, 457:455, 1996.

B. Moore, F. Governato, T. Quinn, J. Stadel, and G. Lake. Resolving the structure of cold dark matter halos. *The Astrophysical Journal*, 499(1):L5, 1998.

J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5(3):e74, 2008.

N. Nakasato, G. Ogiya, Y. Miki, M. Mori, and K. Nomoto. Astrophysical particle simulations on heterogeneous CPU-GPU systems. *arXiv preprint arXiv:1206.1199*, 2012.

National Rail. Maps of the National Rail network of Great Britain. `https://www.nationalrail.co.uk/stations_destinations/maps.aspx`, 2015.

J. F. Navarro, C. S. Frenk, and S. D. M. White. The structure of cold dark matter halos. *The Astrophysical Journal*, 462:563, may 1996.

J. F. Navarro, C. S. Frenk, and S. D. White. A universal density profile from hierarchical clustering. *The Astrophysical Journal*, 490(2):493, 1997.

A. F. Neto, L. Gao, P. Bett, S. Cole, J. F. Navarro, C. S. Frenk, S. D. White, V. Springel, and A. Jenkins. The statistics of ΛCDM halo concentrations. *Monthly Notices of the Royal Astronomical Society*, 381(4):1450–1462, 2007.

S. J. e. a. NF Brazeau, R Verity. COVID-19 Infection Fatality Ratio: Estimates from Seroprevalence, 2020. URL `https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-34-IFR/`.

NHS. COVID-19 Situation Reports. `https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/data-provision-notices-dpns/covid-19-situation-reports`. Accessed: 14 December 2020.

Office for National Statistics. Deaths involving COVID-19 in the care sector, England and Wales. `https : / / www.ons.gov.uk / peoplepopulationandcommunity / birthsdeathsandmarriages / deaths / articles / deathsinvolvingcovid19inthecaresectorenglandandwales / deathsoccurringupto12june2020andregisteredupto20june2020provisional# characteristics-of-care-home-residents-who-died-from-covid-19`, a.

Office for National Statistics. LC1105EW Residence type by sex by age. `https: //www.nomisweb.co.uk/census/2011/lc1105ew`, b.

Office for National Statistics. Deaths registered weekly in England and Wales, provisional. `https : / / www.ons.gov.uk / peoplepopulationandcommunity / birthsdeathsandmarriages / deaths / datasets / weeklyprovisionalfiguresondeathsregisteredinenglandandwales`, c.

Office for National Statistics. Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2019. `https:// www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/ populationestimates / bulletins / annualmidyearpopulationestimates / latest`, d.

Office for National Statistics. UK SIC 2007. `https : / / www.ons.gov.uk / methodology / classificationsandstandards / ukstandardindustrialclassificationofeconomicactivities / uksic2007`, 2007.

Office for National Statistics. QS103EW (Age by single year). `https:// www.nomisweb.co.uk/census/2011/qs103ew`, 2011a.

Office for National Statistics. QS701EW (Method of travel to work). `https: //www.nomisweb.co.uk/census/2011/qs701ew`, 2011b.

Office for National Statistics. DC1104EW (Residence type by sex by age). `https: //www.nomisweb.co.uk/census/2011/dc1104ew`, 2011c.

Office for National Statistics. KS405UK (Communal establishment residents). `https://www.nomisweb.co.uk/census/2011/ks405uk`, 2011d.

Office for National Statistics. KS605EW-KS607EW (Industry by sex). `https://www.nomisweb.co.uk/census/2011/ks605ew`, 2011e.

Office for National Statistics. UK Business Counts - enterprises by industry and employment size band. `https://www.nomisweb.co.uk/datasets/idbrent`, 2011f.

Office for National Statistics. DC2101EW (Ethnic group by sex by age). `https://www.nomisweb.co.uk/census/2011/dc2101ew`, 2011g.

Office for National Statistics. WU01EW (Location of usual residence and place of work by sex). `https://www.nomisweb.co.uk/census/2011/wu01ew`, 2011h.

Office for National Statistics. LC1109EW (Household composition by age by sex). `https://www.nomisweb.co.uk/census/2011/lc1109ew`, 2011i.

Office for National Statistics. KS105EW (Household composition). `https://www.nomisweb.co.uk/census/2011/ks105ew`, 2011j.

Office for National Statistics. Sex by age. `https://www.nomisweb.co.uk/census/2011/lc1117ew`, 2011k.

Office for National Statistics. Output Area (2011) to Major Towns and Cities (December 2015) Lookup in England and Wales. `https://geoportal.statistics.gov.uk/datasets/78ff27e752e44c3194617017f3f15929`, 2015.

Office for National Statistics. Birth characteristics in England and Wales: 2017. `https://www.ons.gov.uk/releases/birthcharacteristicsinenglandandwales2017`, 2017a.

Office for National Statistics. Marriages in England and Wales. `https://www.ons.gov.uk/peoplepopulationandcommunity/`

`birthsdeathsandmarriages/marriagecohabitationandcivilpartnerships/`
`datasets/marriagesinenglandandwales2013`, 2017b.

Office for National Statistics. Deaths involving COVID-19 in the care sector, England and Wales: deaths occurring up to 12 June 2020 and registered up to 20 June 2020 (provisional). `https : / / www.ons.gov.uk /`
`peoplepopulationandcommunity / birthsdeathsandmarriages / deaths /`
`articles / deathsinvolvingcovid19inthecaresectorenglandandwales /`
`deathsoccurringupto12june2020andregisteredupto20june2020provisional`,
2020a.

Office for National Statistics. Coronavirus and the latest indicators for the UK economy and society: 29 October 2020.
`https : / / www.ons.gov.uk / peoplepopulationandcommunity /`
`healthandsocialcare / conditionsanddiseases / bulletins /`
`coronavirustheukeconomyandsocietyfasterindicators / 29october2020`,
2020b.

Office for National Statistics. Coronavirus and key workers in the UK. `https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/`
`earningsandworkinghours / articles / coronavirusandkeyworkersintheuk /`
`2020-05-15`, 2020c.

Office for National Statistics. Coronavirus and the social impacts on Great Britain: 25 September 2020.
`https : / / www.ons.gov.uk / peoplepopulationandcommunity /`
`healthandsocialcare / healthandwellbeing / bulletins /`
`coronavirusandthesocialimpactsongreatbritain / 25september2020`,
2020d.

G. Ogiya, M. Mori, Y. Miki, T. Boku, and N. Nakasato. Studying the core-cusp problem in cold dark matter halos using $N$-body simulations on GPU clusters.

In *Journal of Physics: Conference Series*, volume 454, page 012014. IOP Publishing, 2013.

G. Ogiya, F. C. van den Bosch, O. Hahn, S. B. Green, T. B. Miller, and A. Burkert. DASH: a library of dynamical subhalo evolution. *Monthly Notices of the Royal Astronomical Society*, 485(1):189–202, 2019.

A. O'Hagan. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, 91:1290—1300, 2006.

J. Onions, A. Knebe, F. R. Pearce, S. I. Muldrew, H. Lux, S. R. Knollmann, Y. Ascasibar, P. Behroozi, P. Elahi, J. Han, et al. Subhaloes going notts: the subhalo-finder comparison project. *Monthly Notices of the Royal Astronomical Society*, 423(2):1200–1214, 2012.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

OpenTable. The state of the restaurant industry. `https://www.opentable.com/state-of-industry`, 2020.

J. Ostriker and J.-K. Mark. Rapidly rotating stars. I. the self-consistent-field method. *The Astrophysical Journal*, 151:1075–1088, 1968.

T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. URL `https://doi.org/10.5281/zenodo.3509134`.

J. Penarrubia, A. J. Benson, M. G. Walker, G. Gilmore, A. W. McConnachie, and L. Mayer. The impact of dark matter cusps and cores on the satellite galaxy population around spiral galaxies. *Monthly Notices of the Royal Astronomical Society*, 406(2):1290–1305, 2010.

A. A. Penzias and R. W. Wilson. A measurement of excess antenna temperature at 4080 mc/s. *Astrophysical Journal, vol. 142, p. 419-421*, 142:419–421, 1965.

M. Pollán, B. Pérez-Gómez, R. Pastor-Barriuso, J. Oteo, M. A. Hernán, M. Pérez-Olmeda, J. L. Sanmartín, A. Fernández-García, I. Cruz, N. Fernández de Larrea, M. Molina, F. Rodríguez-Cabrera, M. Martín, P. Merino-Amador, J. León Paniagua, J. F. Muñoz-Montalvo, F. Blanco, R. Yotti, F. Blanco, R. Gutiérrez Fernández, M. Martín, S. Mezcua Navarro, M. Molina, J. F. Muñoz-Montalvo, M. Salinero Hernández, J. L. Sanmartín, M. Cuenca-Estrella, R. Yotti, J. León Paniagua, N. Fernández de Larrea, P. Fernández-Navarro, R. Pastor-Barriuso, B. Pérez-Gómez, M. Pollán, A. Avellón, G. Fedele, A. Fernández-García, J. Oteo Iglesias, M. T. Pérez Olmeda, I. Cruz, M. E. Fernandez Martinez, F. D. Rodríguez-Cabrera, M. A. Hernán, S. Padrones Fernández, J. M. Rumbao Aguirre, J. M. Navarro Marí, B. Palop Borrás, A. B. Pérez Jiménez, M. Rodríguez-Iglesias, A. M. Calvo Gascón, M. L. Lou Alcaine, I. Donate Suárez, O. Suárez Álvarez, M. Rodríguez Pérez, M. Cases Sanchís, C. J. Villafáfila Gomila, L. Carbo Saladrigas, A. Hurtado Fernández, A. Oliver, E. Castro Feliciano, M. N. González Quintana, J. M. Barrasa Fernández, M. A. Hernández Betancor, M. Hernández Febles, L. Martín Martín, L.-M. López López, T. Ugarte Miota, I. De Benito Población, M. S. Celada Pérez, M. N. Vallés Fernández, T. Maté Enríquez, M. Villa Arranz, M. Domínguez-Gil González, I. Fernández-Natal, G. Megías Lobón, J. L. Muñoz Bellido, P. Ciruela, A. Mas i Casals, M. Doladé Botías, M. A. Marcos Maeso, D. Pérez del Campo, A. Félix de Castro, R. Limón Ramírez, M. F. Elías Retamosa, M. Rubio González, M. S. Blanco Lobeiras, A. Fuentes Losada, A. Aguilera, G. Bou, Y. Caro, N. Marauri, L. M. Soria Blanco, I. del Cura González, M. Hernández Pascual, R. Alonso Fernández, P. Merino-Amador, N. Cabrera Castro, A. Tomás Lizcano, C. Ramírez Almagro, M. Segovia Hernández, N. Ascunce Elizaga, M. Ederra Sanz, C. Ezpeleta Baquedano, A. Bustinduy Bascaran, S. Iglesias Tamayo, L. Elorduy Otazua, R. Benarroch Benarroch, J. Lopera Flores, and A. Vázquez de la Villa. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *The Lan-*

*cet*, 396(10250):535–544, Aug. 2020. ISSN 0140-6736. doi: 10.1016/S0140-6736(20)31483-5. URL `https://doi.org/10.1016/S0140-6736(20)31483-5`. Publisher: Elsevier.

C. Power, J. F. Navarro, A. Jenkins, C. S. Frenk, S. D. White, V. Springel, J. Stadel, and T. Quinn. The inner structure of ΛCDM haloes—I. A numerical convergence study. *Monthly Notices of the Royal Astronomical Society*, 338(1):14–34, 2003.

W. Press and M. Davis. How to identify and weigh virialized clusters of galaxies in a complete redshift catalog. *The Astrophysical Journal*, 259:449–473, 1982.

W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery. *Numerical Recipes Example Book (FORTRAN)*. Cambridge University Press, 1992.

Public Health England (PHE). COVID-19 Hospitalisation in England Surveillance System (CHESS) – daily reporting. `https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/03/phe-letter-to-trusts-re-daily-covid-19-hospital-surveillance-11-march-2020.pdf`, 2020. Accessed: 14 December 2020.

F. Riccardo, M. Ajelli, X. D. Andrianou, A. Bella, M. Del Manso, M. Fabiani, S. Bellino, S. Boros, A. M. Urdiales, V. Marziano, M. C. Rota, A. Filia, F. P. D'Ancona, A. Siddu, O. Punzo, F. Trentini, G. Guzzetta, P. Poletti, P. Stefanelli, M. R. Castrucci, A. Ciervo, C. Di Benedetto, M. Tallon, A. Piccioli, S. Brusaferro, G. Rezza, S. Merler, and P. Pezzotti. Epidemiological characteristics of COVID-19 cases in Italy and estimates of the reproductive numbers one month into the epidemic. *medRxiv*, 2020. doi: 10.1101/2020.04.08.20056861. URL `https://www.medrxiv.org/content/early/2020/04/11/2020.04.08.20056861`.

A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The astronomical journal*, 116(3):1009, 1998.

R. J. Rockett, A. Arnott, C. Lam, R. Sadsad, V. Timms, K.-A. Gray, J.-S. Eden, S. Chang, M. Gall, J. Draper, et al. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature medicine*, 26(9):1398–1404, 2020.

V. C. Rubin and W. K. Ford Jr. Rotation of the andromeda nebula from a spectroscopic survey of emission regions. *The Astrophysical Journal*, 159:379, 1970.

R. E. Russell, R. A. Katz, K. Richgels, D. P. Walsh, and E. Grant. A Framework for Modeling Emerging Diseases to Inform Management. *Emerging Infectious Diseases*, 23(1):1–6, 2017. URL `https://dx.doi.org/10.3201/eid2301.161452`.

J. L. Sanders, E. J. Lilley, E. Vasiliev, N. W. Evans, and D. Erkal. Models of distorted and evolving dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 499(4):4793–4813, 2020.

Scientific Advisory Group for Emergencies. Dynamic CO–CIN report to SAGE and NERVTAG, 13 May 2020. `https://www.gov.uk/government/publications/dynamic-co-cin-report-to-sage-and-nervtag-13-may-2020`.

Scientific Advisory Group for Emergencies. Dynamic CO-CIN report to SAGE and NERVTAG - 30 June 2020. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/903395/S0612_Dynamic_CO-CIN_report_to_SAGE_and_NERVTAG.pdf`, 2020.

V. Springel. The cosmological simulation code gadget-2. *Monthly notices of the royal astronomical society*, 364(4):1105–1134, 2005.

V. Springel and S. D. White. Tidal tails in cold dark matter cosmologies. *Monthly Notices of the Royal Astronomical Society*, 307(1):162–178, 1999.

V. Springel, S. D. White, G. Tormen, and G. Kauffmann. Populating a cluster of galaxies–I. Results at z= 0. *Monthly Notices of the Royal Astronomical Society*, 328(3):726–750, 2001a.

V. Springel, N. Yoshida, and S. D. White. GADGET: a code for collisionless and gasdynamical cosmological simulations. *New Astronomy*, 6(2):79–117, 2001b.

V. Springel, J. Wang, M. Vogelsberger, A. Ludlow, A. Jenkins, A. Helmi, J. F. Navarro, C. S. Frenk, and S. D. White. The Aquarius Project: the subhaloes of galactic haloes. *Monthly Notices of the Royal Astronomical Society*, 391(4): 1685–1711, 2008.

V. Springel, R. Pakmor, O. Zier, and M. Reinecke. Simulating cosmic structure formation with the GADGET-4 code. *Monthly Notices of the Royal Astronomical Society*, 506(2):2871–2949, 2021.

J. Stadel, D. Potter, B. Moore, J. Diemand, P. Madau, M. Zemp, M. Kuhlen, and V. Quilis. Quantifying the heart of darkness with GHALO–a multibillion particle simulation of a galactic halo. *Monthly Notices of the Royal Astronomical Society: Letters*, 398(1):L21–L25, 2009.

C. Suen, H. Leung, K. Lam, et al. Feasibility of Reusing Surgical Mask Under Different Disinfection Treatments. *medRxiv*, 2020. doi: https://doi.org/10.1101/ 2020.05.16.20102178. URL `https://www.medrxiv.org/content/early/2020/ 05/20/2020.05.16.20102178`.

E. Toomey, Y. Conway, C. Burton, S. Smith, M. Smalle, X. Chan, A. Adisesh, S. Tanveer, L. Ross, I. Thomson, D. Devane, and T. Greenhalgh. Extended use or reuse of single-use surgical masks and filtering face-piece respirators during the coronavirus disease 2019 (COVID-19) pandemic: A rapid systematic review. *Infect Control Hosp Epidemiol*, 8:1–9, 2020.

UK Department for Education. Attendance in education and early years settings during the coronavirus (COVID-19) outbreak. `https://www.gov.uk/ government/collections/attendance-in-education-and-early-years- settings-during-the-coronavirus-covid-19-outbreak`, 2020.

UK Department for Transport. RAI0201 (City centre peak and all day arrivals and departures by rail on a typical autumn weekday, by city). `https://www.gov.uk/government/statistical-data-sets/rai02-capacity-and-overcrowding`, 2011.

F. C. van den Bosch. Dissecting the evolution of dark matter subhaloes in the Bolshoi simulation. *Monthly Notices of the Royal Astronomical Society*, 468(1): 885–909, 2017.

F. C. van den Bosch and F. Jiang. Statistics of dark matter substructure–ii. comparison of model with simulation results. *Monthly Notices of the Royal Astronomical Society*, 458(3):2870–2884, 2016.

F. C. van den Bosch and G. Ogiya. Dark matter substructure in numerical simulations: a tale of discreteness noise, runaway instabilities, and artificial disruption. *Monthly Notices of the Royal Astronomical Society*, 475(3):4066–4087, 2018.

F. C. van den Bosch, G. Ogiya, O. Hahn, and A. Burkert. Disruption of dark matter substructure: fact or fiction? *Monthly Notices of the Royal Astronomical Society*, 474(3):3043–3066, 2018.

G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

I. Vernon, M. Goldstein, and R. G. Bower. Galaxy Formation: a Bayesian Uncertainty Analysis. *Bayesian Analysis*, 5(4):619–670, 2010a.

I. Vernon, M. Goldstein, and R. G. Bower. Rejoinder for Galaxy Formation: a Bayesian Uncertainty Analysis. *Bayesian Analysis*, 5(4):697–708, 2010b.

I. Vernon, J. Liu, M. Goldstein, J. Rowe, J. Topping, and K. Lindsey. Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions. *BMC Systems Biology*, 12(1): arXiv:1607.06358 [q–bio.MN], 2018.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

H. Ward, C. J. Atchison, M. Whitaker, K. E. C. Ainslie, J. Elliott, L. C. Okell, R. Redd, D. Ashby, C. A. Donnelly, W. Barclay, A. Darzi, G. Cooke, S. Riley, and P. Elliott. Antibody prevalence for SARS-CoV-2 in England following first peak of the pandemic: REACT2 study in 100,000 adults. *medRxiv*, 2020. doi: 10.1101/2020.08.12.20173690. URL `https://www.medrxiv.org/content/early/2020/08/21/2020.08.12.20173690`.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

S. D. White, M. Davist, G. Efstathioui, and C. S. Frenk. Galaxy distribution in a cold dark matter universe. *Nature*, 330(6147):451–453, 1987.

E. J. Williamson, A. J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C. E. Morton, H. J. Curtis, A. Mehrkar, D. Evans, P. Inglesby, J. Cockburn, H. I. McDonald, B. MacKenna, L. Tomlinson, I. J. Douglas, C. T. Rentsch, R. Mathur, A. Y. S. Wong, R. Grieve, D. Harrison, H. Forbes, A. Schultze, R. Croker, J. Parry, F. Hester, S. Harper, R. Perera, S. J. W. Evans, L. Smeeth, and B. Goldacre. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821):430–436, 2020.

YouGov. YouGov COVID-19 behaviour changes tracker: Avoiding going to work.

`https://yougov.co.uk/topics/international/articles-reports/2020/03/` `17/personal-measures-taken-avoid-covid-19`, 2020a.

YouGov. YouGov COVID-19 behaviour changes tracker: Wearing a face mask when in public places. `https://yougov.co.uk/topics/international/articles-` `reports/2020/03/17/personal-measures-taken-avoid-covid-19`, 2020b.

C. Zoellner, R. Jennings, M. Wiedmann, and R. Ivanek. EnABLe: An agent-based model to understand Listeria dynamics in food processing facilities. *Scientific reports*, 9(1):1–14, 2019.

F. Zwicky. The redshift of extragalactic nebulae. *Helvetica Physica Acta*, 6:110–127, 1933.