Universidad Autónoma de Madrid | Escuela de Doctorado

Improving Cosmological Simulations: Semi-analytic Models, the Internal Structure of Haloes, and Machine Learning Techniques

Daniel López Cano,

Programa de Doctorado en Astrofísica

Centro realización Tesis/Institución Donostia International Physics Center (DIPC).

Departamento UAM/Facultad de Ciencias.

Madrid, 2024

Facultad de Física
Departamento de Física Teorica

Improving Cosmological Simulations: Semi-analytic Models, the Internal Structure of Haloes, and Machine Learning Techniques

PhD thesis

Candidate:

Daniel López-Cano

Supervisor:

Prof. Dr. Raul E. Angulo

May 9, 2024

Universidad Autónoma
de Madrid

The universe itself keeps on expanding and expanding
In all of the directions it can whizz
. . .
So remember, when you're feeling very small and insecure
How amazingly unlikely is your birth
And pray that there's intelligent life somewhere out in space
'Cause it's bugger all down here on Earth

"The Galaxy Song" from the 1983 film Monty Python's The Meaning of Life.

Abstract

This compendium thesis compiles three scientific articles elaborated during my doctoral studies. These articles are framed within the field of computational cosmology and explore different aspects related to the analysis and interpretation of cosmological simulations. Cosmological simulations involve computational algorithms for studying the behaviour of multiple particles under the effect of gravity in a cosmological context. Such simulations are instrumental in characterizing the formation processes and to connect the theoretical predictions of cosmological models with the observations made by telescopes that map the positions and properties of hundreds of millions of galaxies. In this document, after a brief introduction to the field of modern cosmology, I will describe the three main articles that constitute this thesis. First, I will explain how it is possible to employ certain models known as "semi-analytical" to populate with galaxies the simulations that only use gravitational interactions to predict the formation of structures. This way it is possible to connect the results from simulations with galaxy observations from space-based or ground-based telescopes. Secondly, I will present a model that we have developed for accurately predicting the internal structure of dark matter halos, which are gravitationally bound structures that generate the potential wells within which galaxies form. This model accurately captures how the internal structure of the halos depends on their formation time, which in turn depends on other properties such as halo mass, the cosmic time at which halos are observed, and the underlying cosmological model that is assumed. Finally, I will explain how to develop a model based on machine learning techniques for predicting regions in the initial conditions of a simulation that end up forming different dark matter halos. This technique makes use of neural networks to capture complex halo formation processes and can be used to make fast predictions, as well as to investigate which aspects of the initial conditions play a role in halo formation. Altogether, these studies contribute to improve the analysis and interpretation of cosmological simulations. Moreover, they show how the use of novel techniques such as machine learning methods can complement traditional methods for studying structure formation processes. These advances are currently of paramount importance as cosmological simulations represent the most important tool employed to interpret galaxy survey observations. In conclusion, the results presented in this work contribute to enhancing our general knowledge about the structure of the Universe and offer a novel perspective from which to approach observational cosmology, paving the way for future research.

Resumen

Esta tesis por compendio consiste en la recopilación de tres artículos científicos elaborados durante el transcurso de mis estudios doctorales. Estos artículos se enmarcan dentro del campo de la cosmología computacional y exploran distintos aspectos relacionados con el análisis y la interpretación de las simulaciones cosmológicas. Las simulaciones cosmológicas se basan en el empleo de algoritmos computacionales para estudiar el comportamiento de múltiples partículas bajo los efectos de la gravedad en un contexto cosmológico. Estas simulaciones sirven para caracterizar la formación de estructuras en el Universo y conectar las predicciones teóricas de modelos cosmológicos con las observaciones llevadas a cabo por telescopios encargados de registrar las posiciones y propiedades de cientos de millones de galaxias. En este documento, tras realizar una breve introducción al campo de la cosmología actual, pasaré a describir los tres trabajos principales que conforman esta tesis. En primer lugar explicaré cómo es posible usar ciertos modelos conocidos como "semi-analíticos" para poblar con galaxias simulaciones que sólo emplean la interacción gravitatoria para describir la formación de estructuras. De esta forma es posible conectar los resultados de simulaciones con los datos sobre posiciones de galaxias observados por telescopios espaciales o terrestres. En segundo lugar presentaré un modelo que hemos desarrollado capáz de predecir con precisión la estructura interna de halos de materia oscura, estructuras ligadas gravitacionalmente que generan los pozos de potencial gravitacional dentro de los cuales se forman las galaxias. Este modelo captura con precisión cómo depende la estructura interna de los halos en función su instante de formación, lo cual depende a su vez de otras propiedades tales cómo la masa de los propios halos, el tiempo cósmico en el cual se observan, y el modelo cosmológico subyacente aque se asume. Por último explicaré cómo es posible emplear un modelo basado en técnicas de aprendizaje automático para predecir las regiones en las condiciones iniciales de una simulación que acaban formando distintos halos de materia oscura. Este técnica emplea redes neuronales para caracterizar procesos complejos de formación de halos y puede ser empleada tanto para realizar predicciones rápidas, como para investigar qué aspectos en las condiciones iniciales juegan un papel relevante en la formación de halos. En conjunto, todos estos trabajos mejoran el análisis y la interpretación de las simulaciones cosmológicas. Además, muestran cómo el uso de ténicas novedosas como elaprendizaje automático pueden complementarse con métodos tradicionales para estudiar procesos de formación de estructuras. Todos estos avances resultan de capital importancia en el momento actual ya que las simulaciones cosmológicas son la herramienta principal empleada para intepretar los datos recogidos por nuevas campañas

observacionales que registran la estructura a gran escala de nuestro universo a través de las posiciones de numerosas galaxias. A modo de conclusión, los métodos presentados en este trabajo ayudan a mejorar nuestro conocimiento en general sobre la estructura del Universo y oferen una perspectiva novedosa desde la cual investigar la situación actual de la cosmología observacional allanando el camino para futuras investigaciones.

Index

Introduction

As I have highlighted in the abstract, the primary research topics I have addressed during my PhD relate to structure formation processes, the study of the internal structure of dark matter (DM) haloes, and the characterization of galaxy populations through cosmological simulations. Throughout my years of doctoral studies, I have also dedicated significant time to understanding fundamental aspects of cosmology that constitute the theoretical backbone upon which different research areas in this field sprout.

In this introduction, I will emphasize some of the most relevant principles that serve as building blocks of cosmology as a whole and make my way to the current open problems I have been working on. My intention is not to create a self-contained manuscript from the most fundamental aspects of cosmology to the current state of the field. Instead, I aim to methodically present core topics I consider essential, from the formulation of general relativity using differential geometry to modern cosmology topics like numerical simulations, large-scale structure surveys, and structure formation theory. My goal is to present an introduction that serves as a structured roadmap, referencing comprehensive sources and covering relevant topics at various levels to outline cosmology's broader landscape.

The introduction is structured as follows: first I will outline some basic concepts of differential geometry that serve as pillars for defining general relativity (0.1). Then I will introduce in a simple way the formulation of general relativity (0.2). Afterwards, I will explain how cosmology arises from general relativity by solving the background homogeneous and isotropic case (0.3). Next, I will focus on topics more relevant to my work related to structure formation processes and the growth of perturbations (0.4). Finally, I will comment on some of the current open problems in cosmology directly relevant to my research (0.5), and contextualize the different projects I have been working on (0.6), thus paving the way to the main chapters of this thesis.

0.1 Fundamentals of differential geometry

I believe that the best way to learn any topic in physics is to first gain a solid understanding of its mathematical foundations. While developing intuition is also crucial, it becomes

increasingly unclear what form this intuition should take as the physical process at hand strays from our everyday classical experience. This is especially true for areas like quantum physics and general relativity where intuition can only flourish after some previous effort in understanding (and practising) with its underlying mathematical footing.

Although the role of a mathematical formalism, in this case differential geometry, might not play a fundamental role in day-to-day calculations (depending on the field), I think it is crucial to have a good grasp of it. Differential geometry is the real theoretical backbone upon which general relativity (and therefore cosmology) is built. Many interesting phenomena can only be truly comprehended after consolidating the basic knowledge of this topic. Understanding differential geometry it is also essential to comprehend how fundamental modifications can yield alternative theories of gravity other than general relativity.

The lectures by Prof. Frederic P. Schuller at the WE-Heraeus International Winter School on Gravity and Light provide an excellent introduction to this topic Schuller (2015). A detailed transcription of these lectures, accompanied by supplementary materials, is available in Dadhley (2015). This series of lectures pivots around a central sentence that encapsulates the formulation of general relativity through differential geometry: "*Spacetime is a four-dimensional topological manifold with a smooth atlas carrying a torsion-free connection compatible with a Lorentzian metric and a time orientation satisfying the Einstein equations*".

To fully grasp the meaning of this sentence, it is necessary to dissect it into smaller chunks and tackle each of them separately. In Appendix A, I present a series of definitions to elucidate what constitutes a "*four-dimensional topological manifold with a smooth atlas*". For a detailed explanation of the sentence's middle portion – "*carrying a torsion-free connection compatible with a Lorentzian metric and a time orientation*" – I direct the reader to Schuller (2015) and Dadhley (2015). The final part of this sentence, "*satisfying the Einstein equations*", is addressed in the following Section 0.2.

Although it is unavoidable to delve into precise mathematical definitions to truly understand this phrase, I will try to give an intuitive explanation of what the different parts of it mean before moving on to the next section. The three-dimensional space that we are familiar with can be regarded along with time as a single mathematical entity comprised by a set of points in four dimensions with nice properties. By "nice properties" I refer to the fact that the points of spacetime behave in relation to each other in such a way that it is possible to represent them in an orderly and smooth manner, somewhat like an elastic fabric. On top of this structure that represents a "four-dimensional smooth manifold" we need to impose additional constraints that describe how objects (for example, tensors) transform when they move from one point of this space to another (this is related to the part "carrying a torsion-free connection"). We also need to specify further requirements to build a physical theory which

is locally compatible with special relativity (in connection with the part "with a Lorentzian metric and a time orientation"). Nevertheless, this construction does not fully constitute what we refer to as the theory of general relativity; it is necessary to state how the presence of energy and its momentum affects the "shape of our well-behaved and soft elastic fabric", and for that we need to postulate Einstein's equations.

0.2 General relativity

The differential geometry framework presented in Section 0.1 (and references therein) specifies how spacetime is defined in the context of general relativity. However, the physical core of general relativity is constituted by the Einstein equations. Einstein's equations describe how the metric, $g_{\mu\nu}$, which characterizes the structure of spacetime, is affected by the matter-energy distribution, described by the stress-energy tensor $T_{\mu\nu}$.

Numerous comprehensive sources provide excellent introductions to general relativity and Einstein's equations, for example, see (Ortín, 2007; Zee, 2013). In this section, I will outline how to derive the equations of motion of a system from its Lagrangian density (using the stationary-action principle) and formulate the Cosmological Einstein equations from the Einstein-Hilbert action.

For any given **Lagrangian density**, $\mathcal{L}$, explicitly dependent on special relativistic fields[2], $\phi^i(x^\mu)$, and their corresponding first derivatives[3], $\partial_\nu\phi^i(x^\mu)$, we can write its corresponding **action** as:

$$\mathcal{S}\left[\phi^i(x^\mu), \partial_\nu\phi^i(x^\mu); x^\mu\right] = \int_\Sigma d^4x \left\{\mathcal{L}\left[\phi^i(x^\mu), \partial_\nu\phi^i(x^\mu); x^\mu\right]\right\}$$

Taking an arbitrary infinitesimal variation, δ_α, and assuming the coordinate variations δx^μ to be zero by hypothesis,

$$\delta_\alpha\mathcal{S} = \int_\Sigma d^4x \left\{\delta_\alpha\mathcal{L}\right\} = \int_\Sigma d^4x \left\{\partial_{\phi^i}\mathcal{L}\ \delta_\alpha\phi^i + \partial_{\partial_\mu\phi^i}\mathcal{L}\ \delta_\alpha\partial_\mu\phi^i \xrightarrow{[\partial_\mu,\delta_\alpha]=0} \ldots\right.$$

$$\ldots \rightarrow \delta_\alpha\mathcal{S} = \int_\Sigma d^4x \left\{\delta_\alpha\phi^i\left[\partial_{\phi^i}\mathcal{L} - \partial_\mu\partial_{\partial_\mu\phi^i}\mathcal{L}\right]\right\} + \int_\Sigma d^4x \left\{\partial_\mu\left[\partial_{\partial_\mu\phi^i}\mathcal{L}\ \delta_\alpha\phi^i\right]\right\}$$

If we now impose that the field variations vanish over the boundary, $[\delta_\alpha\phi^i|_{\partial\Sigma^4}$, the second term cancels out:

$$\int_{\partial\Sigma} d^4x \left\{\partial_\mu\left[\partial_{\partial_\mu\phi^i}\mathcal{L}\delta_\alpha\phi^i\right]\right\} = \left[\partial_{\partial_\mu\phi^i}\mathcal{L}\ \delta_\alpha\phi^i|_{\partial\Sigma^4} = 0.\right.$$

[2]Defined on a four-dimensional smooth manifold whose elements we denote with coordinate map components x^μ.

[3]I will employ the notation $\partial_\mu \coloneqq \frac{\partial}{\partial x^\mu}$. It is possible to consider the more general case where the Lagrangian density, $\mathcal{L}$, also depends on higher derivative terms of the field, $\left\{\phi^i, \partial_\nu\phi^i, \partial_\kappa\partial_\nu\phi^i, \ldots\right.$. However, most physical laws can be obtained considering only the first two terms.

Taking into account the **stationary-action principle**, $\delta_\alpha S = 0$, we obtain the **Euler-Lagrange equations** from the first term:

$$\delta_\alpha \mathcal{S} = \iint d^4x \left\{ \delta_\alpha \phi^i \left[\partial_{\phi^i} \mathcal{L} - \partial_\mu \partial_{\partial_\mu \phi^i} \mathcal{L} \right] \right\} \xrightarrow{\delta_\alpha S = 0} \partial_{\phi^i} \mathcal{L} - \partial_\mu \partial_{\partial_\mu \phi^i} \mathcal{L} = 0.$$

Although not all the equations in physics can be obtained from an action formulation, most of them do. In particular, the equations of motion for the gravitational field, Einstein's equations, can be derived starting with the **Einstein-Hilbert action**,

$$\mathcal{S}_{\mathrm{EH}} \left[g_{\mu\nu}(x^\mu); x^\mu \right] = \frac{c^3}{16\pi G_{\mathrm{N}}} \iint_\Sigma \mathrm{d}^4x \sqrt{-g} R(g_{\mu\nu}),$$

where $G_{\mathrm{N}} = 6.67430(15) \cdot 10^{-11} m^3 kg^{-1} s^{-2}$ (Mohr and Taylor, 2000) is the gravitational constant, and $R(g_{\mu\nu})$ is the Ricci scalar, which depends on the metric as follows:

$$\begin{aligned}
\text{Ricci scalar:} \quad & R \coloneqq g^{\mu\nu} R_{\mu\nu} \\
\text{Ricci curvature:} \quad & R_{\mu\nu} \coloneqq R^\alpha{}_{\mu\alpha\nu} \\
\text{Riemann-Christoffel curvature tensor:} \quad & R^\rho{}_{\sigma\mu\nu} \coloneqq \partial_\mu \Gamma^\rho{}_{\nu\sigma} - \partial_\nu \Gamma^\rho{}_{\mu\sigma} + \Gamma^\rho{}_{\mu\lambda} \Gamma^\lambda{}_{\nu\sigma} - \Gamma^\rho{}_{\nu\lambda} \Gamma^\lambda{}_{\mu\sigma} \\
\text{Christoffel Symbols of the Second Kind:} \quad & \Gamma_{\mu\nu\kappa} \coloneqq \frac{1}{2} \left(\partial_\kappa g_{\mu\nu} + \partial_\nu g_{\mu\kappa} - \partial_\mu g_{\nu\kappa} \right).
\end{aligned}$$

After extensive manipulation (see Ortín, 2007, for an explicit derivation), we can derive the following equation:

$$\delta \mathcal{S}_{\mathrm{EH}} \left[g_{\mu\nu}(x^\mu); x^\mu \right] = \int_\Sigma \mathrm{d}^4x \sqrt{-g} \left\{ \delta g^{\alpha\beta} \left[R_{\alpha\beta} - \frac{1}{2} g_{\alpha\beta} R \right] + \nabla_\alpha \left(g^{\alpha\beta} \delta \Gamma^\kappa_{\kappa\beta} - g^{\beta\kappa} \delta \Gamma^\alpha_{\beta\kappa} \right) \right\}$$

If the covariant derivative[4] gives a term that vanishes when $\delta g_{\mu\nu} = 0$ on the boundary[5], we obtain the **Einstein equations in the vacuum**,

$$G^{\mu\nu} \coloneqq R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R = 0.$$

[4]The covariant derivative ∇_κ, of a tensor r-times contravariant and s-times covariant, $T^{\mu_1 \cdots \mu_r}_{\nu_1 \ldots \nu_s}$, is defined as:

$$\begin{aligned}
\nabla_\kappa T^{\mu_1 \cdots \mu_r}_{\nu_1 \ldots \nu_s} =& \partial_\kappa T^{\mu_1 \cdots \mu_r}_{\nu_1 \ldots \nu_s} + \ldots \\
& \ldots + \Gamma^{\mu_1}_{\alpha\kappa} T^{\alpha \mu_2 \cdots \mu_r}_{\nu_1 \ldots \nu_s} + \ldots + \Gamma^{\mu_r}_{\alpha\kappa} T^{\mu_1 \ldots, \mu_{r-1} \alpha}_{\nu_1 \ldots \nu_s} \\
& \ldots - \Gamma^\alpha_{\nu_1 \kappa} T^{\mu_1 \cdots \mu_r}_{\alpha \nu_2 \ldots \nu_s} - \ldots - \Gamma^\alpha_{\nu_s \kappa} T^{\mu_1 \cdots \mu_r}_{\nu_1 \ldots \nu_{s-1} \alpha}
\end{aligned}$$

[5]At this point it is necessary to point out a very interesting appreciation about the Einstein-Hilbert action that is rarely mentioned in most textbooks but that is carefully addressed in Ortín (2007): "*The Einstein–Hilbert action contains second derivatives of the metric. However, the terms with second derivatives take the form of a total derivative. This means that the original action can in principle be used to obtain equations of motion that are of second order in derivatives of the metric. However, we would have to impose conditions on the derivatives of the metric on the boundary. [...]. The solution to these problems consists in adding a general-covariant boundary term to the original Einstein–Hilbert action.*". This term is known as the **Gibbons–Hawking–York boundary term** and is given by:

$$\mathcal{S}_{\mathrm{GHY}} = \frac{1}{8\pi} \iint_{\partial \mathcal{M}} \mathrm{d}^3y \epsilon \sqrt{h} K.$$

We can now consider what is the effect of matter by including an additional term in the action, $\mathcal{S}_{\rm m}\left[g_{\mu\nu}(x^\mu), \phi^i(x^\mu); x^\mu\right]$, that depends on the matter fields, $\phi^i(x^\mu)$. If we define the **matter stress-energy tensor** as

$$T^{\mu\nu}_{\rm m}\left[\phi^i\right] \equiv \frac{2c}{\sqrt{-g}}\frac{\delta \mathcal{S}_{\rm m}}{\delta g_{\mu\nu}},$$

Where $T^{\mu\nu}_{\rm m}$ satisfies the **continuity equation**: $\nabla_\alpha T^{\mu\alpha}_{\rm m} = 0$. We recover the **Einstein equations in presence of matter**:

$$G^{\mu\nu} = \frac{4\pi G_{\rm N}}{c^4} T^{\mu\nu}_{\rm m}.$$

Finally, let's consider the effect of including a **cosmological constant**, Λ, term in our action,

$$\mathcal{S}_\Lambda\left[g_{\mu\nu}(x^\mu); x^\mu\right] = \frac{c^3}{16\pi G_{\rm N}} \int_\Sigma \mathrm{d}^4x\sqrt{-g}\left[-2\Lambda\right]. \rightarrow \mathcal{S} = \mathcal{S}_{\rm EH} + \mathcal{S}_{\rm m} + \mathcal{S}_\Lambda$$

In this case, we recover the **Cosmological Einsteins's equations** given by

$$G^{\mu\nu} = \frac{8\pi G_{\rm N}}{c^4} T^{\mu\nu}_{\rm m} - g^{\mu\nu}\Lambda. \tag{1}$$

These equations would be central for the rest of this work as they describe the interaction between spacetime and matter. Moreover, they incorporate the effect of a cosmological constant, currently the most widely accepted approach for modelling the observed accelerated expansion of the Universe (Riess et al., 1998).

0.3 The homogeneous universe

In this section I will present the basic assumptions taht are employed in cosmology for describing the evolution of the Universe as a whole on its largest scales. I will employ the machinery from general relativity presented in sections 0.1 & 0.2 to build towards the ΛCDM model, the current cornerstone of modern cosmology. I will derive the main equations that are employed for studying the behaviour of a homogeneous and isotropic universe. A comprehensive introduction to this topic can be found in classic references (Kurki-Suonio, 2024b, 2023; Dodelson and Schmidt, 2020; Baumann, 2022). My intention for this section is to present compact and direct derivations for analyzing dynamical aspects of the homogeneous universe. At the end of this section, I will also comment on some crucial thermodynamical results, but, since will not play a direct role on the main results Iwill present of this thesis, I refer the avid reader to the aforementioned sources for more in depth explanations.

To analyze the behavior of the Universe at its largest scales, we are going to assume that it looks the same at every point (**homogeneity**) and in every direction (**isotropy**).

This assumption captures the behaviour of the Universe as a whole effectively and allows to find simple solution to Einstein's equations (1). Imposing these conditions leads to a restricted form for the metric known as the **Friedmann–Lemaître–Robertson–Walker (FLRW)** metric given by:

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu \quad \text{with} \quad g_{\mu v} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & \frac{R(t)^2}{1-kx^2} & 0 & 0 \\ 0 & 0 & R(t)^2x^2 & 0 \\ 0 & 0 & 0 & R(t)^2x^2\sin^2\theta \end{pmatrix}, \tag{2}$$

where $x^\mu = (ct, x^i)$ are referred to as **comoving coordinates**, $R(t)$ is a parameter that describes how distances evolve with **cosmic time**, t, and k measures the **curvature** of the universe[6].

From this restricted metric, we can directly derive several interesting quantities. If we consider a photon (that follows a **null-geodesic** with $ds^2 = 0$) following a radial trajectory[7], we find that

$$c^2dt^2 = R^2(t)\frac{dx^2}{1-kx^2} \rightarrow \int_{t_{\rm E}}^{t_f} \frac{cdt}{R(t)} = \int_{x_i=0}^{x_f} \frac{dx}{\sqrt{1-kx^2}} = \begin{cases} \frac{\text{arcsinh}\left(\sqrt{|k|}x_f\right)}{\sqrt{|k|}} & \text{k} < 0 \\ x_f & \text{k} = 0 \\ \frac{\arcsin\left(\sqrt{|k|}x_f\right)}{\sqrt{|k|}} & \text{k} > 0 \end{cases} \tag{3}$$

From this expression we can define the **comoving distance**, χ, as the comoving separation between two comoving points. It is constant for objects that move with the Hubble flow (the intrinsic expansion of the universe) and can expressed as

$$\chi = \int_{x_i=0}^{x_f} \frac{dx}{\sqrt{1-kx^2}} = \int_{t_{\rm E}}^{t_0=t_{\rm E}+\Delta t} \frac{cdt}{R(t)}, \tag{4}$$

where x_f is the comoving coordinate reached by a photon that was emitted at time $t_{\rm E}$ from $x_i = 0$ and traveled during a time Δt (as measured by the observer located at $x_i = 0$). Note that independently of the emission time, $t_{\rm E}$, a photon will always arrive to the same comoving coordinate x_f at t_0. We can now define the **proper distance** as $d_{\rm p}(t) \coloneqq \chi\frac{R(t)}{R(t_0)}$. The proper distance, $d_{\rm p}(t)$, indicates the physical separation (measured in your favourite units: $\mathrm{m}, \mathrm{Gpc}, ...$) that corresponds to the comoving distance χ measured at a certain cosmic time t.

Using Equation 3 we can derive the relation between the scale factor, a, and redshift, z (both defined below). We start by considering the trajectories of two photons that are emitted one shortly after the other. We consider that the first photon is emitted from the comoving

[6]The curvature, k, can only take the values $k = +1$ (positive curvature), $k = 0$ (zero curvature, or flat), and $k = -1$ (negative curvature).

[7]Without loss of generality we can set $\theta = 0$.

position $x_{\rm E}$ at $t_{\rm E}$ and reaches its destination, x_0, at t_0. The second photon is emitted shortly after from $x_{\rm E}$ at $t_{\rm E}+\delta t_{\rm E}$ and reaches x_0 at $t_0+\delta t_0$, then,

$$\text{1st phot.}: \int_{x_{\rm E}}^{x_0}\frac{dx}{\sqrt{1-kx^2}}=\int_{t_{\rm E}}^{t_0}\frac{cdt}{R(t)} \quad . \text{ 2nd phot.}: \int_{x_{\rm E}}^{x_0}\frac{dx}{\sqrt{1-kx^2}}=\int_{t_{\rm E}+\delta t_{\rm E}}^{t_0+\delta t_0}\frac{cdt}{R(t)}.$$

Since the comoving spatial integrals are the equivalent, we can combine the temporal part of both equations,

$$\int_{t_{\rm E}}^{t_0}\frac{cdt}{R(t)}=\int_{t_{\rm E}+\delta t_{\rm E}}^{t_0+\delta t_0}\frac{cdt}{R(t)}=\int_{t_{\rm E}}^{t_0}\frac{cdt}{R(t)}-\int_{t_{\rm E}}^{t_{\rm E}+\delta t_{\rm E}}\frac{cdt}{R(t)}+\int_{t_0}^{t_0+\delta t_0}\frac{cdt}{R(t)}\rightarrow ...$$
$$... \rightarrow \int_{t_0}^{t_0+\delta t_0}\frac{cdt}{R(t)}=\int_{t_{\rm E}}^{t_{\rm E}+\delta t_{\rm E}}\frac{cdt}{R(t)}\xrightarrow[\delta t_{\rm E}<<1]{\delta t_0<<1}\frac{c\Delta t_0}{R(t_0)}=\frac{c\Delta t_{\rm E}}{R(t_{\rm E})}.$$

In the derivation's final step, I have considered the limit where the time intervals between the emission and reception of the photons ($\delta t_{\rm E}$ and δt_0, respectively) are so small that the scale factor of the universe can be considered constant. If we now consider that instead of having two different photons we are dealing with two events that correspond to an interval equivalent to the frequency of a particular lightray, such that $c\Delta t_j = c/\nu_j = \lambda_j$, we can rewrite the last equation as

$$\frac{\lambda_0}{R(t_0)}=\frac{\lambda_E}{R(t_{\rm E})}\rightarrow\frac{R(t_{\rm E})}{R(t_0)}=\frac{\lambda_E}{\lambda_0}=\frac{1}{1+\frac{\lambda_0}{\lambda_E}-1}\xrightarrow[z:=\lambda_0/\lambda_E-1]{a(t_{\rm E}):=R(t_{\rm E})/R(t_0)}a(t_{\rm E})=\frac{1}{1+z}. \qquad (5)$$

The **redshift** z measures the ratio of the observed wavelength, λ_0, to the emitted wavelength, $\lambda_{\rm E}$. We denote the **scale factor** by $a(t)$.

From Equation 3 it is also possible to define different characteristic scales relevant in cosmology:

- Since the age of the Universe and the speed of light are both finite, we can define the (comoving) **particle horizon**, $\chi_{\rm p}$, as the maximum (comoving) distance from which we can retrieve information, that is, the past observable universe. The particle horizon, $\chi_{\rm p}$, is defined in terms of Equation 3 as the (comoving) distance associated with a photon that was emitted at decoupling time, $t_i = t_{\rm dec}$, from $x_i = x_{\rm dec}$, that is received by an observer located at $x_f = 0$ at present time $t_f = t_0$. It is common to see the term particle horizon defined in terms of Big Bang time instead of decoupling time, in this case, $t_i = 0$, $x_i = x_{BB}$.

- The (comoving) **event horizon** corresponds to the maximum (comoving) distance a photon would be able to travel if it is emitted from $x_i = 0$ at present time, $t_i = t_0$, and travels for all eternity, $t_f = \infty$. It can be finite or infinite, depending on the behaviour of $R(t)$, which depends on the composition of the universe.

- The (comoving) **Hubble radius** is defined as the (comoving) distance at which recessional velocity equals the speed of light $\chi_H(t) = c/H(t)$. Where $H(t)$ is the Hubble parameter which I will define below.

So far we have discussed the consequences of restricting the metric to the homogeneous and isotropic FLRW case. We now move on to characterize the behaviour of $R(t)$ once a specific form of the stress-energy tensor, $T_{\mu\nu}$, is specified.

Let's consider that the total stress-energy tensor of the universe, $T_{\mu\nu}$, can be expressed as a sum of the individual stress-energy tensors of different perfect fluids, $T^{(i)}_{\mu\nu}$, that is,

$$T_{\mu\nu} = \sum_i T^{(i)}_{\mu\nu}, \quad \text{with} \quad T^{(i)}_{\mu\nu} = \begin{bmatrix} \rho_i & 0 & 0 & 0 \\ 0 & P_i & 0 & 0 \\ 0 & 0 & P_i & 0 \\ 0 & 0 & 0 & P_i \end{bmatrix}. \tag{6}$$

Where ρ_i is the energy density and P_i is the pressure of the "i-th" perfect fluid species.

Considering the continuity equation (see section 0.2), we obtain that each species satisfies:

$$\dot{\rho}_i = -3\frac{\dot{R}}{R}\left(\rho_i + p_i/c^2\right) \tag{7}$$

Where I have employed the notation $\dot{\rho} \coloneqq d\rho/dt$.

We can substitute Equation 6 on Equation 1 assuming a FLRW metric (Equation 2); after extensive manipulations we arrive to the **Friedmann Equations**,

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G_{\mathrm{N}}}{3}\sum_i \rho_i - \frac{kc^2}{R^2} + \frac{\Lambda c^2}{3}, \tag{8}$$

$$\frac{\ddot{R}}{R} = -\frac{4\pi G_{\mathrm{N}}}{3}\sum_i \left[\rho_i + \frac{3p_i}{c^2}\right] + \frac{\Lambda c^2}{3}, \tag{9}$$

We can simplify Equation 8 and Equation 9 even further but we need to take a small thermodynamic detour.

Let's assume that the second law of thermodynamics holds for each fluid component, $T_i dS_i = dU_i + p_i dV$. Where T_i is the temperature associated with the i-th fluid component, $U_i = \rho_i R(t)^3 c^2$ is its internal energy, S_i represents its entropy, and $V = R(t)^3$. Then:

$$T_i\dot{S}_i = \dot{U}_i + p_i\dot{V} = 3R^2\dot{R}\rho_i c^2 + R^3\dot{\rho}_i c^2 + 3p_i R^2\dot{R} \tag{10}$$

We can show that this expression is equal to zero substituting Equation 7. Therefore, $T_i\dot{S}_i = 0$, and since $T_i \neq 0\ \forall i, t\ \rightarrow \dot{S}_i = 0$, hence, entropy is conserved[8].

[8]For the condition $\dot{S}_i = 0$ people commonly say that the expansion of the universe is **"adiabatic"**, However, this term is employed in many different contexts and can mean diverse things, so be carefull out there.

If we now assume that each fluid component can be described as a **barotropic** fluid with $p_i = \omega_i \rho_i c^2$, from Equation 10 we have that

$$dS_i = 0 = dU_i + p_i dV = d\left(R^3 \rho_i c^2\right) + \omega_i \rho_i c^2 d\left(R^3\right) = c^2 R^3 d\rho_i + c^2 \rho_i dR^3 + \omega_i \rho_i c^2 dR^3 = ...$$
$$... = c^2 R^3 d\rho_i + c^2 \rho_i dR^3 \left(1 + \omega_i\right) \rightarrow -c^2 \rho_i dR^3 \left(1 + \omega_i\right) = c^2 R^3 d\rho_i \rightarrow ...$$
$$... \rightarrow -\left(1 + \omega_i\right) \frac{dR^3}{R^3} = \frac{d\rho_i}{\rho_i} \rightarrow -\int_{R(t_0)}^{R(t)} \left(1 + \omega_i\right) \frac{dR^3}{R^3} = \int_{\rho_i(t_0)}^{\rho_i(t)} \frac{d\rho_i}{\rho_i}$$

Assuming that ω_i is a constant we have that

$$-\left(1 + \omega_i\right) \int_{R(t_0)}^{R(t)} \frac{dR^3}{R^3} = \int_{\rho_i(t_0)}^{\rho(t)} \frac{d\rho_i}{\rho_i} \rightarrow \rho_i(t) = \rho_i(t_0)\, a(t)^{-3(1+\omega_i)} \tag{11}$$

This equation relates how the density of a barotropic fluid with constant ω_i evolves as a function of the scale factor. Using this result we can simplify Friedmann's equations after introducing some additional definitions:

- The **Hubble parameter** at time t, $H(t) \coloneqq \dot{R}/R$, indicates the ratio between the recession velocity of an object, that is static in the comoving frame, and its distance to a given observer, which is also at rest in the comoving frame. The **Hubble constant** is defined as $H_0 = H(t_0)$ and its value according to Planck Collaboration et al. (2020a) is $H_0 = (67.4 \pm 0.5)\mathrm{km\,s^{-1}\,Mpc^{-1}} \approx 0.069\mathrm{Gyr}^{-1}$.

- The time dependent **critical density of the universe** is defined as $\rho_{\mathrm{crit}}(t) \coloneqq \frac{3H(t)^2}{8\pi G_{\mathrm{N}}}$. At present time, $\rho_{\mathrm{crit}}(t_0) \equiv \frac{3H_0^2}{8\pi G_{\mathrm{N}}} = 1.8788 \times 10^{-26} h^2\,\mathrm{kg\,m^{-3}} = 2.7754 \times 10^{11} h^2\, M_\odot\,\mathrm{Mpc}^{-3}$, where the **little** h parameter is defined as $h \coloneqq H_0/(100\,\mathrm{km\ s^{-1}Mpc^{-1}})$ (Planck Collaboration et al., 2020a).

- We define the present time **density parameters** for the curvature and the cosmological constant as

$$\Omega_k(t_0) \equiv \Omega_{k,0} \coloneqq -\frac{kc^2}{H_0^2} \quad ; \quad \Omega_\Lambda(t_0) \equiv \Omega_{\Lambda,0} \coloneqq \frac{c^2 \Lambda}{3H_0^2}.$$

  The density parameter associated to a barotropic fluid is

$$\Omega_i(t_0) \equiv \Omega_{i,0} \coloneqq \frac{8\pi G_{\mathrm{N}}}{H_0^2} \rho_i(t_0).$$

  In general, we define the time-dependent density parameter for any component (curvature, cosmological constant, or barotropic fluid) as

$$\Omega_i(t) \coloneqq \Omega_{i,0} \frac{a(t)^{-3(1+\omega_i)}}{\Omega_{r,0} a^{-4} + \Omega_{m,0} a^{-3} + \Omega_{k,0} a^{-2} + \Omega_{\Lambda,0} a^0}.$$

  For curvature we are going to associate the value $\omega_k \coloneqq -1/3$. For the cosmological constant we define that $\omega_\Lambda \coloneqq 0$. Based on statistical mechanic arguments (see

Baumann, 2018), it can be shown that for radiation $\omega_r = 1/3$, and for collisionless matter $\omega_m = 0$. We will also enforce the following definition for all density parameters that would allow us to effectively talk about curvature and cosmological constant densities:

$$\Omega_i(t) \coloneqq \frac{\rho_i(t)}{\rho_{\rm crit}(t)} = \frac{8\pi G_{\rm N}}{H_0^2}\rho_i(t_0)\frac{a(t)^{-3(1+\omega_i)}}{\Omega_{r,0}a^{-4} + \Omega_{m,0}a^{-3} + \Omega_{k,0}a^{-2} + \Omega_{\Lambda,0}a^0}$$

Taking into account the definitions above and substituting the result of Equation 11 in Equations 8 and Equations 9 we obtain the Friedmann Equations written in the most common form:

$$H(a)^2 = H_0^2\left(\Omega_{r,0}a^{-4} + \Omega_{m,0}a^{-3} + \Omega_{k,0}a^{-2} + \Omega_{\Lambda,0}a^0\right) = H_0^2\sum_i \Omega_i(a)E^2(a) \tag{12}$$

$$q = -\frac{\ddot{R}R}{\dot{R}^2} = \frac{1}{2}\Omega_m(t) + \Omega_r(t) - \Omega_\Lambda(t) \tag{13}$$

These equations have some analytical solutions, for example in the case in which only one component is considered. However, to solve them in general way it is necessary to employ a numerical approach.

I would like to add that this derivation can be generalized for considering fluids that transition between a relativistic behaviour (contributing to the radiation term) and a non-relativistic behaviour (contributing to the matter term). This treatment becomes particularly important for describing the effect of neutrinos in the background. In the case of neutrinos it is possible to write the following term (see Lesgourgues and Pastor, 2006; Zennaro et al., 2016, and M. Zennaro & D. López-Cano, in prep., for a detailed derivation):

$$\Omega_\nu(a)E^2(a) = \left(\frac{15}{\pi^4}\right)\Gamma_\nu^4\Omega_\gamma(a)E^2(a)\mathcal{F}(y) = \left(\frac{15}{\pi^4}\right)\Gamma_\nu^4\Omega_{\gamma,0}a^{-4}\mathcal{F}(y).$$

$$\text{With} \quad F(y_i) \equiv \int_0^\infty dx_i \frac{x_i^2}{\sqrt{x_i^2 + y_i^2}}\frac{1}{e^{x_i}+1}. \quad \text{where} \quad x_i = \frac{p_i c}{k_B T_{\nu,i}} \quad \text{and} \quad y_i = \frac{m_{\nu,i}c^2}{k_B T_{\nu,i}}.$$

Were $\Gamma_\nu = 0.71611.$ is the non-instantaneous interaction rate of neutrinos with photons (see Hannestad and Madsen, 1995; Dolgov et al., 1997; Esposito et al., 2000).

In this section I have presented the Friedmann equations 12 & 13 starting from the cosmological Einstein's equations 1. Friedmann equations describe how a homogeneous and isotropic universe evolves with cosmic time depending on its composition[9] The information of the evolution is encoded in the scale factor. In the next section 0.4 I will go beyond the homogeneous framework to investigate how is it possible to model the growth of structures supposing that, from a certain scale downwards, there exists perturbations that imprint

[9]There exist many other interesting phenomena that can be analyzed in the "background" framework (decoupling, nucleosynthesis, time of recombination, etc.). For a comprehensive review check Kurki-Suonio (2024b, 2023); Dodelson and Schmidt (2020); Baumann (2022).

small-scale inhomogeneities. These inhomogeneities (which are originally created due to **cosmic inflation**) grow afterwards naturally due to effect of gravity.

0.4 Growth of perturbations

There exist many approaches for describing the behaviour of perturbations within a homogeneous and isotropic background. One of the most rigorous methods involves adding fluctuations to the metric and to the stress-energy tensor in a perturbative manner using the framework of general relativity. While this approach yields exact solutions for various scenarios using first-order terms, including higher-order terms significantly increases complexity, limiting its applicability. For a detailed explanation of the **GR perturbatition** framework I highly recommend reading the notes by Baumann (2018) and Kurki-Suonio (2024a).

Even though the GR perturbative approach represents the most exhaustive method for dealing with metric fluctuations, there exist certain regimes in which some approximations can be made and still recover accurate predictions. Throughout this section, I will focus on three formalisms that allow us to study how matter perturbations grow in our Universe. This regime is crucial to describe structure formation processes and explain the distribution of observed galaxies, which constitutes the central observational target of large-scale structure surveys (more about LSS surveys in section 0.5).

I first discuss treating perturbations as classical fluid components evolving under **Newtonian gravity** within the expanding background. The main idea for developing this formalism is that, even though the background expansion of the Universe needs to be treated using general relativity, it is possible to accurately approximate the evolution of perturbations employing the Newtonian law of gravity. This approximation is valid as long as we focus on matter perturbations at distances well within the Hubble horizon and not coupled with other components such as radiation, where Newton's law of gravity accurately describes their evolution. To see a detailed derivation of these equations check, e.g., Kurki-Suonio (2023); Dodelson and Schmidt (2020); Baumann (2022), or M. Zennaro & D. López-Cano (in prep.). Here, I present the general equations that cosmological simulation codes used to describe the evolution of perturbations:

$$\left.\begin{aligned} \frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla[(1+\delta)\mathbf{u}] + (1+\delta)\left(\frac{1}{\bar{\rho}}\frac{\partial \bar{\rho}}{\partial t} + 3H\right) &= 0 \quad \text{(Continuity Equation)}, \\ \frac{\partial \mathbf{u}}{\partial t} + a\mathbf{x}\left[\frac{3}{2}H^2 + \frac{\partial H}{\partial t}\right] + H\mathbf{u} + \frac{1}{a}(\mathbf{u}\cdot\nabla)\mathbf{u} &= -\frac{1}{a\rho}\nabla\delta p - \frac{1}{a}\nabla\varphi \quad \text{(Euler Equation)}, \\ \nabla^2\varphi &= 4\pi G\bar{\rho}a^2\delta \quad \text{(Poisson Equation)}, \end{aligned}\right.$$

Where $\mathbf{x}$ represents comoving coordinates and the derivatives ∇ are also taken with respect to the comoving coordinates. The symbol $\delta \coloneqq \rho/\bar{\rho} - 1$ corresponds to the overdensity of a fluid component and the variable $\mathbf{u} \coloneqq \mathbf{v} - aH\mathbf{x}$ is the peculiar velocity. ϕ is the comoving Newtonian potential, and δp represents the pressure perturbations of a fluid component.

These equations are crucial to implementing particle-mesh cosmological simulation codes (e.g. Klypin and Holtzman, 1997; Feng et al., 2016) which are employed to study structure formation processes. I will not investigate these equations any further here. I suggest checking Brandbyge et al. (2017); Fidler et al. (2016) for a more detailed justification regarding why these equations allow to reproduce the evolution of matter perturbations from the perspective of GR.

Next, I discuss the **Top-Hat Spherical Collapse Model**. The key approximation considered in this framework is that, once an overdense region of space decouples from the global background expansion of the universe (due to its gravitational pull), the evolution of this patch can be approximated by the solution to Friedmann's equations (12) of a closed universe with positive curvature only composed by a homogeneous distribution of matter[10]. To obtain information about the collapsed system it is possible to consider its behaviour once it has reached virialization. If we employ the virialization condition and study the linearized solution for the overdensity evolution we reach a very interesting conclusion: The value of the linear overdensity required for this patch to collapse into a halo (reaching virialization) is $\delta_{\rm SC} \approx 1.686$. This back-of-the-envelope calculation provides an approximate value for the linear overdensity required of a region to collapse onto itself due to its own gravitational pull. Even though The top-hat Spherical Collapse Model deals with a very idealized system, it sheds some intuition about the relevant processes that lead to the gravitational collapse and is wildly used by other structure formation theories such as the one discussed below.

Lastly, I introduce the **Extended Press-Schechter (EPS)** model or **Excursion Set Theory**. This framework allows treating analytically the linear growth of perturbations from a statistical standpoint. Thanks to this technique it is possible to qualitatively predict relevant cosmological quantities such as the **Halo Mass Function (HMF)**, or the halo merger and accretion history. The origin of this theory can be traced back to the seminal work of Press and Schechter (1974a) and Bond et al. (1991a), but more recent articles improve upon the original formulation, providing a more robust mathematical justification, and employing this formalism for the prediction of additional observables (e.g. Lacey and Cole, 1993; Sheth and Tormen, 2002; Zentner, 2007; de Simone et al., 2011). I will now briefly summarize the basic assumptions made by this formalism to derive the analytical expression for the HMF.

Let's start by considering a realization of the overdensity field of matter perturbations at

[10] In this scenario we can compute a parametric solution to Friedmans's equations describing how the scale factor evolves with cosmic time.

a certain scale factor, $\delta_{\rm m}(\mathbf{x}, a)$. Given this field, we can compute at each comoving point, $\mathbf{x}$, the "smoothed" value of the field at a characteristic scale, $\rm s$, having previously defined a window function $W(\rm s; \mathbf{r})$:

$$\delta_{\rm m}(s; \mathbf{x}, a) \coloneqq \iiint_{-\infty}^{\infty} W(\rm s; |\mathbf{x} - \mathbf{y}|)\delta_{\rm m}(\mathbf{y}, a) d^3\mathbf{y}.$$

Additionally, we can compute the **variance of the overdensity field** at that scale by averaging across all points in space, i.e.,

$$\sigma^2_{\rm m}(\rm s; a) \coloneqq \iiint_{-\infty}^{\infty} \left|\delta_{\rm m}(s; \mathbf{x}, a)\right|^2 d^3\mathbf{x}.$$

If we consider a Gaussian random overdensity field whose modes evolve independently in time in Fourier space[11] we can impose a sharp k-space window function to obtain random walk "trajectories", $\delta_{\rm m}(s; \mathbf{x}, a)$, as a function of the variance of the field, $\sigma^2_{\rm m}(s; a)$. Factoring out the linear growth factor we obtain that collectively (in the statistical ensemble sense), the different random walks satisfy the diffusion equation at all times

$$\frac{\partial P}{\partial S} = \frac{1}{2}\frac{\partial^2 P}{\partial \delta^2}, \tag{14}$$

where $P(S, \delta)$ represents the probability density function of finding a trajectory at S with a value δ, and we have performed the following notation simplification:

$$S \coloneqq \sigma^2_{\rm m}(\rm s; a), \quad \text{and} \quad \delta \coloneqq \delta_{\rm m}(s; \mathbf{x}, a).$$

Now it is possible to include some additional physical assumptions for solving Equation 14 analytically. According to the top hat spherical collapse approximation, we can consider that halo formation takes place whenever a linear overdensity value is higher than δc. For this reason, we can assume that there exists an absorbing barrier condition at[12] $\delta_{\rm c}(a) \approx 1.686\frac{D(a)}{D(a_0)}$ such that $P(S, \delta \geq \delta_{\rm c}) = 0$. This condition imposes that any trajectory that crosses the threshold value $\delta_{\rm c}(a)$, collapses into a halo with the characteristic mass $M_{\rm h}$ associated with the scale S, hence it cannot contribute any longer to the path probability distribution P for larger values of S. Considering this constraint and assuming as initial condition that $P(\delta_0, S_0) = \delta_{\rm D}(\delta_0)$ we obtain that the probability distribution function (with respect to δ) of trajectories, $\delta_{\rm m}(s; \mathbf{x}, a)$, that have never exceed the threshold value, $\delta_{\rm c}(a)$, prior to S is

$$P(\delta, S; a, \delta_0, S_0) = \frac{1}{\sqrt{2\pi(S - S_0)}}\left\{\exp\left[-\frac{(\delta - \delta_0)^2}{2(S - S_0)}\right] - \exp\left[\left(-\frac{(2(\delta - \delta_{\rm c}(a)) - (\delta - \delta_0))^2}{2(S - S_0)}\right]\right\}\left(\tag{15}$$

[11] This occurs in the linear regime when the growth of modes can be expressed in terms of the primordial amplitude fluctuations in k-space times a multiplicative linear growth factor, $D(a)$. See the references above for a detailed derivation.

[12] The factor $D(a)$ denotes the linear growth factor and needs to be taken into account for the barrier height since we have factored out the time evolution of the overdensity modes in the equation 14.

Therefore the fraction of trajectories that have crossed above the threshold, $\delta c(a)$, before S is

$$F(S; a, \delta_0, S_0) = 1 - \int_{-\infty}^{\delta_c(a)} P(\delta, S; a, \delta_0, S_0)\, d\delta = \mathrm{erf}\left(\frac{\delta_c(a) - \delta_0}{\sqrt{2}(S - S_0)}\right),$$

and the differential probability of first piercing the threshold at S can be expressed as

$$f(S; a, \delta_0, S_0)\, dS = \frac{dF(S; a, \delta_0, S_0)}{dS}\, dS = \ldots$$
$$\ldots = -\frac{d}{dS}\int_{-\infty}^{\delta_c(a)} P(\delta, S; a, \delta_0, S_0)\, d\delta = -\left[\frac{1}{2}\frac{\partial P(\delta, S; a, \delta_0, S_0)}{\partial \delta}\right]_{-\infty}^{\delta_c(a)}$$

where, in the last step, we have commuted the integral operation with the derivative with respect to S, and afterwards we have employed Equation 14. Finally, after substituting in the last equation the result from Equation 15 we obtain that

$$f(S; a, \delta_0, S_0)\, dS = \frac{\delta_c(a) - \delta_0}{\sqrt{2\pi}(S - S_0)^{3/2}} \exp\left[-\frac{(\delta_c(a) - \delta_0)^2}{2(S - S_0)}\right] dS$$

To obtain the HMF predicted by EPS we need to consider that, within a finite volume, V, that contains a total (matter) mass, $M_\mathrm{m} \coloneqq \bar{\rho}_\mathrm{m} V$, the fraction of mass contained in haloes of characteristic mass, M_h, is

$$\frac{N_\mathrm{h} M_\mathrm{h}}{M_\mathrm{m}} = \frac{N_\mathrm{h} M_\mathrm{h}}{\bar{\rho}_\mathrm{m} V} = f(S_\mathrm{h}; a, 0, 0)\ \frac{dS}{dM}\ dM,$$

where S_h corresponds to the characteristic Lagrangian scale associated with the characteristic halo mass, M_h, and is determined by the specific window function chosen. The ratio N_h/V is known as the HMF predicted by EPS and corresponds to the differential number of halos of per unit volume that have a characteristic halo mass M_h. Rearranging the terms of the last equation we obtain

$$\frac{N_\mathrm{h}}{V} = \frac{dn_\mathrm{h}}{dM_\mathrm{h}} dM_\mathrm{h} = \sqrt{\frac{2}{\pi}} \frac{\bar{\rho}}{M_\mathrm{h}^2} \frac{\delta_c(a)}{\sigma_\mathrm{m}^2(\mathrm{h}; a)}\ \frac{d\log\sigma_\mathrm{m}}{d\log M_\mathrm{h}}\ \exp\left[-\frac{\delta_c(a)^2}{2\sigma_\mathrm{m}^2(\mathrm{h}; a)}\right] dM_\mathrm{h} \qquad (16)$$

The EPS formalism is a powerful analytical approximation for describing halo formation from the initial density field fluctuations. It is of particular relevance for this work and some of its results play a relevant role in the chapters 2 and 3 of this thesis.

0.5 Current state: ΛCDM, LSS surveys, simulations, haloes, machine learning

Until this point in the introduction, I have introduced the fundamental theory describing the universe's behaviour at the background level (section 0.3) and outlined some models to characterize structure formation processes (section 0.4). However, I have not yet discussed

the connection between these theories and actual astronomical observations. This section will explain how major observational discoveries have shaped our understanding of cosmology and established the ΛCDM paradigm.

A pivotal observational evidence that marks the birth of cosmology as a consolidated scientific discipline is the discovery of the Cosmic Microwave Background (CMB) radiation. Detectable from every direction in the sky, this radiation has a temperature of $T_{\mathrm{CMB}} = 2.72548 \pm 0.00057\,K$ (Fixsen, 2009) and originates in the early universe ($z_{\gamma,dec} \approx 1090$). The CMB radiation originates from a process known as photon decoupling that occurs when the interaction rate between photons and matter (electrons in particular) fell below the universe's expansion rate. The CMB radiation's detection in the 1960s through radio experiments (Penzias and Wilson, 1965), and its interpretation within the cosmological context, triggered posterior dedicated efforts for cosmological studies, leading to the development of the ΛCDM model.

The ΛCDM model provides a robust description of multiple astronomical observations, it is based on general relativity and considers the following main components: ordinary matter, cold dark matter, and dark energy. The term cold dark matter (CDM) refers to all non-baryonic elements of the universe (that are not visible) that satisfy $\omega_{\mathrm{DM}} \approx 0$ (Peebles, 1982). The Dark energy component is modelled in the ΛCDM model through a cosmological constant term. The addition of this term to the standard cosmological model occurred throughout the nineties, and the work by Riess et al. (1998) stands out in particular since it helped to consolidate the dark energy term by providing direct evidence regarding the accelerated expansion of the Universe.

One of the most notable accomplishments of the ΛCDM model has been to accurately describe CMB anisotropies. These small matter perturbations imprinted on the CMB signal were generated in the primordial Universe by the quantum fluctuations of the inflaton field. They correspond to the initial matter overdensity perturbations from which later structures such as galaxies and galaxy clusters grew due to gravity. NASA's COBE mission first detected the CMB anisotropies in 1992 (Smoot, 1999). NASA's WMAP observatory (Bennett et al., 2013) and ESA's Planck experiment provided more precise measurements in the 2010s Planck Collaboration et al. (2020a), offering the most accurate cosmological parameter estimates for the ΛCDM model to date.

Despite all the ΛCDM model's successes, several phenomena and astronomical measurements do not align perfectly with it (see Perivolaropoulos and Skara, 2022, for a review). Testing the ΛCDM model in all possible regimes is necessary to probe the microscopic nature of dark matter and dark energy and to test the theory of general relativity on large scales.

Over the last decades, large-scale structure (LSS) surveys have emerged as a promising

avenue for cosmological studies. These experiments aim to map the positions of as many galaxies as possible in the sky. By studying the distribution of galaxies in LSS surveys, it is possible to test cosmological theories by characterizing the statistical properties of their distribution. Many LSS surveys have been conducted, with more currently collecting data or planned to start in the coming years (Alam et al., 2017a; Euclid Collaboration et al., 2022; DESI Collaboration et al., 2016; Ivezić et al., 2019, for example).

LSS surveys generate petabytes of data on galaxy properties and positions. This information is crucial for investigating astrophysical effects and studying cosmological models. To do so it is necessary to compare the observed galaxy distributions with the predictions from models that correctly capture cosmological and astrophysical processes. However developing such models is challenging since galaxies form in high-density regions of the matter field, where very difficult to model non-linear processes occur. Currently, numerical simulations are the most commonly used method to study galaxy formation in cosmological studies.

Cosmological simulations are numerical algorithms that accurately capture non-linear structure formation processes. They predict the matter distribution across cosmological scales, allowing for a comparison between observed galaxy distributions and theoretical models. These simulations typically solve Newton's equations for a set of tracer particles that evolve in an expanding background, effectively capturing non-linear gravitational processes without needing full general relativistic treatment at all scales (see Angulo and Hahn, 2022a, for a review).

Despite considerable advancements in the field of cosmological simulations, limitations in computing power and data storage prevent us from running complete forward predictions that can be directly compared with observational galaxy catalogues. The simulations need to be extensive enough to cover large cosmological volumes while also incorporating detailed modelling of all the astrophysical processes that influence the final galaxy properties. To address these challenges, numerous strategies have been developed to create simulations that are faster to compute and require less storage space, however, there is still significant potential for improvement and numerous research groups work to enhance the performance of cosmological simulations.

In the remainder of this section, I briefly discuss various aspects related to improving cosmological simulation results. Specifically, I will focus on topics most relevant to the work presented in the central chapters of this thesis.

One of the main challenges in cosmological simulations is to reduce their computational cost for simulating the formation and evolution of galaxies. To accurately reproduce this process it is necessary to simulate the behaviour of normal ("baryonic") matter and compute the gravitational interactions at the same time. To incorporate these baryonic processes, it is

common to include in simulations hydrodynamical recipes. However, solving these equations is computationally expensive and limits the potential simulated volume to scales not large enough for comparison with galaxy survey data. Various alternatives exist to circumvent this problem, most involving running a gravity-only simulation and including galaxies in a post-processing step. Common techniques include Halo Occupation Distribution (HOD) models (e.g. Berlind et al., 2003), (sub-)halo abundance matching techniques (e.g. Vale and Ostriker, 2004), and semi-analytic models (e.g. Knebe et al., 2018b).

Including galaxies in gravity-only simulations during a post-processing step requires understanding how to link Dark Matter (DM) haloes with the corresponding galaxies they may host. The most important property for linking DM haloes with galaxies is their mass. However, mass alone is insufficient for accurately matching host haloes with galaxies. It's necessary to consider additional properties related to the internal structure of DM haloes. Navarro et al. (1996, 1997) pointed out that, along with halo mass, halo concentration is sufficient to model the internal mass distribution of haloes. Understanding how halo concentration varies with mass, redshift, and cosmology is essential for describing matter distribution at small scales and for correctly linking galaxies with haloes.

Finally, the last problem I want to discuss is that, despite the current efficiency achieved by gravity-only simulations, they remain significantly expensive to execute, requiring millions of CPU hours for covering large volumes with sufficient mass resolution. One of the most substantial advancements in computational science over the last decade is related to the development of artificial intelligence and machine learning techniques. These methods have significantly impacted cosmology lately, offering fast and accurate emulation of various processes. Recent works have employed machine learning to accelerate calculations (He et al., 2019; Giusarma et al., 2019; Alves de Oliveira et al., 2020; Wu et al., 2021; Jamieson et al., 2022), perform likelihood-free inference (Hahn et al., 2023), and use machine learning frameworks as a tool for interpreting halo properties (Lucie-Smith et al., 2018, 2019, 2020; Chacón et al., 2022; Betts et al., 2023).

0.6 About this thesis

This thesis explores the connection between theoretical cosmological models and Large-Scale Structure (LSS) observations through the use of cosmological simulations. These simulations model the Universe's structure formation processes and the large-scale distribution of galaxies. Enhancing the speed and reliability of cosmological simulations is crucial for bridging galaxy observations with the theoretical modelling side.

As of September 2020 (when I started my doctoral studies), a number observational LSS campaigns have concluded, unveiling tensions between early Universe probes and late

Universe observations, particularly regarding the σ_8 and H_0 parameters (see Verde et al., 2019, for a review). The upcoming survey campaigns at the time (such Euclid and DESI, both currently operational) had amplified the community's interest in advancing cosmological simulation techniques for analyzing and interpreting the anticipated influx of high-quality observations of millions of galaxies. Next, I categorize some of the main challenges for enhancing cosmological simulations; I will mainly focus on aspects that have been particularly relevant for my work:

- Improving the modelization of structure formation processes: From the developement of the first cosmological simulation codes more than twenty years have passed. During this time both the efficiency and accuracy of these codes has significantly improved. Alongside, the development of analytical models that link the initial conditions with the final halo properties has helped to speed up prediction tasks.

- Improving our understanding of baryonic processes: Another big challenge for performing realistic simulations is the necessity of including baryonic process to accurately reproduce galaxy formation processes. There currently exist several approaches (some of them mentioned in section 0.5), each of with certain advantages and limitations that try to balance the accuracy of the predictions with the computational cost required to execute them.

- Accelerating simulations: Originally, paralelization techniques helped to improve greatly the speed of simulations. Over the last years, machine learning algorithms and GPU acceleration have transformed the field of cosmological simulations reducing significantly the computational time required to make predictions using emulators and other techniques.

My work during the Ph.D. has focused on these three aspects, contributing to the development of next-generation cosmological simulators from multiple angles. These advancements aim to provide a more accurate description of the Universe's matter distribution and observed galaxy populations. This thesis compiles three papers that have already being published in specialized scientific jounals. I have adapted these articles into separate self-contained chapters that constitute the main body of this work. Here I summarize the contents of each chapter and contextualize the problem each one focuses on.

- Chapter 1 is based on the article titled "UNITSIM-Galaxies: data release and clustering of emission-line galaxies" (Knebe et al., 2022). This article is currently published in the journal Monthly Notices of the Royal Astronomical Society (MNRAS) and can currently be accessed through the following link . This work describes how is it possible to develop a mock galaxy catalog to simulate the expected galaxy

distribution observed by the Euclid mission. By employing gravity-only simulations and a semi-analytical model, we generate a galaxy mock catalog and post-process it to predict the expected H_α flux for galaxies detected by the Euclid survey. My contribution to this article, in which I appear as second author, has been crucial. I have played a key role developing the codes and methods necessary for generating, analyzing and interpreting the results presented here.

- Chapter 2 compiles the paper "The cosmology dependence of the concentration-mass-redshift relation" (López-Cano et al., 2022). This article is currently published in the journal Monthly Notices of the Royal Astronomical Society (MNRAS) and can currently be accessed through the following link . The focus of this article is to investigate the concentration of halos, a parameter that determines their internal structure. This chapter explains how is it possible to model the cosmology dependence of the concentration parameter combining the Exceursion Set Theory formalism with a relation empirically derived from multiple gravity only simulations.

- Chapter 3 describes the work "Characterizing Structure Formation through Instance Segmentation" (López-Cano et al., 2023). This article is currently published in the journal Astronomy & Astrophysics (A&A) and can currently be accessed through the following link . It showcases machine learning's ability to identify features in the initial conditions that lead to the formation of dark matter haloes. This work illustrates the potential of machine learning techniques to describe structure formation processes.

- The conclusion section 3 provides a summary of my contributions, their significance in cosmology, and their potential impact on future research.

Chapter 1

UNITSIM-Galaxies: data release and clustering of emission-line galaxies

New surveys such as ESA's Euclid mission are planned to map with unprecedented precision the large-scale structure of the Universe by measuring the 3D positions of tens of millions of galaxies. It is necessary to develop theoretically modelled galaxy catalogues to estimate the expected performance and to optimise the analysis strategy of these surveys. We populate two pairs of $(1h^{-1}\mathrm{Gpc})^3$ volume dark-matter-only simulations from the UNIT project with galaxies using the SAGE semi-analytic model of galaxy formation, coupled to the photoionisation model GET_EMLINES to estimate their Hα emission. These catalogues represent a unique suite that includes galaxy formation physics and – thanks to the fixed-pair technique used – an effective volume of $\sim (5h^{-1}\mathrm{Gpc})^3$, which is several times larger than the Euclid survey. We present the performance of these data and create five additional emission-line galaxy (ELG) catalogues by applying a dust attenuation model as well as adjusting the flux threshold as a function of redshift in order to reproduce Euclid-forecast dN/dz values. As a first application, we study the abundance and clustering of those model Hα ELGs: for scales greater than $\sim 5h^{-1}\mathrm{Mpc}$, we find a scale-independent bias with a value of $b \sim 1$ at redshift $z \sim 0.5$, that can increase nearly linearly to $b \sim 4$ at $z \sim 2$, depending on the ELG catalogue. Model galaxy properties, including their emission-line fluxes (with and without dust extinction) are publicly available.

1.1 Introduction

During the last few decades, numerous projects have been aimed at creating large cartographic maps of galaxies, such as 2dFGRS (Cole et al., 2005), SDSS (Alam et al., 2017a; Eisenstein et al., 2005), WiggleZ (Drinkwater et al., 2010; Parkinson et al., 2012), BOSS (Dawson et al., 2013; Alam et al., 2017b), eBOSS (Dawson et al., 2016; Alam et al., 2021a) or DES (The Dark Energy Survey Collaboration, 2005; Abbott et al., 2018). They

have been carried out with the objective of trying to better understand the large-scale structure of the Universe, to estimate the different parameters that regulate the formation of structures, to determine the expansion history of the Universe, to study how galaxies form, to reconstruct their star formation histories, and to impose constraints upon different models that currently exist for dark energy and for alternative theories of gravity. While advances have certainly been made, all these grand topics remain open areas of investigation, and likely will for years to come.

New surveys such as Euclid (Laureijs et al., 2011; Amendola et al., 2013), the Nancy Grace Roman Space Telescope (Spergel et al., 2013, 2015), the Dark Energy Spectroscopic Instrument (DESI, Collaboration et al., 2016), and the 4-metre Multi-Object Spectroscopic Telescope (4MOST, de Jong et al., 2012) are planned to map with unprecedented precision the large-scale structure of the Universe by measuring the 3D positions of tens of millions of galaxies. These missions are expected to start operating in the coming years, providing the scientific community with wider, deeper, and more accurate data, which may be used to impose stronger constraints upon theoretical models and to provide more accurate estimates for some of the aforementioned parameters relevant in cosmology. Some of these forthcoming missions (e.g., Euclid) will focus on conducting spectroscopic surveys of galaxies using near-infrared grisms in order to determine the positions of galaxies by observing their emission lines such as Hα. The wavelength of the observed emission lines will serve to determine the redshifts of the detected objects. Such observations have already been undertaken in the past. There are, for instance, the High-z Emission Line Survey (HiZELS, Geach et al., 2008) and the Wide Field Camera 3 Infrared Spectroscopic Parallels survey (WISP, Atek et al., 2010). The WISP survey, for instance, has been used by Colbert et al. (2013) to measure the number density evolution of Hα emitters; Sobral et al. (2016) employed the HiZELS data (and additional follow-up observations) to quantify the evolution of the Hα luminosity function. But all previous efforts lack the volumes to be probed by future missions.

Observational campaigns need to be complemented by cosmological simulations: a cornerstone of large-scale structure analysis. Cosmological simulations inform and validate galaxy clustering models. They are also used to test and optimise different estimators and analysis pipelines, to estimate covariance matrices, and to compare with measurements from data. Smaller scales (i.e. below 1 Mpc) are known to contain many more Fourier modes than larger ones and hence constraining power. However, they are heavily affected by the physics of galaxy formation. Since the spatial volumes that the aforementioned surveys seek to study are notoriously large, it is still necessary to rely on dark-matter-only simulations in which galaxies are introduced in post-processing either by halo occupation distribution (HOD, e.g. Berlind et al., 2003, as well as the Euclid *Flagship* mock galaxy catalogue), (sub-)halo abundance matching (SHAM, e.g. Vale and Ostriker, 2004) or semi-analytic models (SAM)

(e.g. the MultiDark-Galaxies,[1] Knebe et al., 2018b). While there are efforts to push the limits of 'full physics' hydrodynamical simulations to larger and larger volumes (e.g. Lee et al., 2020), it still remains more feasible to match the volumes that missions like Euclid will cover with gravity-only simulations.

The demand for large volumes modelled with sufficiently high resolution is also the reason why, during the last years, alternatives to running such demanding simulations have been explored. For instance, the technique developed by Angulo and Pontzen (2016b) dramatically reduces the variance arising from the sparse sampling of wavemodes in cosmological simulations. The method uses two simulations that are 'fixed' and 'paired', i.e. the initial Fourier mode amplitudes are fixed to the ensemble average power spectrum and their phases are shifted by π. This approach has been adopted by the UNIT collaboration[2] (Chuang et al., 2019) where it has been shown that the effective volume of such fixed-and-paired simulations can be several times larger than the actual volume simulated: in Chuang et al. (2019) we have shown that the original four $(1h^{-1}\mathrm{Gpc})^3$ simulations correspond to a total effective volume of ca. $(5h^{-1}\mathrm{Gpc})^3$, i.e. ~ 7 times of the survey volume of Euclid or DESI. We use the same two pairs of simulations for our study here. Our simulations include the large scales with an accuracy greater than expected by these surveys, and here we have populated them with galaxies using a semi-analytical model that includes all the relevant physical processes for galaxy formation. In terms of galaxy clustering statistics, each pair can be as precise on (non-)linear scales as an average over approximately 150 traditional simulations. They therefore are suitable to statistically study matter–galaxy interplay and galaxy clustering alongside its bias.

In this work we present and use galaxy catalogues for simulations that were generated by applying the SAGE semi-analytic model (Croton et al., 2016) to the aforementioned gravity-only UNIT simulations. These SAGE galaxies have then been processed with the GET_EMLINES code (Orsi et al., 2014) in order to obtain emission-line galaxies (ELGs). Using the resulting ELG catalogues, we study the predicted number density evolution of Hα emitters and compare it to other theoretical models as well as observational data. We also generate additional ELG catalogues by imposing certain flux threshold and/or even apply a dust attenuation model. All catalogues are used to study the clustering of our Hα galaxies and their linear bias with respect to the dark matter field, a quantity first studied by Kaiser (1984) for Abell clusters and developed in theoretical detail by Bardeen et al. (1986). The bias is a key parameter and a result of not only halo formation but also the varied physics of galaxy formation that can cause the spatial distribution of baryons to differ from that of dark matter. The bias connects the observed statistics to theoretical predictions and has recently

[1] Galaxy catalogues based upon three distinct SAMs can be downloaded from CosmoSim.

[2] `http://www.unitsims.org`

been the target of many theoretical studies in light of ELGs (e.g. Geach et al., 2012; Cochrane et al., 2017; Merson et al., 2019; Tutusaus et al., 2020). Our results add to these and may be used to make forecasts for Euclid and related studies for which both the abundance and bias of Hα ELGs is an input.

There already exist previous works based upon the UNIT simulations and the modelling of ELGs in them (Zhai et al., 2021, 2019). However, the important difference to our work is that in those papers only one of the UNIT simulations has been used, as opposed to all four here. Further, Zhai et al. applied a completely different modelling for the ELGs, namely the Galacticus semi-analytic model (Benson, 2012), coupled to the CLOUDY photoionisation code (Ferland et al., 2013) for the calculation of emission line properties. Further, their dust model was tuned as a function of redshift to match observations of the Hα luminosity function in the redshift range $z \in [0.8, 2.3]$. And while Zhai et al. also studied galaxy clustering in the later work, they have not investigated the bias. Our work therefore extends those previous studies and should be viewed as complementary. We further have made our galaxy catalogues publicly available.

The structure of this article is as follows. In Section 1.2 the methods used to generate the ELG catalogues are presented, namely the N-body UNIT simulations (Section 1.2.1), the SAGE semi-analytic model (Section 1.2.2) and the emission-line modelling (Section 1.2.3). Next, in Section 1.3, we present a series of figures to validate the galaxy catalogues generated by SAGE by comparing key properties with observational results. Then in Section 1.4 we examine the validity of the modelling for the emission lines of the galaxies. Afterwards, in Section 1.5, the results obtained by studying the two-point correlation function and the bias obtained for the ELGs in the Euclid range of redshifts will be presented. Finally, in Section 1.6, the conclusions derived from this work will be outlined.

1.2 The Methods

1.2.1 The UNIT Simulations

As a basis for this work, four gravity-only simulations that have been developed within the UNIT project have been employed. The names for the two pairs of simulations that we use throughout this work are UNITSIM1 (U1), UNITSIM1-Inverted Phase (U1IP), UNITSIM2 (U2), and UNITSIM2-Inverted Phase (U2IP). The procedure followed for generating these simulations as well as an analysis of the resulting correlation properties is discussed in Chuang et al. (2019). For this particular study we have used the two pairs of simulations in which the code Gadget (Springel et al., 2001b) has been used to study the behavior of a total of 4096^3 particles in a volume of $1h^{-3}\text{Gpc}^3$ per simulation, thus obtaining a mass resolution of

$1.2 \times 10^9 h^{-1}\mathrm{M}_\odot$ per simulation particle.

In Chuang et al. (2019) it is also explained how the ROCKSTAR halo catalogues and the corresponding CONSISTENTTREES merger trees have been generated for each of the gravity-only simulations using the publicly available codes from Behroozi et al. (2012). All the data corresponding to the UNIT simulations are publicly available at `http://www.unitsims.org`. By making the galaxy catalogues and their emission-line properties available too, this work further adds to the community.

1.2.2 Semi-analytic galaxy modelling via SAGE

SAGE (Semi-Analytic Galaxy Evolution, Croton et al., 2016) is a modular, publicly available[3] semi-analytic model of galaxy formation, branched from the Munich family of models (specifically from Croton et al., 2006). Haloes (in this case, from the UNIT simulations) are initially seeded with 'hot' gas based on the cosmic baryon fraction (modulo a reionization factor at higher redshift and in low-mass haloes). Cooling/accretion of this gas onto the central galaxy is based on the two-mode (hot and cold) model of White and Frenk (1991). Star formation in the disc occurs once the gas is above a critical average surface density (see Kennicutt, 1989; Kauffmann, 1996). Metals are immediately injected and gas recycled into the inter-stellar medium (ISM), where a constant mass-loading factor is also applied to reheat gas out of the disc, some of which will end up in an ejected component if the energy budget allows it. A parametrized fraction of the ejected gas (connected to the virial velocity) is reincorporated into the halo on a dynamical time-scale. Satellite galaxies are tracked in the merger trees until merged or unresolved. Once their subhaloes become unresolved, satellites are either disrupted (where their baryons are placed in intracluster reservoirs) or immediately merged with the central, dependent on how long they survived as a satellite. SAGE, therefore, does not have orphan galaxies. Mergers and disc instabilities trigger starbursts, drive stars into the bulge, and cause gas to be accreted onto the central black hole. This triggers quasar-mode active galactic nuclei (AGN) feedback, which reheats gas from the disc. When galaxies have sufficiently (super)massive black holes, cooling is also suppressed by radio-mode AGN activity (both past and present), modelled by a phenomenological 'heating' radius that can only grow with time, within which gas cannot cool.

This is the same SAGE model that was also applied to the MultiDark simulation MDPL2 (Knebe et al., 2018b). The model was calibrated for that simulation by fitting visually first the $z = 0$ stellar mass function (Baldry et al., 2008), and secondarily using the stellar metallicity–mass relation (Tremonti et al., 2004), baryonic Tully–Fisher relation (Stark et al.,

[3] `https://github.com/darrencroton/sage`

2009), black hole–bulge mass relation (Scott et al., 2013), and cosmic star formation rate density (Somerville et al., 2001). The model has not been re-calibrated here as both the UNIT and MDPL2 simulations were run with the same cosmological parameters (Planck Collaboration et al., 2015) and have the same box size. However, the mass resolution is marginally better for UNITSIM, due to the 20 per cent larger number of particles. For the general performance of the SAGE model we refer the reader to the results presented in Knebe et al. (2018b), as the calibration plots change minimally when going from MPDL2 to UNIT (see also Fig. 1.1, in this paper). The calibration does not include constraints for emission-line galaxies.

For a more detailed description of the model we refer the reader to Croton et al. (2016) and section 2.4 of Knebe et al. (2018b).

1.2.3 Emission-line galaxy modelling

Once we have populated the dark matter haloes from the UNIT simulations with the semi-analytic galaxies generated by SAGE we obtain values for the intensity of the most relevant emission lines such as Hα, [OIII]4959, [OIII]5007, [NII]6548 and [NII]6584 for each of the model galaxies. In this study we focus on the Hα line – with a particular focus on the Euclid mission. The other emission lines are left for future work.

GET_EMLINES code. In order to reproduce the intensity of Hα emission lines of our galaxies, we have used the method presented in Orsi et al. (2014), i.e. the publicly available GET_EMLINES code.[4] This code is based on the algorithm MAPPINGS-III described in Groves et al. (2004) and Allen et al. (2008), which relates the ionization parameter of gas in galaxies, q, to their cold-gas metallicity $Z_{\rm cold}$ as:

$$q(Z) = q_0 \left(\frac{Z_{\rm cold}}{Z_0} \right)^{-\gamma}, \tag{1.1}$$

where q_0 is the ionisation parameter of a galaxy that has cold gas metallicity Z_0 and γ is the exponent of the power law. We adpoted the suggested values of $q_0 = 2.8 \times 10^7\,{\rm cm\,s^{-1}}$ and $\gamma = 1.3$, which were found to yield Hα luminosities for star-forming galaxies in good agreement with observations (Orsi et al., 2014). Cold gas metallicity is defined as the ratio between the cold gas mass in metals to the total cold-gas mass:

$$Z_{\rm cold} = \frac{M_{Z,{\rm cold}}}{M_{\rm cold}}. \tag{1.2}$$

The other relevant component is the star formation rate (SFR).[5] Note that SAGE provides this quantity averaged over the previous time-step in the merger trees, despite this interval

[4] `https://github.com/aaorsi/get_emlines`

[5] Total SFR in SAGE is the sum of the `SfrDisk` and `SfrBulge` fields.

being broken into sub-time-steps in the code. But the model ideally requires as inputs the instantaneous SFR and cold gas metallicity of galaxies. However, Favole et al. (2020) have shown that for galaxies that are not too bright the differences are negligible. To be able to properly compare our results to observations, we convert the luminosities to fluxes and also apply a dust extinction to the luminosities of the model galaxies.

Please note that when applying the GET_EMLINES code to the SAGE catalogues,[6] we rejected all galaxies with a star formation rate equal to zero. One might be inclined to therefore claim that our emission-line galaxies are 'star-forming galaxies', but usually a threshold on the specific star formation of order 0.01/Gyr (and hence clearly larger than 0) is assumed to separate 'passive' and 'star-forming' galaxies. Therefore, our ELGs are based upon SAGE galaxies that do form stars, but also include 'passive' galaxies in the conventional sense.

Dust extinction. We use here a Cardelli extinction law implemented following Favole et al. (2020)[7], but we also summarize it here. The attenuation from interstellar dust is added to the intrinsic Hα luminosity using:

$$L(\lambda_j)^{\rm att} = L(\lambda_j)^{\rm intr} 10^{-0.4 A_\lambda(\tau_\lambda^z,\theta)}, \tag{1.3}$$

where the attenuation coefficient, as a function of the galaxy optical depth τ_λ^z and the dust scattering angle θ, is defined as (Osterbrock, 1989; Draine, 2003; Izquierdo-Villalba et al., 2019; Favole et al., 2020):

$$A_\lambda(\tau_\lambda^z,\theta) = -2.5\log_{10}\frac{1-\exp(-a_\lambda\,\sec\theta)}{a_\lambda\,\sec\theta}. \tag{1.4}$$

In Eq. (1.4), $a_\lambda = \sqrt{1-\omega_\lambda}\tau_\lambda^z$, and ω_λ is the dust albedo. We assume $\cos\theta = 0.30$ and $\omega_\lambda = 0.56$, meaning that the scattering is not isotropic but forward-oriented, and about 60 per cent of the extinction is caused by scattering.

The galaxy optical depth is defined as (Hatton et al., 2003; De Lucia and Blaizot, 2007):

$$\tau_\lambda^z = \left(\frac{A_\lambda}{A_V}\right)_{Z_\odot}\left(\frac{Z_{\rm cold}}{Z_\odot}\right)^{1.6}\left(\frac{\langle N_H\rangle}{2.1\times10^{21}{\rm atoms\ cm^{-2}}}\right)\Bigg(\tag{1.5}$$

in terms of the cold gas metallicity $Z_{\rm cold}$ defined in Eq. (1.2) and the extinction curve for solar metallicity: $Z_\odot = 0.0134$ (Asplund et al., 2009). We assume the Cardelli et al. (1989) extinction law:

$$\left(\frac{A_\lambda}{A_V}\right)\Bigg(= a(x) + b(x)/R_V, \tag{1.6}$$

[6] We are using the plural here when referring to the catalogues as we will always have at our disposal the four catalogues coming from the two pairs of UNIT simulations.

[7] https://github.com/gfavole/dust

where $x \equiv \lambda^{-1}$, $R_V \equiv A_V/E(B-V) = 3.1$ is the ratio of total to selective extinction for the diffuse interstellar medium in the Milky Way, and

$$\begin{aligned} a(x) =& 1 + 0.17699\,y - 0.50447\,y^2 - 0.02427\,y^3 + \\ & 0.72085\,y^4 + 0.01979\,y^5 - 0.77530\,y^6 + 0.32999\,y^7, \\ b(x) =& 1.41338\,y + 2.28305\,y^2 + 1.07233\,y^3 - 5.38434\,y^4 \\ & - 0.62251\,y^5 + 5.30260\,y^6 - 2.09002\,y^7, \end{aligned} \tag{1.7}$$

with $y = (x - 1.82)$. The quantity $\langle N_H \rangle$ in Eq. (1.5) is the mean hydrogen column density defined as (Hatton et al., 2003; De Lucia and Blaizot, 2007):

$$\langle N_H \rangle = \frac{M_{\rm cold}^{\rm disc}}{1.4\,m_{\rm p}\,\pi\,(1.68 R_{1/2}^{\rm disc})^2}\ \text{atoms cm}^{-2}, \tag{1.8}$$

where $M_{\rm cold}^{\rm disc}$ is the cold-gas mass of the disc, $m_{\rm p} = 1.67 \times 10^{-27}$ kg is the proton mass, and $R_{1/2}^{\rm disc}$ is the half-mass radius of the disc.

We caution that emission lines are expected to be more attenuated than the continuum, e.g. De Barros et al. (2016), which is the model used here.

1.3 The SAGE galaxies

The aim of this section is to validate how well our theoretically modelled SAGE galaxies perform with respect to the quantities that enter into the calculation of the emission-line properties. This involves a) stellar mass, b) star formation rates, c) metallicities, and d) disc lengths. We will further focus on redshifts in the range $z \in [1, 2]$ and compare to observational data where possible. For comparisons of other properties to observations and the calibration plots, respectively, we refer the reader to Knebe et al. (2018b) where SAGE has been applied to the MultiDark simulation MDPL2. Note that in this Section we are using the complete SAGE galaxy catalogue, not restricting any results to ELGs. However, we provide in the Appendix all the corresponding plots for our model ELGs.

1.3.1 Stellar Mass Function

The stellar mass function (SMF) is one of the most significant properties that can be inferred from galaxy surveys since this function represents the number of galaxies in stellar-mass bins, normalized to the volume of the survey/simulation and to the bin width. Its simplicity yet fundamental importance resides in the fact that the SMF is often employed for calibrating semi-analytic models such as SAGE used here.

In the main panel of Fig. 1.1 the results obtained for the SMF computed from the SAGE galaxies modelled over the UNITSIM1 simulation are presented for three different redshifts

$z = [0.0, 1.710, 2.695]$. Together with the results obtained from our simulation, a series of observational results obtained for a range of redshifts similar to those simulated are also represented in the same figure. The compilation for redshift $z = 0$ is taken from the so-called 'CARNage calibration' data set described in great detail in section 3.3 and appendix A of Knebe et al. (2018a)[8]. The observations for the higher redshifts are taken from Davidzon et al. (2017) and are based on the UltraVISTA near-infrared survey of the COSMOS field. In the bottom panel of Fig. 1.1 the variation in SMF between UNITSIM1 and the three other UNIT simulations is shown, i.e. the y-axis represents[9]

$$\delta(\mathrm{U}_i, \mathrm{U}_j) = \frac{\mathrm{SMF}(\mathrm{U}_i)}{\mathrm{SMF}(\mathrm{U}_j)} - 1, \tag{1.9}$$

where U_i refers to the one of our four UNIT simulations (and U_j to another, different one).

For all the simulations conducted, the results produced for the SMF qualitatively follow the observational trends. This outcome is in line with previous results such as those presented in Favole et al. (2020) and Asquith et al. (2018). The results obtained at redshift $z = 0$ agree almost seamlessly with the observational data. This is readily explained by the fact that the SAGE model was pre-calibrated to very similar data. When studying the behavior at higher redshifts (which is a prediction of the model) certain discrepancies start to show up. For stellar masses below $10^{11}h^{-1}\mathrm{M}_\odot$ the SMF calculated for the SAGE galaxies exceeds the observational points, while the opposite is true for masses higher than $10^{11}h^{-1}\mathrm{M}_\odot$. This is related to the condition that getting both the SMF at $z = 0$ *and* the cosmic star-formation history to simultaneously agree with the observations demands that stars that should have been formed in haloes below this simulation's resolution limit must instead be formed as extra stars in the haloes that are resolved. This inevitably leads to resolved high-z galaxies having too much stellar mass (and star-formation rates that are too high) in the model. It also changes how galaxies acquire stellar mass through mergers (as fewer mergers are resolved), which might help explain why there are too few galaxies with $M_* > 10^{11}h^{-1}\mathrm{M}_\odot$ at higher z in the model. Additionally, the deviations observed here for high redshifts – especially at the low-mass end – are similarly found when studying the SMF produced by other semi-analytic models, as extensively discussed in Asquith et al. (2018). Our explanation is hence generic and not only applies to SAGE. Therefore, despite the discrepancies seen in Fig. 1.1, the results obtained are reasonably accurate for us to say that the modelled SAGE galaxies fairly depict the behaviour of the SMF that could be expected in the redshift range for which Euclid is planned to operate.

[8]The 'CARNage calibration' set is available for download from `http://popia.ft.uam.es/public/CARNageSet.zip`.

[9]Note that we use the same strategy for presenting the variations across the four UNIT simulations in practically all plots.

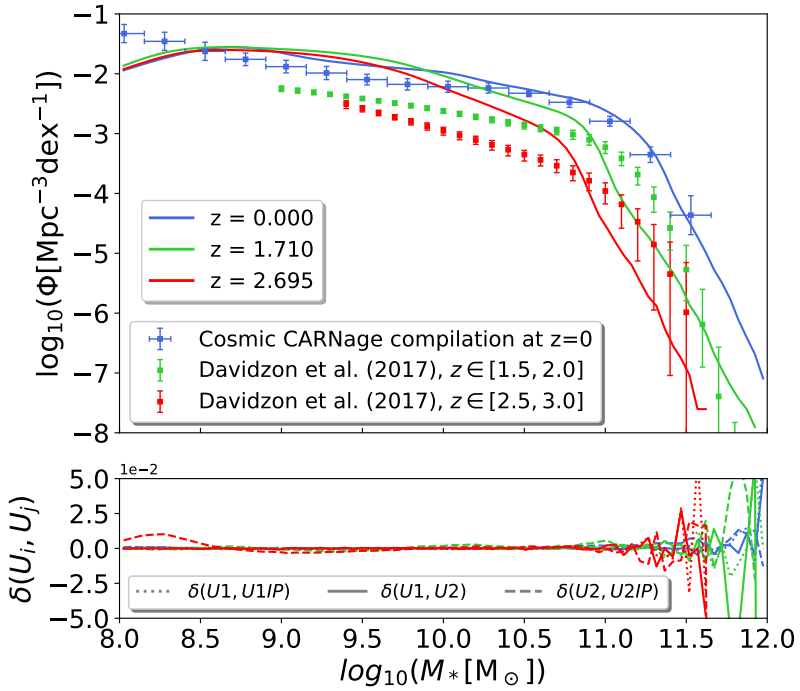

Figure 1.1: Stellar mass function. In the upper panel we compare the results for the modelled galaxies at various redshifts (solid lines) to observational data (points with error bars). The lower panel shows the fractional difference of U1 to the other UNIT simulations. Note that the $z = 0$ SMF has been used to calibrate the SAGE model whereas the results for higher redshift are a prediction of the model.

Another important aspect worth mentioning in this section is that due to resolution limitations in our simulations, galaxies whose stellar mass is lower than $10^9 h^{-1}\mathrm{M}_\odot$ have not been considered. Please refer to Knebe et al. (2018a,b) for a justification of this threshold, but we can also see in Fig. 1.1 how the number of galaxies starts to decline for stellar masses below that threshold due to numerical limitations. Therefore, to produce the results presented in the following sections we will discard all those galaxies whose mass is inferior to this threshold. This is not a cause for concern in this work though, as the vast majority of relevant ELGs have stellar masses above this threshold (see Appendix B).

1.3.2 Star Formation

With respect to the star formation (SF) in galaxies, which is also used as an input to the GET_EMLINES code, we only present the relation between specific star formation (i.e. SF per unit stellar mass) and stellar mass at redshift $z \sim 2$. We find that SAGE makes a prediction for this relation that is in excellent agreement with the observations of Daddi et al. (2007): in the main panel of Fig. 1.2 the specific SF rate (sSFR) of U1-SAGE galaxies is plotted against the stellar mass M_* for redshift $z = 2.028$. We show both the contours of a 2D histogram of this scatter plot as well as the median of the values obtained for sSFR within a series of bins along the x-axis. As is customary, in the bottom panel of the Fig. 1.2

the variations between simulations with respect to the other UNITSIM-SAGE galaxies have been represented. When comparing our results to observational data extracted from Daddi et al. (2007), we find sufficient agreement, at least within the 1σ regions. Though not explicitly shown here, we also confirm that our SAGE results are in excellent agreement with observational data for the sSFR (as provided by Elbaz et al., 2011) as a function of stellar mass at redshift $z = 0$. These results, in turn, are also compatible to those shown in Favole et al. (2020) for redshift $z = 0.1$.

For a comparison of the star formation rate (SFR) function to observational data at redshift $z = 0.14$ and the redshift evolution of the cosmic star formation rate density, we refer the reader to Knebe et al. (2018b). While the SFR function is compatible with the observational data at low redshift – as seen for the MultiDark galaxies and also confirmed for the UNITSIM galaxies (though not explicitly presented here) – it is worth mentioning that for SFR values greater than $\sim 10^{1.6} h^{-1}\mathrm{M}_\odot/\mathrm{yr}$, the number of galaxies generated with SAGE seems to underestimate the observed number (see fig. 2 in Knebe et al., 2018b). As we will see later in Section 1.4.2 this is going to leave an imprint on the abundance of (dust-attenuated) ELGs, especially at high redshifts. We finally like to remark again that the relation between sSFR and stellar mass as shown here is a prediction of the SAGE model.

Based on these results, we can say that our galaxies sufficiently reproduce the behaviour of the sSFR that would be expected for a sample of real galaxies in Euclid's operating range of redshifts.

1.3.3 The mass–metallicity relation

Another aspect of galaxies to be considered for the emission-line modelling is the chemical composition, since – depending on the fraction of metals that a galaxy may contain – its SFR may be substantially modified due to the fact that a higher metal content favours cooling mechanisms. This property is explicitly taken into account by the GET_EMLINES code and has to be provided as an input, respectively.

Since SF is regulated by the collapse of cold gas clouds, in Fig. 1.3 we study the relation that exists between the total mass of metals contained in such clouds and the total mass of cold gas in a given galaxy throughout the parameter $\mathcal{Z}$ which is calculated as (Favole et al., 2020; Knebe et al., 2018b):

$$\mathcal{Z} = 8.69 + \log_{10}(Z_{\mathrm{cold}}) - \log_{10}(Z_\odot) \tag{1.10}$$

where Z_{cold} was previously defined in Eq. (1.2), and we use the same $Z_\odot = 0.0134$ as already in Eq. (1.5). Note that this quantity $\mathcal{Z}$ is meant to be a proxy for $12 + \log(\mathrm{O/H})$.

In the main panel of Fig. 1.3 we show the correlation between $\mathcal{Z}$ and stellar mass as contours alongside the median (solid blue line) for redshift $z \sim 1$. The lower panel shows

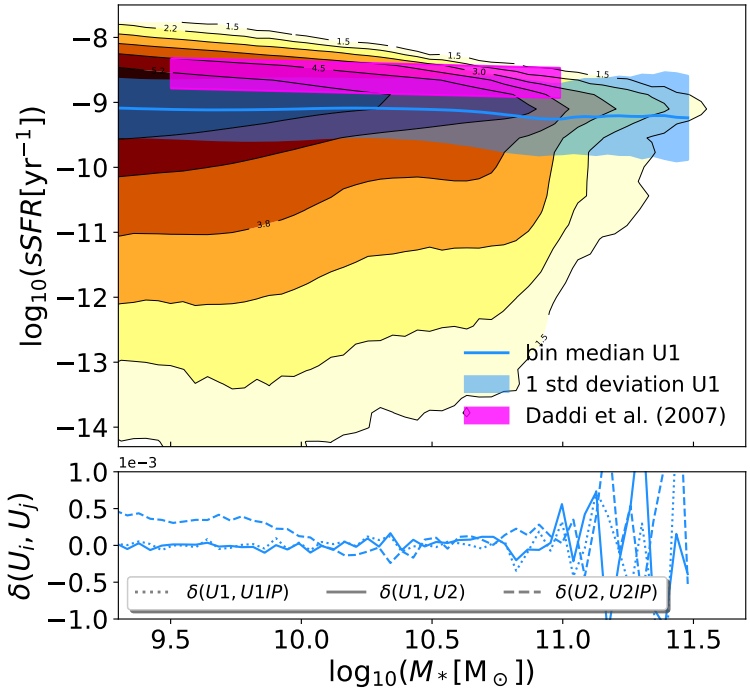

Figure 1.2: Specific star formation rate vs. stellar mass at redshift $z \sim 2$ in comparison to observations by Daddi et al. (2007). We show both the median (solid blue line) and the contour levels of the scatter relation. This relation is a prediction of the SAGE model.

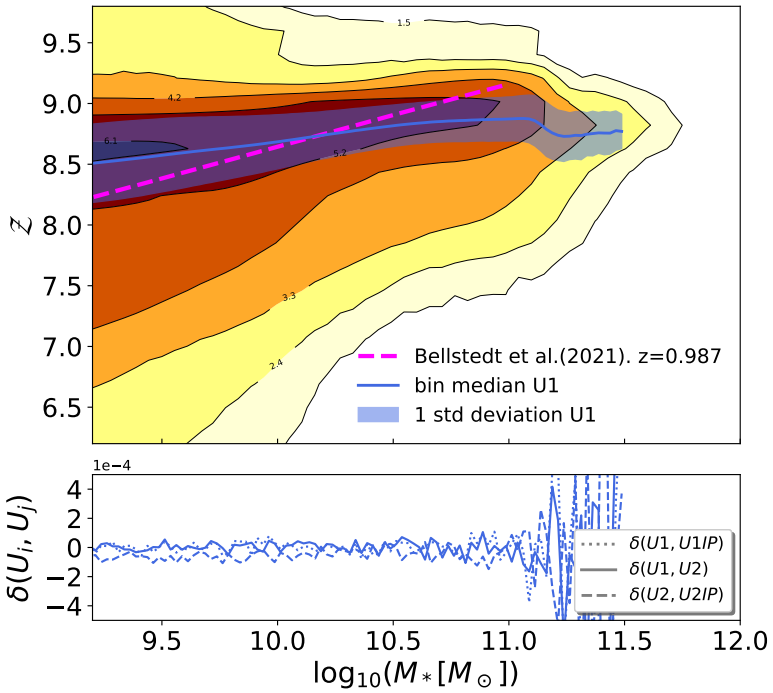

Figure 1.3: Cold gas metallicity vs. stellar mass. The relation shown here for redshift $z \sim 1$ is a prediction of the SAGE model. We also show the relation as found in Bellstedt et al. (2021) (dashed line).

again the fractional difference with respect to to the other UNITSIM model. The relation is as expected, i.e. larger mass galaxies have larger metallicities, with a strength comparable to the one observed for lower redshifts. This relation – as observed at redhift ~ 0.1 – is used during the calibration of the SAGE model; its extension to $z \sim 1$ shown here nevertheless is a clear prediction. We also show the relation as expected from observations by using the best-fit function presented in Bellstedt et al. (2021, eq. 6). This fitting function was obtained by applying the spectral-energy-distribution-fitting code `ProSpect` to galaxies from the Galaxy and Mass Assembly (GAMA) survey at $z < 0.06$; comparing with observations of gas-phase metallicity over a large range of redshifts, they then showed that their best-fit evolving mass–metallicity relationship is consistent with observations at all epochs and hence used here by us at redshift $z \sim 1$. We only show the Bellstedt et al. function out to $M_* = 10^{11} h^{-1} \mathrm{M}_\odot$ which was their limit for obtaining the best-fit parameters. The predictions of the SAGE model are in fair agreement with the Bellstedt et al. function. If one were to extrapolate the Bellstedt results, we would find a deficit of cold gas metallicity for the highest mass galaxies with $M_* > 10^{11} h^{-1} \mathrm{M}_\odot$. Even though there is no observational data in that regime, one possible explanation could be that the cold gas in those galaxies comes from mergers rather than accretion/cooling. I.e. AGN feedback might have shut off cooling entirely, so enriched gas in the circumgalactic medium will not get back to the inter-stellar medium. Instead, we might just be seeing the low-metallicity gas from now-cannibalised low-mass galaxies dominating most of the cold gas in the galaxy. But it yet remains unclear if the the drop in metallicity *predicted* for SAGE galaxies at $M_* = 10^{11} h^{-1} \mathrm{M}_\odot$ will also be seen in observations. While the redshift $z \sim 1$ is relevant for the Euclid mission, it also appears to be important to verify the mass-metallicity relation for even higher redshifts as it plays an important role in the estimation of emission lines. The Bellstedt et al. (2021) function can also be used to obtain results at, for instance, $z = 2$. There also exists a best-fit relation derived from actual observations at $z = 2.2$ (Maiolino et al., 2008, eq. 2 together with table 5). We refrain from showing the corresponding plot here, but confirm that our SAGE galaxies reproduce those two observations equally well as seen here for $z = 1$.

1.3.4 The disc size–mass relation

The last relevant quantity to validate for our SAGE galaxies is the size of the disc. While it is not important for GET_EMLINES it nevertheless enters into our dust attenuation model via Eq. (1.8). We therefore show in Fig. 1.4 the correlation of the effective disc radius (i.e. exponential scale radius, as calculated by SAGE) with stellar mass at redshift $z = 1.22$. For comparison we use the best-fit relation as reported by Yang et al. (2021, eq. 1) for late-type galaxies at redshift $z = 1.25$ and as derived from the complete

Hubble Frontier Fields data set. While the agreement is very good for higher mass galaxies, the disc sizes predicted by SAGE for galaxies with mass $M_* < 10^{10} h^{-1} \mathrm{M}_\odot$ are systematically larger than the observed ones. However, this does not significantly affect our results here as our ELGs preferentially have stellar masses $M_* > 10^{10.5} h^{-1} \mathrm{M}_\odot$ (see Fig. B.2).

Given all the results presented throughout this particular section, with the majority even being predictions of the SAGE model, we are confident that our UNITSIM-SAGE galaxies meet all the requirements to be used for the emission-line modelling, which is discussed in great detail in the following section.

1.4 SAGE's Emission-Line Galaxies (ELGs)

The results presented in the previous section indicate that our SAGE model galaxies are in sufficient agreement with a range of observations, in particular those properties that are used as an input for the model that calculates spectral emission lines. Here we now focus on the ELGs and contrast additional properties with a set of observations.[10] To this extent, we start with generating two distinct ELG catalogues, constructed from the full list of SAGE galaxies: one set will be obtained by simply applying the GET_EMLINES code (*RawELGs*) and another one by additionally modelling dust extinction (*DustELGs*). These value-added properties are included in the publicly available catalogues. However, in order to compare to existing observations and to make predictions for Euclid, we apply a redshift-independent flux cut of $F_{\mathrm{cut}} = 2 \times 10^{-16}$ erg s^{-1} cm^{-2}, which corresponds to the limit of the Euclid satellite.

1.4.1 The luminosity function of Hα-ELGs

We start with comparing the Hα luminosity functions (LFs) – as obtained by GET_EMLINES – at various redshifts of interest to observational data. The results can be viewed in Fig. 1.5 for the two base catalogues *RawELGs* and *DustELGs* at $z = 0.49, 0.987, 1.48$, and 2.23. For the first two redshifts we contrast our theoretical LFs to observations as found in Colbert et al. (2013). The data are taken from their table 3, where we removed again the [NII] contamination; as the data have not been corrected for dust extinction, they are best compared against our *DustELGs*. For the latter two redshifts, the observations from Sobral et al. (2016) are used. We used the data as provided in their table 4, noting that here they corrected for dust extinction, and hence those curves should be compared against our *RawELGs*. We actually find that our ELGs match the observations fairly well, though there are some obvious

[10] The same validation plots as shown in Section 1.3 for the SAGE galaxies can be found for the ELGs in Appendix B.

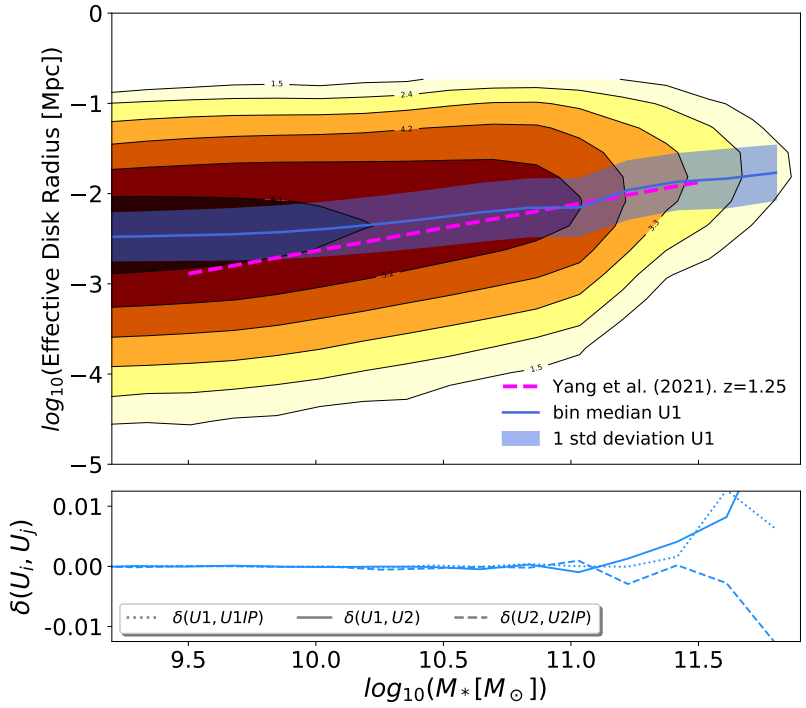

Figure 1.4: Effective disc radius as a function of stellar mass (contours and blue solid line with 1σ error region). This is a prediction of the SAGE model. We also show the relation as reported for late-type galaxies in Yang et al. (2021) at $z = 1.25$ (dashed line).

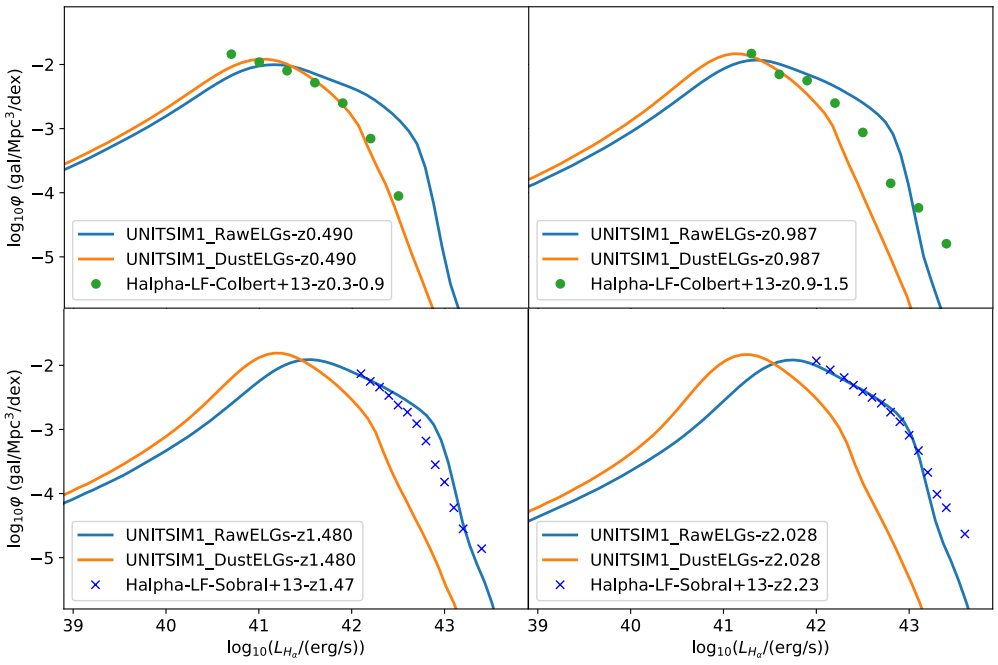

Figure 1.5: Evolution of the Hα luminosity function for our *RawELGs* (blue) and *DustELGs* catalogues (orange) ELGs in comparison to observational data: the Colbert et al. (2013) data (red points) are best compared to *DustELGs* whereas the Sobral et al. (2016) data (blue crosses) to *RawELGs* (see main text for details). Redshifts are (clockwise starting in the upper left panel) $z = 0.490, 0.987, 1.480,$ and 2.018.

discrepancies at redshift $z \sim 1$: both *DustELGs* and even *RawELGs* do not provide enough high-luminosity ELGs. This eventually translates into a too-low (integrated) abundance, as we will see below. But we are not too concerned about that as the relevant redshift range for Euclid is $z \in [0.9, 1.8]$, and the match of our *RawELGs* galaxies with the Sobral et al. (2016) observations is rather good for $z \sim 1.5$, i.e. the centre of that interval.

In the Introduction we mentioned that Zhai et al. (2019) also model Hα ELGs using the Galacticus SAM coupled to the single UNITSIM1 simulation. But their catalogue was constructed such that the SAM parameters were tuned to best reproduce – amongst other properties – the Hα LFs, and in particular the observed ones shown here for redshifts $z = 1.48$ and 2.23 (see their fig. 1). They accomplish this by – in practice – adjusting $A_\lambda(\tau_\lambda^z, \theta)$ (as also found in our Eq. (1.3)) as a free parameter, tuning it until they match the observed Hα LF at a given redshift. Our value for $A_\lambda(\tau_\lambda^z, \theta)$ is based upon physical properties of the underlying galaxies whose values change as a function of redshift (leading to an implicit redshift dependence of our dust model). Meaning, we actually use a physically motivated A_λ and hence the LFs seen here are a clear prediction of our modelling.

1.4.2 Abundance evolution of flux-selected Hα-ELGs

We show in Fig. 1.6 the redshift evolution of the number density for our *RawELGs* and *DustELGs* catalogues, after applying the redshift-independent flux cut of $F_{\rm cut} = 2 \times 10^{-16}$ erg s^{-1} cm^{-2}, in comparison to observational data from Colbert et al. (2013) and Bagley et al. (2020). We also show two of the three models of Pozzetti et al. (2016, P16). By fitting to observed luminosity functions from existing Hα surveys, P16 build three distinct models for the Hα number density evolution. Different fitting methodologies, functional forms for the luminosity function, subsets of the empirical input data, and treatment of systematic errors were considered to explore the robustness of the results. Functional forms and model parameters were made available[11] (and are being used here), along with the counts and redshift distributions up to $z \sim 2.5$ for a range of limiting fluxes bracketing the sensitivity of Euclid. Their models are named 'Pozzetti model #1, #2, and #3', with model #1 being the most optimistic and model #3 the most pessimistic for Euclid.[12] Both these models are shown here, also for a flux cut of 2×10^{-16} erg s^{-1} cm^{-2}.

We can see in Fig. 1.6 how, for $z < 1$, our *DustELGs* follow the same trends as the P16 models, but show a substantial lack of objects at higher redshift. By comparison, our *RawELGs* clearly overpredict the abundance of ELGs for the applied redshift-independent

[11] The P16 data can be downloaded from here: `http://www.bo.astro.it/~pozzetti/Halpha/Halpha.html`

[12] P16 called the models that way themselves, based upon the fact that if you have more galaxies, you reduce the shot-noise. Hence, Pozzetti model #1 is more optimistic for Euclid's figure-of-merits than #3, as we will have smaller error bars in the cosmological parameters.

Table 1.1: Average number density and flux cuts as a function of redshift z (first column). Columns 2–3 top table (*RawELGs*) and bottom table (*DustELGs*) list the mean and standard deviation (across the four UNIT simulations) of the number density of ELGs with an applied redshift-independent flux cut of $F_{\rm cut} = 2 \times 10^{-16}$ erg s^{-1} cm^{-2}. Columns 4–5 and 6–7 give the target number density (taken from table 3 in P16) and average flux cut applied to reach it (the standard deviation is smaller than the reported accuracy and hence left out for clarity) for *RawELGs-Poz1* and *RawELGs-Poz3*, respectively (top table). The bottom table provides the same information for *DustELGs-Poz1* and *DustELGs-Poz3*.

z	*RawELGs* ($F_{\rm cut} = 2$)		*RawELGs-Poz1*		*RawELGs-Poz3*	
	$<dN/dz>$	σ	dN/dz	$<F_{\rm cut}>$	dN/dz	$<F_{\rm cut}>$
0.490	24652	47	9946	6.441	–	–
0.987	22015	89	7353	4.864	3779	7.080
1.220	17709	94	5097	4.600	2518	6.300
1.321	15809	98	4281	4.452	2148	5.759
1.425	13988	77	3447	4.343	1817	5.353
1.650	10277	57	2253	3.930	1279	4.564
2.028	5294	38	1006	3.330	616	3.687

z	*DustELGs* ($F_{\rm cut} = 2$)		*DustELGs-Poz1*		*DustELGs-Poz3*	
	$<dN/dz>$	σ	dN/dz	$<F_{\rm cut}>$	dN/dz	$<F_{\rm cut}>$
0.490	15262	30	9946	2.85	–	–
0.987	3238	12	7353	1.37	3779	1.88
1.220	957	7	5097	1.13	2518	1.50
1.321	577	4	4281	1.05	2148	1.35
1.425	370	3	3447	0.98	1817	1.22
1.650	153	3	2253	0.84	1279	1.00
2.028	35	1	1006	0.67	616	0.77

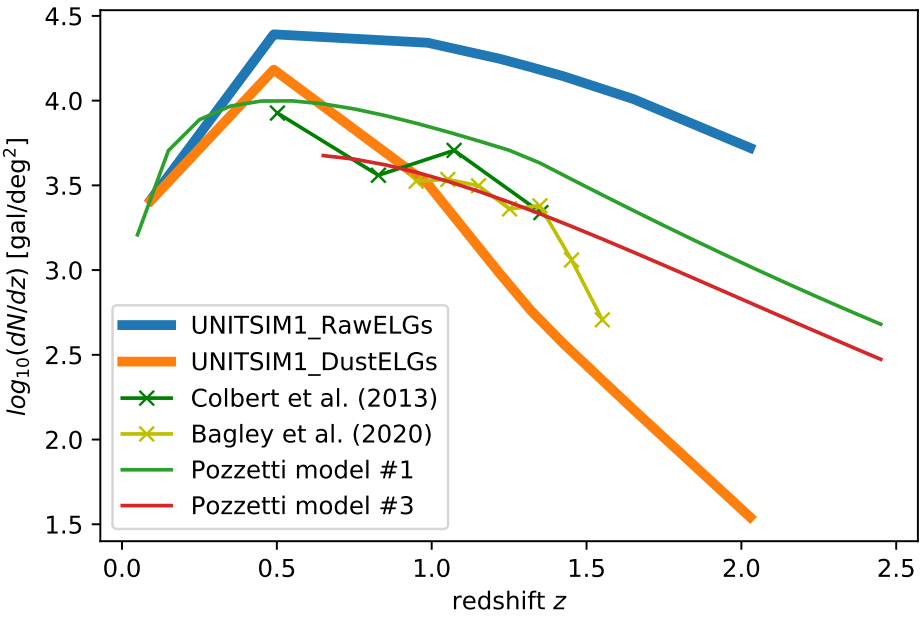

Figure 1.6: Redshift evolution of the number density of dust-attenuated ELGs (*DustELGs*, orange) and the initial *RawELGs* catalogue (i.e. no dust modelling, blue), both for a redshift-independent flux cut at 2×10^{-16} erg s^{-1} cm^{-2}, in comparison to the observational data of Colbert et al. (2013) and Bagley et al. (2020). We also show model #1 and #3 of Pozzetti et al. (2016) for the same flux threshold (not to be confused with our catalogues *PozMod1* and *PozMod3* that were designed to match these number densities). Only UNITSIM1 ELGs are shown for clarity.

flux cut (at least for $z > 0.5$). A similar discrepancy between semi-analytic galaxies and the P16 models can also be seen in fig. 5 of P16, where their three models are compared against the results from two other SAMs. It should also be mentioned that a more recent study of the observed number density evolution of Hα ELGs indicates a possible decline beyond redshift $z \sim 1.4$ (Bagley et al., 2020, lower right panel of their fig. 7), although it is not as pronounced as the dip found for our *DustELGs*. To highlight this we have added those data points[13] to our plot, too. While there is agreement between the observations of Bagley et al. and P16's model #3 in the redshift range $z \in [1, 1.5]$, the observational data drop more steeply at higher redshifts and are more in line with our *DustELGs* prediction. However, Bagley et al. (2020) also say that their higher redshift points are in the region were the sensitivity of their instrument could be degraded.

This discrepancy between ours and the Pozzetti ELG number densities is also reflected in Table 1.1, where we list as a function of redshift the number density of ELGs in our

[13]The Bagley et al. (2020) data are based upon completeness-corrected measurements of the blended Hα and NII fluxes, while our fluxes include only Hα. We have therefore 'corrected' the Bagley et al. (2020) data points – as obtained with PlotDigitizer – by reversing their adjustment to model #3 of P16 to account for the combined fluxes. This was done by finding the shift needed to bring the digitized data points into the same kind of agreement with the original Pozzetti model #3, as seen in the lower right panel of Bagley's fig. 7 for the blended Pozzetti model #3.

reference *RawELGs* and *DustELGs* catalogues (as averaged over the four UNIT simulations, also providing the standard deviation). While we have to acknowledge that both our *RawELGs* and *DustELGs* do not reproduce the predictions of P16, we also have to remark again that it yet remains unclear what the correct abundance evolution dN/dz should look like.

1.4.3 Flux-adjusted catalogues

Taking the models of P16 as the reference, we now construct four additional catalogues that are designed to match the P16 dN/dz curves as shown in Fig. 1.6. We take *RawELGs* as the starting point and adjust the flux threshold until reaching the target dN/dz values as given by P16's models #1 and #3, providing us with the two models *RawELGs-Poz1* and *RawELGs-Poz3*. We use the same approach for *DustELGs*, providing two more models: *DustELGs-Poz1* and *DustELGs-Poz3*. We used this methodology with all four UNITSIM catalogues. The means of the required flux cuts to our data are listed in columns 4–7, and 10–13 of Table 1.1 (we omit error estimates as they are below the reported accuracy). The remaining columns – 2, 3, 8, and 9 – are the mean number densities (and its standard deviation) for the *RawELGs* and *DustELGs* catalogues, respectively, when using a redshift independent flux threshold of $F_{\rm cut} = 2 \times 10^{-16}$ erg s^{-1} cm^{-2}. Our methodology for constructing ELGs eventually leaves us with six distinct catalogues[14]

1. **RawELGs**: directly coming from GET_EMLINES (with a flux threshold of $F_{\rm cut} = 2 \times 10^{-16}$ erg s^{-1} cm^{-2} across all redshifts, when used here),

2. **RawELGs-Poz1**: variable flux threshold applied to *RawELGs* to match the number density of Pozzetti's model #1 at each redshift,

3. **RawELGs-Poz3**: variable flux threshold applied to *RawELGs* to match the number density of Pozzetti's model #3 at each redshift,

4. **DustELGs**: passing the *RawELGs* ELGs through our dust model (with a flux threshold of $F_{\rm cut} = 2 \times 10^{-16}$ erg s^{-1} cm^{-2} across all redshifts, when used here),

5. **DustELGs-Poz1**: variable flux threshold applied to *DustELGs* to match the number density of Pozzetti's model #1 at each redshift,

6. **DustELGs-Poz3**: variable flux threshold applied to *DustELGs* to match the number density of Pozzetti's model #3 at each redshift,

where we note that all the ELGs are, by construction, a subset of the full SAGE catalogue used in the previous section. Likewise, the four additional '*ELGs-Poz*' catalogues are

[14] We need to state here again that the public versions of *RawELGs* and *DustELGs* are *not* subjected to any flux cut: they contain all ELGs as provided by GET_EMLINES.

sub-sets of the public *RawELGs* and *DustELGs*, respectively.

Instead of introducing a redshift-dependent flux cut – which might be considered counter-intuitive, as Euclid will have a fixed flux threshold – we could have also taken the *RawELGs* model as the starting point and tuned our dust extinction parametersuntil we match the P16 dN/dz values, akin to what Zhai et al. (2019) have done. But finding the best possible dust model is beyond the scope of this work and hence we prefer to adhere to the former approach. The main idea here is to restrict the model ELGs to the brightest ones that are still observable. And we have seen in Fig. 1.5 that applying the dust model basically just shifts the LF towards lower luminosities, especially at high redshift and for the brightest ELGs (e.g. Sobral et al., 2016). Therefore, adjusting the luminosity threshold will still select the brightest galaxies. Moreover, one could also re-calibrate SAGE, the GET_EMLINES code or choose a different dust model beyond a Cardelli law, all of which can affect the number density of ELGs. But exploring all these possibilities is beyond the scope of the present work. We prefer to work with minimal variations to the existing models and codes.

We also like to emphasize that our '*-Poz1*' and '*-Poz3*' models are *not* the two models #1 and #3 of P16. They are ELG catalogues where we adjusted the number densities to match those of model #1 and #3 of Pozzetti, respectively. We did this to correct for the mismatch of ELGs with respect to the Pozzetti models seen in Fig. 1.6. We further refrain from showing their abundance evolution as they match – by construction – the curves from P16.

Given the results presented in this section, we conclude that our UNITSIM-SAGE-ELGs provide a fair sample and can be used for further analysis. The *RawELGs* and *DustELGs* galaxies will serve as the two base catalogues, with the four additional catalogues acting as our best predictions for Euclid. As a particular application we employ them now for a study of galaxy clustering and the related bias.

1.5 Clustering of ELGs

Quantifying the clustering of galaxies is one of the main objectives of ongoing and upcoming galaxy surveys such as the Euclid satellite mission. Clustering measurements probe the fluctuations of the underlying dark matter from the positions of galaxies, and they encode geometric, model-dependent cosmological information. Using the positions of our theoretical UNITSIM ELGs, we now study the two-point correlation function $\xi_{\rm ELGs}(r)$ and its redshift evolution. We further use the positions of 10^7 randomly selected dark matter particles from the total 4096^3 particles present in each of the UNIT gravity-only simulations

to calculate $\xi_{\rm DM}(r)$.[15] This allows us to also infer the bias that we define here as

$$b(r) = \sqrt{\frac{\xi_{\rm ELGs}(r)}{\xi_{\rm DM}(r)}} \tag{1.11}$$

between both populations and study its evolution across redshift. The bias b, i.e. the statistical relation between the distribution of galaxies and matter, needs to be taken into account when interpreting galaxy surveys; it describes how galaxies trace the underlying dark matter distribution. The biased galaxy formation scenario (e.g. Dekel and Rees, 1987) implies that galaxies are not uniformly distributed in the Universe, but primarily form in the peaks of the matter density field. Galaxies are therefore biased tracers of it, sampling only the overdense regions (see Desjacques et al., 2018a, for a recent review). The particular bias of ELGs, i.e. a sub-class of all galaxies, will be of greatest importance for surveys such as Euclid.

All two-point correlation functions (2PCFs) have been obtained with the CUTE[16] software presented in Alonso (2012). In addition, for the results that we will present throughout this section, we have taken the average of the values computed for the 2PCF over the four simulations UNITSIM1, UNITSIM1-InvertedPhase, UNITSIM2, and UNITSIM2-InvertedPhase.

In the top panel of Fig. 1.7 we present the 2PCF computed for the *RawELGs* (dashed lines) and dark matter (solid lines). The lower panel of the same figure shows the bias $b(r)$ defined via Eq. (1.11). In order to better verify the scale-dependence of the bias, we also calculate the 'average' bias

$$\langle b \rangle = \frac{1}{N_{\rm bin} - 1} \sum_{2}^{N_{\rm bin}} b_i \,, \tag{1.12}$$

where $N_{\rm bin}$ is the number of bins and $b_i = b(r_i)$ the value of the bias in distance bin r_i. This average bias $\langle b \rangle$ is shown as a dashed horizontal line in the lower panel of Fig. 1.7. Note that we exclude the first bin in this calculation since for such small distances the bias is certainly scale-dependent (see Fig. 1.10 below). It is also obvious that the data for this particular model become rather noisy at high redshifts due to the very low number of objects above the reference flux cut of $F_{\rm cut} = 2 \times 10^{-16}$ erg s^{-1} cm^{-2} (see Table 1.1). But we can nevertheless appreciate that for distances $r \gtrsim 5h^{-1}{\rm Mpc}$ the bias is remarkably constant, something we will quantify in more detail below.

An equivalent analysis has been conducted for our other ELG catalogues, but we decided to only show here in Fig. 1.8 the results for the bias and not also the 2PCFs. Once more we can see that we get fairly noisy results at redshift $z = 2.028$ due to the reduced number of

[15] We confirm that the resulting 2PCFs have converged and will not change when using more particles. Further, this number of dark matter particles is comparable to the number of ELGs, at least at redshifts $z \leq 1$.

[16] `https://github.com/damonge/CUTE`

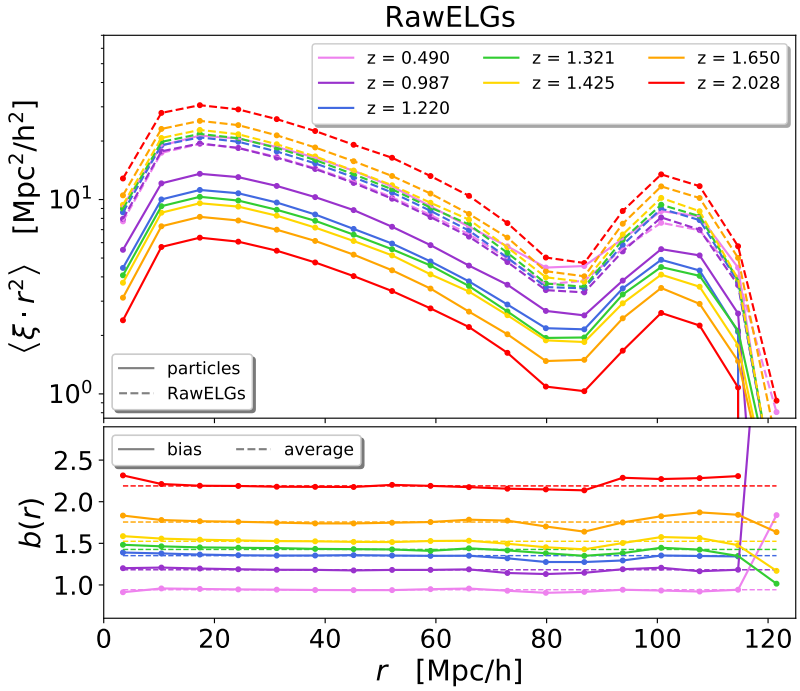

Figure 1.7: Top panel: 2PCF of SAGE ELGs with flux greater than 2×10^{-16} erg s^{-1} cm^{-2} (*RawELGs* galaxies, dashed lines) and collisionless trace particles (solid lines) for various redshifts. Bottom panel: associated bias as defined by Eq. (1.11).

galaxies at that redshift. We also observe that at scales $\sim 120h^{-1}\mathrm{Mpc}$ the bias behaves more erratic, which can be explained by the fact that the 2PCF crosses zero at $r \sim 130h^{-1}\mathrm{Mpc}$ (Sánchez et al., 2008; Prada et al., 2011): taking the numerical ratio between two numbers close to zero then introduces noise. But the most important point is that the bias of ELGs (at least for $z < 2$) in all our catalogues remains constant on scales $r \in [5, 100]h^{-1}\mathrm{Mpc}$ (in line with the findings of, for instance, Abbott et al., 2018). Below $5h^{-1}\mathrm{Mpc}$ it is obvious that the mixture contribution between the one- and two-halo terms will introduce non-linear effects which in turn will cause the bias to no longer behave independently with scale. On larger scale we have already seen above that the zero-crossing of the 2PCF is introducing noise and hence the results for the bias are expected to be affected by this, too. We further note that the bias clearly is a function of redshift. But this is also expected, as the mass of the haloes hosting ELGs will change with redshift (see Fig. B.1 in the Appendix). Not only that, but haloes of the same mass or luminosity at different redshifts will also have a different bias. It therefore only appears natural that the bias increases with redshift as, for instance, modelled analytically by Basilakos et al. (2008) or found in other cosmological simulations (e.g. Merson et al., 2019; Tutusaus et al., 2020).

Fig. 1.9 now quantifies the evolution of the average bias $\langle b \rangle$ (obtained from the results presented in Fig. 1.7 and Fig. 1.8) as a function of redshift for all our catalogues. This figure is accompanied by Table 1.2 that lists the plotted values. We find that for all our galaxies the

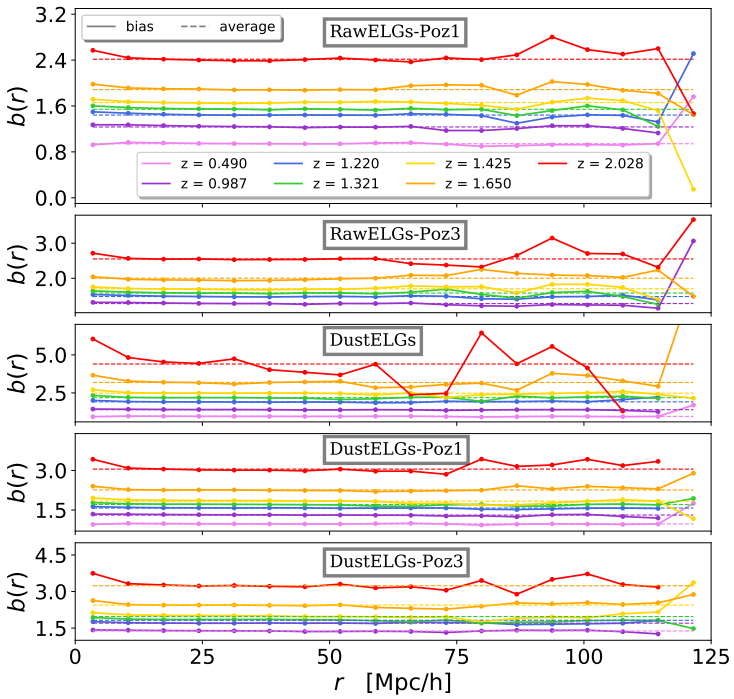

Figure 1.8: The bias for the *RawELGs-Poz1*, *RawELGs-Poz3*, *DustELGs*, *DustELGs-Poz1*, and *DustELGs-Poz3* galaxies (in that order from top to bottom), using the same r-range and redshift colouring as for Fig. 1.7.

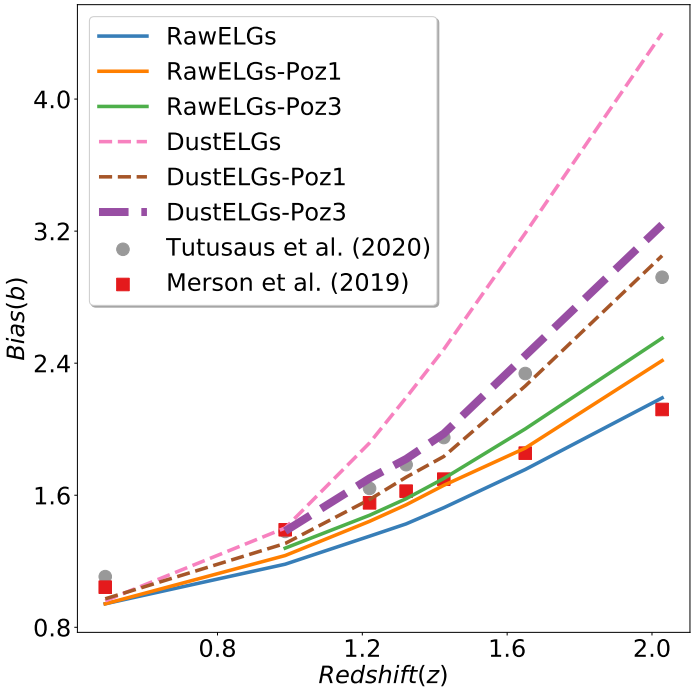

Figure 1.9: Bias values averaged for scales larger than 5 $h^{-1}\mathrm{Mpc}$ computed for our six models. The grey points are the best fit $b(z)$ for the bias found in Euclid's *Flagship* simulation (Tutusaus et al., 2020), and the red squares the results as reported by Merson et al. (2019).

Table 1.2: Bias values averaged for scales larger than $5h^{-1}\mathrm{Mpc}$ computed for all our ELG catalogues. The values listed here correspond to the lines presented in Fig. 1.9.

z	*RawELGs*	*RawELGs Poz1*	*RawELGs Poz3*	*DustELGs*	*DustELGs Poz1*	*DustELGs Poz3*
0.490	0.94	0.94	–	0.96	0.97	–
0.987	1.18	1.23	1.28	1.40	1.31	1.38
1.220	1.35	1.44	1.48	1.92	1.57	1.70
1.321	1.42	1.54	1.58	2.19	1.71	1.82
1.425	1.53	1.66	1.70	2.48	1.83	1.97
1.650	1.76	1.89	2.00	3.19	2.26	2.45
2.028	2.19	2.42	2.55	4.40	3.05	3.24

bias systematically increases with redshift, despite showing different growth rates, especially for the two base catalogues *RawELGs* and *DustELGs*. We also acknowledge that the strength of this $b(z)$ relation for our four '*-Poz*' galaxies – especially the ones based upon *DustELGs* – is in excellent agreement with the relation presented in Tutusaus et al. (2020, eq. 11), shown as circles in Fig. 1.9. The $b(z)$ function given in Tutusaus et al. is derived from studying the bias in the Euclid *Flagship* simulation,[17] which is also just based upon dark matter. But the way in which the dark matter haloes are populated with galaxies is quite distinct to our approach: they have applied a Halo Occupation Distribution (HOD) that does not take into account the merger trees of the haloes (for a comparison of these two different techniques see, for instance, Knebe et al., 2015, 2018a).[18] Merson et al. (2019) also forecast the redshift evolution of the linear bias for Hα-emitting galaxies in a similar redshift range. Their data are shown here as squares. Like Tutusaus et al. (2020), they also used a HOD for which they calibrated the dust attenuation to reproduce observed Hα counts. Merson et al. (2019) now predict lower biases than Tutusaus et al. (2020) and our dust-based '*-Poz*' galaxies, more in line with the results we obtain for our *RawELGs* catalogue and its derivates. The comparison of these three different $b(z)$ predictions for ELGs indicates that the theoretical models have not yet converged. There are degeneracies and uncertainties that still require more detailed and refined investigations before any final conclusion could be drawn. But we finally remark that our findings for the redshift evolution of the bias $b(z)$ are also in agreement with those of Favole et al. (2017, right panel of their fig. 6), who used a SHAM model. However, in their work, the bias increases more mildly, as the SDSS redshift range studied there is very much reduced compared to ours.

[17] `https://www.euclid-ec.org/?page_id=4133`

[18] While there is no reference paper for this galaxy catalogue, we nevertheless like to mention that it is based upon the MICE HOD (Carretero et al., 2015). The clustering is fit to SDSS galaxies as a function of magnitude and colour at low redshift. Then, most of the properties are assumed to depend on redshift only via their SEDs/color evolution, allowing for correlations between many observables.

We further recognize in Fig. 1.9 that the bias is sensitive to the particulars of our modelling, especially at high redshift. This certainly relates to how we treat the dust extinction and select the observable ELGs from *RawELGs*, respectively. But this is known and can also be appreciated when comparing the bias predictions from Tutusaus et al. (2020) and Merson et al. (2019) where similar discrepancies are seen. We particularly notice the degeneracy between dust modelling and flux selection: first applying our extinction prescription and then matching a preset dN/dz by varying the flux threshold always leads to larger bias than not employing a dust model at all. Even though we argued before that the dust-attenuated luminosities – as seen in Fig. 1.5 – are a shifted version of the raw values (at least for luminous ELGs; see also Sobral et al., 2016, where a constant luminosity offset was applied to model dust extinction), here we realize that their relation is not that simple. But we have clearly seen that fixing the abundance of Hα ELGs, the differences substantially reduce. Nevertheless, we like to stress again that designing a new dust extinction model is beyond the scope of this work and hence we leave a more detailed study of this to a future work. Note that in this work we primarily aim at presenting the publicly available data, discussing its scope and possible limitations.

So far we have mainly focused on large scales, but to conclude this section we also present how the bias varies for *small* scales. In Fig. 1.10 we present the bias $b(z)$ for various redshifts and all our catalogues out to $r \approx 20h^{-1}\mathrm{Mpc}$ using logarithmic binning. We observe that for redshifts $z < 2$ the bias remains constant down to scales $r \approx 3h^{-1}\mathrm{Mpc}$ and then starts to mildly drop. It is actually around this distance that we expect the contribution from the one-halo term to start to become relevant. However, this behaviour weakens for higher redshifts and possibly reverses for $z = 2$. Something similar has also been observed by Nuza et al. (2012, fig. 10) for BOSS CMASS galaxies, but there the inversion was already seen at redshift $z \approx 0.53$ (and one needs to bear in mind that CMASS galaxies and ELGs are not directly comparable as they are different types of galaxies, where the latter are mostly star-forming and the former could be dominated by passive galaxies).

1.6 Conclusions

Realistic simulations are a necessary tool to optimise and validate the methodology that will be used to extract cosmological constraints from future surveys. Indeed, they are used to estimate the theoretical error budget on surveys (for example, for the eBOSS-ELG analysis, see Alam et al., 2021b).

In this work we have employed the UNIT simulations, which model the evolution of dark matter within a $1h^{-1}$Gpc box at a mass resolution of $1.2 \times 10^9 h^{-1}\mathrm{M}_\odot$ per particle

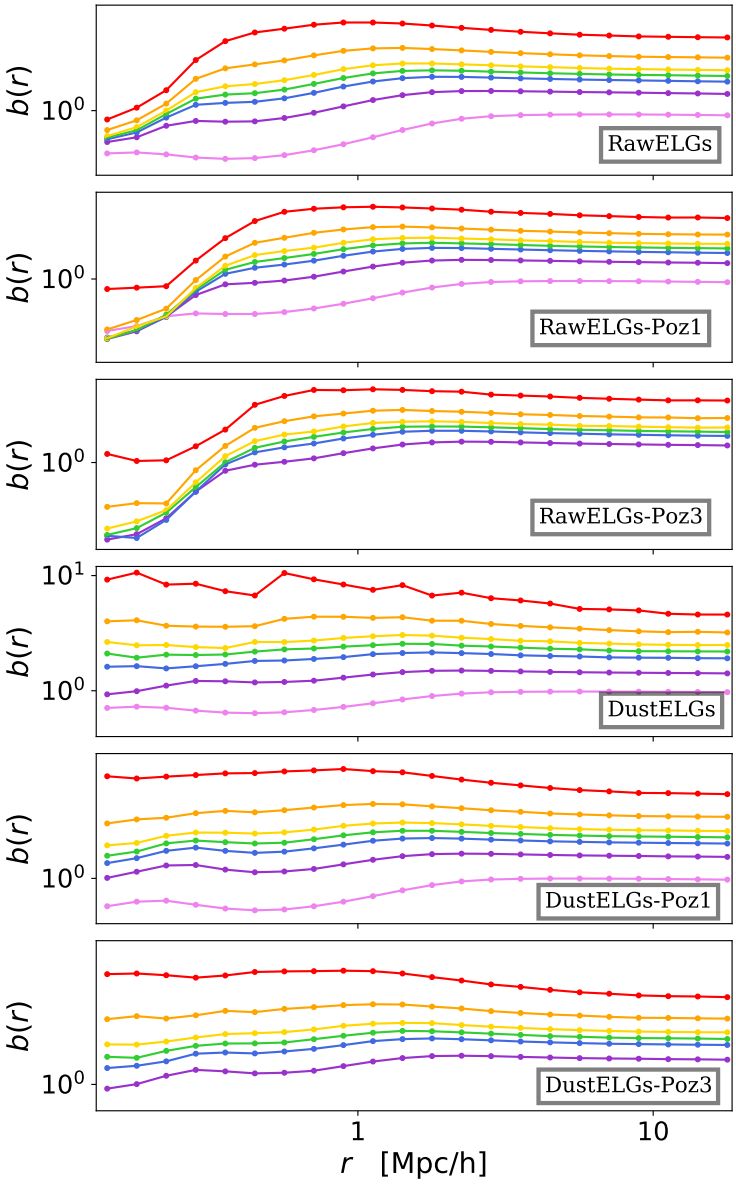

Figure 1.10: The bias $b(r)$ for all six models in logarithmic r bins, focusing on the small scales up to $r < 20h^{-1}\mathrm{Mpc}$, using the same redshift colouring as for Fig. 1.7.

(Chuang et al., 2019). Given the large volume of these simulations together with the fixed-and-paired technique of Angulo and Pontzen (2016b) that enhances the effective volume of the simulations, the resulting galaxy mocks that we have produced represent a unique resource for model testing based on a semi-analytic model of galaxy formation. We used our ELG catalogues to make predictions for the galaxy statistics that the Euclid experiment is expected to obtain for redshifts between $0.9 < z < 1.8$. Note that the simulations presented here cover an effective survey volume of about seven times the effective survey volume of Euclid (Chuang et al., 2019). And having the galactic physics included is key, since the complicated relation between haloes and galaxies can modify the clustering of ELGs significantly, even at scales used to put cosmological constraints when working in Fourier space (see, for instance, Gonzalez-Perez et al., 2020; Avila et al., 2020).

For this work we have generated six synthetic catalogues of emission-line galaxies of which the two base ones (i.e. *RawELGs* and *DustELGs*, without any flux cuts applied) are publicly available. The galaxies were first obtained by applying the semi-analytic galaxy formation model SAGE (Croton et al., 2016) to the gravity-only UNIT simulations. They were then subjected to the emission line modelling with the GET_EMLINES code (Orsi et al., 2014) and an additional dust attenuation model (following Favole et al., 2020). This left us with the two base ELG catalogues *RawELGs* and *DustELGs*, in addition to the general SAGE galaxy catalogues. As argued throughout Section 1.3, the properties associated with our UNITSIM-SAGE galaxies reproduce observed properties of galaxies with $0 \leq z < 2$. Here we have focused on those properties that are most relevant for the construction of ELGs catalogues, i.e. stellar mass, star formation rate, metalicity, and disc size. In particular, we find that the (evolution of the) mass–metallicity relation agrees sufficiently well with observations. However, we have seen in Knebe et al. (2018b) that the SAGE model underpredicts the number of galaxies with high SFRs. This then affects the abundance of our (dust-attenuated) ELGs as seen in Fig. 1.6. While we presented the validation plots in the main body of the paper only for the full set of SAGE galaxies, the corresponding plots for the *RawELGs* and *DustELGs* ELGs can be found in Appendix B.

In Section 1.4 we adjusted the number densities of our two base UNITSIM-ELG sets by applying distinct flux thresholds to them (using the Euclid-models as given by P16), eventually comparing the redshift evolution of their abundance to observations. When studying the density of galaxies per deg^2 with fluxes greater than 2×10^{-16} erg s^{-1} cm^{-2} as a function of redshift we observe that the density obtained for the raw ELG galaxies is above both the observations and other theoretical modelling. That means that some additional selection needs to be applied to end up with a more realistic ELG catalogue. We have addressed this in several ways. We first applied a dust-attenuation (a Cardelli law, following Favole et al., 2020), which led to a possible underestimation of the expected density of

galaxies dN/dz observed from redshift $z \sim 1.4$ onwards. Nevertheless, the most recent study by Bagley et al. (2020) suggests that the observational value for dN/dz could be closer to our results than predicted by P16. We also designed additional catalogues where we instead varied the flux threshold for the selection of galaxies from the *RawELGs* catalogue; those fluxes were adjusted to reproduce number densities as predicted by P16.

The linear bias is a key parameter to understand the cosmological power of Euclid and can help construct forecasts that inform the optimisation of both observational and analysis strategies. The bias of Hα galaxies may be particularly relevant for forecasts on studies such as primordial non-Gaussianities or relativistic effects. We therefore studied the clustering of all our six samples listed in Table 1.1: two with the Euclid flux cut applied and four in which the flux cuts are adjusted to follow the predictions by two of the models presented in P16. We measure the linear bias as a function of redshift by averaging $\xi_{\rm ELGs}/\xi_{\rm DM}$ for scales $r > 5h^{-1}{\rm Mpc}$. For the samples whose abundances are matched to the to P16 predictions, we find a $b(z)$ in line with that reported in Tutusaus et al. (2020) for the Euclid *Flagship* simulation (and mildly in agreement with the same results reported by Merson et al., 2019). This is striking, as the *Flagship* mock was constructed following a very different methodology (Carretero et al., 2015). Additionally, we report the clustering at small scales, that becomes scale-dependent. These measurements can be used to test the robustness of different large-scale structure models to extract cosmological information from the small scales, that have the highest signal-to-noise ratio but at the same time are the most difficult to model.

We close with the remark that an improved dust attenuation modelling might be the most physical approach for choosing the ELGs so that the observed dN/dz will be recovered. This would, however, only affect the catalogues that are based upon *DustELGs*; it will leave *RawELGs* untouched, which is the primary ELG catalogue made available publicly. Therefore, while we have shown throughout this work that the particulars of the dust extinction have an effect on the the results, the published data contain all that is required for the community to apply their favourite post-processing models for dust and emission lines from star-forming regions. Or put differently, the base catalogue *RawELGs* is certainly inclusive, i.e. a superset of the ELGs of interest. A better understanding of the process of selecting observable ELGs from that base catalogue and developing an improved dust attenuation model will be left for a future work. The public data can already been used for a great variety of studies and have extensive applications like, for example, informing Halo Occupation Distribution models. Indeed, we will study the properties of Hα ELG HOD models in a follow-up paper.

Chapter 2

The cosmology dependence of the concentration-mass-redshift relation

The concentrations of dark matter haloes provides crucial information about their internal structure and how it depends on mass and redshift – the so-called concentration-mass-redshift relation, denoted $c(M, z)$. We present here an extensive study of the cosmology-dependence of $c(M, z)$ that is based on a suite of 72 gravity-only, full N-body simulations in which the following cosmological parameters were varied: σ_8, $\Omega_{\rm M}$, $\Omega_{\rm b}$, $n_{\rm s}$, h, M_ν, w_0 and $w_{\rm a}$. We characterize the impact of these parameters on concentrations for different halo masses and redshifts. In agreement with previous works, and for all cosmologies studied, we find that there exists a tight correlation between the characteristic densities of dark matter haloes within their scale radii, r_{-2}, and the critical density of the Universe at a suitably defined formation time. This finding, when combined with excursion set modelling of halo formation histories, allows us to accurately predict the concentrations of dark matter haloes as a function of mass, redshift, and cosmology. We use our simulations to test the reliability of a number of published models for predicting halo concentration and highlight when they succeed or fail to reproduce the cosmological $c(M, z)$ relation.

2.1 Introduction

Cosmological simulations have revealed that the spherically-averaged density profiles of dark matter (DM) haloes exhibit a high degree of self-similarity across a wide range of masses, redshifts, and cosmologies (Navarro et al. 1996; Huss et al. 1999; Bode et al. 2001; Bullock et al. 2001; Neto et al. 2007; Macciò et al. 2008; Knollmann et al. 2008; Wang and White 2009; Hellwing et al. 2013; Ludlow and Angulo 2017; Brown et al. 2020 and Angulo and Hahn 2022b for a review). The most popular analytic expression used to describe these

profiles is the NFW profile (Navarro et al., 1996, 1997), written as

$$\rho_{\rm NFW} = \frac{4\,\rho_{-2}}{r/r_{-2}\,(r/r_{-2}+1)^2}, \tag{2.1}$$

where r_{-2} is the characteristic radius at which the profile's logarithmic slope is equal to -2, and $\rho_{-2} = \rho_{\rm NFW}(r_{-2})$ is the corresponding density. These two parameters fully specify the NFW profile and therefore completely describe the structure of dark matter haloes. It is however common practice to recast these parameters in terms of the halo's virial mass[1], $M_{200,m}$, and concentration, $c = r_{200,m}/r_{-2}$, an approach we follow in this paper; in what follows we refer to these quantities simply as M_{200} and r_{200}.

As simulations grew in volume and simultaneously achieved higher mass and spatial resolution, it became clear that simulated halo profiles exhibit slight but systematic departures from the NFW shape. As discussed in Navarro et al. (2004, see also Gao et al. 2008; Ludlow et al. 2011; Dutton and Macciò 2014; Child et al. 2018), simulated halo profiles are better described by the Einasto (1965) profile, which can be written

$$\rho_{\rm E} = \rho_{-2} \exp\left\{-\frac{2}{\alpha}\left[\left(\frac{r}{r_{-2}}\right)^{\alpha} - 1\right]\right\} \tag{2.2}$$

where r_{-2} and ρ_{-2} have the same meaning as in Eq. (2.1), and α is a shape parameter that can be tailored to better-fit individual haloes. For $\alpha \approx 0.18$, Eq. (2.2) resembles the NFW profile over a wide range of scales.

Neglecting the slight deviations between simulated halo density profiles and the NFW profile, the values of c and M_{200} are sufficient to determine their structure. This led to numerous studies of the relationship between halo mass and concentration, and how it changes as a function of redshift and cosmology (the so-called concentration-mass-redshift relation, often denoted $c(M, z)$). These studies paint a clear picture of the structure of CDM haloes: at fixed redshift, their concentrations, on average, decrease with increasing mass, and at fixed mass, on average, decrease with increasing redshift (e.g. Bullock et al., 2001; Dolag et al., 2004; Prada et al., 2012; Ludlow et al., 2012, 2013; Bhattacharya et al., 2013; Kwan et al., 2013; Ludlow et al., 2014; Correa et al., 2015; Ludlow et al., 2016; Diemer and Joyce, 2019; Brown et al., 2020; Ragagnin et al., 2021). Although the exact physical mechanism that sets the concentration of a halo is not known, numerous studies have convincingly demonstrated that it is closely connected to its assembly history (Navarro et al., 1996, 1997; Bullock et al., 2001; Wechsler et al., 2002; Zhao et al., 2003; Dolag et al., 2004; Ludlow et al., 2014, 2016; Diemer and Joyce, 2019).

A number of studies have also addressed the mass and redshift dependence of α, the Einasto shape parameter in Eq. (2.2). For example, Gao et al. (2008, see also Dutton and

[1]We define the virial mass $M_{200,\rm m}$ of a DM halo as the total mass enclosed by a sphere of radius $r_{200,\rm m}$, centered on the halo particle with the minimum potential energy, that encloses a mean density of $200 \times \rho_{\rm m}$, where $\rho_{\rm m} = \Omega_{\rm m}\rho_{\rm c}$ is the mean matter density and $\rho_{\rm c} = 3\,H^2/8\pi G$ is the critical density of the universe.

Macciò 2014; Child et al. 2018) demonstrated that the average value of α increases with both halo mass and redshift in a manner that can be neatly described by a single relation between α and peak height[2], $\nu(M, z)$. Ludlow et al. (2013, see also Ludlow and Angulo 2017) showed that, like the halo concentration, α is intimately linked to the assembly histories of dark matter haloes.

Given the approximate self-similarity of halo structure, the ability to accurately predict halo concentrations has numerous applications, including estimating merger rates of primordial black holes (e.g. Mandic et al., 2016), predicting the lensing signal associated with haloes (e.g. Bartelmann et al., 2002; Fedeli et al., 2007; Mandelbaum et al., 2008; Amorisco et al., 2021) and their substructure (e.g. Despali et al., 2018), and to estimate the gamma ray signal potentially produced by dark matter annihilation (e.g. Sánchez-Conde and Prada, 2014; Okoli et al., 2018). Another potential application – indeed, the one that motivated this work – is to improve the performance of cosmological rescaling algorithms (Angulo and White, 2010; Contreras et al., 2020) that can be used to transmute a template N-body simulation carried out with a set of cosmological parameters to a synthetic simulation consistent with another cosmology. Whether existing models can appropriately account for the cosmology dependence of halo concentrations has not been rigorously tested.

The aim of this work is therefore to study the dependence of the $c(M, z)$ relation on cosmology, and to test the extent to which it can be reproduced by published models for predicting halo concentrations. To do so, we ran a large suite of gravity-only simulations in which the cosmological parameters were systematically varied with respect to the best-fit Planck Collaboration et al. (2020b) results. In Section 2 we present our suite of cosmological simulations along with their associated halo and merger tree catalogs (§§2), explain our approach to discarding unrelaxed haloes (§§2), and outline how we measure halo concentrations (§§2). In Section 2 we present the $c(M, z)$ relations obtained for different cosmologies (§§2) and study the relation between the internal structure of haloes and their formation histories. In §§2 we compare the performance of different published models for predicting the mass- and redshift-dependence of halo concentration, focusing on their ability to reproduce the cosmology-dependence of the $c(M, z)$ relation. In Section 2 we discuss how accurate predictions for halo concentration can lead to improved accuracy when applied to a cosmological scaling algorithm. In Section 2 we provide a few concluding remarks.

[2]The peak height, a dimensionless mass parameter, is defined as $\nu(M, z) = \delta_c/\sigma(M, z)$, where $\sigma(M, z)$ is the variance of the matter density perturbations linearly extrapolated to redshift z, and δ_c is the critical density for gravitational collapse, usually estimated from the spherical collapse model for which $\delta_c \approx 1.686$ (e.g. Peebles, 1980).

2.2 Numerical simulations and analysis

Below we describe the pertinent details of the numerical simulation used in this work, and discuss our analysis algorithms and techniques.

2.2.1 Numerical simulations

Our results are inferred from a suite of DM-only simulations in which we modify the values of different cosmological parameters. We define these parameters below.

1. σ_8: The root mean square of matter density perturbations averaged in spheres of radius $R = 8h^{-1}$Mpc and linearly extrapolated to $z = 0$.

2. $\Omega_{\rm m}$: The dimensionless matter density parameter, $\Omega_{\rm m} \equiv \rho_{\rm m}/\rho_{\rm c} = 8\pi G\rho_{\rm m}/3H^2$, which is the ratio of the total matter density, $\rho_{\rm m}$, and the critical density, $\rho_{\rm c}$. Note that $\Omega_{\rm m}$ includes contributions from both DM and baryons, i.e. $\Omega_{\rm m} = \Omega_{\rm cdm} + \Omega_{\rm b}$. For runs in which $\Omega_{\rm m}$ is varied, we only modify the value of $\Omega_{\rm cdm}$ (keeping $\Omega_{\rm b}$ fixed) and adjust the value of $\Omega_{\rm DE}$ (the cosmic dark energy density) to maintain a flat cosmology. Note that neutrinos do not contribute to this definition of $\Omega_{\rm m}$.

3. $n_{\rm s}$: The scalar spectral index of the primordial density fluctuation power spectrum, $P(k) \propto k^{n_{\rm s}-1}$.

4. w_0 and $w_{\rm a}$: The dynamical dark energy parameters used in the Chevallier-Polarski-Linder (CPL) parameterization (Chevallier and Polarski, 2001; Linder, 2003). When $w_0 = -1$ and $w_{\rm a} = 0$ the dark energy contribution to the background expansion is consistent with a cosmological constant, see Eq. (2.3).

5. M_ν: The sum of the individual masses for the three neutrino species, which is related to the neutrino density parameter by $\Omega_\nu = M_\nu/[(93.14\text{eV})h^2]$ (with M_ν expressed in eV)[3]. When we increase the value of Ω_ν we reduce the value of $\Omega_{\rm cdm}$ by the same

[3]The first Friedmann equation can be written in terms of the neutrino density parameter, Ω_ν, as (Zennaro et al., 2017):

$$H^2(a) = H_0^2 \left[\left(\Omega_{\rm cdm,0} + \Omega_{\rm b,0} \right) a^{-3} + \Omega_\nu(a) E^2(a) + \right. \\ \left. + \Omega_{\rm DE,0} a^{-3(1+w_0+w_{\rm a})} e^{3aw_a} \right] \left(\right. \tag{2.3}$$

where we have also included the Chevallier-Polarski-Linder (CPL) parameterization (Chevallier and Polarski, 2001; Linder, 2003) of dynamical dark energy component whose equation of state is $w(z) = w_0 + w_a z/(1+z)$ (Linden and Virey, 2008).

amount in order to maintain a flat cosmology. When we vary M_ν in our simulations we kept fixed the value of the power spectrum initial amplitude, A_s, therefore varying M_ν will result in different values for σ_8 at $z = 0$.

6. h: The dimensionless Hubble-Lemaitre parameter, which sets the value of the Hubble-Lemaitre constant, i.e. $H_0 = 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ at $z = 0$.

7. Ω_b: The baryon density parameter, $\Omega_\mathrm{b} \equiv \rho_\mathrm{b}/\rho_\mathrm{c}$. Changes to Ω_b are compensated by changing Ω_cdm such that Ω_m remains constant.

Our suite of simulations is designed around four reference runs, which we refer to as *Nenya*, *Narya*, *Vilya* and *The One*. All reference simulations share a number parameters – specifically, $\sigma_8 = 0.9$, $M_\nu = 0.0$ eV, $w_0 = -1.0$, $w_a = 0.0$, and $L_\mathrm{box} = 512\,h^{-1}$Mpc are the same for all of them – but other parameters are varied as described in Table 2.1. Along with these reference runs, we carried out 32 additional simulations divided in 8 groups (with 4 simulations in each group) according to the cosmological parameter that was varied. For the runs in a given group we uniformly vary a particular cosmological parameter so that it spans a 5σ or 10σ region (depending on the parameter) around the best-fit parameter values provided by Planck Collaboration et al. (2020b). For the case of the Hubble-Lemaitre parameter, h, we explore values that span a $4\,\sigma$ region around the best-fit value obtained from low-redshift supernovae data Riess et al. (2016).

The selection of these cosmologies was motivated by the criteria set forth in Contreras et al. (2020), and we have generated them, in part, to serve as a follow-up of the runs presented in that work (some of our simulations have, in fact, already been used in other studies, e.g. Contreras et al. 2021; Zennaro et al. 2021; Pellejero-Ibanez et al. 2022). The objective of Contreras et al. (2020) was to test the performance of cosmology-rescaling algorithms (which we explain in more detail in Section 2). We therefore designed our simulation suite

Table 2.1: Our four "reference" simulations (*Nenya*, *Narya*, *Vilya* and *The One*) share the following cosmological parameters: $\sigma_8 = 0.9$, $M_\nu = 0$, $w_0 = -1$ and $w_\mathrm{a} = 0$. The parameters listed below have been varied.

Name	Ω_m	n_s	h	Ω_b	$m_\mathrm{DM}\,[h^{-1}\mathrm{M}_\odot]$
Nenya	0.315	1.01	0.60	0.050	$10^{9.51}$
Narya	0.360	1.01	0.70	0.050	$10^{9.57}$
Vilya	0.270	0.92	0.65	0.060	$10^{9.44}$
The One	0.307	0.96	0.68	0.048	$10^{9.5}$

Table 2.2: The values of the cosmological parameters that are modified for each simulation. All runs have the same cosmological parameters as those used for one of the four reference simulations listed in Table 2.1 but with one parameter modified to match the values listed below. For example, the run referred to in the upper-left entry adopts cosmological parameters consistent with the *Nenya* simulation, but with a lower value of the rms density fluctuation amplitude, i.e. $\sigma_8 = 0.730$.

Ref - σ_8	Ref - $\Omega_{\mathbf{m}}$	Ref - $n_{\mathbf{s}}$	Ref - w_0
Nenya 0.730	*Nenya* 0.23	*The One* 0.920	*Nenya* -0.70
The One 0.770	*Nenya* 0.27	*The One* 0.940	*Nenya* -0.85
Nenya 0.815	*Narya* 0.36	*The One* 0.965	*Nenya* -1.15
Nenya 0.860	*Narya* 0.40	*Narya* 0.990	*Nenya* -1.30

Ref - $w_{\mathbf{a}}$	Ref - M_ν	Ref - h	Ref - Ω_b
Nenya -0.30	*Nenya* 0.1 eV	*Nenya* 0.65	*Nenya* 0.040
Nenya -0.15	*Nenya* 0.2 eV	*Narya* 0.70	*Nenya* 0.045
Nenya 0.15	*Nenya* 0.3 eV	*Narya* 0.75	*Nenya* 0.055
Nenya 0.30	*Nenya* 0.4 eV	*Narya* 0.80	*Nenya* 0.060

in such a way that each run can be compared to a rescaled simulation obtained from one of our four reference runs. This is why we modified only one cosmological parameter per simulation, while keeping all others fixed with respect to the values used for one of the reference simulations. The various runs are listed in Table 2.2, where the column headers indicate the cosmological parameter that was modified, and the prefix indicates the reference model.

All simulations were carried out using a lean version of `L-Gadget3` (see Springel et al., 2008; Angulo et al., 2012) and evolved the DM density field using $\mathrm{N_{DM}} = 1536^3$ equal-mass DM particles; they all employed the same softening length: $\epsilon = 5\,h^{-1}\mathrm{kpc}$. All simulation volumes are approximately $V_{\mathrm{box}} \approx (512\,h^{-1}\mathrm{Mpc})^3$, but vary slightly from run to run[4]. The slight variation in box size, along with changes to Ω_{m}, result in small differences in the DM particle masses between simulations. Our lowest-mass resolution run has $m_{\mathrm{DM}} = 10^{10.01}\,h^{-1}\mathrm{M}_\odot$ (*Extreme high-ns*), and our highest mass-resolution run has $m_{\mathrm{DM}} = 10^{9.41}\,h^{-1}\mathrm{M}_\odot$ (*Extreme low-h*); the particle masses of all other simulations falls within this range. We use a version of NgenIC (Springel, 2015) that employs second-order Lagrangian Perturbation Theory (2LPT) to generate the initial conditions at $z = 49$ for each simulation.

For simulations including massive neutrinos, we created initial conditions according to the cold matter power spectrum obtained using the scale dependent backscaling technique

[4]Slight differences in the box size between the various runs ensures that variations of our reference models, when the cosmology-rescaling algorithm is employed to match the corresponding reference cosmology, will have a volume of exactly $V_{\mathrm{box}} = (512\,h^{-1}\mathrm{Mpc})^3$. Selecting the simulation volumes this way simplifies the comparison between the N-body simulations and the results obtained from the cosmology-rescaling algorithm (see Contreras et al., 2020)

described in Zennaro et al. (2017). We then evolved these simulations using a version of `L-Gadget3` that incorporates the neutrino implementation of Ali-Haïmoud and Bird (2013), where neutrino perturbations were solved on a grid, employing a linear response function that is sensitive to the non-linearities developed in the cold matter distribution.

To reduce cosmic variance, we followed the approach of Angulo and Pontzen (2016a) and carried out paired-phase counterparts of each of our simulations, which doubles the total number of simulations used in our analysis. We differentiate the two simulations within each of the fixed-paired doublets with the suffixes "- 0" and "- π". For more information regarding the fixing and pairing technique and how it reduces cosmic variance in cosmological simulations see Angulo and Pontzen (2016a), Chuang et al. (2019), Knebe et al. (2021) and Maion et al. (2022).

We identify haloes and subhaloes in our simulations using a Friends-of-Friends algorithm (Davis et al., 1985), with linking length $b = 0.2$, and a modified version of `SUBFIND` (Springel et al., 2001a). As discussed in Contreras et al. (2020), our implementation of `SUBFIND` is able to robustly identify substructure haloes by considering their prior evolution.

We construct merger trees by linking haloes and subhaloes between consecutive snapshots, starting from the first snapshot in which a particular halo is identified. We then progress through subsequent snapshots and determine which halo or subhalo is its most likely descendant. To do so, we track its 15 most-bound particles between snapshots and identify all (sub)haloes in which these particles end up; these constitute a set of possible descendants. We identify the most likely "true" descendant by considering which (sub)halo candidate has the highest score based on the number of particles it inherits weighted by their rank ordered binding energy with respect to the original (sub)halo. This approach constitutes a slight modification to the method used by Angulo et al. (2012) where only the inherited number of (most-bound) particles is considered but not their binding energies.

2.2.2 Halo dynamical state and relaxedness

In this work we analyze the $c(M, z)$ relations of "relaxed" DM haloes. We discard unrelaxed haloes from our analysis because their density distribution is likely to deviate from spherical symmetry, and as such be ill fit by simple analytic profiles such as NFW or Einasto. Considering only relaxed haloes biases the median concentrations to higher values in mass bins where a large number of haloes are expected to be out of equilibrium, particularly high-mass bins (unrelaxed haloes typically have larger values of r_{-2} than relaxed ones of the same mass). However, excluding unrelaxed haloes is crucial for our analysis because it

allows us to quantify the connection between the inner structure of haloes and their formation histories, and eliminates the possibility of dynamical processes such as mergers biasing our results).

Following Ludlow et al. (2012), we consider a halo unrelaxed if its half-mass formation lookback time (since identification), i.e. $t_{\rm h} = t_{\rm lb}(z_{\rm h}) - t_{\rm lb}(z_0)$, is less than a crossing time, $t_{\rm cross} = 2\, r_{200}/V_{200}$. Following Neto et al. (2007), we also discard haloes for which the distance between their center of mass and the position of the gravitational potential minimum is greater than $0.07\, r_{200}$ as well as those whose substructure mass fraction (i.e. the mass contained in subhaloes within r_{200} of the host halo) exceeds $0.1\, M_{200}$.

2.2.3 Analysis of halo density profiles

Much of our analysis focuses on the median mass-concentration-redshift relations obtained from the best-fit density profiles of well-resolved haloes in our simulations, which we initially compute using logarithmically spaced mass bins of width $\Delta \log M_{200} = 0.1$ that span the range $M_{200} \in (10^{13}, 10^{15.2})\ h^{-1}{\rm M}_\odot$. Following previous works (e.g., Gao et al., 2008; Dutton and Macciò, 2014; Child et al., 2018; Ludlow et al., 2019; Brown et al., 2020), we then discard bins corresponding to haloes with fewer 5000 particles within their virial radius, r_{200}, as well as those containing fewer than 50 haloes, the latter to avoid excessive noise in the relations.

To compute the concentrations of haloes we fit each of their spherically-averaged density profiles to Einasto's formula, i.e. Eq. (2.2), but fix the value of α according to the $\alpha - \nu$ relation obtained by Gao et al. (2008), i.e.

$$\alpha = 0.155 + 0.0095\, \nu\, (M, z)^2. \tag{2.4}$$

When fitting the density profiles we discard radial bins that are below the resolution limit, $r_{\rm min}$. We follow Power et al. (2003) and define $r_{\rm min}$ as the radius at which relaxation time is equal to the circular orbital time at the virial radius, i.e. $t_{\rm relax}(r_{\rm min}) = t_{\rm circ}(r_{200})$ (see also Zhang et al., 2019; Ludlow et al., 2019). This yields the following condition:

$$\frac{t_{\rm relax}\ (r_{\rm min}\,)}{t_{\rm circ}\ (r_{200})} = \frac{\sqrt{200}}{8}\frac{N\,(r_{\rm min}\,)}{\ln N\,(r_{\rm min}\,)}\left[\frac{\rho_{\rm c}(z_0)}{\rho_{\rm enc}\,(r_{\rm min}\,)}\right]^{\frac{1}{2}} = 1, \tag{2.5}$$

where $\rho_{\rm c}(z_0)$ is the critical density of the universe at the halo identification redshift z_0, and $N(r_{\rm min}\,)$ and $\rho_{\rm enc}(r_{\rm min}\,)$ are the enclosed number of particles and enclosed density at $r_{\rm min}$, respectively. We also discard radial bins for which $r > r_{\rm max} = 0.8 r_{200}$, where density profiles can be sensitive to local departures from equilibrium (see, e.g. Ludlow et al., 2020). When carrying out our fits, we restrict the best-fit value of r_{-2} to the range $r_{\rm min} \leq r_{-2} \leq r_{\rm max}$.

Although we have excluded unrelaxed haloes from our analysis, we nonetheless encounter a great diversity in profile shapes, and for a number of them the best-fit value of r_{-2} is equal to $r_{\rm min}$ or $r_{\rm max}$. For these cases, the true value of r_{-2} is likely outside the resolved radial range and our estimate of r_{-2} therefore represents a lower or upper limit. We surmount this problem by discarding all mass bins in which more than 30 per cent of haloes have either $r_{-2} = r_{\rm min}$ or $r_{-2} = r_{\rm max}$, which ensures that such poorly-fit systems do not bias the median concentrations used in our analysis. We have employed a simulation with higher resolution (more than 3 times the number of particles and 50 per cent smaller force softening) to verify that this procedure yields robust values for the median concentrations.

In Fig. 2.1 we show the median $z_0 = 0$ density profiles (weighted by a factor of r^2) for haloes of different virial mass in the *The One* $- \pi$ simulation. Halo masses are logarithmically-spaced and span the range $M_{200} \in (10^{13}, 10^{15.2})$ $h^{-1}{\rm M}_{\odot}$. The filled circles correspond to radial bins with $r_{\rm min} \leq r \leq r_{\rm max}$. By plotting $\log_{10}(\rho r^2)$, the value of r_{-2} is readily apparent as the radius of the "peak" of each best-fit profile. In addition to the median density profiles, we present their best Einasto fits (with α computed using Eq. (2.4); solid lines).

2.3 Results

2.3.1 Cosmology dependence of the mass-concentration-redshift relation

In Fig. 2.2 we plot using connected circles the $c(M)$ relations obtained from our suite simulations at $z_0 = 0$. The results are split into different panels according to the cosmological parameter that was varied. With a total of 72 simulations, Fig. 2.2 represents, to our knowledge, the most extensive analysis to date of the cosmology-dependence of the mass-concentration relation. For each cosmology, we plot the average concentration of haloes in each mass bin after combining the fixed amplitude and inverted-phase simulations (all the results presented henceforth correspond to averages of our fixed-amplitude and inverted-phase simulations). For completeness, in Appendix E we present the concentration-mass relations at $z_0 = 0.5$.

In agreement with previous findings, Fig. 2.2 shows that the concentrations of relaxed DM haloes decreases as a function of halo mass for all cosmological models studied. This is consistent with interpretation that structure forms hierarchically, i.e. low-mass haloes typically form before more massive ones, and that the concentrations of haloes are correlated with their formation times.

The results also illustrate how varying different cosmological parameters affects the

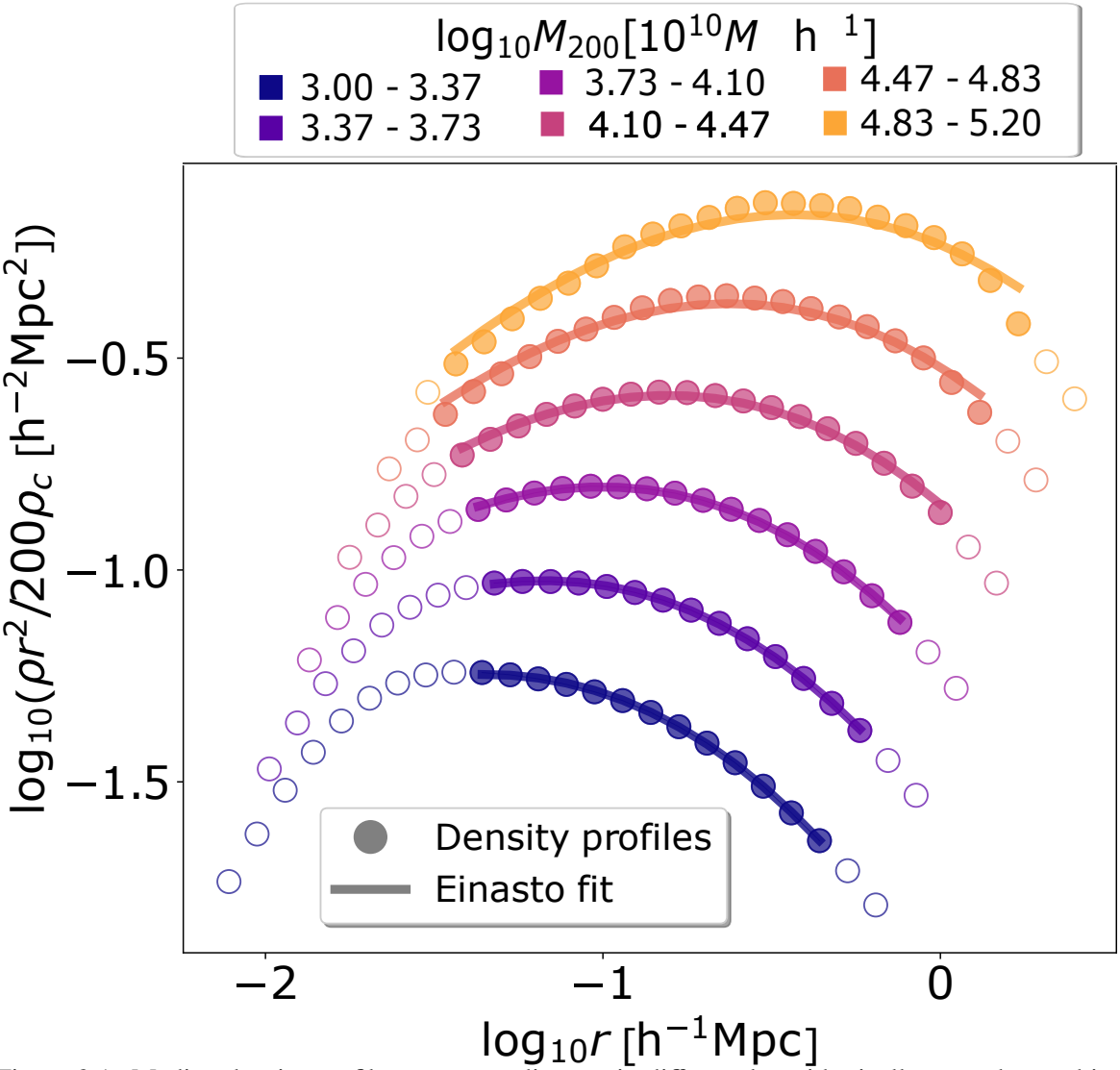

Figure 2.1: Median density profiles corresponding to six different logarithmically-spaced mass bins spanning the range $M_{200} \in (10^{13}, 10^{15.2})\ h^{-1}\mathrm{M}_{\odot}$. All haloes were identified in the *The One* $-\ \pi$ simulation at $z_0 = 0$ (circles). The filled circles correspond to the "resolved" radii used when carrying out our fits, i.e. they correspond to radial bins satisfying $r_{\mathrm{min}} \leq r \leq r_{\mathrm{max}}$ (see subsection 2 for details). The thick solid lines show to the best-fit Einasto profiles with the values for α computed using Eq. (2.4). Different colors distinguish the different median virial masses, M_{200}, which are indicated in the legend in units of $\log_{10} M_{200}[10^{10} M_{\odot} \mathrm{h}^{-1}]$.

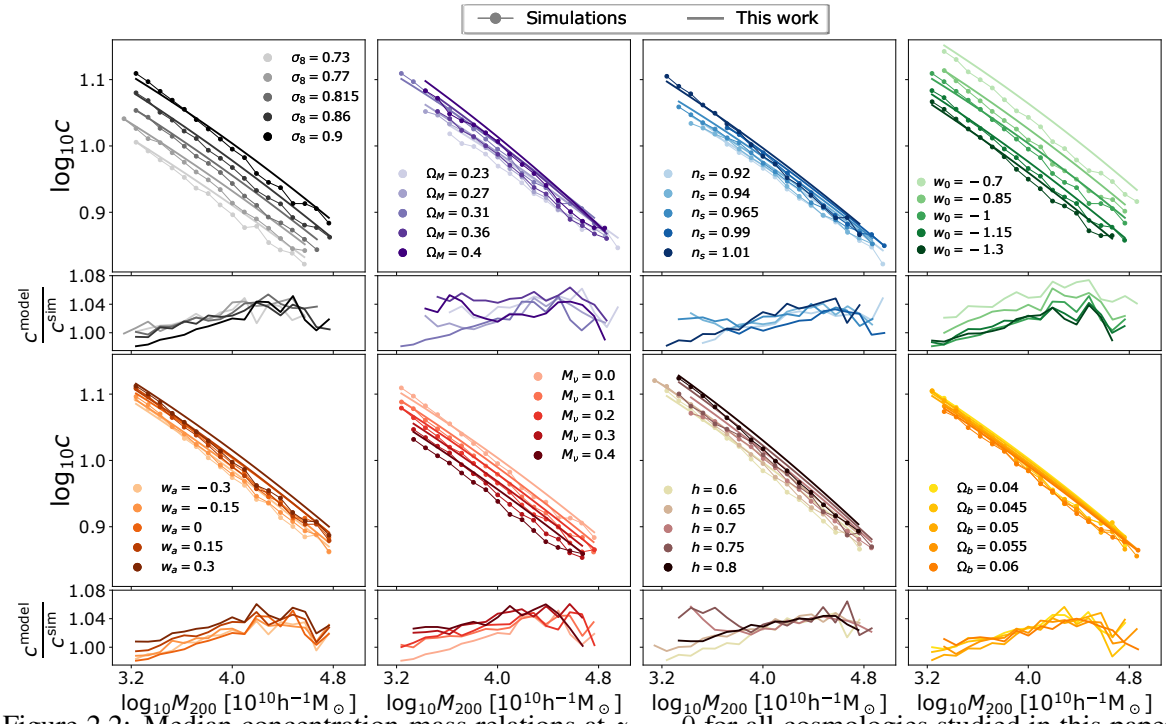

Figure 2.2: Median concentration-mass relations at $z_0 = 0$ for all cosmologies studied in this paper (see Table 2.2). Simulation results are shown as connected colored circles; the solid lines plotted in the panels of the first and third rows show the relations that are predicted by the model presented in this work, a modified version of the L16 model (using $A = 493$; see Section 2 for details). From top-to-bottom and left-to-right, the cosmological parameters varied are, σ_8, $\Omega_{\rm m}$, $n_{\rm s}$, w_0, $w_{\rm a}$, M_ν, h and $\Omega_{\rm b}$. The simulation results correspond to the average of the median concentrations obtained for phase-0 and phase-π simulations. The solid lines plotted in the smaller "residual" panels (second and fourth rows) correspond to the ratio of the concentrations predicted by the L16 model and the concentrations measured in our simulations.

concentration-mass relation. For example, regardless of halo mass, increasing the value of σ_8 leads to higher concentration. This is because higher σ_8 implies higher linear fluctuation amplitudes at fixed mass, and so earlier average formation times.

Higher values of w_0 also increase concentrations at all masses. This is because w_0 alters the growth histories of haloes through the dark energy term in Eq. (2.3). Specifically, higher w_0 leads to earlier halo formation times since (for a fixed value of σ_8) the increased contribution of dark energy to the universal expansion history demands that the haloes of a given mass form earlier, which in turn increases their concentration.

As a final example, consider the impact of $\Omega_{\rm b}$. For the runs plotted in the lower-right panel of Fig. 2.2, $\Omega_{\rm b}$ contributes at least 4 per cent and at most 6 per cent of the critical density of the universe. Such a small contribution from baryons implies that the matter component in all our runs is dominated by cold dark matter. As such, the formation histories–and as a consequence, the concentrations–of haloes are largely insensitive to $\Omega_{\rm b}$, at least over the range of values studied here.

2.3.2 The relationship between the characteristic densities of haloes and their formation histories

As pointed out in Section 2, there are a number models that aim to accurately predict the $c(M, z)$ relation, as well as its dependence on cosmological parameters. Many are based on empirical fits to results obtained from large suites of simulations (e.g. Dutton and Macciò, 2014; Diemer and Joyce, 2019), while others are based on physical models that relate the concentrations of haloes to their collapse histories (e.g. Navarro et al., 1996, 1997; Bullock et al., 2001; Wechsler et al., 2002; Gao et al., 2008; Ludlow et al., 2013, 2014; Correa et al., 2015; Ludlow et al., 2016). One model in particular, that of Ludlow et al. (2016, L16 hereafter), has been shown to reproduce the mass-concentration relation for a variety of cosmological models, including cold and warm dark matter models that adopt sharply truncated power spectra (Ludlow et al., 2016; Wang et al., 2020; Richardson et al., 2022). The L16 model is based on the assumption (see appendix D for more details) that the enclosed density within a halo scale radius, $\langle \rho_{-2} \rangle \equiv \langle \rho(r_{-2}) \rangle$, is directly proportional to the critical density of the universe at the time when its characteristic mass, i.e. $M_{-2} \equiv M(< r_{-2})$, had first assembled into progenitors more massive than $0.02 \times M_0$, where M_0 is the present-day mass of the halo. The redshift evolution of the mass fraction collapsed in such progenitors (i.e. those with masses exceeding $0.02 \times M_0$) defines the halo's "collapsed mass history" (hereafter CMH for short). Below we test whether this result also holds for the various cosmologies explored in our simulation suite.

To do so, we use the simulated profiles to determine the mass M_{-2} enclosed by the best-fit

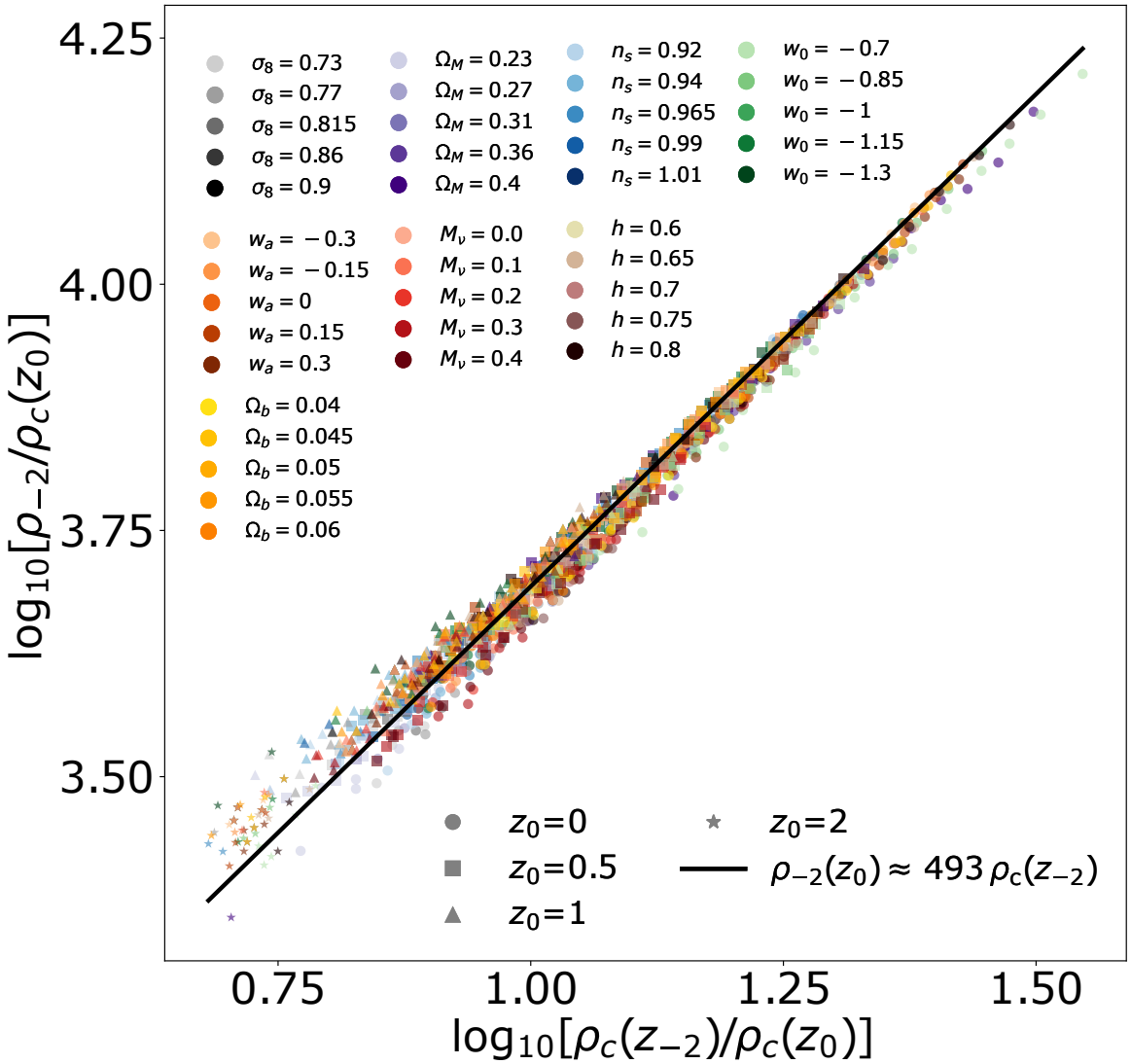

Figure 2.3: Relation between the median values of $\langle\rho_{-2}\rangle$ and $\rho_c(z_{-2})$ computed for relaxed haloes identified $z_0 = 0$, 0.5, 1 and 2 in our suite of simulations (the points correspond to median values obtained for equally-spaced logarithmic mass bins). The color of the points indicate the cosmological model and match the colors used for Fig. 2.2. The shapes of the points indicate the redshift: $z_0 = 0$ as circles, $z_0 = 0.5$ as squares, $z_0 = 1$ as triangles, and $z_0 = 2$ as stars. The solid black line is the best fit to all the points: $\langle\rho_{-2}\rangle \approx 493\,\rho_c(z_{-2})$.

scale radius r_{-2}, and then define $\langle\rho_{-2}\rangle = 3M_{-2}/4\pi r_{-2}^3$. The formation time, z_{-2}, is defined as the redshift at which the halo's CMH first exceeds M_{-2} , which is obtained by interpolating along the CMH that we calculated using each halo's merger tree.

In Fig. 2.3 we plot the relation between $\langle\rho_{-2}\rangle$ and $\rho_c(z_{-2})$ for all the simulations described in Section 2, and for redshifts $z_0 = 0, 0.5, 1, 2$ (distinguished using different symbols). Each point corresponds to the average $\langle\rho_{-2}\rangle$ and $\rho_c(z_{-2})$ calculated for the same mass bins used to construct Fig. 2.2. Fig. 2.3 reveals an approximate power-law relation between $\langle\rho_{-2}\rangle$ and $\rho_c(z_{-2})$ that is largely independent of cosmology, halo mass and redshift. Note too that the relation plotted has a "natural" slope very close to 1, i.e. $\langle\rho_{-2}\rangle \propto \rho_c(z_{-2})$. The solid line shows the best-fit relation: $\langle\rho_{-2}\rangle \approx 493\,\rho_{\rm c}(z_{-2})$.

The existence of a tight relation between $\langle\rho_{-2}\rangle$ and $\rho_c(z_{-2})$ suggests that the concentrations of haloes – regardless of mass, redshift, or cosmology – can be predicted if an accurate model for the CMHs of haloes can be found. We investigate this next.

2.3.3 Predicted formation times based on the extended Press-Schechter formalism

In Fig. 2.4 we show the median CMHs of haloes of different mass identified at $z_0 = 0$ in the *The One* $-\,\pi$ simulation (solid lines; note that these are the same mass bins used to construct the density profiles plotted in Fig. 2.1). The outsized squares indicate the average halo formation times, z_{-2}, for the different mass bins. The dashed curves show, for comparison, the CMHs predicted by the extended Press-Schechter (EPS) formalism (Bond et al., 1991b; Lacey and Cole, 1993) for haloes of the same present day mass, see Eq. (D.2).[5] The open triangles show the values of z_{-2} associated with these EPS-collapsed mass histories (the latter referred to henceforth as EPS-CMHs). Note that the measured and predicted formation times agree quite well, as do the overall shapes of the CMHs.

In Fig. 2.5 we test how accurately the EPS model describes the formation times of haloes in our simulations (after a suitable modification to account for the impact of massive neutrinos in EPS, see Appendix D). Here we plot the relative difference between the EPS-predicted formation redshifts, expressed as $\rho_{\rm c}^{\rm EPS}(z_{-2})$, and the formation redshifts measured directly from the simulated CMHs, i.e. $\rho_{\rm c}^{\rm CMH}(z_{-2})$. Each point corresponds to the median values of these quantities in bins of halo mass, and are plotted at different redshifts, which increase from the top to bottom panels. To help with visualization, we have applied a small horizontal shift to the values in each mass bin so that results obtained for different cosmological parameters can be easily distinguished. Our results show that the EPS-predicted formation redshifts

[5] In order to predict the CMHs for DM haloes using the EPS formalism, we adopt a critical density for gravitational collapse of $\delta_{\rm sc} = 1.46$. This minimizes the typical difference between the predicted halo collapse redshifts and those measured in our simulations (see Fig. 2.5).

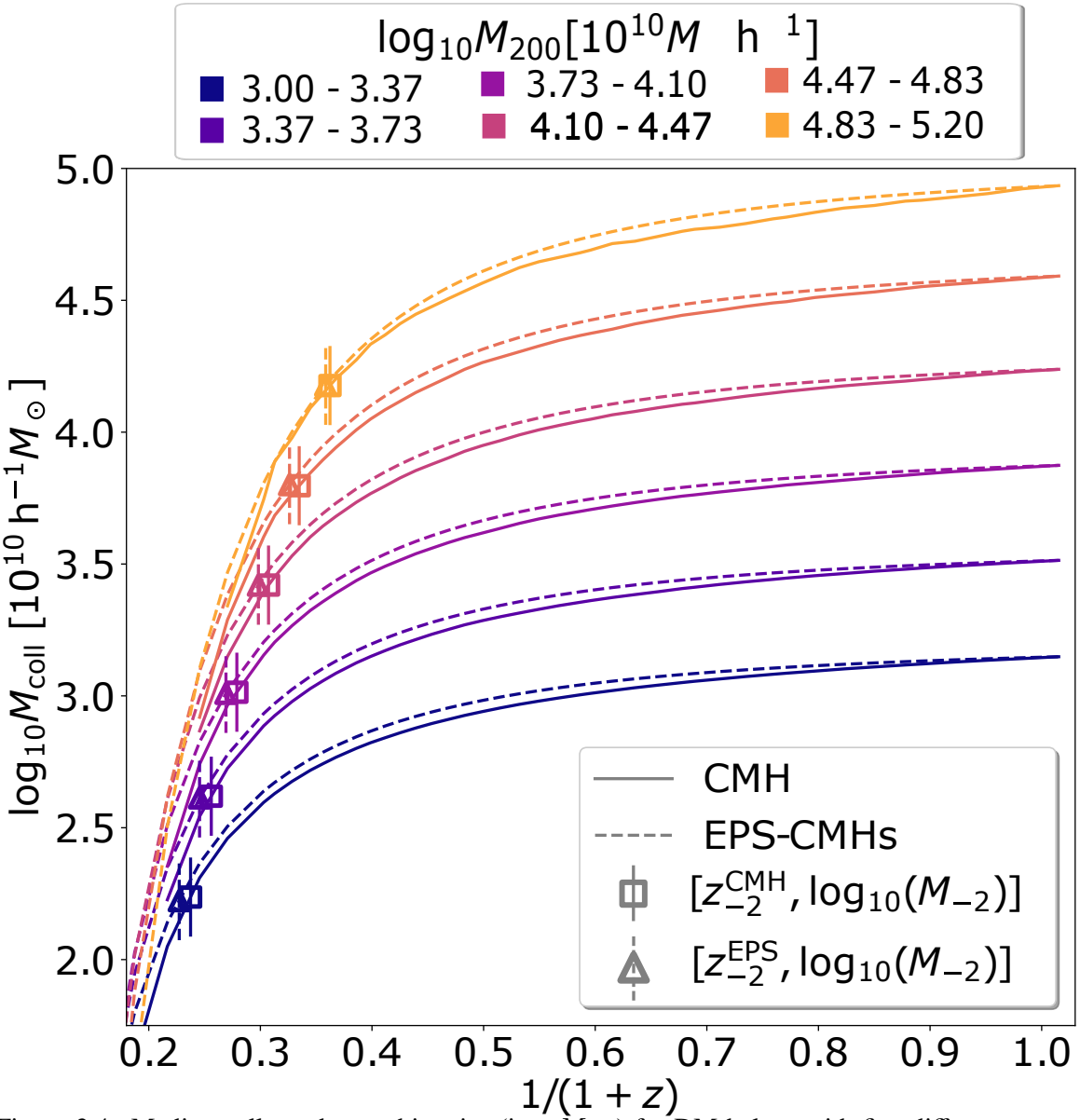

Figure 2.4: Median collapsed mass histories (i.e. $M_{\rm coll}$) for DM haloes with five different masses identified in the *The One* simulation at redshift $z_0 = 0$ (solid lines; note that these are the same haloes whose density profiles are plotted in Fig. 2.1). Results are plotted as a function of scale factor, a. The formation time (defined as the point at which $M_{\rm coll} = M_{-2}$) for each halo is marked with a square crossed by a solid vertical segment (used for clarity) on top of the CMH. The dashed colored lines correspond to the CMHs predicted by the extended Press-Schechter theory (EPS-CMHs) and have been computed using Eq. (D.2) with $\delta_{sc} = 1.46$ and $f = 0.02$. The open triangles crossed by dashed vertical segments indicate the formation times obtained from the EPS-CMHs.

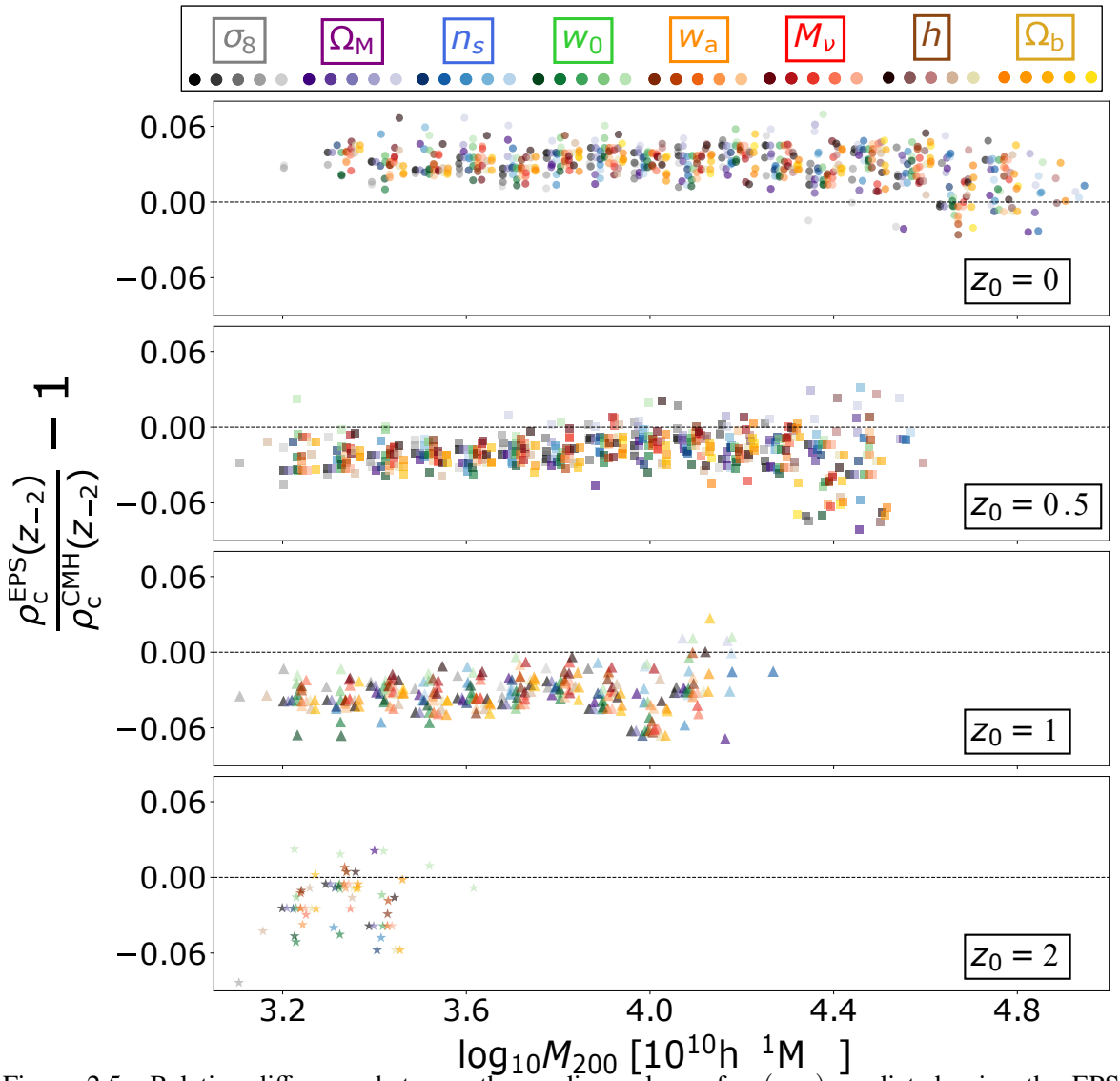

Figure 2.5: Relative difference between the median values of $\rho_c(z_{-2})$ predicted using the EPS formalism, and measured from the simulated CMHs, plotted as a function of M_{200}. The results are split into four panels corresponding to redshifts $z_0 = 0, 0.5, 1$, and 2. We present the results for all available cosmologies in our suite of simulations and color-code the points accordingly to the simulation they belong to following the color scheme adopted for Fig. 2.2. The values associated with different cosmologies are systematically shifted with respect to the mass-bin-centers for better visualization.

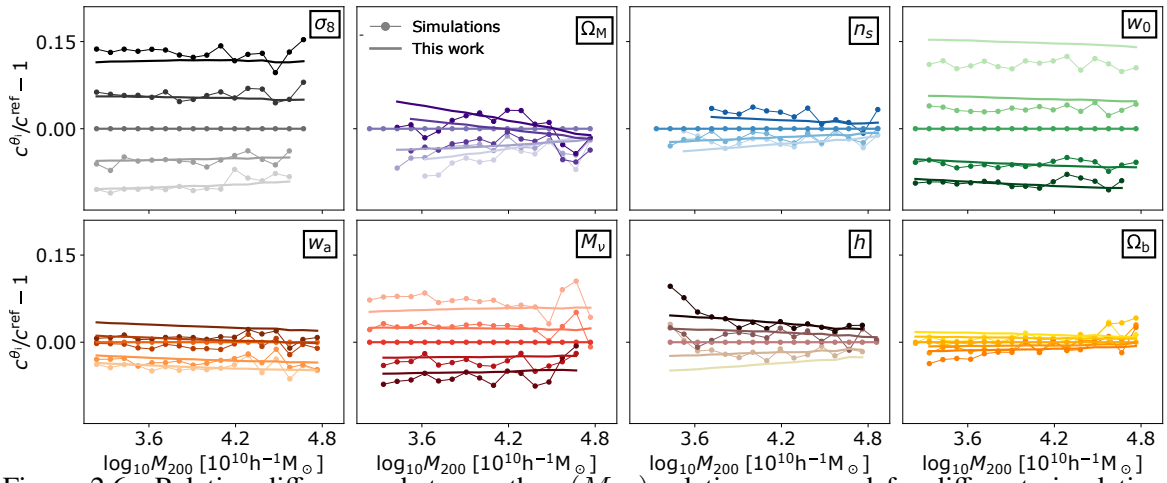

Figure 2.6: Relative differences between the $c(M, z)$ relation measured for different simulations (connected circles). The simulation taken as reference to compute the relative differences in each panel is the one with the intermediate value of the cosmological parameter that is varied. The plotting conventions match those used for Fig. 2.2. The solid lines correspond to the relative differences between the predictions for the concentration computed using the re-calibrated L16 model.

agree well with the simulated ones, with residuals that show no clear systematic dependency on cosmology or on mass. But the residuals do exhibit a slight redshift dependence, but it remains below about 6 per cent for all models, mass bins and redshifts analyzed. Such small differences between the predicted and measured formation times of haloes do not significantly impact our ability to accurately model halo concentrations based on EPS CMHs, and we conclude that the CMHs of cold dark matter haloes can reliably modelled using the EPS formalism for a wide range of cosmological models.

2.3.4 Model predictions for the mass-concentration-redshift relation

We follow L16 and use the power-law relation between $\langle\rho_{-2}\rangle$ and $\rho_c(z_{-2})$ presented in Fig. 2.3, together with EPS-predicted formation times to predict the cosmology-dependence of the $c(M, z)$ relation. The results are plotted in Fig. 2.2 as solid colored lines, which agree well with the results of our simulations.

Fig. 2.6 further explores the extent to which the L16 model captures the correct cosmology- and mass-dependence of the $c(M, z)$ relation. The plot is organized to match Fig. 2.2, with each panel showing results obtained from runs that vary a particular cosmological parameter; all results are plotted at $z_0 = 0$. The various connected circles show the relative differences between the median concentrations in each simulation measured with respect to those obtained from the run that was carried out with the intermediate value of the relevant cosmological parameter. The solid lines show the predictions of the L16 model, which reproduces the cosmology-dependence of concentration-mass relation rather well.

In Fig. 2.7 we compare how well our measurements for the $c(M, z)$ relation can be

reproduced by various other published concentration models. To produce Fig. 2.7 we select the values for the concentration measured for each mass bin (considering separately the simulations in each subpanel of Fig. 2.6), then, we obtain the gradient of the concentration with respect to the cosmological parameter that is varied, $dc/d\theta$, by fitting the selected points to a straight line. We repeat the process for all mass bins. The results obtained are then normalized by dividing by the interval spanned in each subpanel by the cosmological parameter that is been varied, $\Delta\theta$ (connected circles). We repeat this operation employing the predictions for the concentration provided by the re-calibrated L16 model (solid lines), the Prada et al. (2012) model ("P12", dotted lines), the Child et al. (2018) model ("C18", dot-dot-dashed), the Diemer and Joyce (2019) model ("DJ19", dashed lines), the Ragagnin et al. (2021) model ("R21", dash-dash-dot-dot lines), and the Brown et al. (2022) model ("B22", dashed-dotted lines).

The model that best captures the dependence of concentration on cosmology is our implementation of L16. Nevertheless, it is important to point out that the comparison of our results with the predictions provided by P12, C18, DJ19, R21, and B22 is somewhat unfair since, for instance, P12 aims to predict the concentrations for all haloes (including unrelaxed ones) and the model of R21 is calibrated using a set of hydrodinamical simulations. Regardless, it is important to note that the P12, C18, DJ19, R21, and B22 models predict that halo concentrations do not depend on w_0 or w_a, whereas L16 provides reasonably accurate predictions for the concentration dependence of these parameters. These results are not unexpected. The P12 and B22 models depend only on the shape of the (smoothed) density fluctuation power spectrum, but not on the assembly histories of haloes. Their predictions are therefore insensitive to the expansion history of the universe. The models of C18 and R21 are based on empirical fits to the simulated concentration-mass-redshift relation that are also insensitive to the expansion history of the universe, and therefore cannot recover its impact halo concentrations. The DJ19 model, however, does consider the slope of the growth factor (instead of the full merger history of haloes) when predicting halo concentrations, but this is largely insensitive to w_0 and w_a, particularly at low redshifts.

2.4 Application of the L16 model to scaling algorithms

In this section we illustrate how the L16 model can be used in studies that require a theoretical model capable of producing accurate concentration predictions. We will provide, as an example, the performance of the scaling algorithms (briefly summarized in the next paragraph), where the results substantially improve when including a concentration correction.

The scaling algorithm is a method developed by Angulo and White (2010) which allows

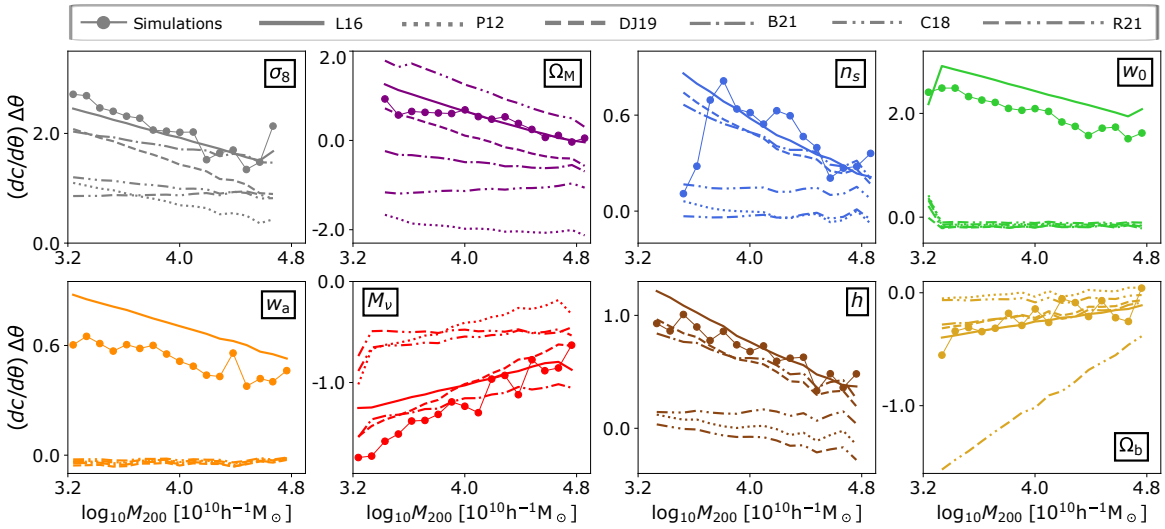

Figure 2.7: Linear dependence of the concentration on different cosmological parameters as a function of M_{200}. The results have been computed as described in subsection §§2. The connected dots correspond to the results derived from our simulations. Different lines correspond to the results derived from different models: re-calibrated L16 model (solid lines), Prada et al. 2012 (P12; dotted lines), Child et al. 2018 (C18; dot-dot-dashed lines), Diemer and Joyce 2019 (DJ19; dashed lines), Ragagnin et al. 2021 (R21; dash-dash-dot-dot lines), and Brown et al. 2022 (B22; dashed-dotted lines).

one to rapidly generate mock or synthetic cosmological simulations from a "template" N-body simulation. The mock simulation that the algorithm generates contains the DM particles of the original simulation displaced to new positions in such a way that its density field accurately reproduces that of an actual N-body simulation executed using different cosmological parameters from those of the original N-body simulation. Zennaro et al. (2019) extended the cosmology-rescaling technique to provide predictions when considering a hot component of arbitrary mass, such as neutrinos.

Contreras et al. (2020) showed that very accurate predictions for the halo clustering can be achieved by including a concentration correction on top of the standard scaling algorithm. The concentration correction modifies the position of DM particles within haloes to match halo concentrations in the target cosmology.

Fig. 2.8 illustrates how concentration corrections improve the accuracy of the power spectrum corresponding to a scaled simulation generated using the scaling algorithm. To generate this figure we employ the set of simulations presented in §2 in which we vary the total neutrino mass, M_ν, from $0.0\,\mathrm{eV}$ to $M_\nu = 0.4\,\mathrm{eV}$, keeping all other cosmological parameters (those of *Nenya*) fixed, see Table 2.2.

We apply a scaling algorithm to the N-body simulation with $M_\nu = 0\,\mathrm{eV}$ to produce mock simulations that mimic the behaviour of runs with $M_\nu = 0.1, 0.2, 0.3, 0.4\,\mathrm{eV}$. We first scale the $M_\nu = 0\,\mathrm{eV}$-simulation to the target cosmologies (changing M_ν to $0.1, 0.2, 0.3, 0.4\,\mathrm{eV}$

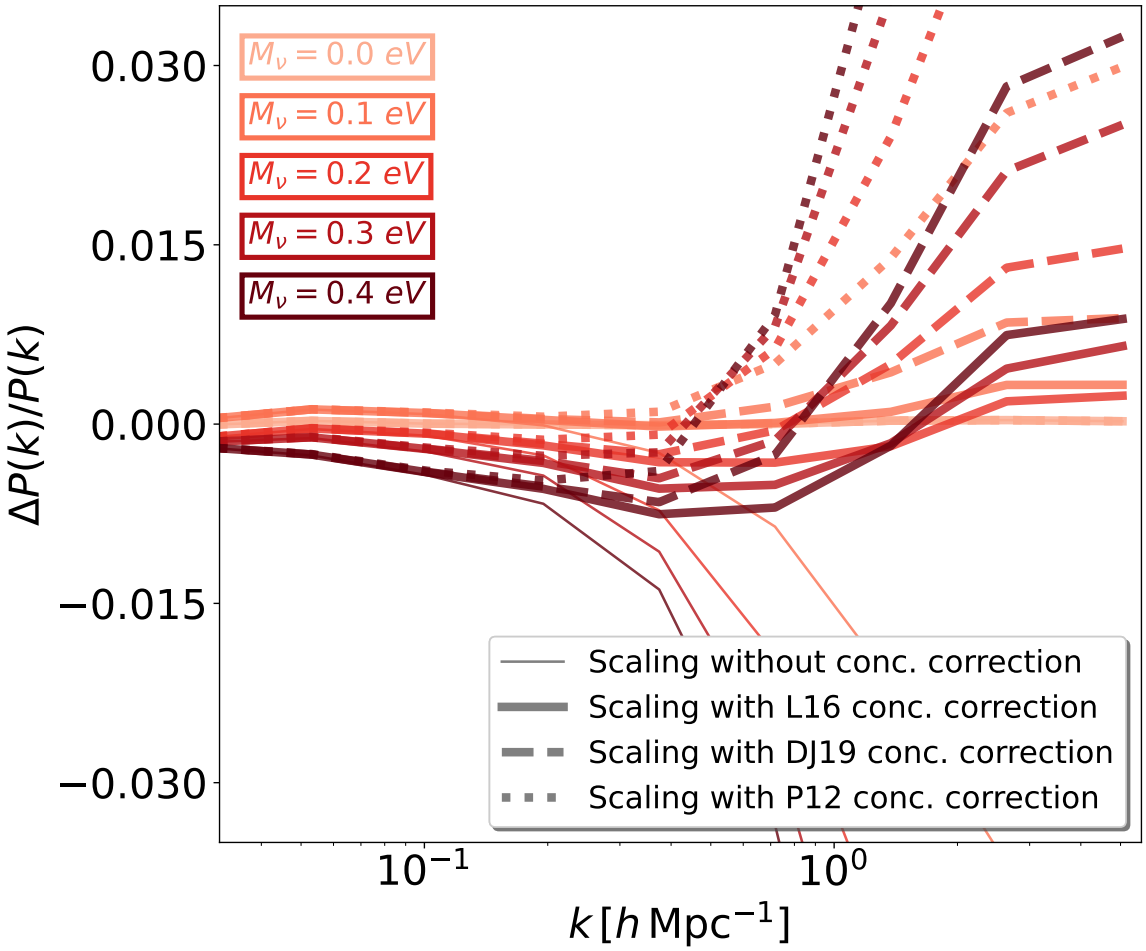

Figure 2.8: Relative difference between the power spectrum obtained from a gravity-only reference simulation and the power spectrum obtained from the corresponding rescaled simulation, i.e., $\Delta P(k)/P(k) = P_{\rm scaled}(k)/P_{\rm N-body}(k) - 1$. We focus on models with different neutrino masses, and choose our reference simulation to be the one with $M_\nu = 0\,eV$. The thin solid lines correspond to the case without applying any concentration correction and the other line styles correspond to concentration corrections obtained using three different concentration models: the re-calibrated L16 model (Eq. (D.1); thick solid lines), the Prada et al. 2012 model (P12; dotted lines), and the model of Diemer and Joyce 2019 (DJ19; dashed lines). The different shades of red correspond to different neutrino masses. The Nyquist frequency corresponding to these set of simulations is $\log_{10} k_{\rm Ny}\,[h\,\mathrm{Mpc}^{-1}] \approx 0.97$, and the number of points selected to compute the power spectrum is sufficiently large so that aliasing effects are not important.

subsequently) without considering concentration corrections, then, we repeat the process employing three different concentration models – the re-calibrated L16 model, the model presented in Prada et al. (2012) and the one from Diemer and Joyce (2019) – to provide the predictions for the concentration corrections.

We compute the power spectrum for the original N-body simulations and the scaled simulations with and without concentration correction. The relative differences between the original and scaled power spectra $\Delta P(k)/P(k) = P_{\rm scaled}(k)/P_{\rm N-body}(k) - 1$ are plotted in Fig. 2.8 as a function of scale. The thin solid lines correspond to the comparison with respect to mock simulations scaled without concentration corrections; the remaining lines correspond to the comparison with scaled simulations in which we have considered the concentration corrections associated with different concentration models: the re-calibrated L16 model (solid lines), Prada et al. 2012 (P12 – dotted lines) and Diemer and Joyce 2019 (DJ19 – dashed lines).

In Fig. 2.8 one can appreciate that the power spectra of the scaled simulations with concentration corrections are closer to the power spectra of the original N-body simulations in comparison with the case without concentration corrections. Different concentration models produce different levels of concentration corrections in the scaling technique which can be observed at the power spectrum level. The re-calibrated L16 model (i.e. Eq. (D.1)) yields the most accurate predictions for the power spectrum when compared to the other models. In the most extreme scenario, when $M_\nu = 0.4\,\mathrm{eV}$, the relative difference between the power spectrum from the rescaled simulation (for the L16 model) and the original simulation at $k \approx 4\,h\,\mathrm{Mpc}^{-1}$ is less than 1%; for the other concentration models the relative differences at this scale are at least twice as large.

2.5 Conclusions

In this paper, we carried out an extensive analysis of the cosmology dependence of the mass-concentration-redshift relation, $c(M, z)$, for dynamically relaxed dark matter haloes. Our results were based on a large suite of gravity-only simulations in which we systematically varied the following cosmological parameters: σ_8, $\Omega_{\rm M}$, $\Omega_{\rm b}$, $n_{\rm s}$, h, M_ν, w_0 and $w_{\rm a}$. Each parameter was varied linearly across a range that spans a 5 to 10σ region (depending on the parameter; see Table 2.2 and Table 2.1) surrounding the best-fit value obtained by Planck Collaboration et al. (2020b).

In agreement with previous work, we find that, regardless of the cosmological parameter varied, the concentrations of DM haloes, on average, decrease with increasing halo mass at fixed redshift (Fig. 2.2), as well as with increasing redshift at fixed halo mass (Fig. E.1). For the range of parameter values we considered, concentrations are most sensitive to changes in

σ_8, the rms amplitude of linear density fluctuations; they are least sensitive to changes in $\Omega_{\rm b}$, the baryon density parameter. This result is not surprising given the strong dependence of halo formation times on σ_8 and their weak dependence on $\Omega_{\rm b}$.

In general, our results agree with previous studies showing that the structure of dark matter haloes is strongly correlated with their formation histories (e.g. Ludlow et al., 2014, 2016; Lucie-Smith et al., 2022). Specifically, we find that halo concentrations, when expressed in terms of the enclosed density within the halo scale radius, i.e. $\langle\rho_{-2}\rangle$, correlate strongly with the critical density at their formation time z_{-2}, i.e. $\rho_c(z_{-2})$. Indeed, when the latter is defined as the point at which the "collapsed mass history" (CMH; defined as the mass in collapsed progenitors larger than a fraction $f = 0.02$ of the halo's present day mass) first exceeds the halo's characteristic mass, i.e. $M_{-2} = M(< r_{-2})$, we find an approximately linear relation between the two densities that may be accurately approximated by

$$\langle\rho_{-2}\rangle = 493 \times \rho_c(z_{-2}). \tag{2.6}$$

This simple relation holds for all cosmologies, redshifts, and masses studied. This is a somewhat surprising result and the most important finding of our paper: The relation between nonlinear halo structure and formation time is universal hinting that it may be a fundamental consequence of gravitational dynamics and collapse. This universality implies that our predictions for the concentration-mass relation should be valid even for cosmologies and halo masses outside the range considered here.

We showed that equation 2.6, when combined with an accurate model for halo CMHs based on extended Press-Schechter theory (see Fig. 2.4 and appendix D), can be used to make accurate prediction for the mass-, cosmology- and redshift-dependence of halo concentrations (Fig. 2.6) even when considering dynamical dark energy and massive neutrinos. We compared our predictions for the $c(M, z)$ relation with other published models (Fig. 2.7) and verified that they more accurately capture its cosmology dependence.

Our results confirm and extend those originally obtained by Ludlow et al. (2016) and suggest that equation 2.6 can be used to accurately predict the concentrations of DM haloes in a wide range of scenarios. This can be very useful in many areas of cosmology, e.g., to improve cosmological rescaling algorithms (see Contreras et al., 2020, Fig. 2.8 and Section 2).

Chapter 3

Characterizing structure formation through instance segmentation

Dark matter haloes form from small perturbations to the almost homogeneous density field of the early universe. Although it is known how large these initial perturbations must be to form haloes, it is rather poorly understood how to predict which particles will end up belonging to which halo. However, it is this process that determines the Lagrangian shape of proto-haloes and is therefore essential to understand their mass, spin and formation history. We present a machine learning framework to learn how the proto-halo regions of different haloes emerge from the initial density field. We develop one neural network to distinguish semantically which particles become part of *any* halo and a second neural network that groups these particles by halo membership into different instances. This instance segmentation is done through the Weinberger method, in which the network maps particles into a pseudo-space representation where different instances can be distinguished easily through a simple clustering algorithm. Our model reliably predicts the masses and Lagrangian shapes of haloes object-by-object, as well as other properties like the halo-mass function. We find that our model extracts information close to optimal by comparing it to the degree of agreement between two N-body simulations with slight differences in their initial conditions. We publish our model open-source and suggest that it can be used to inform analytical methods of structure formation by studying the effect of systematic manipulations of the initial conditions.

3.1 Introduction

Dark matter (DM) haloes are the primary structures in the universe within which galaxies form and evolve. Acting as gravitational anchors, they play a pivotal role in connecting theoretical cosmology with empirical observations from galaxy surveys. Given their significance in cosmology, a comprehensive understanding of DM haloes and their behaviour

is paramount. Currently, our most detailed insights into their formation and properties come from N-body simulations (see Frenk and White, 2012, for a review). These computationally intensive simulations model the interactions of vast numbers of particles, pinpointing the regions of the density field where gravitational collapse leads to the formation of DM haloes (e.g. Angulo and Hahn, 2022a). Therefore, understanding the formation and behaviour of DM haloes is essential to bridge the gap between theoretical models and observational data.

However, providing quick and accurate predictions (based on the initial conditions of a simulation) remains a challenging task for physically-motivated models. An accurate model for halo formation must be able to capture the nonlinear growth of density fluctuations. Previous analytical or semi-analytical models for halo formation, such as the top-hat spherical collapse (Gunn and Gott, 1972; Gunn, 1977; Peebles, 1980), the Press-Schechter / Excursion Set Theory (Press and Schechter, 1974b; Bond et al., 1991b; Lacey and Cole, 1993), or ellipsoidal collapse approaches (e.g. Sheth et al., 2001; Sheth and Tormen, 2002), qualitatively reproduce the behaviour of the halo-mass function and the merging rate of haloes, however, they fail on predicting these quantities accurately (e.g. Jiang and van den Bosch, 2014). Further, N-body simulations show the formation of "peak-less" haloes, that cannot be accounted for by any of these methods (Ludlow and Porciani, 2011).

Traditional analytical methods have provided foundational insights into the process of halo formation, but they struggle to capture the full complexity of it. Machine Learning (ML) techniques have emerged as a promising alternative, capable of capturing intricate non-linear dynamics inherent to the gravitational collapse of structures. ML algorithms can be trained on N-body simulations to emulate the results of much more expensive calculations. Previous studies have trained ML models to map initial positions and velocities of particles to their final states (He et al., 2019; Giusarma et al., 2019; Alves de Oliveira et al., 2020; Wu et al., 2021; Jamieson et al., 2022) and to predict the distribution of non-linear density fields (Rodríguez et al., 2018; Perraudin et al., 2019; Schaurecker et al., 2021; Zhang et al., 2023; Schanz et al., 2023).

Further, ML has been used to predict and gain insights into the formation of haloes. Some studies utilized classification methods to anticipate if a particle will become part of a halo (Lucie-Smith et al., 2018; Chacón et al., 2022; Betts et al., 2023), or to predict its final mass category (Lucie-Smith et al., 2019). In Lucie-Smith et al. (2020) a regressor network is trained to predict the final halo mass for the central particle in a given simulation crop. The work by Bernardini et al. (2020) demonstrates how ML-segmentation techniques can be applied to predict halo Lagrangian regions. In Berger and Stein (2019) a semantic segmentation network is trained to predict Peak-Patch-haloes. In Lucie-Smith et al. (2023) a network is trained to predict the mass of haloes when provided with a Lagrangian region centred on the centre-of-mass of proto-halo patches and is then used to study assembly bias

when exposed to systematic modifications of the initial conditions.

While interesting qualitative insights have been obtained in these studies, it would be desirable to develop a model that accurately predicts halo membership at a particle level, surpassing some of the limitations from previous works. An effective model should predict particles forming realistic N-body halos, improving upon previous models restricted to simpler halo definitions (e.g. Berger and Stein, 2019, where Peak-patch haloes are targeted). Additionally, an ideal model should be able to predict disconnected Lagrangian halo patches, overcoming the limitations of methods like the watershed technique used in Bernardini et al. (2020), which can only handle simply connected regions. Furthermore, particles within the same halo should share consistent mass predictions, avoiding having different halo mass estimates for particles belonging to the same halo.

We present a general ML framework to predict the formation of haloes from the initial linear fields. We create a ML model designed to forecast the assignment of individual particles from the initial conditions of an N-body simulation to their respective haloes. To do so we train two distinct networks, one for conducting semantic segmentation and another for instance segmentation. These two networks together conform what is known as a panoptic-segmentation model. Our model effectively captures the dynamics of halo formation and offers accurate predictions. We provide the models used in this study for public access through our GitHub repository: `https://github.com/daniellopezcano/instance_halos`.

The rest of this paper is organized as follows: In Section 3, we define the problem of identifying different Lagrangian halo regions from the initial density field (§§3), introduce the panoptic segmentation method (§§3), present the loss function employed to perform instance segmentation (§§3), describe the simulations used for model training (§§3), asses the level of indetermination for the formation of proto-haloes (§§3), outline the CNN architecture (§§3), and explain our training process (§§3). In Section 3, we present the outputs of our semantic model (§§3) and our instance segmentation approach (§§3). We investigate how our model reacts to changes in the initial conditions in §§3 & §§3, and study how the predictions of our model are affected when varying the cosmology §§3. We conclude with a summary and final thoughts in Section 3.

3.2 Methodology

We aim to predict the formation of DM haloes provided an initial density field. To comprehensively address this problem, we divide this section into distinct parts. In §§3, we explain the problem of predicting halo-collapse and discuss the most general way to phrase it. In §§3, we introduce the panoptic segmentation techniques and explain how they can

be employed to predict halo formation. We divide §§3 into two separate parts: semantic segmentation and instance segmentation. In §§3 we describe the loss function employed to perform instance segmentation. In §§3, we present the suite of simulations generated to train and test our models. In §§3 we assess the level of indetermination of proto-halo formation. In §§3 we explain how to build a high-performance model employing convolutional neural networks. Finally, in §§3 we present the technical procedure followed to train our models.

3.2.1 Predicting structure formation

The goal of this work is to develop a machine-learning framework to predict the formation of haloes from the initial conditions of a given universe. Different approaches are possible to define this question in a concrete input/output setting. We want to define the problem in a way that is as general as possible so that our model can be used in many different contexts.

The input of the model will be the linear density field discretized to a three-dimensional grid δ_{ijk}. A slice through such a linear density field is shown in the top panel of Figure 3.1 and represents how our universe looked in early times, e.g., $z \gtrsim 100$. Beyond the density field, we also provide the linear potential field ϕ_{ijk} as an input. The information included in the potential is in principle degenerate with the density field if the full universe is specified. However, if only a small region is provided, then the potential contains additional information of e.g. the tidal field sourced by perturbations outside of the region considered.

The model shall predict which patches of the initial density field become part of which haloes at later times. Concretely, we want it to group the N^3 initial grid cells (corresponding, e.g., to particles in a simulation) into different sets so that each set contains exactly all particles that end up in the same halo at a later time. Additionally, there has to be one special extra set that contains all remaining particles that do not become part of any halo:

$$\text{Input:} \quad \delta_{ijk}, \phi_{ijk} \tag{3.1}$$

$$\text{Output:} \quad \overbrace{\{\mathrm{id}_A, \mathrm{id}_B, ...\}}^{\text{halo 1}}, \overbrace{\{\mathrm{id}_C, \mathrm{id}_D, ...\}}^{\text{halo2}}, ..., \overbrace{\{\mathrm{id}_E, \mathrm{id}_F, ...\}}^{\text{outside of haloes}}, \tag{3.2}$$

This task is called in the ML literature an *instance segmentation* problem. Note that it is different from typical classification problems since (A) the number of sets depends on the considered input and (B) the sets have no specific order. In practice, it is useful to define the different sets by assigning different number-labels to them. For example, one possible set of particles belonging to the same halo can be given the label "1", another set the label "2", and so forth. These number-labels do not have a quantitative meaning and are permutation invariant, for example, interchanging the label "1" with "2" yields the same sets.

We show such labelling of the initial space in the bottom panel of Fig. 3.1. In this case, the labels were inferred by the membership to haloes in an N-body simulation that employs

the initial conditions depicted in the top panel of Fig. 3.1 (see Sec. 3). Our goal is to train a model to learn this instance segmentation into halo sets by training it on the output from N-body simulations.

We note that other studies have characterised the halo-formation processes through a slightly different prediction problem. For example, Lucie-Smith et al. (2020) trains a neural network to predict the final halo masses directly at the voxel level. While their approach offers insights into halo formation, our method provides a broader perspective: halo masses can be inferred easily through the size of the corresponding sets, but other properties can be inferred as well – for example the Lagrangian shapes of haloes which are important to determine their spin (White, 1984). Furthermore, our approach ensures the physical constraint that particles that become part of the same halo are assigned the same halo mass.

3.2.2 Panoptic Segmentation

The proposed problem requires first to segment the particles semantically into two different classes (halo or non-halo) and then to classify the particles inside the halo class into several different instances. The combination of such semantic plus instance segmentation is sometimes referred to as *panoptic segmentation*. Several strategies have been proposed to solve such panoptic segmentation problems (Kirillov et al., 2016; Bai and Urtasun, 2016; Arnab and Torr, 2017; De Brabandere et al., 2017; Kirillov et al., 2018, 2023) and they usually operate in two-steps:

1. **Semantic segmentation**: The objective of this task is to predict, for each voxel in our initial conditions (representing a tracer particle in the N-body code), whether it will be part of a DM halo at $z = 0$. This task is a classification problem, and we will employ the balanced cross-entropy (BaCE) loss (Xie and Tu, 2015) to tackle it:

$$\mathcal{L}_{\text{BaCE}}\left(\mathbf{Y}, \hat{\mathbf{Y}}\right) = -\beta \mathbf{Y} \log \hat{\mathbf{Y}} - (1-\beta)\,(1-\mathbf{Y}) \log(1-\hat{\mathbf{Y}}) \tag{3.3}$$

   Here, $\mathbf{Y}$ represents the ground truth data vector, each entry corresponds to a voxel and is equal to 1 if the associated particle ends up being part of a halo; otherwise, its value is 0. $\hat{\mathbf{Y}}$ contains the model predictions, with each entry representing the probability that this particle ends up in a halo. The parameter β handles the class imbalance and is calculated as the number of negative samples divided by the total number of samples. We measure β using our training simulations (see §§3) and obtain a value of $\beta = 0.5815$[1]. After training our network, we need to choose a semantic threshold to

[1]The value of β depends on many properties such as the cosmological parameters chosen for the simulations, the redshift, or the mass resolution. We would need to retrain our network and recompute the value of β to obtain reliable predictions in different scenarios.

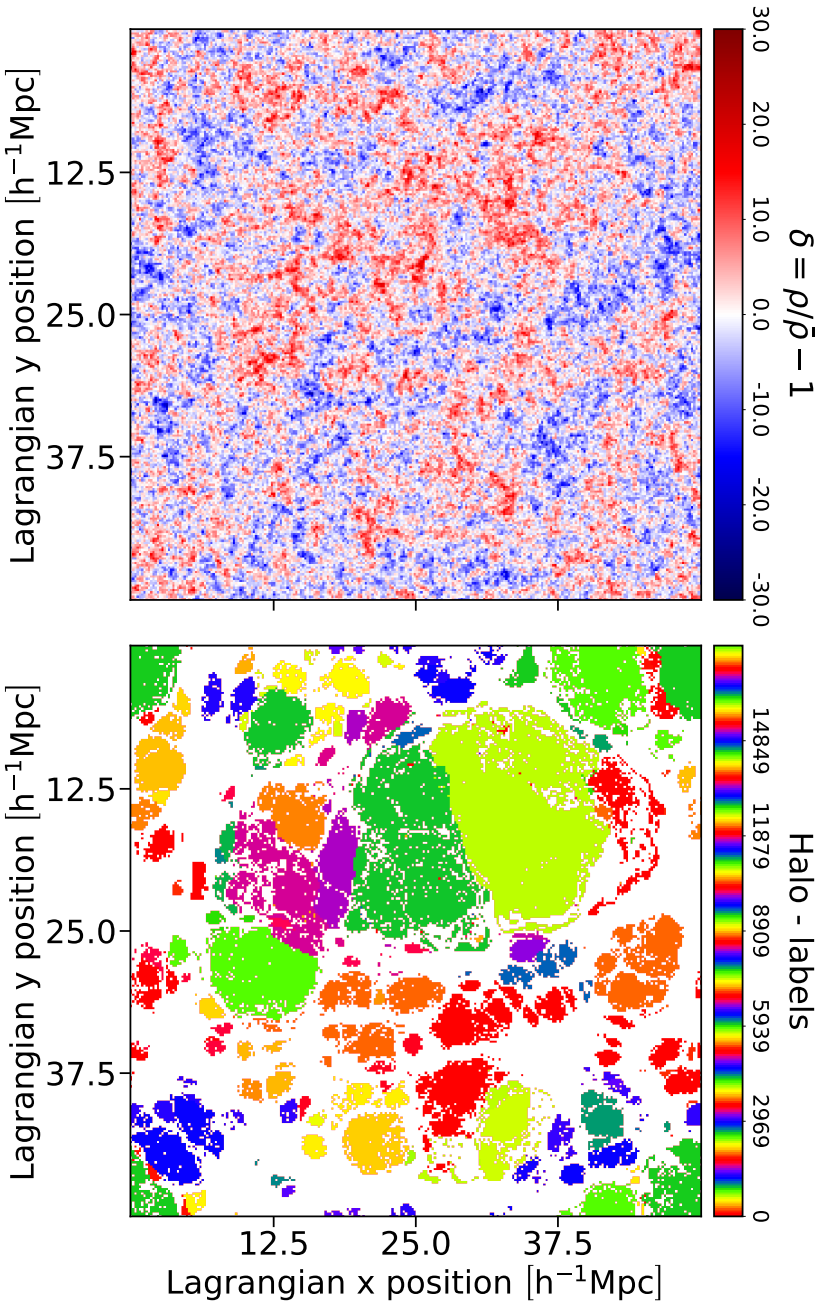

Figure 3.1: Example of the prediction problem considered in this article. **Top panel**: Slice of the three-dimensional initial density field of an N-body simulation. Each voxel (represented here as a pixel) corresponds to a particle that can become part of a halo at later times. **Bottom panel**: Regions in the initial condition space (same slice as the top panel) that are part of different DM haloes at redshift $z = 0$. Pixels coloured in white do not belong to any halo. Pixels with different colours belong to different haloes

. In this work, we present a machine-learning approach to predict the formation of haloes (as in the bottom panel) from the initial condition field (top panel).

generate the final semantic predictions. This threshold is calibrated to ensure that the fraction of predicted particles belonging to haloes is equal to $1 - \beta$, resulting in a value of 0.589 (refer to Appendix H for an in-depth explanation).

2. **Instance segmentation**: The objective of this task is to recognize individual haloes (instances) by identifying which particles (from those that are predicted to be part of a DM halo) belong to the same object and separating them from others.

   Instance segmentation tasks are not conventional classification problems and tackle the problems of having a varying number of instances and a permutational-invariant labelling. To our knowledge, there is no straightforward way to phrase the problem of classifying each voxel into a flexible number of permutable sets through a differentiable loss function. Typical approaches train a model to predict a related differentiable loss and then apply a postprocessing step on top of it. Unfortunately, this leads to the loss function not directly reflecting the true objective.

   Various approaches have been proposed to tackle this problem (Kirillov et al., 2016; Bai and Urtasun, 2016; Arnab and Torr, 2017; De Brabandere et al., 2017; Kirillov et al., 2018, 2023). A popular method is the watershed technique (Kirillov et al., 2016; Bai and Urtasun, 2016). This method uses a network to predict semantic segmentation and the borders of different instances (Deng et al., 2018) and then applies a watershed algorithm to separate different instances in a post-processing step. However, the watershed approach comes with several limitations:

   - It cannot handle the identification of disconnected regions belonging to the same instance, a problem known as occlusion.
   - It is necessary to select appropriate threshold values for the watershed post-processing step to generate the final instance map. These parameters are typically manually chosen to match some particular metric of interest, but might negatively impact the prediction of other properties. For instance, in Bernardini et al. (2020), they apply the watershed technique to predict Lagrangian halo regions identified with the `HOP` algorithm (Eisenstein and Hut, 1998). However, they choose the watershed threshold to reproduce the halo-mass-function, which does not ensure that the Lagrangian halo regions are correctly predicted.
   - The watershed approach would struggle to identify the borders of Lagrangian halo regions since they are difficult to define. In Fig. 3.1 it can be appreciated that the borders of halo regions are very irregular. There also exist points in the "interior" of these regions which are "missing" and make it particularly complex to define the border of a halo.

Despite all the challenges presented by the watershed approach, in Section F, we apply this method to predict the formation of FoF-haloes and discuss how the border-prediction problem can be addressed.

An approach that offers greater flexibility for grouping arbitrarily arranged particles was presented by De Brabandere et al. (2017). We will follow this approach through the remainder of this work. The main idea behind this method, which we will refer to as the "Weinberger approach"[2], is to train a model to produce a "pseudo-space representation" for all the elements of our input space (i.e., voxels/particles in the initial conditions). An ideal model would map voxels belonging to the same instance close together in the pseudo-space while separating them from voxels belonging to different instances. Consequently, the pseudo-space distribution would consist of distinct clouds of points, each representing a different instance (see Fig. 3.2). The postprocessing step required to generate the final instance segmentation in the Weinberger approach is a clustering algorithm which operates on the pseudo-space distributions.

3.2.3 Weinberger loss

The Weinberger approach possesses some advantages over other instance segmentation techniques: First of all, the loss function more closely reflects the instance segmentation objective; that is, to classify different instances into a variable number of permutationally invariant sets. Secondly, the approach is more flexible and makes fewer assumptions, for example, it can handle occlusion cases and does not need to assume the existence of well-defined instance borders.

In Fig. 3.2, we schematically illustrate the effects of the individual components of the Weinberger loss. Each point in this figure represents a pseudo-space embedding of an input voxel. The colours indicate the assigned labels based on the ground truth. Points sharing the same colour belong to the same instance (according to the ground truth), whereas different colours depict separate instances. The "centre of mass" for each cluster is computed and indicated with coloured crosses as "cluster centres". The Weinberger loss is constituted by three separate terms:

- **Pull force,** Eq. (3.4):

$$L_{pull} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N_c}\max\left(0, \left(\|\mu_{\mathbf{c}} - \mathbf{x_i}\| - \delta_{\text{Pull}}\right)^2\right) \tag{3.4}$$

[2]The loss function employed by De Brabandere et al. (2017) to perform instance segmentation is inspired by a loss function originally proposed by Weinberger and Saul (2009) in the context of contrastive learning as a triplet-loss function.

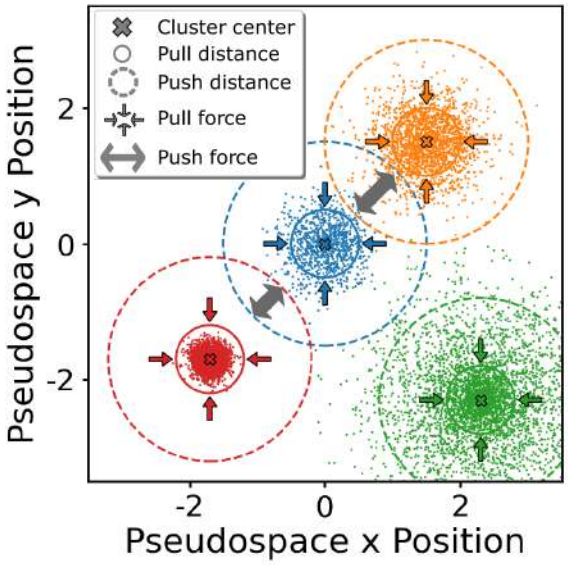

Figure 3.2: Example of a two-dimensional pseudo-space employed to separate different instances according to the Weinberger loss. Coloured points represent individual points mapped into the pseudo-space. The centres of the clusters are presented as coloured crosses. Coloured arrows depict the influence of the pull force term, only affecting points outside the δ_{Pull} range of their corresponding cluster centre. Grey arrows show the influence of the push force that manifests if two cluster centres are closer than the distance $2 \cdot \delta_{\text{Push}}$

Given a certain instance c (where C is the total number of instances), a point i belonging to that set, whose pseudo-space position is $\mathbf{x_i}$, will feel an attraction force proportional to the distance to the instance centre $\mu_c = \sum_{i=1}^{N_c} \mathbf{x_i}/N_c$, where N_c is the number of members associated with the instance c. Points closer than δ_{Pull} (which is a hyperparameter of the Weinberger loss) from the instance centre will not experience any pull force. The pull force is represented in Fig. 3.2 as coloured arrows pointing towards the instance centres outside the solid-line circles, which symbolize the distance δ_{Pull} to the instance centres.

- **Push force,** Eq. (3.5):

$$L_{\text{push}} = \frac{1}{C(C-1)} \sum_{\substack{c_A=1 \\ c_A \neq c_B}}^{C} \sum_{c_B=1}^{C} \Big(\max\left(\left(2\delta_{\text{Push}} - \|\mu_{\mathbf{c_A}} - \mu_{\mathbf{c_B}}\| \right)^2, 0 \right) \Big(\tag{3.5}$$

Two instances A and B will repel each other if the distance between their instance centres in the pseudo-space, μ_{c_A} and μ_{c_B}, is smaller than $2\delta_{\text{Push}}$ (a hyperparameter of the Weinberger loss). The force they feel is proportional to the distance between them. In Fig. 3.2 the push force is represented as grey arrows. The dashed circles represent the distance δ_{Push} to the instance centres.

- **Regularization force,** Eq. (3.6):

$$L_{\text{reg}} = \frac{1}{C}\sum_{c=1}^{C}\Big(\|\mu_{\mathbf{c}}\| \tag{3.6}$$

To avoid having an arbitrarily big pseudo-space distribution all instance centers will feel an attraction towards the pseudo-space origin.

The overall effect of these forces on the total Weinberger loss is written as:

$$\mathcal{L}_{\text{Wein}} = c_{\text{Pull}} \cdot L_{\text{Pull}} + c_{\text{Push}} \cdot L_{\text{Push}} + c_{\text{Reg}} \cdot L_{\text{Reg}} \tag{3.7}$$

Where c_{Pull}, c_{Push}, and c_{Reg} are hyperparameters that regulate the strength of the different components.

Minimizing Eq. (3.7) ensures that the pseudo-space mapping produces instance clusters separated from each other. A model trained effectively will predict pseudo-space distributions with points corresponding to the same instances being grouped together and distinctly separated from other instances. In an ideal scenario in which the Weinberger loss is zero, all points are closer than δ_{Pull} to their corresponding cluster centres, and clusters are at least $2\delta_{\text{Push}}$ apart. However, realistically, the Weinberger loss won't be exactly zero, necessitating a robust clustering algorithm for accurate instance map predictions.

In Appendix G we describe the clustering algorithm that we have developed to robustly identify the different instance maps. In our clustering algorithm we first compute the local density for each point in our pseudo-space based on a nearest neighbors calculation. We then identify groups as descending manifolds of density maxima surpassing a specified persistence ratio threshold. Particles are assigned to groups according to proximity and density characteristics. We merge groups selectively, ensuring that the persistence threshold is met. The algorithm relies on three key hyper-parameters for optimal performance: N_{dens}, N_{ngb} and p_{thresh}. This approach effectively segments the pseudo-space distribution of points, even when perfect separation is not achieved, thus enhancing the reliability of predicted instance maps.

3.2.4 Dataset of Simulations

We generate twenty N-body simulations with different initial conditions to use as training and validation sets for our panoptic segmentation model. Our simulations are carried out using a lean version of `L-Gadget3` (see Springel et al., 2008; Angulo et al., 2012, 2021). For each of these simulations, we evolve the DM density field employing $\mathrm{N_{DM}} = 256^3$ DM particles in a volume of $V_{\text{box}} = (50\,h^{-1}\mathrm{Mpc})^3$, resulting in a DM particle-mass of

$m_{\rm DM} = 6.35 \cdot 10^8\, h^{-1}{\rm M}_\odot$. All our simulations employ the same softening length: $\epsilon = 5\, h^{-1}{\rm kpc}$, and share the cosmological parameters derived by Planck Collaboration et al. (2020b), that is, $\sigma_8 = 0.8288$, $n_{\rm s} = 0.9611$, $h = 0.6777$, $\Omega_{\rm b} = 0.048252$, $\Omega_{\rm m} = 0.307112$, and $\Omega_\Lambda = 0.692888$. Our suite of simulations is similar to the one employed in Lucie-Smith et al. (2020).

We use a version of the `NgenIC` code (Springel, 2015) that uses second-order Lagrangian Perturbation Theory (2LPT) to generate the initial conditions at $z = 49$. We employ a different random seed for each simulation to sample the Gaussian random field that determines the initial density field. We identify haloes at redshift $z = 0$ in our simulations using a Friends-of-Friends algorithm (Davis et al., 1985), with linking length $b = 0.2$. In this work, we will only consider haloes formed by 155 particles or more, corresponding to $M_{\rm FoF} \gtrapprox 10^{11}\, h^{-1}{\rm M}_\odot$. We use 18 of these simulations to train our model and keep 2 of them to validate our results.

3.2.5 Assessing the level of indetermination

In addition to the training and test sets, we run a set of simulations to establish a target accuracy for our model. These simulations test to what degree small sub-resolution changes of the initial density field can affect the final Lagrangian halo regions.

Structure formation simulations resolve the initial conditions of a considered universe only to a limited degree and exhibit therefore an inherent degree of uncertainty. (1) The numerical precision of simulations is limited (e.g. to 32bit floating point numbers) and therefore any results that depend on the initial conditions beyond machine precision are inherently uncertain. For example, Genel et al. (2019) show that changes in the initial displacement of N-body particles at the machine-precision level can lead to differences in the final locations of particles as large as individual haloes. (2) The initial discretization can only resolve the random perturbations of the Gaussian random field down to a minimum length scale of the mean-particle separation. If the resolution of a simulation is increased, then additional modes enter the resolved regime and act as additional random perturbations. Such additional perturbations may induce some random changes in the halo assignment of much larger-scale structures.

A good model should learn all aspects of structure formation that are certain and well resolved at the considered discretization level. However, there is little use in predicting aspects that are under-specified and may change with resolution levels. Therefore, we conduct an experiment to establish a baseline of how accurate our model shall be.

We run two additional $N = 256^3$ simulations with initial conditions generated by `MUSIC` code (Hahn and Abel, 2011). For these simulations we keep all resolved modes fixed (up

to the Nyquist frequency of the 256^3 grid), but we add to the particles different realisations of perturbations that would be induced by the next higher resolution level. We do this by selecting every 2^3th particle from two initial condition files with 512^3 particles and with different seeds at the highest level ("level 9" in `MUSIC`). Therefore, the two simulations differ only in the random choice of perturbations that are unresolved at the 256^3 level. We refer to these two simulations as the "baseline" simulations.

In Fig. 3.3 we show a slice of the Lagrangian halo patches at $z = 0$ through these simulations (left and right panels respectively). The colour map in this Figure represents the masses of the halo that each particle becomes part of, which correspond to the size of the corresponding halo-set. We colour each pixel (which corresponds to a certain particle) according to the mass of the halo that it belongs to. We can appreciate that the outermost regions of the Lagrangian regions are particularly affected while the innermost parts remain unchanged. Notably, in certain instances, significant changes appear due to the merging of haloes in one of the simulations where separate haloes are formed in the other (black-circled regions).

Throughout this article, we will use the degree of correspondence between the baseline simulations as a reference accuracy level. We consider a model close to optimal if the difference between its predictions and the ground truth is similar to the differences observed between the two baseline simulations. A lower accuracy than this would mean that a model has not optimally exploited all the information that is encoded in the initial conditions. A higher accuracy than this level is not desirable, since it is not useful to predict features that depend on unresolved aspects of the simulation and may be changed by increasing the resolution level.

3.2.6 V-Net Architecture

V-nets are state-of-the-art models, product of many advances in the field of ML over the last decades (Fukushima, 1980; Lecun et al., 1998; Krizhevsky et al., 2012; Szegedy et al., 2014; Long et al., 2014; Ronneberger et al., 2015; He et al., 2015). They are a particular kind of convolutional neural network (CNN) developed and optimized to efficiently map between volumetric inputs and volumetric outputs. V-nets are formed by two separate modules: the encoder (or contracting path) which learns how to extract large-scale abstract features from the input data; and the decoder (or up-sampling path) that translates the information captured by the encoder to voxel-level predictions (also making use of the information retained in the "skipped connections"). We train V-nets to minimize the loss functions presented in §§3 and §§3. We now explain the technical characteristics of how we have implemented a V-net architecture in TensorFlow (Abadi et al., 2015) (see Fig. 3.4 for a schematic representation

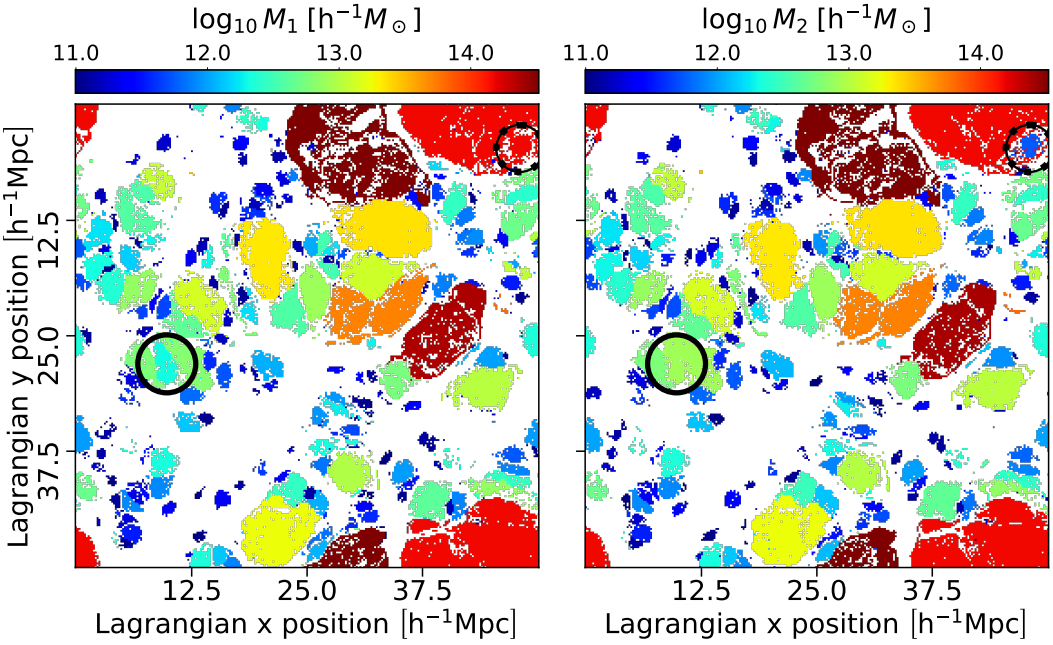

Figure 3.3: Slice of the Lagrangian halo regions of the two "baseline" simulations (left and right panels respectively). These simulations only differ in sub-resolution perturbations to the initial conditions and their level of agreement sets a baseline for the desired accuracy of our models. The colours employed for both panels represent the mass of the halo associated with each particle for the different Lagrangian halo patches. Circled regions highlight Lagrangian patches whose associated mass significantly changes between the two simulations.

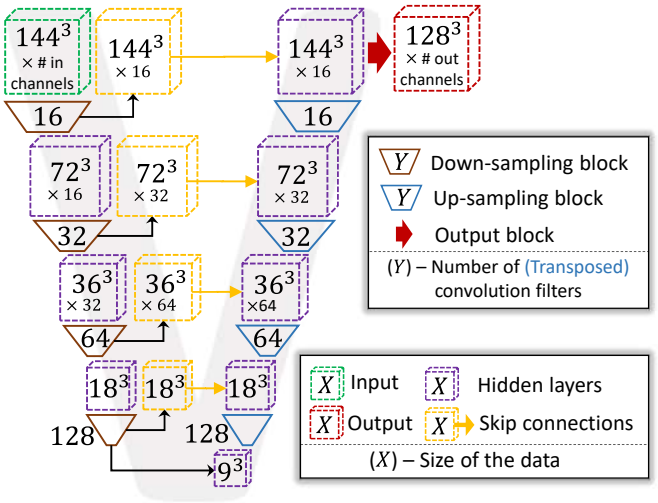

Figure 3.4: Flowchart of the particular V-Net architecture we have implemented. The network can take as input multiple channels with dimensions of 144^3 (top left green cube) and generates predictions for the central voxels with dimensions 128^3 (top right red cube). The flowchart illustrates the encoder and decoder paths, along with other distinctive features of the network. Notably, the hidden layers and skip connections are represented by purple and yellow cubes, with their respective dimensions annotated at their centres. The down-sampling and up-sampling blocks are shown as brown and purple trapezoids, in their centres we indicate the number of filters employed for the convolution (or transposed convolution) operations.

of our network architecture):

- Input: Our network is designed to accept as input 3D crops consisting of 144^3 voxels.[3] For the results presented in Section 3, we employ two input channels for the semantic segmentation model, corresponding to the initial density field and the displacement potential, which is defined through Poisson's equation as:

$$\delta(\vec{q}) = \vec{\nabla}^2 \phi(\vec{q}) \tag{3.8}$$

  For the instance segmentation model, we include three additional input channels corresponding to the Lagrangian positions of particles. This is necessary since the network has to be able to map different haloes with the same density (and potential) structure at different locations in the initial field to different locations in the pseudo space.

- Encoder / contractive / down-sampling / down-scaling path: This module consists of consecutive down-scaling blocks that reduce the number of voxels per dimension by half at each level of the network. The purpose of the down-scaling path is to enlarge the network's field of view, enabling per-voxel predictions that take into account distant regions of the field. Achieving this would be impractical using large convolution kernels, as they would consume excessive memory. Within each down-sampling block, we apply three consecutive convolution operations followed by a Leaky-ReLu activation function. The number of convolution filters in a contractive block doubles with each level of compression to improve the performance of the model. For each level, the latent maps computed before the final convolution (the one used to reduce the data size) are temporarily stored to serve as a skip connection for the up-scaling path. In Fig. 3.4 we show the dimensions of the latent maps computed at each level of the contractive path; the deepest level of our network has a size of $9^3 \times 128$.

- Decoder / up-sampling / up-scaling path: This path operates opposite to the contractive path; each up-scaling block doubles the number of voxels per dimension, ultimately recovering an image with the same dimensions as the original input (see Fig. 3.4). The up-sampling path facilitates the extraction of smaller-scale features that influence the final per-voxel predictions. Within an up-sampling block, the final convolution is substituted with a transposed convolution operation, that allows doubling the output size per dimension.

- Output: The final module of our network takes as input the latent maps with dimensions $144^3 \times 16$. The functionality of this module varies depending on the task at hand. For

[3] Ideally, we would prefer to accept as input 256^3 voxels (corresponding to the full simulation box). However, our GPU resources, though powerful (specifically, an `NVIDIA QUADRO RTX 8000` with 48 GB of memory), are insufficient to accommodate such an input size while maintaining a reasonably complex network architecture.

semantic segmentation, a single convolution operation is performed, resulting in a latent map of $144^3 \times 1$. This map is subsequently cropped to $128^3 \times 1$, and finally, a sigmoid activation function is applied. In the case of instance segmentation, we have decided to work in a three-dimensional pseudo-space, hence, we employ a convolution with three filters to obtain $144^3 \times 3$ maps, which are afterwards cropped to $128^3 \times 3$. In both cases, the final cropping operation is implemented to enhance model performance by focusing on the central region of the image.

The V-Net architecture we have implemented is a state-of-the-art model that encompasses over $3 \cdot 10^6$ trainable parameters.

3.2.7 Training

We train our segmentation networks using a single Nvidia Quadro RTX 8000 GPU card. As mentioned in §§3, we employ 18 simulations for training, dividing the training process into separate stages for the semantic and instance models.

To ensure robust training and enhance the diversity of training examples without needing to run more computationally expensive simulations, we apply the following data augmentation operations each time we extract a training sample from our simulation suite:

1. Select one of the training simulation boxes at random.
2. Select a random voxel as the center of the input/output regions.
3. Extract the input (144^3) and target (128^3) fields of interest by cropping the regions around the central point, considering the periodic boundary conditions of the simulations.
4. Randomly transpose the order of the three input grid dimensions q_x, q_y, q_z.
5. Randomly chose to flip the axes of the input fields.

To train our semantic and instance segmentation networks we minimize the respective loss functions – Eq. (3.3) and Eq. (3.7) – employing the Adam optimizer implemented in `TensorFlow` (Abadi et al., 2015). We train our models for over 80 epochs, each epoch performs mini-batch gradient descent using 100 batches, and each batch is formed by 2 draws from the training simulations. We deliberately choose a small batch size to avoid memory issues and ensure the network's capability to handle large input and output images (144^3 and 128^3 respectively). Selecting a small batch size induces more instability during training; we mitigate this issue by using the clip normalization operation defined in `TensorFlow` during the backpropagation step.

The hyper-parameter β in the Balanced Cross-Entropy Eq. (3.3) is determined by computing the ratio of negative samples to the total number of samples in the training data. The value of β measured in different training simulations lies in the interval $[0.575, 0.5892]$. There exists a slight predominance of voxels/particles that do not collapse into DM haloes with mass $M_{\text{FoF}} \gtrapprox 10^{11}\, h^{-1}\text{M}_{\odot}$ at $z = 0$ considering the Planck Collaboration et al. (2020b) cosmology. We fix the hyper-parameter β in Eq. (3.3) to the mean value $\beta = 0.5815$.

Regarding the hyper-parameters in the Weinberger loss Eq. (3.7), we adopt the values presented in De Brabandere et al. (2017), as we have observed that varying these parameters does not significantly affect our final results. The specific hyper-parameter values are the following: $c_{\text{Pull}} = 1$, $\delta_{\text{Pull}} = 0.5$, $c_{\text{Push}} = 1$, $\delta_{\text{Push}} = 1.5$, and $c_{\text{Reg}} = 0.001$. We have conducted a hyper-parameter optimization for the clustering algorithm described in Appendix G and found the following values: $N_{\text{dens}} = 20$, $N_{\text{ngb}} = 15$ and $p_{\text{thresh}} = 4.2$ (see Table 3.2).

Our semantic and instance models are designed to predict regions comprising 128^3 particles due to technical limitations regarding GPU memory. To overcome this limitation and enable the prediction of larger simulation volumes, we have developed an algorithm that seamlessly integrates sub-volume crops. For our semantic model, we serially concatenate sub-volume predictions to cover the full simulation box. For our instance network, we propose the method described in Appendix I. In summary, this method works as follows: we generate two overlapping lattices. Both lattices cover the entire simulation box, but the second one is shifted with respect to the first one (its sub-volume centres lay in the nodes of the first one). The overlapping regions between the lattices are employed to determine whether instances from different crops should merge or not. We have verified that this procedure is robust by checking that the final predictions are not sensitive to the particular lattice choice.

We train our semantic and instance networks separately. The semantic predictions are not employed at any stage during the training process of the instance model. To compute the instance loss, Eq. (3.7) is evaluated using the true instance maps and the pseudo-space positions. The semantic predictions are only employed once both models have been trained. We use the semantic predictions to mask out pseudo-space particles not belonging to haloes. Then, the clustering algorithm described in Appendix G is applied to identify clusters of particles in the pseudo-space (which yields the final proto-halo regions).

Table 3.1: Hyper-parameters employed in our instance segmentation pipeline.

δ_{Pull}	δ_{Push}	c_{Pull}	c_{Push}	c_{Reg}	N_{dens}	N_{ngb}	p_{thresh}
0.5	1.5	1	1	0.001	20	15	4.2

3.3 Model Evaluation

In this section, we test the performance of our models for semantic segmentation (§§3) and instance segmentation (§§3). We use the two simulations reserved for validation to generate the results presented in this section.

3.3.1 Semantic Results

In Fig. 3.5, we compare the predictions of the semantic segmentation network with the halo segmentation found in the validation simulation. The leftmost panel illustrates a slice of the ground truth. Voxels/particles of the initial conditions belonging to a DM halo at $z = 0$ are shown in red; blue voxels represent particles not belonging to a DM halo at $z = 0$.

The central panel of Fig. 3.5 displays the probabilistic predictions from our semantic model for the same slice. The colour map indicates the probability assigned to each pixel for belonging or not to a DM halo. Voxels with a white colour have a 50% predicted probability of belonging to a halo. The neural network tends to smooth out features, assigning uncertain probabilities to regions near halo borders, while consistently assigning high probabilities to inner regions and low probabilities to external regions. In the ground truth it is possible to observe that some interior particles within proto-haloes are predicted to belong to the background. We refer to these as "missing voxels". One of the consequences of the smoothing effect of our network is to ignore these missing voxels, predicting a homogeneous probability of collapse in the interior regions of proto-haloes. The missing voxels in the Lagrangian structure seem to be a feature very sensitive to the initial conditions impossible to capture accurately at a voxel level. This is supported by the fact that the missing voxels also change significantly in the baseline simulations (see Fig. 3.3).

The rightmost panel of Fig. 3.5 shows the pixel-level error map for the same slice. We select a semantic threshold value equal to 0.589 to generate these results. We choose this value for the semantic threshold so that the total predicted number of particles that belong to a halo matches the number of collapsed voxels in the validation simulations. In Appendix H we further analyze the sensitivity of our semantic results to the value chosen for the semantic threshold. We use different colours to represent the corresponding classes of the confusion matrix: Green corresponds to true positive (TP) cases, blue to true negatives (TN), black to false negatives (FN), and red to false positives (FP).

Some regions are particularly challenging to predict for the network, likely due to their sensitivity to changes in the initial conditions. For example, in the rightmost panel of Fig. 3.5, it is easy to appreciate many FN regions that appear as black string-like structures surrounding TP collapsed regions. These FN cases likely correspond to particles infalling into the halo at $z = 0$, identified as part of the FoF group despite not having completed the

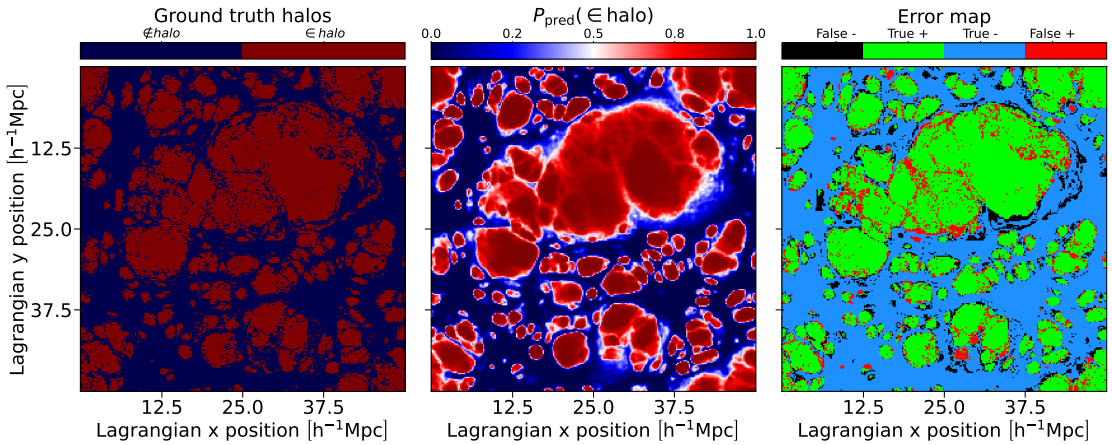

Figure 3.5: Slice through the predictions of our semantic segmentation network applied to a validation simulation. **Left panel**: Ground truth representation showing in red the voxels/particles belonging to a DM halo at $z = 0$ and in blue those particles that do not belong to a DM halo. **Central panel**: Probabilistic predictions of the semantic network with colour-coded probabilities for halo membership. **Right panel**: Pixel-level error map indicating true positive (green), true negative (blue), false negative (black), and false positive (red) regions resulting after applying a semantic threshold of 0.589 to our predicted map. The network effectively captures complex halo boundaries and exhibits high validation accuracy (acc $= 0.86$) and F_1-score ($\mathrm{F}_1 = 0.83$).

first pericentric passage. Capturing this behaviour might be particularly challenging for the network since the exact shape of these "first-infall" regions is more sensitive to small changes in the initial condition and can also be influenced by distant regions of the proto-haloes that do not completely fit within the field-of-view of our network (which can occur for very massive proto-halos). Also, we can appreciate FP regions that appear between the FN string-like regions and the TPs corresponding to the central Lagrangian regions of haloes. Additionally, the boundaries of the largest haloes may be especially difficult to predict for the network, since they only fit partially into the field of view.

The results presented in Fig. 3.5 suggest, upon visual inspection, that our model accurately captures many of the complex dynamics that determine halo collapse. To rigorously assess the performance of our model we need to quantify the results obtained from our semantic network and compare them with the differences between the baseline simulations, as discussed in Section 3.

In Table 3.2 we present the values of some relevant metrics that we can employ to evaluate the performance of our semantic network (we have considered the semantic threshold of 0.589). In particular, we study the behaviour of five different metrics: True Positive Rate $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, True Negative Rate $\mathrm{TNR} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$, Positive Predictive Value $\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, Accuracy ACC and the F_1-score (which is a more representative score than the accuracy when considering unbalanced datasets), see Eq. (3.9):

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \quad ; \quad \mathrm{F_1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3.9}$$

Table 3.2 also contains the scores measured using the baseline simulations. Our model returns values for all the metrics very close to the optimal target from the baseline simulations. This demonstrates the reliability of our model in predicting the well-specified aspects of halo collapse. See Appendix H for a more detailed discussion about the performance of our semantic model and the relation between the selected semantic threshold with the results contained in Table 3.2.

In addition to the optimal case, we compare our semantic model with the explicit implementation of the excursion set theory from ExSHalos (Voivodic et al., 2019). The ExSHalos code grows spheres around the density peaks in the Lagrangian density field until the average density inside crosses a specified barrier for the first time. The barrier shape is motivated by the ellipsoidal collapse (Sheth et al., 2001; de Simone et al., 2011) with three free parameters that were fitted to reproduce the mean mass function of our simulations. In Table 3.2 we include the semantic metrics measured with the ExSHalos results. While ExSHalos can describe halo formation to some degree, there exist some aspects that go beyond the spherical excursion set paradigm which are better captured by our semantic model. A more detailed analysis of the results obtained with ExSHalos is presented in Appendix J.

In Fig. 3.6 we compare the values of the predicted TPR as a function of ground truth halo mass ($\mathrm{TPR_{Pred}}$, solid green line), with the TPR values measured from the baseline simulations ($\mathrm{TPR_{base}}$, solid black line). It is possible to perform this comparison for the TPR because, in the ground truth data, we retain information about the mass of the FoF-haloes associated with each DM particle. Therefore, we can compute the fraction of TP cases in different ground-truth-mass-bins by selecting the voxels according to the mass associated with them in the ground truth.

In Fig. 3.6, the values for $\mathrm{TPR_{base}}$ increase with halo mass, indicating that particles that end up in lower-mass haloes are more sensitive to small-scale changes in the initial conditions,

Table 3.2: Performance metrics of our semantic segmentation model, along with the ExSHalos results, compared against the optimal target accuracy estimated from the baseline simulations. The table presents True Positive Rate (TPR), True Negative Rate (TNR), Positive Predictive Value (PPV), and Negative Predictive Value (NPV).

Type	TPR	TNR	PPV	ACC	$\mathrm{F_1}$
ExSHalos	0.518	0.845	0.707	0.708	0.598
Pred.	0.838	0.883	0.838	0.864	0.838
Optimal	0.887	0.914	0.882	0.903	0.884

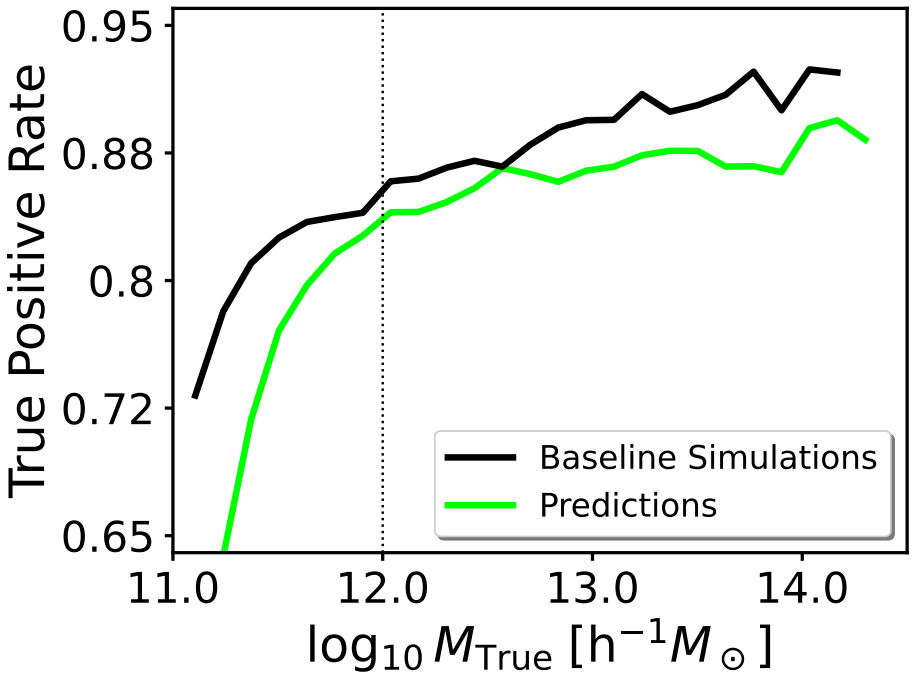

Figure 3.6: True Positive Rate expressed as a function of the halo mass associated with the ground truth voxels. We present the results measured from the model predictions (solid bright green line) in comparison to the optimal target accuracy from the baseline simulations (solid black line). The vertical dotted line at $10^{12}\,h^{-1}\mathrm{M}_\odot$ marks the point where model predictions start to differ from the baseline results.

consequently, harder to predict accurately. Our network's predictions follow a similar trend, albeit with some discrepancies. The model seems to under-predict the number of particles that end up in haloes with masses lower than $M_{\text{True}} \lessapprox 10^{12}\,h^{-1}\mathrm{M}_\odot$ (dotted vertical black line in Fig. 3.6). This indicates that our model tends to under-predict the number of pixels that are identified as TPs in the lower mass end. For haloes whose mass is greater than $10^{12}\,h^{-1}\mathrm{M}_\odot$, our model returns accurate predictions to a good degree over a broad range, extending more than two orders of magnitude in halo mass.

In this subsection, we have demonstrated that our semantic model extracts most of the predictable aspects of halo formation by comparing our results with the baseline simulations (which only differ in unresolved aspects of the initial conditions). We now employ the predictions of our semantic network to generate the final results using our instance segmentation model.

3.3.2 Instance Results

We provide some examples of our instance predictions in Fig. 3.7. The left column displays the ground truth masses of halo Lagrangian regions extracted from the simulation results (analogous to Fig. 3.3); the right column shows the predictions obtained from our

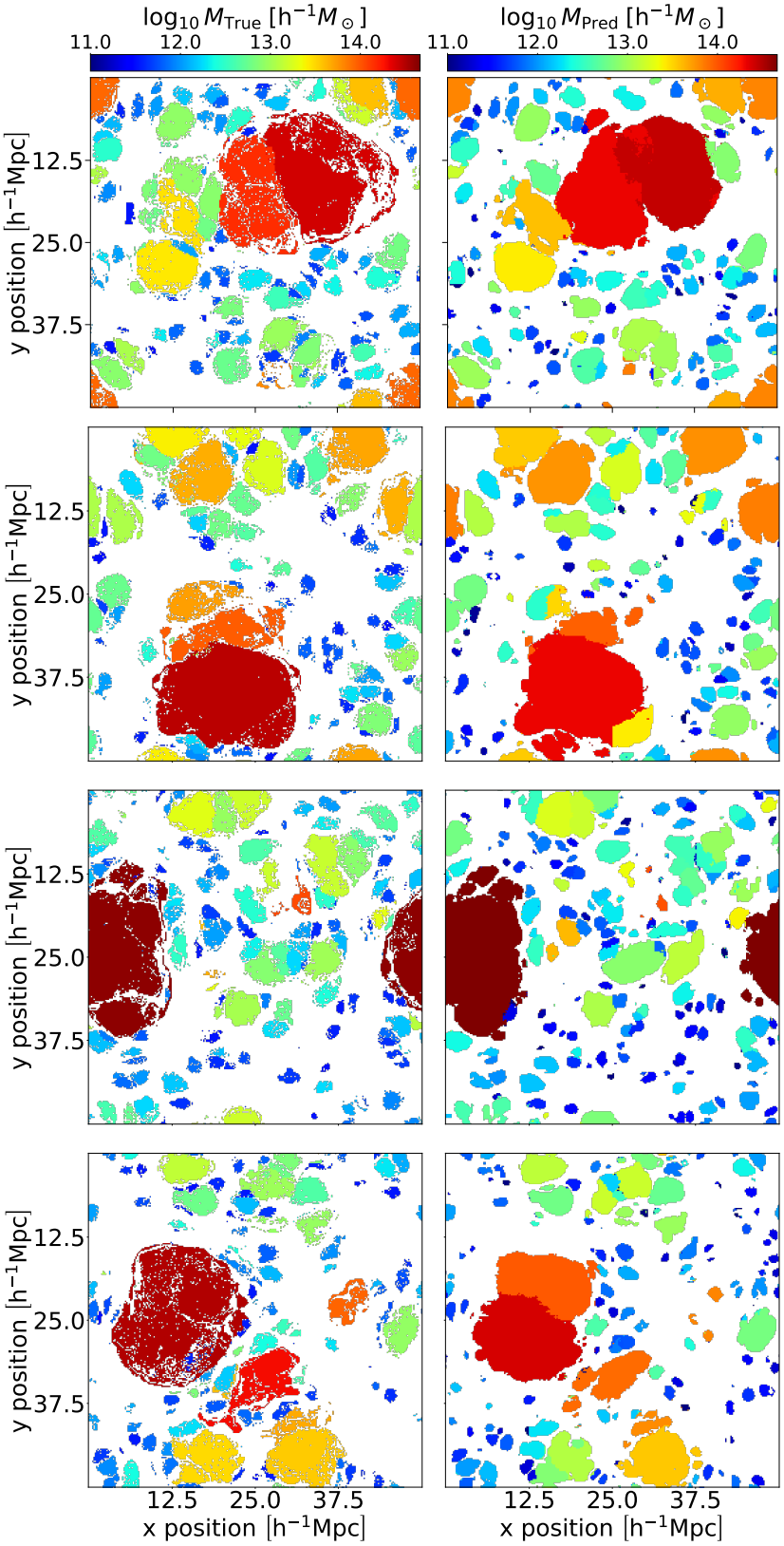

Figure 3.7: Examples of the instance segmentation results obtained with our model. **Left column**: ground truth masses obtained using N-body simulations. **Right column**: predicted masses obtained using our instance segmentation pipeline. The model can predict the Lagrangian patches of haloes, although some small differences – e.g. regarding the connectivity of haloes – exist.

segmentation pipeline. The way in which we compute halo masses from the instance predictions is by counting the number of particles/voxels that have been assigned to the same label and multiplying that by the particle mass of our simulations, $m_{\rm DM} = 6.35 \cdot 10^8\, h^{-1}{\rm M}_\odot$.

The shapes of the halo contours are well-captured thanks in part to the semantic predictions. The instance segmentation pipeline successfully distinguishes the different haloes that have formed, and in most cases, correctly separates neighbouring haloes. This is not a trivial task since the size of halo Lagrangian regions varies across several orders of magnitude. Therefore, the instance segmentation pipeline must correctly separate wildly different particle groupings in the pseudo-space. Fig. 3.7 shows that our instance segmentation pipeline correctly identifies different Lagrangian halo regions for the majority of cases. However, we note that differences arise on the one hand for very small haloes that are close to the resolution limit and on the other hand for very large haloes that are larger than the field of view of the network.

In Fig. 3.8, we present a comparison between the ground truth halo masses and the predicted masses associated with the particles/voxels in our validation set. To generate these results we apply the following procedure: We select all the ground truth voxels/particles that end up in FoF-haloes and study the predictions associated with them. We can associate a predicted mass for all the voxels that belong to a DM halo. In these cases, we can compare the predicted mass values ($M_{\rm Pred}$) with the ground truth masses ($M_{\rm True}$) at a voxel level. This comparison is shown in the main panel of Fig. 3.8 as black violin plots ("violins" henceforth). The mass range covered by the black violins goes from $M_{\rm True} = 10^{11}\, h^{-1}{\rm M}_\odot$, corresponding to the minimum mass of haloes (155 particles), to $M_{\rm True} \approx 10^{14.7}\, h^{-1}{\rm M}_\odot$, which is the mass of the most massive halo identified in the validation simulations. The number of high-mass haloes is smaller than small-mass ones and therefore the higher-mass end of the violin plot exhibits more noise. We can appreciate that the median predictions (black dots) correctly reproduce the expected behaviour (ground truth) for several orders of magnitude.

The voxels identified as part of a halo in the ground truth, but not in the predicted map, are false negative (FN) cases. For these occurrences, we can study the dependence of the False Negative Rate (FNR) as a function of the ground truth halo mass (solid black line on the top panel of Fig. 3.8; analogous to 3.6). We can also study the reciprocal case in which a voxel is predicted to be part of a halo (hence, it has an associated $M_{\rm Pred}$) but the ground truth voxel is not collapsed. These cases correspond to False positives (FP) but to make a comparison as a function of mass we can only express it in terms of the predicted mass. Therefore, we show as a dashed black line in the top panel of Fig. 3.8 the false discovery rate,

$$\text{FDR} = \frac{[\text{FP}|M_{\rm Pred}]}{[\text{TP}|M_{\rm Pred}] + [\text{FP}|M_{\rm Pred}]} \quad . \tag{3.10}$$

We compare our results with those obtained from the baseline simulations. In the main

panel of Fig. 3.8 we present the corresponding violin plots from the baseline simulations with green lines. The range that the green violins span is smaller than the black violins since the most massive halo identified in the baseline simulations has a mass of $M_{\rm True} \approx 10^{14.4}\, h^{-1}{\rm M}_{\odot}$. In the top panel, the solid and dashed green lines represent the FPR and FDR respectively. As expected, the FPR and FDR coincide in the case of the baseline simulations. The top panel results demonstrate that our predictions are comparable to those of the baseline simulations (as pointed out in Fig. 3.6) over most of the considered mass range. However, they get progressively worse for masses below $M_{\rm True} \lessapprox 10^{12}\, h^{-1}{\rm M}_{\odot}$ (vertical dotted black line), deviating from the baseline trend. This indicates that our model struggles to capture the correct behaviour of lower-mass haloes but it produces accurate predictions for higher-mass ones. When comparing the violin plot distributions of our model with the baseline simulations we appreciate that we obtain similar (but slightly broader) contours. Being able to achieve a similar scatter as in the baseline simulations indicates that our model can capture the well-resolved aspects of halo formation. We want to emphasize that precise predictions for halo masses are not directly enforced through the training loss, but are a side product, consequence of precisely reproducing halo Lagrangian patches. The scatter broadens for smaller halo mass and the network loses accuracy in these cases, sometimes associating smaller haloes close to a big Lagrangian patch to its closest more massive neighbour.

In the main panel of Fig. 3.8, we include the violin plot lines presented in Lucie-Smith et al. (2020) (blue violin lines). In this study, a neural network was trained to minimize the difference between predicted and true halo-masses at the particle level using as inputs the initial density field or the potential. The focus of Lucie-Smith et al. (2020) is to examine how different features of the initial conditions influence mass predictions within a framework that mirrors analytical models.

The comparison between our methodology and Lucie-Smith et al. (2020) in Fig. 3.8 highlights the differing outcomes that arise from the unique objectives and constraints each model employs. While both models ultimately predict halo masses, we suggest that our approach benefits from the rigid operator that groups particles together and assigns them the same halo mass. Therefore, analytical approaches towards predicting the formation of structures may benefit from knowing about the fate of neighbouring particles. Since in excursion set formalisms, this is only possible to a limited degree, this increases the motivation for considering alternative approaches, like the one proposed by Musso and Sheth (2023a).

In Appendix J, we include a comparison of our instance model with the predictions of ExSHalos (Voivodic et al., 2019). In Fig. J.1, we show a map-level comparison between the Lagrangian shapes of friends-of-friends proto-haloes and ExSHalos predictions. The shapes of proto-haloes predicted by the ExSHalos implementation are limited to sphere-like volumes, which affects its flexibility and, consequently, its accuracy (see Table 3.2). While

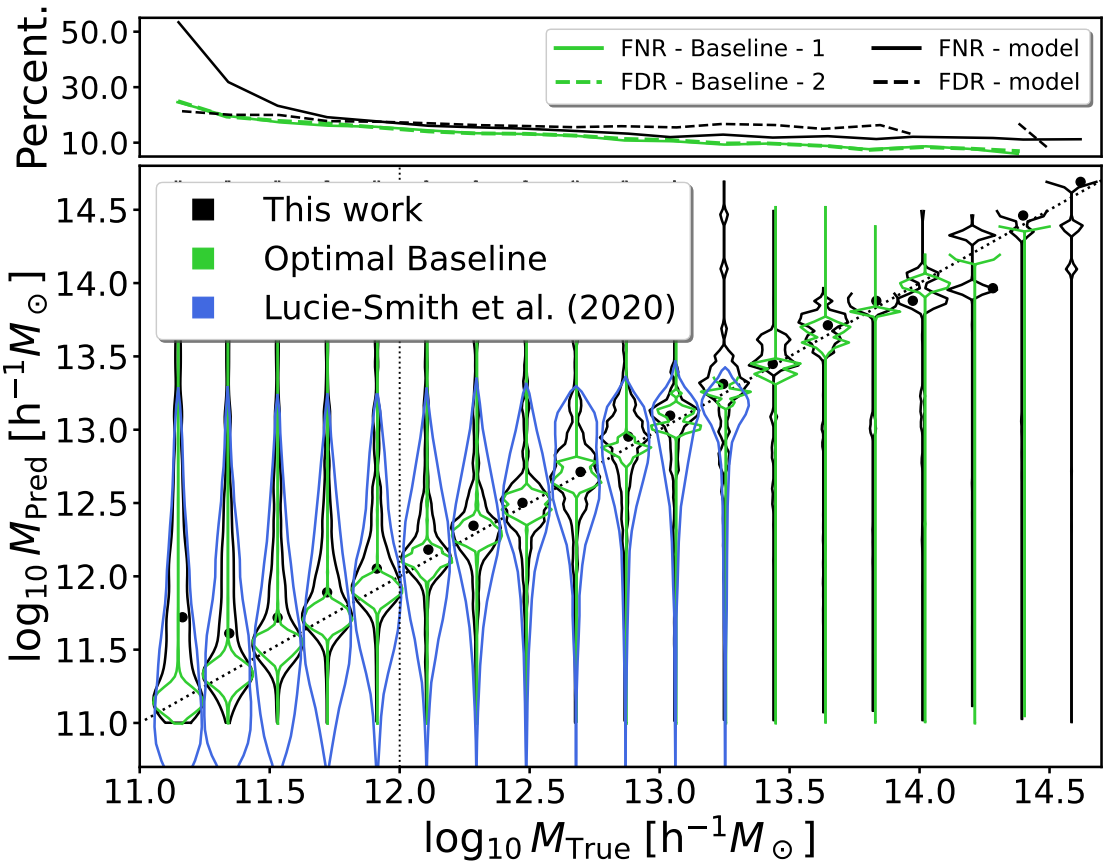

Figure 3.8: "Violin plot", visualizing the distribution of predicted halo masses (at a voxel level) for different ground-truth mass bins. The black violin plots show the results obtained with our instance segmentation model. Green violin plots show the agreement between the two baseline simulations – representing an optimal target accuracy. The blue violin plots in the main panel show the results presented in (Lucie-Smith et al., 2020). The solid black line in the top panel shows the false negative rate, FNR, as a function of the ground truth halo mass. The dashed black line represents the fraction of predicted collapsed pixels that are not actually collapsed as a function of predicted halo mass (false discovery rate, FDR). The green lines on the top panel correspond to the analogous results obtained from the baseline simulations. The model predicts haloes accurately object-by-object for masses $M \gtrsim 10^{12} M_\odot/h$.

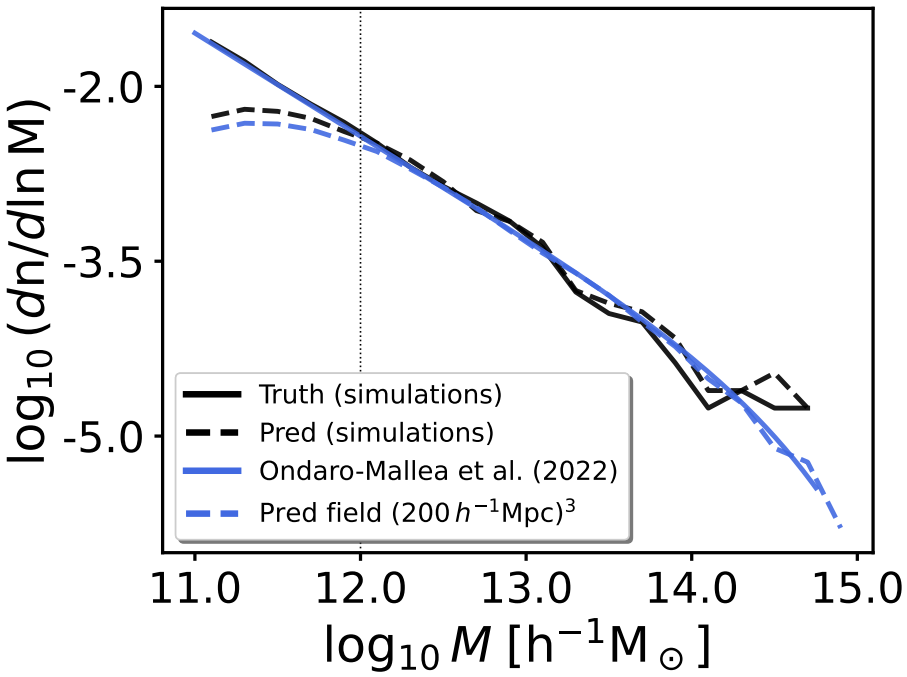

Figure 3.9: Halo-mass-function (HMF) computed using our N-body simulations reserved for validation (solid black line). The dashed black line represents the predicted HMF using the Lagrangian halo regions obtained with our instance segmentation pipeline. The solid blue line shows the HMF prediction from (Ondaro-Mallea et al., 2022). The dashed blue line corresponds to the HMF obtained after evaluating our model in a simulation with 1024^3 particles and $V_{\rm box} = (200\,h^{-1}{\rm Mpc})^3$.

ExSHALOS correctly replicates the halo mass function of friends-of-friends haloes, it struggles to reproduce particle-level mass predictions, as shown in the violin plot in Fig. J.2.

In Fig. 3.9 we present the halo-mass-function (HMF) computed using the validation simulations (solid black line). The dashed black line shows the predicted HMF computed using the results of our instance segmentation pipeline. We can appreciate that our predictions reproduce the N-body results over a range that spans more than two orders of magnitude. Our results improve upon the prediction mass range for the HMF of previous similar approaches (Berger and Stein, 2019; Bernardini et al., 2020). This is despite the fact that Bernardini et al. (2020) select their hyper-parameters to reproduce the HMF; while in Berger and Stein (2019) they reproduce the HMF corresponding to `Peak Patch` haloes (Stein et al., 2019), instead of the HMF associated with FoF haloes. In Fig. 3.9 we also include a solid blue line representing the theoretical HMF predictions using the model by Ondaro-Mallea et al. (2022). We compare this result with the HMF associated with the haloes predicted by our model using the density and potential fields of a realization with 1024^3 particles and a volume of $V_{\rm box} = (200\,h^{-1}{\rm Mpc})^3$. Both lines show a good agreement in the $10^{12} - 10^{15}\,h^{-1}{\rm M}_\odot$ range.

We conclude that our semantic plus instance segmentation pipeline correctly reproduces

the Lagrangian halo shapes of FoF-haloes spanning a mass range between $10^{12}\,h^{-1}\mathrm{M}_{\odot}$ and $10^{14.7}\,h^{-1}\mathrm{M}_{\odot}$. We have tested the accuracy of our results employing different metrics (presented in several tables and figures). Inferred quantities from our predicted Lagrangian halo regions, such as the predicted halo masses, correctly reproduce the trends computed using N-body simulations and improve upon the results presented in previous studies.

3.4 Experiments

In this section, we test how our network reacts to systematic modifications to the input density field and potential and how well it generalizes to scenarios that lie beyond the trained domain. Therefore, we analyze the response to large-scale density perturbations, to large-scale tidal fields and to changes in the variance of the density field.

3.4.1 Response to large scale densities

We study the response of the haloes to a large-scale over-density such as typically considered in separate universe simulations (Wagner et al., 2015a; Lazeyras et al., 2016; Li et al., 2014). We add a constant δ_ϵ to the input density field $\delta(\vec{q})$ so that the new density field $\delta_*(\vec{q})$ is given by

$$\delta_*(\vec{q}) = \delta(\vec{q}) + \delta_\epsilon, \tag{3.11}$$

and to maintain consistency with Poisson's equation, see Eq. (3.8), we add a quadratic term to the potential:

$$\phi_*(\vec{q}) = \phi(\vec{q}) + \frac{\delta_\epsilon}{6}(\vec{q} - \vec{q}_0)^2 \tag{3.12}$$

where $\vec{q}_0$ is an arbitrary (and irrelevant) reference point (Stücker et al., 2021a), which we choose to be in the centre of our considered domain. Note that we break the periodic boundary conditions here, so it is difficult to do this operation for the whole box, but instead we consider it only for a smaller region to avoid boundary effects.

We show how haloes respond to this modification in Fig. 3.10. The middle panel shows the predicted masses associated with the particles/voxels (in a similar way to Fig. 3.7) for the reference field, $\delta_\epsilon = 0$. The upper and lower panels show the results of including a constant term to the initial over-density field of $\delta_\epsilon = -0.5$ and $\delta_\epsilon = 0.5$, respectively.

Increases in the background density lead to more mass collapsing onto haloes, thus generally increasing the Lagrangian volume of haloes. Furthermore, it leads in many cases to previously individual haloes merging into one bigger structure. This is qualitatively consistent with what is observed in separate universe simulations (e.g. Dai et al., 2015; Wagner et al.,

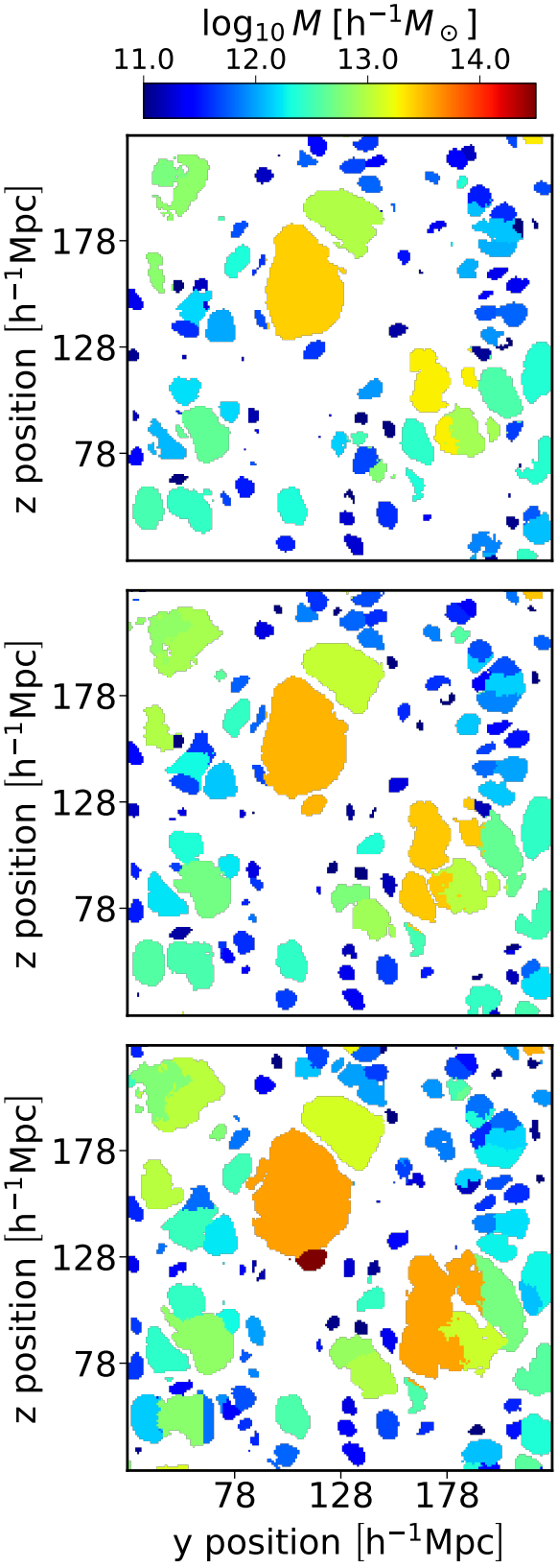

Figure 3.10: Response of proto-haloes to large-scale over-densities. The three panels show over-densities of $\delta_\epsilon = -0.5$, 0 and 0.5 respectively. A larger large-scale density tends to increase the Lagrangian volume of haloes and leads to additional mergers in some cases.

2015b; Barreira et al., 2019; Jamieson and Loverde, 2019; Terasawa et al., 2022; Artigas et al., 2022).

To evaluate quantitatively whether the model has learned the correct response to large-scale density perturbations, we test whether it recovers the same halo bias that has been measured in previous studies (Desjacques et al., 2018b, for a review). In separate universe experiments, the linear bias parameter can be inferred as the derivative of the halo mass function with respect to the large-scale density:

$$b_{1L}(M) = \frac{1}{n_h(M)} \frac{\partial n_h(M)}{\partial \delta_\epsilon} \tag{3.13}$$

Therefore, (Lazeyras et al., 2016) used the halo mass function measured in separate universe simulations with different large-scale densities δ_ϵ to measure the bias parameters through a finite differences approach. While our qualitative experiment from Figure 3.11 follows this in spirit, it is difficult to do the same measurement here, since the addition of the quadratic potential term in equation (3.12) breaks the periodic boundary conditions and makes it difficult to measure the mass function reliably over a large domain. Therefore, we instead adopt an approach to infer the bias from the unperturbed $\delta_\epsilon = 0$ case. (Paranjape et al., 2013) shows that the Lagrangian bias parameter can be measured by considering the (smoothed) linear over-density at the Lagrangian location of biased tracers δ_i:

$$b_{1L} = \frac{1}{N} \sum \frac{\delta_i}{\sigma^2} \tag{3.14}$$

where the sum goes over N different tracers (e.g. all haloes in a given mass bin) and where $\sigma^2 = \langle \delta^2 \rangle$ is the variance of the (smoothed) linear density field. Since this measurement should give meaningful results only on reasonably large scales, we smooth the Lagrangian density field with a Gaussian kernel with width $\sigma_r = 6h^{-1}\mathrm{Mpc}$. We measure the smoothed linear density δ_i at the Lagrangian centre of mass of each halo patch and then we measure the bias by evaluating equation (3.14) in different mass bins.

We show the resulting b_{1L} as a function of mass in Figure 3.11. The blue solid and dashed lines show the bias parameters measured in an $L = 50h^{-1}$Mpc box for the simulated versus predicted halo patches respectively. These two seem consistent, showing that the model has correctly learned the bias relation that is captured inside of the training set. However, this ($L = 50h^{-1}$Mpc) relation is not consistent with the well-measured relation from larger scale simulations, indicated as a black solid line adopted from (Lazeyras et al., 2016). This is because very massive haloes $M \gg 10^{14}h^{-1}M_\odot$ do not form in simulations of such a small volume, but they are important to get the correct bias of smaller mass haloes, since wherever a large halo forms, no smaller halo can form. Our network has never seen such large scales, so it is questionable whether it has any chance of capturing the large-scale bias correctly. However, it might be that what it has learned in the small-scale simulation transfers to larger

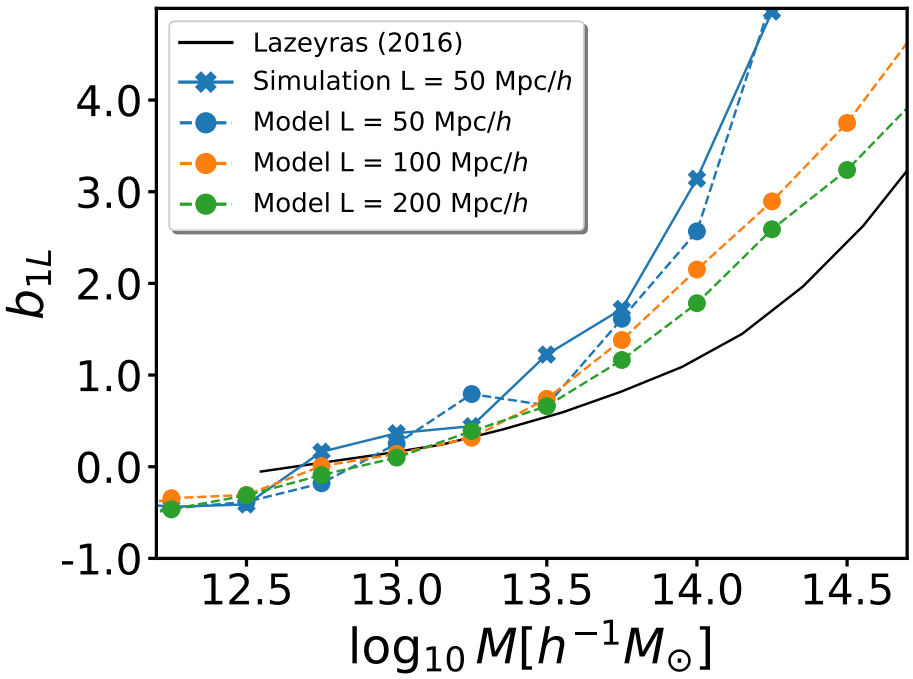

Figure 3.11: Linear Lagrangian bias parameter b_{1L} for the haloes, measured for different boxsizes L and comparing simulation and model. The model agrees well with the simulation at the $L = 50h^{-1}$Mpc scale, but both are inconsistent with the true large-scale bias relation from (Lazeyras et al., 2016) due to effects from the limited size of the simulation volume. Evaluation on larger boxes moves the prediction closer to the known relation, but some deviation is maintained.

scales. To test this, we evaluate the network on two larger boxes, $L = 100h^{-1}$Mpc and $L = 200h^{-1}$Mpc, shown as orange and green lines in Figure 3.11. These cases match the true bias relation better, but still show some significant deviation e.g. at $M \sim 10^{14}h^{-1}M_\odot$. Therefore, we conclude that the network generalizes only moderately well to larger scales and halo masses. Improved performance could possibly be achieved by extending the training set to larger simulations and by increasing the field of view of the network.

3.4.2 Response to large scale tidal fields

In a second experiment, we want to study the response of haloes to purely anisotropic changes of the initial conditions, by adding a large-scale tidal field. We, therefore, aim to emulate a modification similar to the ones considered in anisotropic separate universe simulations (Schmidt et al., 2018; Stücker et al., 2021b; Masaki et al., 2020; Akitsu et al., 2021). We modify the input potential through the term

$$\phi_*(\vec{q}) = \phi(\vec{q}) + \frac{1}{2}(\vec{q} - \vec{q}_0)^T T(\vec{q} - \vec{q}_0) \tag{3.15}$$

$$T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\lambda_z & 0 \\ 0 & 0 & \lambda_z \end{pmatrix} \tag{3.16}$$

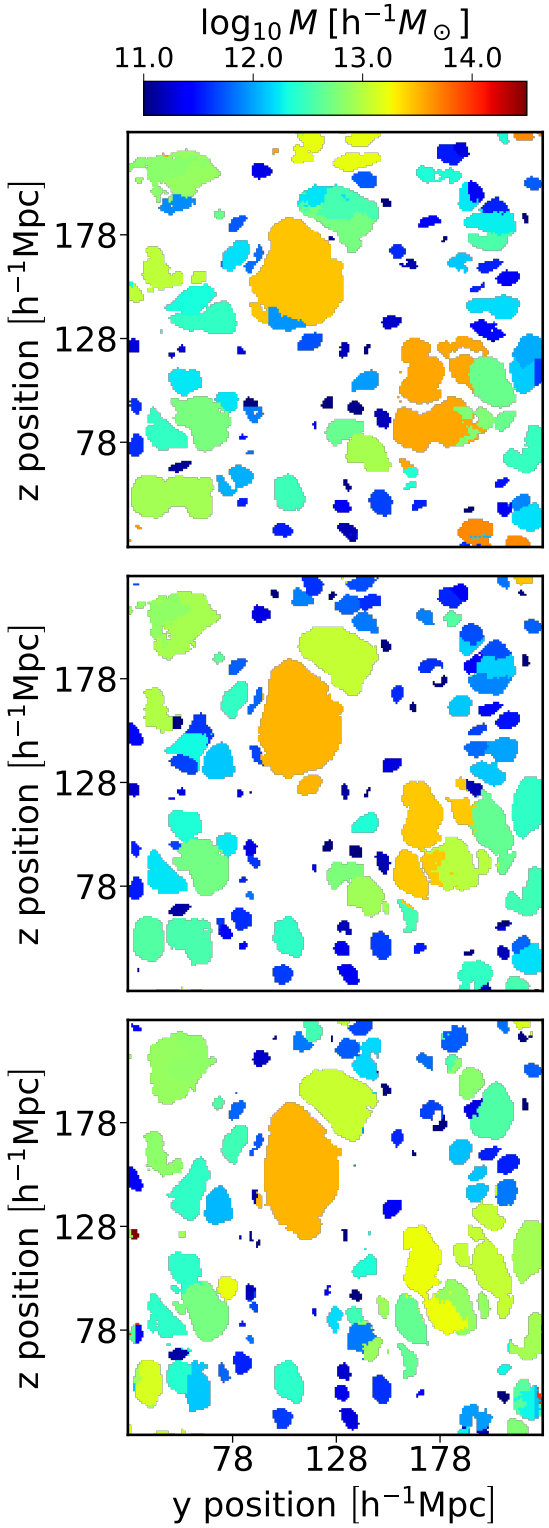

Figure 3.12: Response of proto-halo regions towards a large-scale tidal field. The different panels show the cases with $\lambda_z = -0.5$, 0 and 0.5 – corresponding to a stretching tidal field, no tidal field and a compressing tidal field in the vertical direction respectively. A negative (stretching) tidal field delays infall and shrinks the proto-halo patches in the corresponding direction, whereas a positive (compressing) tidal field facilitates infall and extends the proto-halo patches.

Since we are considering a trace-free tidal tensor, we do not need to include any modifications to the initial density field. The results of introducing the tidal field are presented Fig. 3.12. In the upper panel in which we have imposed a value of $\lambda_z = -0.5$, the regions of typical proto-haloes are slightly reduced in the z-direction and extended in the y-direction. Further, in some cases haloes merge additionally in the y-direction while separating in the z-direction. In the bottom panel with $\lambda_z = 0.5$ we observe the opposite behaviour, with proto-halo shapes elongated in the z-direction and reduced in the y-direction. These observations are consistent with the naive expectation: A positive λ_z means a contracting tidal field in the z-direction, which facilitates infall in this direction, whereas a negative λ_z delays the infall. Therefore, proto-haloes appear extended in the direction where the tidal field has a contracting effect. This should not be confused with the response of the halo shapes in Eulerian space which has the opposite behaviour – reducing the halo's extent in the direction where the tidal field is contracting (Stücker et al., 2021b). Therefore, a large-scale tidal field effects that *the direction from which more material falls in, is the direction where the final halo is less extended*.

However, by comparing Figures 3.10 and 3.12, we note that the effect of modifying the eigenvalues of the tidal tensor (while keeping the trace fixed) is much less significant than modifying its trace δ by a similar amount. Modifying δ leads to strong differences in the abundance and the masses of haloes whereas the modifications to the tidal field strongly affect the shapes, but has a much smaller effect on typical masses – if at all.

Our investigation into the role of anisotropic features in the initial conditions complements the findings of Lucie-Smith et al. (2020). They find that anisotropic features of the initial conditions do not significantly enhance halo mass predictions when compared to predictions based on spherical averages. Therefore, they conclude that including anisotropic features would not significantly improve the mass predictions that can be obtained within excursion set frameworks. This observation is consistent with masses not changing significantly when applying a large-scale tidal field. However, we find that anisotropic features are in general important for the formation of structures since they affect which particles become part of which halo.

Finally, we note that the response of the Lagrangian shape of haloes is particularly interesting in the context of tidal torque theory (White, 1984). To predict the angular momentum of haloes, tidal torque theory requires knowledge of both the tidal tensor and the Lagrangian inertia tensor of haloes. Further, it has been argued that the misalignment of tidal field and Lagrangian inertia tensor is a key factor for predicting galaxy properties (Moon and Lee, 2023). Our experiments show that modifications of the tidal tensor itself also trigger modifications of the Lagrangian shape. Precisely understanding this relation would be relevant to correctly predict halo spins from the initial conditions. Note that such responses are inherently absent in most density-based structure formation models (e.g. Press

and Schechter, 1974b; Bond et al., 1991b; Sheth and Tormen, 2002), but could possibly be accounted for by recently proposed approaches based on the Lagrangian potential (Musso and Sheth, 2021a, 2023b).

3.4.3 Response to changes in the variance of the density field

We now study whether our model can generalize to scenarios different from the training set by investigating how it responds to variations in σ_8, deviating 30% from the original Planck Collaboration et al. (2020b) cosmology. We aim to discern if the network, trained on a singular variance setting, has gained enough insight into halo formation to anticipate outcomes considering different values for the variance of the initial density field. These modifications only affect the initial conditions which are fully visible to the network, so it could be possible that the network correctly extrapolates to these scenarios.

In Fig. 3.13 we show how the HMF reacts to changes in σ_8 in comparison to the measured mass functions from Ondaro-Mallea et al. (2022) (solid lines) as a benchmark. Our predictions for the HMF (dashed lines) are generated by taking the average results of 10 different boxes, each one spanning $L = 50h^{-1}$Mpc, with σ_8 values set to 0.5802 (blue lines), 0.8288 (black lines), and 1.077 (red lines). The model's predictions reveal a discrepancy with the anticipated HMF behaviour beneath the threshold of $\sim 10^{12.7}h^{-1}M_{\odot}$ for both $\sigma_8 \approx 0.5802$, and $\sigma_8 \approx 1.077$. This discrepancy is attributed to the model's training on datasets characterized by the specific σ_8 from Planck Collaboration et al. (2020b). The model's ability to extrapolate to different variances remains limited. At higher masses, however, the network's predictions correspond more closely with the expected HMF. This partial alignment suggests that the network possesses some degree of generalization capability. Nonetheless, for reliable application across varying cosmologies, incorporating these scenarios into the training set is essential.

3.5 Discussion & Conclusions

We present a novel approach to understand and predict halo formation from the initial conditions employed in N-body simulations. Benchmark tests indicate that our model can predict Lagrangian FoF-halo regions for simulations efficiently, taking around 7 minutes in a GPU for a simulation with 256^3 particles in a volume of $50h^{-1}\mathrm{Mpc}$. For those interested in leveraging or further enhancing our work, we have made our codes publicly available: `https://github.com/daniellopezcano/instance_halos`.

Our model consists of a semantic network that reliably recognizes regions in Lagrangian space where haloes form, and an instance segmentation network, that identifies individual haloes from the semantic output. Our predictions accurately reproduce simulation results

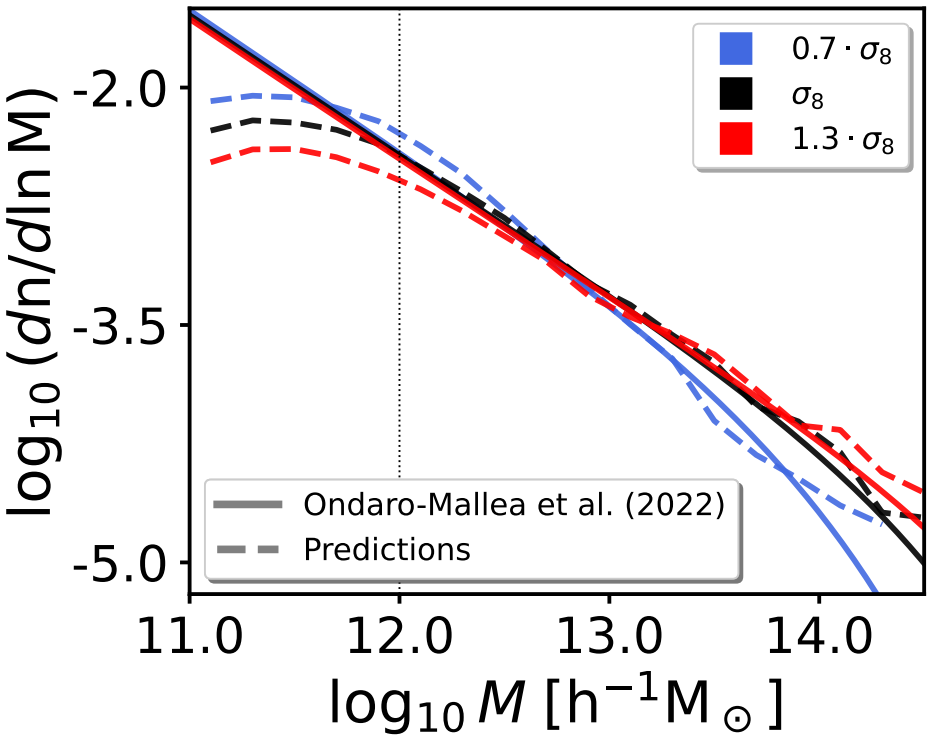

Figure 3.13: Comparison of HMF predictions with variations in the cosmological parameter σ_8. Solid lines represent HMF predictions from (Ondaro-Mallea et al., 2022). Dashed lines indicate our model's predictions. Blue and red curves correspond to scenarios with $\sigma_8 = 0.5802$ and $\sigma_8 = 1.077$ respectively. Black lines show the results for $\sigma_8 = 0.8288$ (our reference cosmology).

and outperform traditional analytical, semi-analytical techniques, and prior ML methods.

The foundation for our instance segmentation model is the Weinberger approach, first introduced by De Brabandere et al. (2017). This technique lets us develop a more general framework for identifying Lagrangian halo patches than previous attempts. Employing the Weinberger loss approach, we bypass some limitations of other instance segmentation methods, like the watershed technique employed by Bernardini et al. (2020). With our approach, we manage to predict the complicated Lagrangian shapes of haloes that are formed in N-body simulations. This is notably more difficult than the predictions of spherical Peak-Patch-haloes that were considered by Berger and Stein (2019).

Additionally, we quantify in how far halo formation is indetermined by the resolved scales of the initial conditions, to establish an optimal performance limit of machine learning methods. We infer this limit by comparing two simulations which only differ in their initial conditions realization on scales beyond the resolution level. We find an agreement between our model predictions and reference simulations similar to the agreement between the two 'baseline' simulations. This shows that our model extracts information encoded in the initial conditions close to optimal. We suggest that such reference experiments may also be used as a baseline in other ML studies to establish whether information is extracted optimally.

Upon evaluating our semantic model, we measure an accuracy of 0.864 and an F_1-score of 0.838. Compared to the baseline simulations, which have an accuracy of 0.903 and an F_1-score of 0.884, our model results stand remarkably close, demonstrating its capability to predict halo regions nearly matching N-body simulations' natural variability.

We also assess our instance segmentation network using various metrics. As depicted in Fig. 3.8, our model closely aligns with the baseline across a broad mass range, outperforming previous methods like Lucie-Smith et al. (2020). We speculate that our approach benefits from the physical constraint that different particles that belong to the same halo are assigned the same halo mass. Moreover, the halo mass function (HMF) predictions in Fig. 3.9 closely match the true ground truth values across three orders of magnitude. The visual representations in Fig. 3.7 reinforce our model's precision, faithfully replicating Lagrangian halo patch positions and shapes.

We have tested through experiments how the network reacts to systematic modifications of the initial conditions. We find that the network correctly captures the response to density perturbations at the finite boxsize provided in the training set. However, it struggles to generalize to larger boxsizes and to cosmologies with different amplitudes of the density field σ_8. This can easily be improved by increasing the diversity of the training set.

Further, we have found that our network utilizes information from the potential field that is not encoded in the density field of any finite region. Modifications to a large-scale tidal field are consistent with the same linear density field, but do affect the potential landscape. Our network predicts that such tidal fields affect the Lagrangian shape of haloes in an anisotropic manner which is consistent with the intuitive expectation of how a tidal field accelerates and decelerates the infall anisotropically.

We have demonstrated the robustness of our model in its current applications and we believe it could find potential utility in several other scenarios like crafting emulated merger trees, aiding separate-universe style experiments (e.g. Lazeyras et al., 2016; Stücker et al., 2021b) and informing the development of analytical methods for halo formation (e.g. Musso and Sheth, 2021b, 2023a). Other works such as `MUSCLE-UPS` (Tosone et al., 2021) can also benefit from our semantic predictions alone by informing their algorithm about which particles will collapse into haloes.

Additionally, our model can be used to help understand the development of spin and intrinsic alignments in haloes and galaxies by establishing how tidal fields modify the Lagrangian shapes of haloes. This is a vital ingredient to predict the spin of haloes through tidal torque theory (White, 1984). Also, we can employ our model to predict changes in the Lagrangian regions of halos in combination with the "splice" technique presented by Cadiou et al. (2021). We believe this approach can provide new insights regarding how modifications in the environment of haloes at initial conditions can affect their final properties.

We encourage experts in these fields to use our open-source code as a basis for tackling and exploring these and other related problems.

The models we have presented in this paper can be easily extended to characterize other properties of halos. One possible extension of the model would be to include an additional spatial dimension to our instance network's output to predict final halo concentrations. In this extension of our model, each particle would have associated a concentration prediction whose average (over all particle members of the same halo) would be trained to minimize the mean square error with respect to the true halo concentration.

The findings presented in this work are promising but there exist some aspects of our models that would benefit from further investigation. For instance, extending our methodology to understand other halo properties beyond mass would be a logical next step. It would also be interesting to test our model's performance under a wider variety of simulation conditions, including variations in cosmology and redshift. An additional avenue of exploration might involve delving into capturing intricate structural details, specifically the gap features in the predicted Lagrangian halo regions. Generative Adversarial Networks (GANs) are tools that have demonstrated potential in reproducing data patterns in the context of cosmological simulations (e.g. Rodríguez et al., 2018; Villaescusa-Navarro et al., 2021; Schaurecker et al., 2021; Robles et al., 2022; Nguyen et al., 2023; Zhang et al., 2023). Hence, employing a GAN-like approach might help recreate these gap features, further improving our model's ability to mimic the structures of haloes found in N-body simulations.

In conclusion, this study showcases the potential of machine learning for facilitating the study of halo formation processes in the context of cosmological N-body simulations. We provide a fast model that exploits the available information close to optimally. We hope our approach serves as a useful tool for researchers working with N-body simulations, opening avenues for future advancements.

Summary and Conclusions

The main objective of my thesis has been to progress upon the current methods used to model cosmological theories and to bridge the gap with large-scale structure observations. I have focused on improving the quality of several techniques derived from cosmological simulations and on advancing strategies that allow for more accurate cosmological analyses. The results of my work comprise a set of tools for handling and interpreting simulation data and can be applied to analyze present and future galaxy survey observations. This effort to bring together theoretical models and astronomical observations is of paramount importance to advance our understanding regarding the structure and evolution of our Universe.

In the process of doing my thesis, I have improved several aspects related to the computational efficiency of cosmological simulations. This contribution represents a useful asset to the community, as it allows a broader exploration of cosmological models with varying initial conditions and physical assumptions.

My research has led to several findings concerning different areas:

- The development of realistic galaxy mock catalogs in the context of the UNITSIM-Galaxies project proves the effectiveness of semi-analytical models for populating large cosmological volumes highlighting the value of employing high-resolution gravity-only simulations for survey forecast problems. This combination provides a robust framework for generating galaxy populations with their corresponding physical properties; it also offers insights into galaxy formation processes and clustering mechanics within the large-scale structure of the Universe.

- The exploitation of analytical models such as excursion set theory to accurately predict the internal structure of dark matter haloes. This research illustrates how simple structure formation theories can help us devise fast methods that approximate complex features within haloes and allow for a better understanding of non-linear processes.

- The incorporation of machine learning techniques, specifically instance segmentation methods, in combination with cosmological simulations for capturing the complicated mechanisms that determine halo formation. With this approach, we explore the potential of ML architectures to generate fast and accurate predictions exploiting GPU

acceleration and how these frameworks can be employed to interpret the relevant intervening processes that play a role in complex physical systems.

All these advances weave together into a global cohesive narrative related to understanding distinct aspects of structure formation. The synergies between developing accurate mock galaxy catalogues, capturing with precision internal halo properties, and the exploitation of machine learning algorithms represent a multifaceted approach integrating different aspects crucial to understanding the complex nature of structure formation processes and robustly describing galaxy survey observations. Altogether my work underscores the importance of combining different state-of-the-art techniques, from analytical prescriptions to numerical methods, for improving our knowledge of complicated physical processes that occur in the context of cosmology. The methodologies I have developed provide new tools to accurately simulate the universe's evolution.

There exist many possibilities for continuing and extending my research in the future, here are some selected ideas:

- Refinement of machine learning algorithms to capture broader aspects of cosmological simulations beyond just halo formation, employing neural networks developed in combination with existing codes.
- Expansion of current neural network architectures to model additional halo properties beyond Lagrangian shapes.
- As the field progresses towards larger and more accurate hydrodynamical simulations, machine learning can serve as both an accelerator and an interpretive tool for analyzing the effects of baryonic processes.
- Exploration of the synergies between semi-analytical galaxy formation models and machine learning to create more realistic mock catalogues for upcoming galaxy surveys, thereby providing critical insights into galaxy formation and evolution.
- Enhancement of the scalability of cosmological simulations to enable the analysis of larger volumes with higher resolution.
- Investigation into the integration of novel data analysis methodologies, such as emulators and contrastive learning techniques to extract cosmological information from observations.
- Further application of the excursion set theory in novel contexts, such as the study of cosmic filaments and voids, to better understand their properties and the role they play in conforming cosmic web structures.

In the long run, the implications of my work extend beyond immediate advancements in cosmological simulations. The broader field of observational cosmology moves as a whole towards more ambitious objectives, trying to better understand the microscopic properties of the different components conforming our Universe. This encompasses close objectives such as determining the masses of neutrinos and other more ambitious goals for constraining the equation of state and exact properties of dark matter and dark energy and testing possible scenarios for gravity beyond general relativity.

Thanks to having participated in projects of different nature during my PhD. I have acquired a broad view of the current cosmological landscape that involves the integration of observational data, theoretical models, and advanced computational tools. The advent of next-generation of galaxy surveys and the developments in the field of machine learning have the potential to revolutionize traditional techniques for data exploration and analysis.

This thesis represents a step forward towards addressing several of the current challenges in cosmology, mainly related to connecting numerical simulations with survey observations, however, there are still many challenges ahead of us. The questions raised by my research and the solutions proposed here encourage a broader dialogue within the scientific community bridging theoretical developments, astronomical observations, and computational implementations, all of them aligning towards advancing our understanding of cosmology as a whole.

In conclusion, the contributions of this thesis to the field of cosmology extend beyond the specifics of simulations, Its roots lie in the need to gain a better understanding of our universe and the complicated processes taking place in it. As I look into the future (from the precarious stability that science provides) I hope to have contributed, if only slightly, to push towards unveiling some of the most fundamental principles that constitute the pillars of physics and our understanding of the universe in general.

Resumen y Conclusiones

El objetivo principal de mi tesis ha sido hacer progresar los métodos que actualmente se utilizan para modelar teorías cosmológicas y reducir la distancia que las separa de las observaciones astronómicas acerca de las estructuras a gran escala de nuestro Universo. Me he centrado en mejorar la calidad de varias técnicas derivadas de simulaciones cosmológicas y en mejorar ciertas estrategias que permiten realizar análisis cosmológicos más precisos. Los resultados de mi investigación comprenden un conjunto de herramientas que pueden ser empleadas para interpretar los datos de simulaciones computacionales y para analizar observaciones sobre la distribución de galaxias. Este esfuerzo por aunar modelos teóricos y observaciones astronómicas es de suma importancia para avanzar en nuestra comprensión de la estructura y evolución del Universo.

Durante la realización de mi tesis, he mejorado varios aspectos relacionados con la eficiencia computacional de las simulaciones cosmológicas. Esta contribución resulta de capital importancia para la comunidad científica, ya que permite una exploración más amplia de modelos cosmológicos con distintas condiciones iniciales y diferentes suposiciones físicas.

Mi investigación me ha llevado a varias conclusiones relativas a distintos ámbitos:

- El desarrollo de catálogos realistas de galaxias simuladas en el contexto del proyecto *UNITSIM-Galaxies* demuestra la eficacia de los modelos semianalíticos para poblar grandes volúmenes cosmológicos. En este trabajo se destaca el valor de emplear simulaciones de alta resolución basadas únicamente en la gravedad para predecir la distribución de galaxias a nivel observacional. La combinación de estas herramientas proporciona un marco robusto con el cual generar poblaciones sintéticas de galaxias y sus correspondientes propiedades físicas; también ofrece información acerca de los procesos de formación de galaxias y los mecanismos mediante los cuales las galaxias se agrupan dentro de la estructura a gran escala del Universo.

- Emplear modelos analíticos como la “Excursion Set Theory” resulta de gran utilidad para predecir con precisión la estructura interna de los halos de materia oscura. Esta investigación ilustra cómo las teorías simples de formación de estructuras pueden ayudarnos a idear métodos rápidos que aproximen características complejas dentro de los halos y permitan una mejor comprensión de los procesos gravitacionales no lineales

que llevan a su formación.

- La incorporación de técnicas de aprendizaje automático, en concreto métodos de segmentación de instancias, en combinación con simulaciones cosmológicas puede ayudar a captar los complicados mecanismos que determinan la formación de halos. Con este método exploramos el potencial de las arquitecturas de aprendizaje automático para generar predicciones rápidas y precisas explotando la aceleración por *GPUs*. También estudiamos cómo estas técnicas pueden emplearse para interpretar los complejos procesos físicos que intervienen en la formación de estructuras.

Todos estos avances se encuentran íntimamente relacionados entre sí y están relacionados con distintos aspectos que mejoran nuestra comprensión sobre los porcesos de formación de estructuras en el Universo. La relación entre el desarrollo de catálogos simulados precisos de galaxias, la capacidad para capturar con precisión las propiedades internas de los halos, y la explotación de algoritmos de aprendizaje automático, conforman un enfoque polifacético que empuja nuestra comprensión sobre los procesos de formación de halos y la predicción de la distribución de galaxias. En conjunto, mi trabajo subraya la importancia de combinar diferentes técnicas vanguardistas, desde prescripciones analíticas hasta métodos numéricos, para mejorar nuestro conocimiento de los complicados procesos físicos que tienen lugar en el contexto de la cosmología. Las metodologías que he desarrollado proporcionan nuevas herramientas para simular con precisión la evolución del universo.

Existen muchas posibilidades de continuar y ampliar mi investigación en el futuro; menciono a continuación algunas ideas seleccionadas:

- Perfeccionamiento de algoritmos de aprendizaje automático para captar aspectos más amplios de las simulaciones cosmológicas.

- Ampliación de las arquitecturas actuales de redes neuronales para modelar propiedades adicionales de los halos (más allá de sus formas lagrangianas).

- A medida que el campo avanza hacia simulaciones hidrodinámicas más grandes y precisas, el aprendizaje automático puede servir, tanto como herramienta de aceleración, como para interpretar y analizar el efecto de los bariones.

- Exploración de las sinergias entre los modelos semianalíticos de formación de galaxias y el aprendizaje automático para crear catálogos simulados más realistas para los próximos sondeos de galaxias, proporcionando así conocimientos fundamentales sobre la formación y evolución de las galaxias.

- Mejora de la escalabilidad de las simulaciones cosmológicas para permitir el análisis de volúmenes más grandes con mayor resolución.

- Investigación sobre la integración de nuevas metodologías de análisis de datos, como emuladores y técnicas de "Contastive Learning" para extraer una mayor cantidad de información cosmológica de las observaciones.

- Amplaición de la "Excursion Set Theory" para estudiar la formación de filamentos y vacíos cósmicos y comprender sus propiedades en el contexto de la estructura a gran escala.

Las implicaciones de mi trabajo a largo plazo van más allá de lograr mejorar las simulaciones cosmológicas en sí mismas. En su conjunto, el campo en el cual se enmarca la cosmología observacional, avanza hacia tratar de comprender las propiedades mircroscópicas de los distintos componentes que conforman nuestro Universo. Estas metas comprenden desde objetivos más realistas y cercanos en el tiempo como tratar de determinar las masas de los neutrinos, hasta otros objetivos más ambiciosos como el de restringir la ecuación de estado y las propiedades de la materia oscura y la energía oscura, o el de explorar otros escenarios para la teoría de la gravedad distintos al de la relatividad general.

Gracias a haber estado involucrado en proyectos de distinta naturaleza durante mi doctorado, he adquirido una amplia visión del campo de la cosmología donde es necesario integrar datos observacionales con modelos teóricos haciendo uso de herramientas computacionales avanzadas. La llegada de la próxima generación de experimentos para la recogida de datos sobre posiciones de galaxias, y los avances en el campo del aprendizaje automático, tienen el potencial de revolucionar las técnicas tradicionales empleadas para la exploración y el análisis de datos.

Esta tesis supone un paso adelante necesario para abordar distintos retos de la cosmología observacional y computacional actual; en particular sobre nuestro conocimiento a cerca de la conexión entre las simulaciones numéricas y las observaciones de galaxias. Las preguntas planteadas por mi trabajo y las soluciones propuestas fomentan un diálogo más amplio dentro de la comunidad científica, tendiendo puentes entre los desarrollos teóricos, las observaciones astronómicas y las implementaciones computacionales, todos ellos alineados para avanzar en nuestra comprensión de la cosmología en su conjunto.

En conclusión, las aportaciones de esta tesis al campo de la cosmología van más allá de las especificidades de las simulaciones, sus raíces se encuentran en la necesidad de comprender mejor nuestro Universo y los complicados procesos que tienen lugar en él. Mirando hacia el futuro (desde la precaria estabilidad que proporciona la ciencia) espero haber contribuido, aunque sea humildemente, a desvelar algunos de los principios fundamentales que constituyen los cimientos de la física y que conforman nuestra concepción del Universo.

Agradecimientos

A la hora de escribir la tesis, una de las tradiciones que más sentido tiene a día de hoy es la de escribir una sección de agradecimientos en la cual apreciar el cariño y el apoyo de todas las personas que te han rodeado durante estos años. Si bien es cierto que he afrontado la escritura de esta tesis como un ejercicio burocrático impersonal y frío, es bonito poder dedicar unas palabras sinceras de agradecimiento a quienes me han acompañado este tiempo y me han brindado su apoyo de forma directa o indirecta.

En primer lugar, quiero agradecer el arropo absoluto de toda mi familia. Saber que hay gente que te quiere y que va a estar ahí, incondicionalmente, en cualquier momento y para lo que necesites, es algo que he dado por hecho durante toda mi vida, pero no por ello tiene menos valor, sino todo lo contrario, es algo que agradecer y tener siempre presente.

Gracias a todos. Gracias de verdad. Gracias a mi padre, gracias a mi madre, gracias a mis abuelos, y a mis tías y a mis tíos, gracias a mis primas, a mis primos, y a todos los sobrinos.

Gracias por las navidades y carnavales en Camuñas, por los veranos en Zarzalejo, por las comidas de cumpleaños, por las noches de pútrida y por los penaltis en el patio.

Gracias por las horas de pesca en la roca lisa, por los viajes a mil ciudades y sus respectivos museos, gracias por subir y bajar conmigo a la Machota, a la cueva de Castrola, a la Mesa de los Tres Reyes, al Teide, y a los lagos de Colomers. Gracias por ayudarme a hacer los deberes en el estudio, gracias por los mojicones a tiempo, la pasión por leer, las tardes lluviosas mirando las tortugas de Atocha, los paseos por el barrio, el Retiro, y el Capricho. Sobre todo, gracias por jugar conmigo, con cualquier cosa, todo el día.

Gracias por sacarme a dar de comer a las ardillas en el parque, por dejarme ayudar a pegar pajitas, gracias por jugar a chocar coches, por las patatas fritas, por hacer los deberes con el brasero encendido, subir a ver al loro Paco, y por ir a echar la quiniela.

Gracias por las excursiones a Villafranca, por hacerme cosquillas, por los juguetes del Carrefour, por enseñarme a hacer quesadas, por compartir derrotas del Getafe pasando un frío horrible con una Coca-Cola entre huevo y huevo, por jugar a la petanca en Quero, por las tortillas, croquetas, gachas, y canelones. Gracias por regarme en verano, por coger aceitunas, uvas, y azafrán, gracias por cantar rancheras en la máquina de coser, gracias por los peluches y por ayudarme a arreglar el telescopio, gracias por las tardes tórridas de piscina, por Alvin y las ardillas, y por la compañía en días duros.

Gracias por pasar horas y horas hablando conmigo sobre cualquier cosa, insuflándome pasión, interés y curiosidad; se podría decir que el hecho de que esté escribiendo esto ahora se debe a todos esos ratos. Gracias por los viajes en verano, por las tardes de juegos de mesa y las comidas ricas. Gracias por hacerme de rabiar y por espabilarme en general, gracias por los días en parques de atracciones, por ir a ver al Joventut de Badalona, y por arrojarme a una poza. Gracias por jugar a la escoba, por ir a escalar, por llevarme por primera vez a visitar un laboratorio. Gracias por dejar que aún os gane al básquet, y por correr conmigo mi primera San Silvestre.

Tengo muchas más cosas que agradeceros, pero estoy alcanzando mi límite de sensiblerías y aún me faltan muchas personas a las que mencionar, así que con esto os tendréis que conformar. En cualquier caso, espero que al leer esto haya conseguido llegaros un poco a la patata.

En segundo lugar, quiero agradecer los buenos ratos a todos mis amigos de un sitio y de otro.

Gracias a Christian, Irene y David por las mil rutas, las caídas en bici, los cletinos amelicanos, ir a coger setas, y arreglar el baño de Cala Reona.

Gracias a Paco, Ángel, Juan, Edgar y Javi por los buenos ratos de cervezas, las vacaciones en Oropesa, los partidos en el Plata y Castañar, y por aguantar mis chapas de física en el instituto.

Gracias a Jaime, Sopena, Pablo, Carlos, Gabi, Eli, Bea Frío, Julia, Bea Mentirosa, Gerardo, De Witt e Íñigo por haber compartido tantos momentos conmigo durante el grado y el máster. Gracias por pasarnos la vida en peceras (que conste que no lo echo de menos), por todas las guerras de aviones en la ruidoteca, por las cosas bien óptimas, la ploteada extrema, la invasión post-sangría, y las comidas en psicología. Gracias por las tardes de juegos de mesa, por no matarnos en GMV, por Cayo Lelio, por la astucia topológicamente protegida, y porque el gravitón tenga spin dos. Gracias por los viajes en tren por la mañana, los miércoles de cine y por escapar de los pavos reales. Gracias por todas las vacaciones, los apuntes y por las risas.

Gracias a todos los amigos que he hecho en Donosti por formar una gran familia dinámica tanto en el tiempo como en el espacio (en gran parte por la tremenda estabilidad que proporciona la ciencia).

Gracias a toda la gente que he conocido en el DIPC: Dani, Rodrigo, Sara, Lucía, Francisco, Matteo, Giovanni, Markos, Lurdes, Marcos, Pina, Leire, Xabi, Martin, Jorge, Cris, Adri, Jonás, Sara, Rodrigo, Antonio, Toni, Marian, Elena, Carol, Kate, Matteo, Nischal, Irene, Sergio, Yetli, Raúl, Nils, Aarón, David, Fer, Nuria, Ying, y seguro que más gente de la cual me acordaré una vez esto ya esté enviado y no se pueda modificar. Gracias también a toda la gente que he conocido a través de amigos en común del BCBL: Ihintza, Irene, Jose, Asier,

Chiara, Coco, Hana, Tomas, Suhail, Sandy, Marta, Bia, Laura, Pablo, Inés, Giulia, Giorgio, Abraham, Vicente, Katerina, Marco... idem con respecto a posibles ausentes. Gracias por las tardes en la playa, por ir conmigo nadando a Santa Clara, por las cervezas en Rebotika, Ikatz, el indio, Polvorín, Mala Gissona... Gracias por las cenas y comidas en casas de unos y de otros, los partidos de basket, jam sessions, rutas por la montaña, sesiones de pintura, tardes de surf... Gracias en general por hacer que estos años de doctorado en Donosti hayan sido geniales. Sé que cuando vuelva aquí la mayoría de vosotros no estaréis (muchos de hecho ya no estáis), pero sé que allá donde estéis tendré un lugar en el mundo que podré visitar gratis gorroneando cama.

Por último, quiero agradecer a todas las personas con las que he estado trabajado estos años su profesionalidad y la amabilidad con la que me habéis tratado. Gracias, Raúl, Jens, Chirag, Daniel y Alexander. Ha sido un placer poder trabajar con científicos tan perspicaces y brillantes como vosotros, me habéis ayudado a madurar como científico y sacar adelante mi doctorado. Gracias a vosotros he aprendido muchas cosas que seguro que me acompañarán a lo largo del resto de mi carrera profesional y estoy seguro de que colaboraremos en el futuro en proyectos interesantes.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., Aleksić, J., Allam, S., Allen, S., Amara, A., Annis, J., Asorey, J., Avila, S., Bacon, D., Balbinot, E., Banerji, M., Banik, N., Barkhouse, W., Baumer, M., Baxter, E., Bechtol, K., Becker, M. R., Benoit-Lévy, A., Benson, B. A., Bernstein, G. M., Bertin, E., Blazek, J., Bridle, S. L., Brooks, D., Brout, D., Buckley-Geer, E., Burke, D. L., Busha, M. T., Campos, A., Capozzi, D., Carnero Rosell, A., Carrasco Kind, M., Carretero, J., Castander, F. J., Cawthon, R., Chang, C., Chen, N., Childress, M., Choi, A., Conselice, C., Crittenden, R., Crocce, M., Cunha, C. E., D'Andrea, C. B., da Costa, L. N., Das, R., Davis, T. M., Davis, C., De Vicente, J., DePoy, D. L., DeRose, J., Desai, S., Diehl, H. T., Dietrich, J. P., Dodelson, S., Doel, P., Drlica-Wagner, A., Eifler, T. F., Elliott, A. E., Elsner, F., Elvin-Poole, J., Estrada, J., Evrard, A. E., Fang, Y., Fernandez, E., Ferté, A., Finley, D. A., Flaugher, B., Fosalba, P., Friedrich, O., Frieman, J., García-Bellido, J., Garcia-Fernandez, M., Gatti, M., Gaztanaga, E., Gerdes, D. W., Giannantonio, T., Gill, M. S. S., Glazebrook, K., Goldstein, D. A., Gruen, D., Gruendl, R. A., Gschwend, J., Gutierrez, G., Hamilton, S., Hartley, W. G., Hinton, S. R., Honscheid, K., Hoyle, B., Huterer, D., Jain, B., James, D. J., Jarvis, M., Jeltema, T., Johnson, M. D., Johnson, M. W. G., Kacprzak, T., Kent, S., Kim, A. G., King, A., Kirk, D., Kokron, N., Kovacs, A., Krause, E., Krawiec, C., Kremin, A., Kuehn, K., Kuhlmann, S., Kuropatkin, N., Lacasa, F., Lahav, O., Li, T. S., Liddle, A. R., Lidman, C., Lima, M., Lin, H., MacCrann, N., Maia, M. A. G., Makler, M., Manera, M., March, M., Marshall, J. L., Martini, P., McMahon, R. G., Melchior, P., Menanteau, F., Miquel, R., Miranda, V., Mudd, D., Muir, J., Möller, A., Neilsen, E., Nichol, R. C., Nord, B., Nugent, P., Ogando, R. L. C., Palmese, A., Peacock, J., Peiris, H. V., Peoples, J., Percival,

W. J., Petravick, D., Plazas, A. A., Porredon, A., Prat, J., Pujol, A., Rau, M. M., Refregier, A., Ricker, P. M., Roe, N., Rollins, R. P., Romer, A. K., Roodman, A., Rosenfeld, R., Ross, A. J., Rozo, E., Rykoff, E. S., Sako, M., Salvador, A. I., Samuroff, S., Sánchez, C., Sanchez, E., Santiago, B., Scarpine, V., Schindler, R., Scolnic, D., Secco, L. F., Serrano, S., Sevilla-Noarbe, I., Sheldon, E., Smith, R. C., Smith, M., Smith, J., Soares-Santos, M., Sobreira, F., Suchyta, E., Tarle, G., Thomas, D., Troxel, M. A., Tucker, D. L., Tucker, B. E., Uddin, S. A., Varga, T. N., Vielzeuf, P., Vikram, V., Vivas, A. K., Walker, A. R., Wang, M., Wechsler, R. H., Weller, J., Wester, W., Wolf, R. C., Yanny, B., Yuan, F., Zenteno, A., Zhang, B., Zhang, Y., Zuntz, J., and Dark Energy Survey Collaboration (2018). Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. *Phys. Rev. D Physical Review D: Particles, Fields, Gravitation & Cosmology*, 98(4):043526.

Abbott, T. M. C., Abdalla, F. B., Allam, S., Amara, A., Annis, J., Asorey, J., Avila, S., Ballester, O., Banerji, M., Barkhouse, W., and et al. (2018). The dark energy survey: Data release 1. *The Astrophysical Journal Supplement Series*, 239(2):18.

Akitsu, K., Li, Y., and Okumura, T. (2021). Cosmological simulation in tides: power spectra, halo shape responses, and shape assembly bias. *Journal of Cosmology and Astroparticle Physics*, 2021(4):041.

Alam, S., Ata, M., Bailey, S., Beutler, F., Bizyaev, D., Blazek, J. A., Bolton, A. S., Brownstein, J. R., Burden, A., Chuang, C.-H., Comparat, J., Cuesta, A. J., Dawson, K. S., Eisenstein, D. J., Escoffier, S., Gil-Marín, H., Grieb, J. N., Hand, N., Ho, S., Kinemuchi, K., Kirkby, D., Kitaura, F., Malanushenko, E., Malanushenko, V., Maraston, C., McBride, C. K., Nichol, R. C., Olmstead, M. D., Oravetz, D., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Pellejero-Ibanez, M., Percival, W. J., Petitjean, P., Prada, F., Price-Whelan, A. M., Reid, B. A., Rodríguez-Torres, S. A., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martín, J. A., Saito, S., Salazar-Albornoz, S., Samushia, L., Sánchez, A. G., Satpathy, S., Schlegel, D. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Sheldon, E. S., Simmons, A., Slosar, A., Strauss, M. A., Swanson, M. E. C., Thomas, D., Tinker, J. L., Tojeiro, R., Magaña, M. V., Vazquez, J. A., Verde, L., Wake, D. A., Wang, Y., Weinberg, D. H., White, M., Wood-Vasey, W. M., Yèche, C., Zehavi, I., Zhai, Z., and Zhao, G.-B. (2017a). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 470(3):2617–2652.

Alam, S., Ata, M., Bailey, S., Beutler, F., Bizyaev, D., Blazek, J. A., Bolton, A. S., Brownstein, J. R., Burden, A., Chuang, C.-H., Comparat, J., Cuesta, A. J., Dawson, K. S., Eisenstein,

D. J., Escoffier, S., Gil-Marín, H., Grieb, J. N., Hand, N., Ho, S., Kinemuchi, K., Kirkby, D., Kitaura, F., Malanushenko, E., Malanushenko, V., Maraston, C., McBride, C. K., Nichol, R. C., Olmstead, M. D., Oravetz, D., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Pellejero-Ibanez, M., Percival, W. J., Petitjean, P., Prada, F., Price-Whelan, A. M., Reid, B. A., Rodríguez-Torres, S. A., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martín, J. A., Saito, S., Salazar-Albornoz, S., Samushia, L., Sánchez, A. G., Satpathy, S., Schlegel, D. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Sheldon, E. S., Simmons, A., Slosar, A., Strauss, M. A., Swanson, M. E. C., Thomas, D., Tinker, J. L., Tojeiro, R., Magaña, M. V., Vazquez, J. A., Verde, L., Wake, D. A., Wang, Y., Weinberg, D. H., White, M., Wood-Vasey, W. M., Yèche, C., Zehavi, I., Zhai, Z., and Zhao, G.-B. (2017b). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 470(3):2617–2652.

Alam, S., Aubert, M., Avila, S., Balland, C., Bautista, J. E., Bershady, M. A., Bizyaev, D., Blanton, M. R., Bolton, A. S., Bovy, J., Brinkmann, J., Brownstein, J. R., Burtin, E., Chabanier, S., Chapman, M. J., Choi, P. D., Chuang, C.-H., Comparat, J., Cousinou, M.-C., Cuceu, A., Dawson, K. S., de la Torre, S., de Mattia, A., Agathe, V. d. S., des Bourboux, H. d. M., Escoffier, S., Etourneau, T., Farr, J., Font-Ribera, A., Frinchaboy, P. M., Fromenteau, S., Gil-Marín, H., Le Goff, J.-M., Gonzalez-Morales, A. X., Gonzalez-Perez, V., Grabowski, K., Guy, J., Hawken, A. J., Hou, J., Kong, H., Parker, J., Klaene, M., Kneib, J.-P., Lin, S., Long, D., Lyke, B. W., de la Macorra, A., Martini, P., Masters, K., Mohammad, F. G., Moon, J., Mueller, E.-M., Muñoz-Gutiérrez, A., Myers, A. D., Nadathur, S., Neveux, R., Newman, J. A., Noterdaeme, P., Oravetz, A., Oravetz, D., Palanque-Delabrouille, N., Pan, K., Paviot, R., Percival, W. J., Pérez-Ràfols, I., Petitjean, P., Pieri, M. M., Prakash, A., Raichoor, A., Ravoux, C., Rezaie, M., Rich, J., Ross, A. J., Rossi, G., Ruggeri, R., Ruhlmann-Kleider, V., Sánchez, A. G., Sánchez, F. J., Sánchez-Gallego, J. R., Sayres, C., Schneider, D. P., Seo, H.-J., Shafieloo, A., Slosar, A., Smith, A., Stermer, J., Tamone, A., Tinker, J. L., Tojeiro, R., Vargas-Magaña, M., Variu, A., Wang, Y., Weaver, B. A., Weijmans, A.-M., Yèche, C., Zarrouk, P., Zhao, C., Zhao, G.-B., and Zheng, Z. (2021a). Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Cosmological implications from two decades of spectroscopic surveys at the Apache Point Observatory. *Phys. Rev. D Physical Review D: Particles, Fields, Gravitation & Cosmology*, 103(8):083533.

Alam, S., de Mattia, A., Tamone, A., Ávila, S., Peacock, J. A., Gonzalez-Perez, V., Smith, A., Raichoor, A., Ross, A. J., Bautista, J. E., Burtin, E., Comparat, J., Dawson, K. S., du Mas des Bourboux, H., Escoffier, S., Gil-Marín, H., Habib, S., Heitmann, K., Hou,

J., Mohammad, F. G., Mueller, E.-M., Neveux, R., Paviot, R., Percival, W. J., Rossi, G., Ruhlmann-Kleider, V., Tojeiro, R., Vargas Magaña, M., Zhao, C., and Zhao, G.-B. (2021b). The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: N-body mock challenge for the eBOSS emission line galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 504(4):4667–4686.

Ali-Haïmoud, Y. and Bird, S. (2013). An efficient implementation of massive neutrinos in non-linear structure formation simulations. *Monthly Notices of the Royal Astronomical Society*, 428(4):3375–3389.

Allen, M. G., Groves, B. A., Dopita, M. A., Sutherland, R. S., and Kewley, L. J. (2008). The mappings iii library of fast radiative shock models. *The Astrophysical Journal Supplement Series*, 178(1):20–55.

Alonso, D. (2012). CUTE solutions for two-point correlation functions from large cosmological datasets. *ArXiv e-prints*.

Alves de Oliveira, R., Li, Y., Villaescusa-Navarro, F., Ho, S., and Spergel, D. N. (2020). Fast and Accurate Non-Linear Predictions of Universes with Deep Learning. *arXiv e-prints*, page arXiv:2012.00240.

Amendola, L., Appleby, S., Bacon, D., Baker, T., Baldi, M., Bartolo, N., Blanchard, A., Bonvin, C., Borgani, S., and et al. (2013). Cosmology and fundamental physics with the euclid satellite. *Living Reviews in Relativity*, 16(1).

Amorisco, N. C., Nightingale, J., He, Q., Amvrosiadis, A., Cao, X., Cole, S., Etherington, A., Frenk, C. S., Li, R., Massey, R., and Robertson, A. (2021). Halo concentration strengthens dark matter constraints in galaxy-galaxy strong lensing analyses. *arXiv e-prints*, page arXiv:2109.00018.

Andres-San Roman, J. A., Gordillo-Vazquez, C., Franco-Barranco, D., Morato, L., Fernández-Espartero, C. H., Baonza, G., Tagua, A., Vicente-Munuera, P., Palacios, A. M., Gavilán, M. P., et al. (2023). Cartocell, a high-content pipeline for 3d image analysis, unveils cell morphology patterns in epithelia. *bioRxiv*, pages 2023–01.

Angulo, R. E. and Hahn, O. (2022a). Large-scale dark matter simulations. *Living Reviews in Computational Astrophysics*, 8(1):1.

Angulo, R. E. and Hahn, O. (2022b). Large-scale dark matter simulations. *Living Reviews in Computational Astrophysics*, 8(1):1.

Angulo, R. E. and Pontzen, A. (2016a). Cosmological N-body simulations with suppressed variance. *Monthly Notices of the Royal Astronomical Society: Letters*, 462(1):L1–L5.

Angulo, R. E. and Pontzen, A. (2016b). Cosmologicaln-body simulations with suppressed variance. *Monthly Notices of the Royal Astronomical Society: Letters*, 462(1):L1–L5.

Angulo, R. E., Springel, V., White, S. D. M., Jenkins, A., Baugh, C. M., and Frenk, C. S. (2012). Scaling relations for galaxy clusters in the Millennium-XXL simulation. *Monthly Notices of the Royal Astronomical Society*, 426(3):2046–2062.

Angulo, R. E., Springel, V., White, S. D. M., Jenkins, A., Baugh, C. M., and Frenk, C. S. (2012). Scaling relations for galaxy clusters in the Millennium-XXL simulation. *Monthly Notices of the Royal Astronomical Society*, 426(3):2046–2062.

Angulo, R. E. and White, S. D. M. (2010). One simulation to fit them all – changing the background parameters of a cosmological N-body simulation. *Monthly Notices of the Royal Astronomical Society*, 405(1):143–154.

Angulo, R. E., Zennaro, M., Contreras, S., Aricò, G., Pellejero-Ibañez, M., and Stücker, J. (2021). The BACCO simulation project: exploiting the full power of large-scale structure for cosmology. *Monthly Notices of the Royal Astronomical Society*, 507(4):5869–5881.

Arnab, A. and Torr, P. H. S. (2017). Pixelwise Instance Segmentation with a Dynamically Instantiated Network. *arXiv e-prints*, page arXiv:1704.02386.

Artigas, D., Grain, J., and Vennin, V. (2022). Hamiltonian formalism for cosmological perturbations: the separate-universe approach. *Journal of Cosmology and Astroparticle Physics*, 2022(2):001.

Asplund, M., Grevesse, N., Sauval, A. J., and Scott, P. (2009). The Chemical Composition of the Sun. *Annual Review of Astron and Astrophysics*, 47:481–522.

Asquith, R., Pearce, F. R., Almaini, O., Knebe, A., Gonzalez-Perez, V., Benson, A., Blaizot, J., Carretero, J., Castander, F. J., Cattaneo, A., et al. (2018). Cosmic carnage ii: the evolution of the galaxy stellar mass function in observations and galaxy formation models. *Monthly Notices of the Royal Astronomical Society*, 480(1):1197–1210.

Atek, H., Malkan, M., McCarthy, P., Teplitz, H. I., Scarlata, C., Siana, B., Henry, A., Colbert, J. W., Ross, N. R., Bridge, C., Bunker, A. J., Dressler, A., Fosbury, R. A. E., Martin, C., and Shim, H. (2010). The WFC3 Infrared Spectroscopic Parallel (WISP) Survey. *The Astrophysical Journal*, 723(1):104–115.

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., de Mattia, A., Zhao, C., Raichoor, A., Tamone, A., Alam, S., Bautista, J., Bianchi, D., Burtin, E., Chapman, M. J., Chuang, C. H., Comparat, J., Dawson, K., Divers, T., du Mas des Bourboux, H., Gil-Marin, H., Mueller, E. M., Habib, S., Heitmann, K., Ruhlmann-Kleider, V., Padilla, N., Percival, W. J., Ross, A. J., Seo, H. J., Schneider, D. P., and Zhao, G. (2020). The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: exploring the halo occupation distribution model for emission line galaxies. *Monthly Notices of the Royal Astronomical Society*, 499(4):5486–5507.

Bagley, M. B., Scarlata, C., Mehta, V., Teplitz, H., Baronchelli, I., Eisenstein, D. J., Pozzetti, L., Cimatti, A., Rutkowski, M., Wang, Y., and Merson, A. (2020). HST Grism-derived Forecasts for Future Galaxy Redshift Surveys. *The Astrophysical Journal*, 897(1):98.

Bai, M. and Urtasun, R. (2016). Deep Watershed Transform for Instance Segmentation. *arXiv e-prints*, page arXiv:1611.08303.

Baldry, I. K., Glazebrook, K., and Driver, S. P. (2008). On the galaxy stellar mass function, the mass-metallicity relation and the implied baryonic mass function. *Monthly Notices of the Royal Astronomical Society*, 388:945–959.

Bardeen, J. M., Bond, J. R., Kaiser, N., and Szalay, A. S. (1986). The statistics of peaks of Gaussian random fields. *The Astrophysical Journal*, 304:15–61.

Barreira, A., Nelson, D., Pillepich, A., Springel, V., Schmidt, F., Pakmor, R., Hernquist, L., and Vogelsberger, M. (2019). Separate Universe simulations with IllustrisTNG: baryonic effects on power spectrum responses and higher-order statistics. *Monthly Notices of the Royal Astronomical Society*, 488(2):2079–2092.

Bartelmann, M., Perrotta, F., and Baccigalupi, C. (2002). Halo concentrations and weak-lensing number counts in dark energy cosmologies. *Astronomy and Astrophysics*, 396:21–30.

Basilakos, S., Plionis, M., and Ragone-Figueroa, C. (2008). The Halo Mass-Bias Redshift Evolution in the ΛCDM Cosmology. *The Astrophysical Journal*, 678(2):627–634.

Baumann, D. (2018). Tasi lectures on primordial cosmology.

Baumann, D. (2022). *Cosmology*. Cambridge University Press.

Behroozi, P. S., Wechsler, R. H., and Wu, H.-Y. (2012). The rockstar phase-space temporal halo finder and the velocity offsets of cluster cores. *The Astrophysical Journal*, 762(2):109.

Bellstedt, S., Robotham, A. S. G., Driver, S. P., Thorne, J. E., Davies, L. J. M., Holwerda, B. W., Hopkins, A. M., Lara-Lopez, M. A., López-Sánchez, Á. R., and Phillipps, S. (2021). Galaxy And Mass Assembly (GAMA): The inferred mass–metallicity relation from z=0 to 3.5 via forensic SED fitting. *arXiv e-prints*, page arXiv:2102.11514.

Bennett, C. L., Larson, D., Weiland, J. L., Jarosik, N., Hinshaw, G., Odegard, N., Smith, K. M., Hill, R. S., Gold, B., Halpern, M., Komatsu, E., Nolta, M. R., Page, L., Spergel, D. N., Wollack, E., Dunkley, J., Kogut, A., Limon, M., Meyer, S. S., Tucker, G. S., and Wright, E. L. (2013). Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results. *Astrophysical Journal, Supplement*, 208(2):20.

Benson, A. J. (2012). Galacticus: A semi-analytic model of galaxy formation. *New Astronomy*, 17(2):175–197.

Berger, P. and Stein, G. (2019). A volumetric deep Convolutional Neural Network for simulation of mock dark matter halo catalogues. *Monthly Notices of the Royal Astronomical Society*, 482(3):2861–2871.

Berlind, A. A., Weinberg, D. H., Benson, A. J., Baugh, C. M., Cole, S., Davé, R., Frenk, C. S., Jenkins, A., Katz, N., and Lacey, C. G. (2003). The Halo Occupation Distribution and the Physics of Galaxy Formation. *The Astrophysical Journal*, 593:1–25.

Bernardini, M., Mayer, L., Reed, D., and Feldmann, R. (2020). Predicting dark matter halo formation in N-body simulations with deep regression networks. *Monthly Notices of the Royal Astronomical Society*, 496(4):5116–5125.

Betts, J. C., van de Bruck, C., Arnold, C., and Li, B. (2023). Machine Learning and Structure Formation in Modified Gravity. *arXiv e-prints*, page arXiv:2305.02122.

Bhattacharya, S., Habib, S., Heitmann, K., and Vikhlinin, A. (2013). Dark Matter Halo Profiles of Massive Clusters: Theory versus Observations. *The Astrophysical Journal*, 766(1):32.

Bode, P., Ostriker, J. P., and Turok, N. (2001). Halo Formation in Warm Dark Matter Models. *The Astrophysical Journal*, 556(1):93–107.

Bond, J. R., Cole, S., Efstathiou, G., and Kaiser, N. (1991a). Excursion set mass functions for hierarchical Gaussian fluctuations. *The Astrophysical Journal*, 379:440–460.

Bond, J. R., Cole, S., Efstathiou, G., and Kaiser, N. (1991b). Excursion Set Mass Functions for Hierarchical Gaussian Fluctuations. *The Astrophysical Journal*, 379:440.

Brandbyge, J., Rampf, C., Tram, T., Leclercq, F., Fidler, C., and Hannestad, S. (2017). Cosmological N -body simulations including radiation perturbations. *Monthly Notices of the Royal Astronomical Society*, 466(1):L68–L72.

Brown, S. T., McCarthy, I. G., Diemer, B., Font, A. S., Stafford, S. G., and Pfeifer, S. (2020). Connecting the structure of dark matter haloes to the primordial power spectrum. *Monthly Notices of the Royal Astronomical Society*, 495(4):4994–5013.

Brown, S. T., McCarthy, I. G., Stafford, S. G., and Font, A. S. (2022). Towards a universal model for the density profiles of dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 509(4):5685–5701.

Bullock, J. S., Kolatt, T. S., Sigad, Y., Somerville, R. S., Kravtsov, A. V., Klypin, A. A., Primack, J. R., and Dekel, A. (2001). Profiles of dark haloes: evolution, scatter and environment. *Monthly Notices of the Royal Astronomical Society*, 321(3):559–575.

Cadiou, C., Pontzen, A., Peiris, H. V., and Lucie-Smith, L. (2021). The causal effect of environment on halo mass and concentration. *Monthly Notices of the Royal Astronomical Society*, 508(1):1189–1194.

Cardelli, J. A., Clayton, G. C., and Mathis, J. S. (1989). The relationship between infrared, optical, and ultraviolet extinction. *The Astrophysical Journal*, 345:245–256.

Carretero, J., Castander, F. J., Gaztañaga, E., Crocce, M., and Fosalba, P. (2015). An algorithm to build mock galaxy catalogues using MICE simulations. *Monthly Notices of the Royal Astronomical Society*, 447:646–670.

Chacón, J., Vázquez, J. A., and Almaraz, E. (2022). Classification algorithms applied to structure formation simulations. *Astronomy and Computing*, 38:100527.

Chevallier, M. and Polarski, D. (2001). Accelerating Universes with Scaling Dark Matter. *International Journal of Modern Physics D*, 10(2):213–223.

Child, H. L., Habib, S., Heitmann, K., Frontiere, N., Finkel, H., Pope, A., and Morozov, V. (2018). Halo Profiles and the Concentration-Mass Relation for a ΛCDM Universe. *The Astrophysical Journal*, 859(1):55.

Chuang, C.-H., Yepes, G., Kitaura, F.-S., Pellejero-Ibanez, M., Rodríguez-Torres, S., Feng, Y., Metcalf, R. B., Wechsler, R. H., Zhao, C., To, C.-H., Alam, S., Banerjee, A., DeRose, J., Giocoli, C., Knebe, A., and Reyes, G. (2019). UNIT project: Universe N-body simulations for the Investigation of Theoretical models from galaxy surveys. *Monthly Notices of the Royal Astronomical Society*, 487(1):48–59.

Chuang, C.-H., Yepes, G., Kitaura, F.-S., Pellejero-Ibanez, M., Rodríguez-Torres, S., Feng, Y., Metcalf, R. B., Wechsler, R. H., Zhao, C., To, C.-H., and et al. (2019). Unit project: Universe n-body simulations for the investigation of theoretical models from galaxy surveys. *Monthly Notices of the Royal Astronomical Society*, 487(1):48–59.

Cochrane, R. K., Best, P. N., Sobral, D., Smail, I., Wake, D. A., Stott, J. P., and Geach, J. E. (2017). The H α luminosity-dependent clustering of star-forming galaxies from z $\sim$ 0.8 to $\sim$2.2 with HiZELS. *Monthly Notices of the Royal Astronomical Society*, 469(3):2913–2932.

Colbert, J. W., Teplitz, H., Atek, H., Bunker, A., Rafelski, M., Ross, N., Scarlata, C., Bedregal, A. G., Dominguez, A., Dressler, A., Henry, A., Malkan, M., Martin, C. L., Masters, D., McCarthy, P., and Siana, B. (2013). Predicting Future Space Near-IR Grism Surveys Using the WFC3 Infrared Spectroscopic Parallels Survey. *The Astrophysical Journal*, 779(1):34.

Cole, S., Percival, W. J., Peacock, J. A., Norberg, P., Baugh, C. M., Frenk, C. S., Baldry, I., Bland-Hawthorn, J., Bridges, T., Cannon, R., Colless, M., Collins, C., Couch, W., Cross, N. J. G., Dalton, G., Eke, V. R., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Glazebrook, K., Jackson, C., Jenkins, A., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Madgwick, D., Peterson, B. A., Sutherland, W., and Taylor, K. (2005). The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications. *Monthly Notices of the Royal Astronomical Society*, 362(2):505–534.

Collaboration, D., Aghamousa, A., Aguilar, J., Ahlen, S., Alam, S., Allen, L. E., Prieto, C. A., Annis, J., Bailey, S., Balland, C., Ballester, O., Baltay, C., Beaufore, L., Bebek, C., Beers, T. C., Bell, E. F., Bernal, J. L., Besuner, R., Beutler, F., Blake, C., Bleuler, H., Blomqvist, M., Blum, R., Bolton, A. S., Briceno, C., Brooks, D., Brownstein, J. R., Buckley-Geer, E., Burden, A., Burtin, E., Busca, N. G., Cahn, R. N., Cai, Y.-C., Cardiel-Sas, L., Carlberg, R. G., Carton, P.-H., Casas, R., Castander, F. J., Cervantes-Cota, J. L., Claybaugh, T. M., Close, M., Coker, C. T., Cole, S., Comparat, J., Cooper, A. P., Cousinou, M. C., Crocce, M., Cuby, J.-G., Cunningham, D. P., Davis, T. M., Dawson, K. S., de la Macorra, A., Vicente, J. D., Delubac, T., Derwent, M., Dey, A., Dhungana, G., Ding, Z., Doel, P., Duan, Y. T., Ealet, A., Edelstein, J., Eftekharzadeh, S., Eisenstein, D. J., Elliott, A., Escoffier, S., Evatt, M., Fagrelius, P., Fan, X., Fanning, K., Farahi, A., Farihi, J., Favole, G., Feng, Y., Fernandez, E., Findlay, J. R., Finkbeiner, D. P., Fitzpatrick, M. J., Flaugher, B., Flender, S., Font-Ribera, A., Forero-Romero, J. E., Fosalba, P., Frenk, C. S., Fumagalli, M., Gaensicke, B. T., Gallo, G., Garcia-Bellido, J., Gaztanaga, E., Fusillo, N. P. G., Gerard, T., Gershkovich, I., Giannantonio, T., Gillet, D., de Rivera, G. G., Gonzalez-Perez, V., Gott, S., Graur, O., Gutierrez, G., Guy, J., Habib, S., Heetderks, H., Heetderks, I., Heitmann, K.,

Hellwing, W. A., Herrera, D. A., Ho, S., Holland, S., Honscheid, K., Huff, E., Hutchinson, T. A., Huterer, D., Hwang, H. S., Laguna, J. M. I., Ishikawa, Y., Jacobs, D., Jeffrey, N., Jelinsky, P., Jennings, E., Jiang, L., Jimenez, J., Johnson, J., Joyce, R., Jullo, E., Juneau, S., Kama, S., Karcher, A., Karkar, S., Kehoe, R., Kennamer, N., Kent, S., Kilbinger, M., Kim, A. G., Kirkby, D., Kisner, T., Kitanidis, E., Kneib, J.-P., Koposov, S., Kovacs, E., Koyama, K., Kremin, A., Kron, R., Kronig, L., Kueter-Young, A., Lacey, C. G., Lafever, R., Lahav, O., Lambert, A., Lampton, M., Landriau, M., Lang, D., Lauer, T. R., Goff, J.-M. L., Guillou, L. L., Suu, A. L. V., Lee, J. H., Lee, S.-J., Leitner, D., Lesser, M., Levi, M. E., L'Huillier, B., Li, B., Liang, M., Lin, H., Linder, E., Loebman, S. R., Lukić, Z., Ma, J., MacCrann, N., Magneville, C., Makarem, L., Manera, M., Manser, C. J., Marshall, R., Martini, P., Massey, R., Matheson, T., McCauley, J., McDonald, P., McGreer, I. D., Meisner, A., Metcalfe, N., Miller, T. N., Miquel, R., Moustakas, J., Myers, A., Naik, M., Newman, J. A., Nichol, R. C., Nicola, A., da Costa, L. N., Nie, J., Niz, G., Norberg, P., Nord, B., Norman, D., Nugent, P., O'Brien, T., Oh, M., Olsen, K. A. G., Padilla, C., Padmanabhan, H., Padmanabhan, N., Palanque-Delabrouille, N., Palmese, A., Pappalardo, D., Pâris, I., Park, C., Patej, A., Peacock, J. A., Peiris, H. V., Peng, X., Percival, W. J., Perruchot, S., Pieri, M. M., Pogge, R., Pollack, J. E., Poppett, C., Prada, F., Prakash, A., Probst, R. G., Rabinowitz, D., Raichoor, A., Ree, C. H., Refregier, A., Regal, X., Reid, B., Reil, K., Rezaie, M., Rockosi, C. M., Roe, N., Ronayette, S., Roodman, A., Ross, A. J., Ross, N. P., Rossi, G., Rozo, E., Ruhlmann-Kleider, V., Rykoff, E. S., Sabiu, C., Samushia, L., Sanchez, E., Sanchez, J., Schlegel, D. J., Schneider, M., Schubnell, M., Secroun, A., Seljak, U., Seo, H.-J., Serrano, S., Shafieloo, A., Shan, H., Sharples, R., Sholl, M. J., Shourt, W. V., Silber, J. H., Silva, D. R., Sirk, M. M., Slosar, A., Smith, A., Smoot, G. F., Som, D., Song, Y.-S., Sprayberry, D., Staten, R., Stefanik, A., Tarle, G., Tie, S. S., Tinker, J. L., Tojeiro, R., Valdes, F., Valenzuela, O., Valluri, M., Vargas-Magana, M., Verde, L., Walker, A. R., Wang, J., Wang, Y., Weaver, B. A., Weaverdyck, C., Wechsler, R. H., Weinberg, D. H., White, M., Yang, Q., Yeche, C., Zhang, T., Zhao, G.-B., Zheng, Y., Zhou, X., Zhou, Z., Zhu, Y., Zou, H., and Zu, Y. (2016). The desi experiment part i: Science,targeting, and survey design.

Contreras, S., Angulo, R. E., Zennaro, M., Aricò, G., and Pellejero-Ibañez, M. (2020). 3 per cent-accurate predictions for the clustering of dark matter, haloes, and subhaloes, over a wide range of cosmologies and scales. *Monthly Notices of the Royal Astronomical Society*, 499(4):4905–4917.

Contreras, S., Chaves-Montero, J., Zennaro, M., and Angulo, R. E. (2021). The cosmological dependence of halo and galaxy assembly bias. *Monthly Notices of the Royal Astronomical Society*, 507(3):3412–3422.

Correa, C. A., Wyithe, J. S. B., Schaye, J., and Duffy, A. R. (2015). The accretion history of dark matter haloes - III. A physical model for the concentration-mass relation. *Monthly Notices of the Royal Astronomical Society*, 452(2):1217–1232.

Croton, D. J., Springel, V., White, S. D. M., De Lucia, G., Frenk, C. S., Gao, L., Jenkins, A., Kauffmann, G., Navarro, J. F., and Yoshida, N. (2006). The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. *Monthly Notices of the Royal Astronomical Society*, 365:11–28.

Croton, D. J., Stevens, A. R. H., Tonini, C., Garel, T., Bernyk, M., Bibiano, A., Hodkinson, L., Mutch, S. J., Poole, G. B., and Shattow, G. M. (2016). Semi-Analytic Galaxy Evolution (SAGE): Model Calibration and Basic Results. *Astrophysical Journal, Supplement*, 222:22.

Daddi, E., Dickinson, M., Morrison, G., Chary, R., Cimatti, A., Elbaz, D., Frayer, D., Renzini, A., Pope, A., Alexander, D. M., Bauer, F. E., Giavalisco, M., Huynh, M., Kurk, J., and Mignoli, M. (2007). Multiwavelength Study of Massive Galaxies at z$\sim$2. I. Star Formation and Galaxy Growth. *The Astrophysical Journal*, 670(1):156–172.

Dadhley, R. (2015). The we-heraeus international winter school on gravity and light. `https://richie291.wixsite.com/theoreticalphysics/all-notes`.

Dai, L., Pajer, E., and Schmidt, F. (2015). On separate universes. *Journal of Cosmology and Astroparticle Physics*, 2015(10):059–059.

Davidzon, I., Ilbert, O., Laigle, C., Coupon, J., McCracken, H. J., Delvecchio, I., Masters, D., Capak, P., Hsieh, B. C., Le Fèvre, O., Tresse, L., Bethermin, M., Chang, Y. Y., Faisst, A. L., Le Floc'h, E., Steinhardt, C., Toft, S., Aussel, H., Dubois, C., Hasinger, G., Salvato, M., Sanders, D. B., Scoville, N., and Silverman, J. D. (2017). The COSMOS2015 galaxy stellar mass function . Thirteen billion years of stellar mass assembly in ten snapshots. *Astronomy and Astrophysics*, 605:A70.

Davis, M., Efstathiou, G., Frenk, C. S., and White, S. D. M. (1985). The evolution of large-scale structure in a universe dominated by cold dark matter. *The Astrophysical Journal*, 292:371–394.

Dawson, K. S., Kneib, J.-P., Percival, W. J., Alam, S., Albareti, F. D., Anderson, S. F., Armengaud, E., Aubourg, É., Bailey, S., Bautista, J. E., Berlind, A. A., Bershady, M. A., Beutler, F., Bizyaev, D., Blanton, M. R., Blomqvist, M., Bolton, A. S., Bovy, J., Brandt, W. N., Brinkmann, J., Brownstein, J. R., Burtin, E., Busca, N. G., Cai, Z., Chuang, C.-H., Clerc, N., Comparat, J., Cope, F., Croft, R. A. C., Cruz-Gonzalez, I., da Costa, L. N., Cousinou, M.-C., Darling, J., de la Macorra, A., de la Torre, S., Delubac, T., du Mas

des Bourboux, H., Dwelly, T., Ealet, A., Eisenstein, D. J., Eracleous, M., Escoffier, S., Fan, X., Finoguenov, A., Font-Ribera, A., Frinchaboy, P., Gaulme, P., Georgakakis, A., Green, P., Guo, H., Guy, J., Ho, S., Holder, D., Huehnerhoff, J., Hutchinson, T., Jing, Y., Jullo, E., Kamble, V., Kinemuchi, K., Kirkby, D., Kitaura, F.-S., Klaene, M. A., Laher, R. R., Lang, D., Laurent, P., Le Goff, J.-M., Li, C., Liang, Y., Lima, M., Lin, Q., Lin, W., Lin, Y.-T., Long, D. C., Lundgren, B., MacDonald, N., Geimba Maia, M. A., Malanushenko, E., Malanushenko, V., Mariappan, V., McBride, C. K., McGreer, I. D., Ménard, B., Merloni, A., Meza, A., Montero-Dorta, A. D., Muna, D., Myers, A. D., Nandra, K., Naugle, T., Newman, J. A., Noterdaeme, P., Nugent, P., Ogando, R., Olmstead, M. D., Oravetz, A., Oravetz, D. J., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Parejko, J. K., Pâris, I., Peacock, J. A., Petitjean, P., Pieri, M. M., Pisani, A., Prada, F., Prakash, A., Raichoor, A., Reid, B., Rich, J., Ridl, J., Rodriguez-Torres, S., Carnero Rosell, A., Ross, A. J., Rossi, G., Ruan, J., Salvato, M., Sayres, C., Schneider, D. P., Schlegel, D. J., Seljak, U., Seo, H.-J., Sesar, B., Shandera, S., Shu, Y., Slosar, A., Sobreira, F., Streblyanska, A., Suzuki, N., Taylor, D., Tao, C., Tinker, J. L., Tojeiro, R., Vargas-Magaña, M., Wang, Y., Weaver, B. A., Weinberg, D. H., White, M., Wood-Vasey, W. M., Yeche, C., Zhai, Z., Zhao, C., Zhao, G.-b., Zheng, Z., Ben Zhu, G., and Zou, H. (2016). The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. *Astronomical Journal*, 151(2):44.

Dawson, K. S., Schlegel, D. J., Ahn, C. P., Anderson, S. F., Aubourg, É., Bailey, S., Barkhouser, R. H., Bautista, J. E., Beifiori, A., Berlind, A. A., Bhardwaj, V., Bizyaev, D., Blake, C. H., Blanton, M. R., Blomqvist, M., Bolton, A. S., Borde, A., Bovy, J., Brandt, W. N., Brewington, H., Brinkmann, J., Brown, P. J., Brownstein, J. R., Bundy, K., Busca, N. G., Carithers, W., Carnero, A. R., Carr, M. A., Chen, Y., Comparat, J., Connolly, N., Cope, F., Croft, R. A. C., Cuesta, A. J., da Costa, L. N., Davenport, J. R. A., Delubac, T., de Putter, R., Dhital, S., Ealet, A., Ebelke, G. L., Eisenstein, D. J., Escoffier, S., Fan, X., Filiz Ak, N., Finley, H., Font-Ribera, A., Génova-Santos, R., Gunn, J. E., Guo, H., Haggard, D., Hall, P. B., Hamilton, J.-C., Harris, B., Harris, D. W., Ho, S., Hogg, D. W., Holder, D., Honscheid, K., Huehnerhoff, J., Jordan, B., Jordan, W. P., Kauffmann, G., Kazin, E. A., Kirkby, D., Klaene, M. A., Kneib, J.-P., Le Goff, J.-M., Lee, K.-G., Long, D. C., Loomis, C. P., Lundgren, B., Lupton, R. H., Maia, M. A. G., Makler, M., Malanushenko, E., Malanushenko, V., Mandelbaum, R., Manera, M., Maraston, C., Margala, D., Masters, K. L., McBride, C. K., McDonald, P., McGreer, I. D., McMahon, R. G., Mena, O., Miralda-Escudé, J., Montero-Dorta, A. D., Montesano, F., Muna, D., Myers, A. D., Naugle, T., Nichol, R. C., Noterdaeme, P., Nuza, S. E., Olmstead, M. D., Oravetz, A., Oravetz, D. J., Owen, R., Padmanabhan, N., Palanque-Delabrouille, N., Pan,

K., Parejko, J. K., Pâris, I., Percival, W. J., Pérez-Fournon, I., Pérez-Ràfols, I., Petitjean, P., Pfaffenberger, R., Pforr, J., Pieri, M. M., Prada, F., Price-Whelan, A. M., Raddick, M. J., Rebolo, R., Rich, J., Richards, G. T., Rockosi, C. M., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martin, J. A., Samushia, L., Sánchez, A. G., Sayres, C., Schmidt, S. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Shelden, A., Sheldon, E., Shen, Y., Shu, Y., Slosar, A., Smee, S. A., Snedden, S. A., Stauffer, F., Steele, O., Strauss, M. A., Streblyanska, A., Suzuki, N., Swanson, M. E. C., Tal, T., Tanaka, M., Thomas, D., Tinker, J. L., Tojeiro, R., Tremonti, C. A., Vargas Magaña, M., Verde, L., Viel, M., Wake, D. A., Watson, M., Weaver, B. A., Weinberg, D. H., Weiner, B. J., West, A. A., White, M., Wood-Vasey, W. M., Yeche, C., Zehavi, I., Zhao, G.-B., and Zheng, Z. (2013). The Baryon Oscillation Spectroscopic Survey of SDSS-III. *Astronomical Journal*, 145(1):10.

De Barros, S., Reddy, N., and Shivaei, I. (2016). Dust Attenuation of the Nebular Regions of z ∼2 Star-forming Galaxies: Insight from UV, IR, and Emission Lines. *The Astrophysical Journal*, 820(2):96.

De Brabandere, B., Neven, D., and Van Gool, L. (2017). Semantic Instance Segmentation with a Discriminative Loss Function. *arXiv e-prints*, page arXiv:1708.02551.

de Jong, R. S., Bellido-Tirado, O., Chiappini, C., Depagne, É., Haynes, R., Johl, D., Schnurr, O., Schwope, A., Walcher, J., Dionies, F., Haynes, D., Kelz, A., Kitaura, F. S., Lamer, G., Minchev, I., Müller, V., Nuza, S. E., Olaya, J.-C., Piffl, T., Popow, E., Steinmetz, M., Ural, U., Williams, M., Winkler, R., Wisotzki, L., Ansorge, W. R., Banerji, M., Gonzalez Solares, E., Irwin, M., Kennicutt, R. C., King, D., McMahon, R. G., Koposov, S., Parry, I. R., Sun, D., Walton, N. A., Finger, G., Iwert, O., Krumpe, M., Lizon, J.-L., Vincenzo, M., Amans, J.-P., Bonifacio, P., Cohen, M., Francois, P., Jagourel, P., Mignot, S. B., Royer, F., Sartoretti, P., Bender, R., Grupp, F., Hess, H.-J., Lang-Bardl, F., Muschielok, B., Böhringer, H., Boller, T., Bongiorno, A., Brusa, M., Dwelly, T., Merloni, A., Nandra, K., Salvato, M., Pragt, J. H., Navarro, R., Gerlofsma, G., Roelfsema, R., Dalton, G. B., Middleton, K. F., Tosh, I. A., Boeche, C., Caffau, E., Christlieb, N., Grebel, E. K., Hansen, C., Koch, A., Ludwig, H.-G., Quirrenbach, A., Sbordone, L., Seifert, W., Thimm, G., Trifonov, T., Helmi, A., Trager, S. C., Feltzing, S., Korn, A., and Boland, W. (2012). 4MOST: 4-metre multi-object spectroscopic telescope. In McLean, I. S., Ramsay, S. K., and Takami, H., editors, *Ground-based and Airborne Instrumentation for Astronomy IV*, volume 8446 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 84460T.

de los Reyes, M. A., Ly, C., Lee, J. C., Salim, S., Peeples, M. S., Momcheva, I., Feddersen, J., Dale, D. A., Ouchi, M., Ono, Y., and Finn, R. (2015). The Relationship between Stellar

Mass, Gas Metallicity, and Star Formation Rate for Hα-Selected Galaxies at z $\approx$ 0.8 from the NewHα Survey. *Astronomical Journal*, 149(2):79.

De Lucia, G. and Blaizot, J. (2007). The hierarchical formation of the brightest cluster galaxies. *Monthly Notices of the Royal Astronomical Society*, 375:2–14.

de Simone, A., Maggiore, M., and Riotto, A. (2011). Conditional probabilities in the excursion set theory: generic barriers and non-Gaussian initial conditions. *Monthly Notices of the Royal Astronomical Society*, 418(4):2403–2421.

Dekel, A. and Rees, M. J. (1987). Physical mechanisms for biased galaxy formation. *Nature*, 326(6112):455–462.

Deng, R., Shen, C., Liu, S., Wang, H., and Liu, X. (2018). Learning to predict crisp boundaries. *arXiv e-prints*, page arXiv:1807.10097.

DESI Collaboration, Aghamousa, A., Aguilar, J., Ahlen, S., Alam, S., Allen, L. E., Allende Prieto, C., Annis, J., Bailey, S., Balland, C., Ballester, O., Baltay, C., Beaufore, L., Bebek, C., Beers, T. C., Bell, E. F., Bernal, J. L., Besuner, R., Beutler, F., Blake, C., Bleuler, H., Blomqvist, M., Blum, R., Bolton, A. S., Briceno, C., Brooks, D., Brownstein, J. R., Buckley-Geer, E., Burden, A., Burtin, E., Busca, N. G., Cahn, R. N., Cai, Y.-C., Cardiel-Sas, L., Carlberg, R. G., Carton, P.-H., Casas, R., Castander, F. J., Cervantes-Cota, J. L., Claybaugh, T. M., Close, M., Coker, C. T., Cole, S., Comparat, J., Cooper, A. P., Cousinou, M. C., Crocce, M., Cuby, J.-G., Cunningham, D. P., Davis, T. M., Dawson, K. S., de la Macorra, A., De Vicente, J., Delubac, T., Derwent, M., Dey, A., Dhungana, G., Ding, Z., Doel, P., Duan, Y. T., Ealet, A., Edelstein, J., Eftekharzadeh, S., Eisenstein, D. J., Elliott, A., Escoffier, S., Evatt, M., Fagrelius, P., Fan, X., Fanning, K., Farahi, A., Farihi, J., Favole, G., Feng, Y., Fernandez, E., Findlay, J. R., Finkbeiner, D. P., Fitzpatrick, M. J., Flaugher, B., Flender, S., Font-Ribera, A., Forero-Romero, J. E., Fosalba, P., Frenk, C. S., Fumagalli, M., Gaensicke, B. T., Gallo, G., Garcia-Bellido, J., Gaztanaga, E., Pietro Gentile Fusillo, N., Gerard, T., Gershkovich, I., Giannantonio, T., Gillet, D., Gonzalez-de-Rivera, G., Gonzalez-Perez, V., Gott, S., Graur, O., Gutierrez, G., Guy, J., Habib, S., Heetderks, H., Heetderks, I., Heitmann, K., Hellwing, W. A., Herrera, D. A., Ho, S., Holland, S., Honscheid, K., Huff, E., Hutchinson, T. A., Huterer, D., Hwang, H. S., Illa Laguna, J. M., Ishikawa, Y., Jacobs, D., Jeffrey, N., Jelinsky, P., Jennings, E., Jiang, L., Jimenez, J., Johnson, J., Joyce, R., Jullo, E., Juneau, S., Kama, S., Karcher, A., Karkar, S., Kehoe, R., Kennamer, N., Kent, S., Kilbinger, M., Kim, A. G., Kirkby, D., Kisner, T., Kitanidis, E., Kneib, J.-P., Koposov, S., Kovacs, E., Koyama, K., Kremin, A., Kron, R., Kronig, L., Kueter-Young, A., Lacey, C. G., Lafever, R., Lahav, O., Lambert, A., Lampton,

M., Landriau, M., Lang, D., Lauer, T. R., Le Goff, J.-M., Le Guillou, L., Le Van Suu, A., Lee, J. H., Lee, S.-J., Leitner, D., Lesser, M., Levi, M. E., L'Huillier, B., Li, B., Liang, M., Lin, H., Linder, E., Loebman, S. R., Lukić, Z., Ma, J., MacCrann, N., Magneville, C., Makarem, L., Manera, M., Manser, C. J., Marshall, R., Martini, P., Massey, R., Matheson, T., McCauley, J., McDonald, P., McGreer, I. D., Meisner, A., Metcalfe, N., Miller, T. N., Miquel, R., Moustakas, J., Myers, A., Naik, M., Newman, J. A., Nichol, R. C., Nicola, A., Nicolati da Costa, L., Nie, J., Niz, G., Norberg, P., Nord, B., Norman, D., Nugent, P., O'Brien, T., Oh, M., Olsen, K. A. G., Padilla, C., Padmanabhan, H., Padmanabhan, N., Palanque-Delabrouille, N., Palmese, A., Pappalardo, D., Pâris, I., Park, C., Patej, A., Peacock, J. A., Peiris, H. V., Peng, X., Percival, W. J., Perruchot, S., Pieri, M. M., Pogge, R., Pollack, J. E., Poppett, C., Prada, F., Prakash, A., Probst, R. G., Rabinowitz, D., Raichoor, A., Ree, C. H., Refregier, A., Regal, X., Reid, B., Reil, K., Rezaie, M., Rockosi, C. M., Roe, N., Ronayette, S., Roodman, A., Ross, A. J., Ross, N. P., Rossi, G., Rozo, E., Ruhlmann-Kleider, V., Rykoff, E. S., Sabiu, C., Samushia, L., Sanchez, E., Sanchez, J., Schlegel, D. J., Schneider, M., Schubnell, M., Secroun, A., Seljak, U., Seo, H.-J., Serrano, S., Shafieloo, A., Shan, H., Sharples, R., Sholl, M. J., Shourt, W. V., Silber, J. H., Silva, D. R., Sirk, M. M., Slosar, A., Smith, A., Smoot, G. F., Som, D., Song, Y.-S., Sprayberry, D., Staten, R., Stefanik, A., Tarle, G., Sien Tie, S., Tinker, J. L., Tojeiro, R., Valdes, F., Valenzuela, O., Valluri, M., Vargas-Magana, M., Verde, L., Walker, A. R., Wang, J., Wang, Y., Weaver, B. A., Weaverdyck, C., Wechsler, R. H., Weinberg, D. H., White, M., Yang, Q., Yeche, C., Zhang, T., Zhao, G.-B., Zheng, Y., Zhou, X., Zhou, Z., Zhu, Y., Zou, H., and Zu, Y. (2016). The DESI Experiment Part I: Science,Targeting, and Survey Design. *arXiv e-prints*, page arXiv:1611.00036.

Desjacques, V., Jeong, D., and Schmidt, F. (2018a). Large-scale galaxy bias. *Physics Reports*, 733:1–193.

Desjacques, V., Jeong, D., and Schmidt, F. (2018b). Large-scale galaxy bias. *Physics Reports*, 733:1–193.

Despali, G., Vegetti, S., White, S. D. M., Giocoli, C., and van den Bosch, F. C. (2018). Modelling the line-of-sight contribution in substructure lensing. *Monthly Notices of the Royal Astronomical Society*, 475(4):5424–5442.

Diemer, B. and Joyce, M. (2019). An Accurate Physical Model for Halo Concentrations. *The Astrophysical Journal*, 871(2):168.

Dodelson, S. and Schmidt, F. (2020). *Modern Cosmology*.

Dolag, K., Bartelmann, M., Perrotta, F., Baccigalupi, C., Moscardini, L., Meneghetti, M., and Tormen, G. (2004). Numerical study of halo concentrations in dark-energy cosmologies. *Astronomy and Astrophysics*, 416:853–864.

Dolgov, A. D., Hansen, S. H., and Semikoz, D. V. (1997). Non-equilibrium corrections to the spectra of massless neutrinos in the early universe. *Nuclear Physics B*, 503:426–444.

Draine, B. T. (2003). Interstellar Dust Grains. *Annual Review of Astronomy and Astrophysics*, 41:241–289.

Drinkwater, M. J., Jurek, R. J., Blake, C., Woods, D., Pimbblet, K. A., Glazebrook, K., Sharp, R., Pracy, M. B., Brough, S., Colless, M., and et al. (2010). The wigglez dark energy survey: survey design and first data release. *Monthly Notices of the Royal Astronomical Society*, 401(3):1429–1452.

Dutton, A. A. and Macciò, A. V. (2014). Cold dark matter haloes in the Planck era: evolution of structural parameters for Einasto and NFW profiles. *Monthly Notices of the Royal Astronomical Society*, 441(4):3359–3374.

Einasto, J. (1965). On the Construction of a Composite Model for the Galaxy and on the Determination of the System of Galactic Parameters. *Trudy Astrofizicheskogo Instituta Alma-Ata*, 5:87–100.

Eisenstein, D. J. and Hut, P. (1998). HOP: A New Group-Finding Algorithm for N-Body Simulations. *The Astrophysical Journal*, 498(1):137–142.

Eisenstein, D. J., Zehavi, I., Hogg, D. W., Scoccimarro, R., Blanton, M. R., Nichol, R. C., Scranton, R., Seo, H.-J., Tegmark, M., Zheng, Z., Anderson, S. F., Annis, J., Bahcall, N., Brinkmann, J., Burles, S., Castander, F. J., Connolly, A., Csabai, I., Doi, M., Fukugita, M., Frieman, J. A., Glazebrook, K., Gunn, J. E., Hendry, J. S., Hennessy, G., Ivezić, Z., Kent, S., Knapp, G. R., Lin, H., Loh, Y.-S., Lupton, R. H., Margon, B., McKay, T. A., Meiksin, A., Munn, J. A., Pope, A., Richmond, M. W., Schlegel, D., Schneider, D. P., Shimasaku, K., Stoughton, C., Strauss, M. A., SubbaRao, M., Szalay, A. S., Szapudi, I., Tucker, D. L., Yanny, B., and York, D. G. (2005). Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. *The Astrophysical Journal*, 633(2):560–574.

Elbaz, D., Dickinson, M., Hwang, H., Díaz-Santos, T., Magdis, G., Magnelli, B., Le Borgne, D., Galliano, F., Pannella, M., Chanial, P., et al. (2011). Goods–herschel: an infrared main sequence for star-forming galaxies. *Astronomy & Astrophysics*, 533:A119.

Esposito, S., Miele, G., Pastor, S., Peloso, M., and Pisanti, O. (2000). Non equilibrium spectra of degenerate relic neutrinos. *Nuclear Physics B*, 590(3):539–561.

Euclid Collaboration, Scaramella, R., Amiaux, J., Mellier, Y., Burigana, C., Carvalho, C. S., Cuillandre, J. C., Da Silva, A., Derosa, A., Dinis, J., Maiorano, E., Maris, M., Tereno, I., Laureijs, R., Boenke, T., Buenadicha, G., Dupac, X., Gaspar Venancio, L. M., Gómez-Álvarez, P., Hoar, J., Lorenzo Alvarez, J., Racca, G. D., Saavedra-Criado, G., Schwartz, J., Vavrek, R., Schirmer, M., Aussel, H., Azzollini, R., Cardone, V. F., Cropper, M., Ealet, A., Garilli, B., Gillard, W., Granett, B. R., Guzzo, L., Hoekstra, H., Jahnke, K., Kitching, T., Maciaszek, T., Meneghetti, M., Miller, L., Nakajima, R., Niemi, S. M., Pasian, F., Percival, W. J., Pottinger, S., Sauvage, M., Scodeggio, M., Wachter, S., Zacchei, A., Aghanim, N., Amara, A., Auphan, T., Auricchio, N., Awan, S., Balestra, A., Bender, R., Bodendorf, C., Bonino, D., Branchini, E., Brau-Nogue, S., Brescia, M., Candini, G. P., Capobianco, V., Carbone, C., Carlberg, R. G., Carretero, J., Casas, R., Castander, F. J., Castellano, M., Cavuoti, S., Cimatti, A., Cledassou, R., Congedo, G., Conselice, C. J., Conversi, L., Copin, Y., Corcione, L., Costille, A., Courbin, F., Degaudenzi, H., Douspis, M., Dubath, F., Duncan, C. A. J., Dusini, S., Farrens, S., Ferriol, S., Fosalba, P., Fourmanoit, N., Frailis, M., Franceschi, E., Franzetti, P., Fumana, M., Gillis, B., Giocoli, C., Grazian, A., Grupp, F., Haugan, S. V. H., Holmes, W., Hormuth, F., Hudelot, P., Kermiche, S., Kiessling, A., Kilbinger, M., Kohley, R., Kubik, B., Kümmel, M., Kunz, M., Kurki-Suonio, H., Lahav, O., Ligori, S., Lilje, P. B., Lloro, I., Mansutti, O., Marggraf, O., Markovic, K., Marulli, F., Massey, R., Maurogordato, S., Melchior, M., Merlin, E., Meylan, G., Mohr, J. J., Moresco, M., Morin, B., Moscardini, L., Munari, E., Nichol, R. C., Padilla, C., Paltani, S., Peacock, J., Pedersen, K., Pettorino, V., Pires, S., Poncet, M., Popa, L., Pozzetti, L., Raison, F., Rebolo, R., Rhodes, J., Rix, H. W., Roncarelli, M., Rossetti, E., Saglia, R., Schneider, P., Schrabback, T., Secroun, A., Seidel, G., Serrano, S., Sirignano, C., Sirri, G., Skottfelt, J., Stanco, L., Starck, J. L., Tallada-Crespí, P., Tavagnacco, D., Taylor, A. N., Teplitz, H. I., Toledo-Moreo, R., Torradeflot, F., Trifoglio, M., Valentijn, E. A., Valenziano, L., Verdoes Kleijn, G. A., Wang, Y., Welikala, N., Weller, J., Wetzstein, M., Zamorani, G., Zoubian, J., Andreon, S., Baldi, M., Bardelli, S., Boucaud, A., Camera, S., Di Ferdinando, D., Fabbian, G., Farinelli, R., Galeotta, S., Graciá-Carpio, J., Maino, D., Medinaceli, E., Mei, S., Neissner, C., Polenta, G., Renzi, A., Romelli, E., Rosset, C., Sureau, F., Tenti, M., Vassallo, T., Zucca, E., Baccigalupi, C., Balaguera-Antolínez, A., Battaglia, P., Biviano, A., Borgani, S., Bozzo, E., Cabanac, R., Cappi, A., Casas, S., Castignani, G., Colodro-Conde, C., Coupon, J., Courtois, H. M., Cuby, J., de la Torre, S., Desai, S., Dole, H., Fabricius, M., Farina, M., Ferreira, P. G., Finelli, F., Flose-Reimberg, P., Fotopoulou, S., Ganga, K., Gozaliasl, G., Hook, I. M., Keihanen, E., Kirkpatrick,

C. C., Liebing, P., Lindholm, V., Mainetti, G., Martinelli, M., Martinet, N., Maturi, M., McCracken, H. J., Metcalf, R. B., Morgante, G., Nightingale, J., Nucita, A., Patrizii, L., Potter, D., Riccio, G., Sánchez, A. G., Sapone, D., Schewtschenko, J. A., Schultheis, M., Scottez, V., Teyssier, R., Tutusaus, I., Valiviita, J., Viel, M., Vriend, W., and Whittaker, L. (2022). Euclid preparation. I. The Euclid Wide Survey. *Astronomy and Astrophysics*, 662:A112.

Favole, G., Gonzalez-Perez, V., Stoppacher, D., Orsi, Á., Comparat, J., Cora, S. A., Vega-Martínez, C. A., Stevens, A. R. H., Maraston, C., Croton, D., Knebe, A., Benson, A. J., Montero-Dorta, A. D., Padilla, N., Prada, F., and Thomas, D. (2020). [O II] emitters in MultiDark-Galaxies and DEEP2. *Monthly Notices of the Royal Astronomical Society*, 497(4):5432–5453.

Favole, G., Rodríguez-Torres, S. A., Comparat, J., Prada, F., Guo, H., Klypin, A., and Montero-Dorta, A. D. (2017). Galaxy clustering dependence on the [O II] emission line luminosity in the local Universe. *Monthly Notices of the Royal Astronomical Society*, 472(1):550–558.

Fedeli, C., Bartelmann, M., Meneghetti, M., and Moscardini, L. (2007). Effects of the halo concentration distribution on strong-lensing optical depth and X-ray emission. *Astronomy and Astrophysics*, 473(3):715–725.

Feng, Y., Chu, M.-Y., Seljak, U., and McDonald, P. (2016). FASTPM: a new scheme for fast simulations of dark matter and haloes. *Monthly Notices of the Royal Astronomical Society*, 463(3):2273–2286.

Ferland, G. J., Porter, R. L., van Hoof, P. A. M., Williams, R. J. R., Abel, N. P., Lykins, M. L., Shaw, G., Henney, W. J., and Stancil, P. C. (2013). The 2013 release of cloudy.

Fidler, C., Tram, T., Rampf, C., Crittenden, R., Koyama, K., and Wands, D. (2016). Relativistic interpretation of Newtonian simulations for cosmic structure formation. *Journal of Cosmology and Astroparticle Physics*, 2016(9):031.

Fixsen, D. J. (2009). The Temperature of the Cosmic Microwave Background. *The Astrophysical Journal*, 707(2):916–920.

Franco-Barranco, D., Andrés-San Román, J. A., Gómez-Gálvez, P., Escudero, L. M., Muñoz-Barrutia, A., and Arganda-Carreras, I. (2023). BiaPy: a ready-to-use library for Bioimage Analysis Pipelines. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Franco-Barranco, D., Muñoz-Barrutia, A., and Arganda-Carreras, I. (2021). Stable deep neural network architectures for mitochondria segmentation on electron microscopy volumes. *Neuroinformatics*.

Frenk, C. S. and White, S. D. M. (2012). Dark matter and cosmic structure. *Annalen der Physik*, 524(9-10):507–534.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202.

Gao, L., Navarro, J. F., Cole, S., Frenk, C. S., White, S. D. M., Springel, V., Jenkins, A., and Neto, A. F. (2008). The redshift dependence of the structure of massive Λ cold dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 387(2):536–544.

Geach, J. E., Smail, I., Best, P. N., Kurk, J., Casali, M., Ivison, R. J., and Coppin, K. (2008). HiZELS: a high-redshift survey of Hα emitters - I. The cosmic star formation rate and clustering at z = 2.23. *Monthly Notices of the Royal Astronomical Society*, 388(4):1473–1486.

Geach, J. E., Sobral, D., Hickox, R. C., Wake, D. A., Smail, I., Best, P. N., Baugh, C. M., and Stott, J. P. (2012). The clustering of Hα emitters at z=2.23 from HiZELS. *Monthly Notices of the Royal Astronomical Society*, 426(1):679–689.

Genel, S., Bryan, G. L., Springel, V., Hernquist, L., Nelson, D., Pillepich, A., Weinberger, R., Pakmor, R., Marinacci, F., and Vogelsberger, M. (2019). A Quantification of the Butterfly Effect in Cosmological Simulations and Implications for Galaxy Scaling Relations. *The Astrophysical Journal*, 871(1):21.

Giusarma, E., Reyes Hurtado, M., Villaescusa-Navarro, F., He, S., Ho, S., and Hahn, C. (2019). Learning neutrino effects in Cosmology with Convolutional Neural Networks. *arXiv e-prints*, page arXiv:1910.04255.

Gonzalez-Perez, V., Cui, W., Contreras, S., Baugh, C. M., Comparat, J., Griffin, A. J., Helly, J., Knebe, A., Lacey, C., and Norberg, P. (2020). Do model emission line galaxies live in filaments at z $\sim$ 1? *Monthly Notices of the Royal Astronomical Society*, 498(2):1852–1870.

Groves, B. A., Dopita, M. A., and Sutherland, R. S. (2004). Dusty, radiation pressure–dominated photoionization. ii. multiwavelength emission line diagnostics for narrow-line regions. *The Astrophysical Journal Supplement Series*, 153(1):75–91.

Gunn, J. E. (1977). Massive galactic halos. I. Formation and evolution. *The Astrophysical Journal*, 218:592–598.

Gunn, J. E. and Gott, J. Richard, I. (1972). On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution. *The Astrophysical Journal*, 176:1.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hahn, C., Eickenberg, M., Ho, S., Hou, J., Lemos, P., Massara, E., Modi, C., Moradinezhad Dizgah, A., Régaldo-Saint Blancard, B., and Abidi, M. M. (2023). SIMBIG: mock challenge for a forward modeling approach to galaxy clustering. *Journal of Cosmology and Astroparticle Physics*, 2023(4):010.

Hahn, O. and Abel, T. (2011). Multi-scale initial conditions for cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 415(3):2101–2121.

Hannestad, S. and Madsen, J. (1995). Neutrino decoupling in the early Universe. *Phys. Rev. D Physical Review D: Particles, Fields, Gravitation & Cosmology*, 52(4):1764–1769.

Hatton, S., Devriendt, J. E. G., Ninin, S., Bouchet, F. R., Guiderdoni, B., and Vibert, D. (2003). GALICS- I. A hybrid N-body/semi-analytic model of hierarchical galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 343:75–106.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv e-prints*, page arXiv:1512.03385.

He, S., Li, Y., Feng, Y., Ho, S., Ravanbakhsh, S., Chen, W., and Póczos, B. (2019). Learning to predict the cosmological structure formation. *Proceedings of the National Academy of Science*, 116(28):13825–13832.

Hellwing, W. A., Cautun, M., Knebe, A., Juszkiewicz, R., and Knollmann, S. (2013). DM haloes in the fifth-force cosmology. *Journal of Cosmology and Astroparticle Physics*, 2013(10):012.

Huss, A., Jain, B., and Steinmetz, M. (1999). How Universal Are the Density Profiles of Dark Halos? *The Astrophysical Journal*, 517(1):64–69.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., Angel, J. R. P., Angeli, G. Z., Ansari, R., Antilogus, P., Araujo, C., Armstrong, R., Arndt, K. T., Astier, P., Aubourg, É., Auza, N., Axelrod,

T. S., Bard, D. J., Barr, J. D., Barrau, A., Bartlett, J. G., Bauer, A. E., Bauman, B. J., Baumont, S., Bechtol, E., Bechtol, K., Becker, A. C., Becla, J., Beldica, C., Bellavia, S., Bianco, F. B., Biswas, R., Blanc, G., Blazek, J., Blandford, R. D., Bloom, J. S., Bogart, J., Bond, T. W., Booth, M. T., Borgland, A. W., Borne, K., Bosch, J. F., Boutigny, D., Brackett, C. A., Bradshaw, A., Brandt, W. N., Brown, M. E., Bullock, J. S., Burchat, P., Burke, D. L., Cagnoli, G., Calabrese, D., Callahan, S., Callen, A. L., Carlin, J. L., Carlson, E. L., Chandrasekharan, S., Charles-Emerson, G., Chesley, S., Cheu, E. C., Chiang, H.-F., Chiang, J., Chirino, C., Chow, D., Ciardi, D. R., Claver, C. F., Cohen-Tanugi, J., Cockrum, J. J., Coles, R., Connolly, A. J., Cook, K. H., Cooray, A., Covey, K. R., Cribbs, C., Cui, W., Cutri, R., Daly, P. N., Daniel, S. F., Daruich, F., Daubard, G., Daues, G., Dawson, W., Delgado, F., Dellapenna, A., de Peyster, R., de Val-Borro, M., Digel, S. W., Doherty, P., Dubois, R., Dubois-Felsmann, G. P., Durech, J., Economou, F., Eifler, T., Eracleous, M., Emmons, B. L., Fausti Neto, A., Ferguson, H., Figueroa, E., Fisher-Levine, M., Focke, W., Foss, M. D., Frank, J., Freemon, M. D., Gangler, E., Gawiser, E., Geary, J. C., Gee, P., Geha, M., Gessner, C. J. B., Gibson, R. R., Gilmore, D. K., Glanzman, T., Glick, W., Goldina, T., Goldstein, D. A., Goodenow, I., Graham, M. L., Gressler, W. J., Gris, P., Guy, L. P., Guyonnet, A., Haller, G., Harris, R., Hascall, P. A., Haupt, J., Hernandez, F., Herrmann, S., Hileman, E., Hoblitt, J., Hodgson, J. A., Hogan, C., Howard, J. D., Huang, D., Huffer, M. E., Ingraham, P., Innes, W. R., Jacoby, S. H., Jain, B., Jammes, F., Jee, M. J., Jenness, T., Jernigan, G., Jevremović, D., Johns, K., Johnson, A. S., Johnson, M. W. G., Jones, R. L., Juramy-Gilles, C., Jurić, M., Kalirai, J. S., Kallivayalil, N. J., Kalmbach, B., Kantor, J. P., Karst, P., Kasliwal, M. M., Kelly, H., Kessler, R., Kinnison, V., Kirkby, D., Knox, L., Kotov, I. V., Krabbendam, V. L., Krughoff, K. S., Kubánek, P., Kuczewski, J., Kulkarni, S., Ku, J., Kurita, N. R., Lage, C. S., Lambert, R., Lange, T., Langton, J. B., Le Guillou, L., Levine, D., Liang, M., Lim, K.-T., Lintott, C. J., Long, K. E., Lopez, M., Lotz, P. J., Lupton, R. H., Lust, N. B., MacArthur, L. A., Mahabal, A., Mandelbaum, R., Markiewicz, T. W., Marsh, D. S., Marshall, P. J., Marshall, S., May, M., McKercher, R., McQueen, M., Meyers, J., Migliore, M., Miller, M., Mills, D. J., Miraval, C., Moeyens, J., Moolekamp, F. E., Monet, D. G., Moniez, M., Monkewitz, S., Montgomery, C., Morrison, C. B., Mueller, F., Muller, G. P., Muñoz Arancibia, F., Neill, D. R., Newbry, S. P., Nief, J.-Y., Nomerotski, A., Nordby, M., O'Connor, P., Oliver, J., Olivier, S. S., Olsen, K., O'Mullane, W., Ortiz, S., Osier, S., Owen, R. E., Pain, R., Palecek, P. E., Parejko, J. K., Parsons, J. B., Pease, N. M., Peterson, J. M., Peterson, J. R., Petravick, D. L., Libby Petrick, M. E., Petry, C. E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P. A., Plante, R., Plate, S., Plutchak, J. P., Price, P. A., Prouza, M., Radeka, V., Rajagopal, J., Rasmussen, A. P., Regnault, N., Reil, K. A., Reiss, D. J., Reuter, M. A., Ridgway, S. T., Riot, V. J., Ritz, S., Robinson, S., Roby, W., Roodman, A., Rosing, W., Roucelle, C., Rumore, M. R.,

Russo, S., Saha, A., Sassolas, B., Schalk, T. L., Schellart, P., Schindler, R. H., Schmidt, S., Schneider, D. P., Schneider, M. D., Schoening, W., Schumacher, G., Schwamb, M. E., Sebag, J., Selvy, B., Sembroski, G. H., Seppala, L. G., Serio, A., Serrano, E., Shaw, R. A., Shipsey, I., Sick, J., Silvestri, N., Slater, C. T., Smith, J. A., Smith, R. C., Sobhani, S., Soldahl, C., Storrie-Lombardi, L., Stover, E., Strauss, M. A., Street, R. A., Stubbs, C. W., Sullivan, I. S., Sweeney, D., Swinbank, J. D., Szalay, A., Takacs, P., Tether, S. A., Thaler, J. J., Thayer, J. G., Thomas, S., Thornton, A. J., Thukral, V., Tice, J., Trilling, D. E., Turri, M., Van Berg, R., Vanden Berk, D., Vetter, K., Virieux, F., Vucina, T., Wahl, W., Walkowicz, L., Walsh, B., Walter, C. W., Wang, D. L., Wang, S.-Y., Warner, M., Wiecha, O., Willman, B., Winters, S. E., Wittman, D., Wolff, S. C., Wood-Vasey, W. M., Wu, X., Xin, B., Yoachim, P., and Zhan, H. (2019). LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873(2):111.

Izquierdo-Villalba, D., Bonoli, S., Spinoso, D., Rosas-Guevara, Y., Henriques, B. M. B., and Hernández-Monteagudo, C. (2019). The build-up of pseudo-bulges in a hierarchical universe. *Monthly Notices of the Royal Astronomical Society*, 488(1):609–632.

Jamieson, D., Li, Y., He, S., Villaescusa-Navarro, F., Ho, S., Alves de Oliveira, R., and Spergel, D. N. (2022). Simple lessons from complex learning: what a neural network model learns about cosmic structure formation. *arXiv e-prints*, page arXiv:2206.04573.

Jamieson, D. and Loverde, M. (2019). Separate universe void bias. *Phys. Rev. D Physical Review D: Particles, Fields, Gravitation & Cosmology*, 100(12):123528.

Jiang, F. and van den Bosch, F. C. (2014). Generating merger trees for dark matter haloes: a comparison of methods. *Monthly Notices of the Royal Astronomical Society*, 440(1):193–207.

Kaiser, N. (1984). On the spatial correlations of Abell clusters. *The Astrophysical Journal Letters*, 284:L9–L12.

Kauffmann, G. (1996). Disc galaxies at z=0 and at high redshift: an explanation of the observed evolution of damped Lyalpha absorption systems. *Monthly Notices of the Royal Astronomical Society*, 281:475–486.

Kennicutt, Jr., R. C. (1989). The star formation law in galactic disks. *The Astrophysical Journal*, 344:685–703.

Kirillov, A., He, K., Girshick, R., Rother, C., and Dollár, P. (2018). Panoptic Segmentation. *arXiv e-prints*, page arXiv:1801.00868.

Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., and Rother, C. (2016). InstanceCut: from Edges to Instances with MultiCut. *arXiv e-prints*, page arXiv:1611.08272.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment Anything. *arXiv e-prints*, page arXiv:2304.02643.

Klypin, A. and Holtzman, J. (1997). Particle-Mesh code for cosmological simulations. *arXiv e-prints*, pages astro–ph/9712217.

Knebe, A., Lopez-Cano, D., Avila, S., Favole, G., Stevens, A. R. H., Gonzalez-Perez, V., Reyes-Peraza, G., Yepes, G., Chuang, C.-H., and Kitaura, F.-S. (2021). UNITSIM-Galaxies: data release and clustering of emission-line galaxies. *arXiv e-prints*, page arXiv:2103.13088.

Knebe, A., Lopez-Cano, D., Avila, S., Favole, G., Stevens, A. R. H., Gonzalez-Perez, V., Reyes-Peraza, G., Yepes, G., Chuang, C.-H., and Kitaura, F.-S. (2022). UNITSIM-Galaxies: data release and clustering of emission-line galaxies. *Monthly Notices of the Royal Astronomical Society*, 510(4):5392–5407.

Knebe, A., Pearce, F. R., Gonzalez-Perez, V., Thomas, P. A., Benson, A., Asquith, R., Blaizot, J., Bower, R., Carretero, J., Castander, F. J., Cattaneo, A., Cora, S. A., Croton, D. J., Cui, W., Cunnama, D., Devriendt, J. E., Elahi, P. J., Font, A., Fontanot, F., Gargiulo, I. D., Helly, J., Henriques, B., Lee, J., Mamon, G. A., Onions, J., Padilla, N. D., Power, C., Pujol, A., Ruiz, A. N., Srisawat, C., Stevens, A. R. H., Tollet, E., Vega-Martínez, C. A., and Yi, S. K. (2018a). Cosmic CARNage I: on the calibration of galaxy formation models. *Monthly Notices of the Royal Astronomical Society*, 475(3):2936–2954.

Knebe, A., Pearce, F. R., Thomas, P. A., Benson, A., Blaizot, J., Bower, R., Carretero, J., Castander, F. J., and et al. (2015). nIFTy cosmology: comparison of galaxy formation models. *Monthly Notices of the Royal Astronomical Society*, 451:4029–4059.

Knebe, A., Stoppacher, D., Prada, F., Behrens, C., Benson, A., Cora, S. A., Croton, D. J., Padilla, N. D., Ruiz, A. N., Sinha, M., Stevens, A. R. H., Vega-Martínez, C. A., Behroozi, P., Gonzalez-Perez, V., Gottlöber, S., Klypin, A. A., Yepes, G., Enke, H., Libeskind, N. I., Riebe, K., and Steinmetz, M. (2018b). MULTIDARK-GALAXIES: data release and first results. *Monthly Notices of the Royal Astronomical Society*, 474(4):5206–5231.

Knollmann, S. R., Power, C., and Knebe, A. (2008). Dark matter halo profiles in scale-free cosmologies. *Monthly Notices of the Royal Astronomical Society*, 385(2):545–552.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Kurki-Suonio, H. (2023). Cosmology ii lecture notes. University of Helsinki.

Kurki-Suonio, H. (2024a). Cosmological perturbation theory. University of Helsinki. Lecture course material.

Kurki-Suonio, H. (2024b). Cosmology i lecture notes. University of Helsinki.

Kwan, J., Bhattacharya, S., Heitmann, K., and Habib, S. (2013). Cosmic Emulation: The Concentration-Mass Relation for wCDM Universes. *The Astrophysical Journal*, 768(2):123.

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

Laureijs, R., Amiaux, J., Arduini, S., Augueres, J. L., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., Garilli, B., Gondoin, P., Guzzo, L., Hoar, J., Hoekstra, H., Holmes, R., Kitching, T., Maciaszek, T., Mellier, Y., Pasian, F., Percival, W., Rhodes, J., Criado, G. S., Sauvage, M., Scaramella, R., Valenziano, L., Warren, S., Bender, R., Castander, F., Cimatti, A., Fèvre, O. L., Kurki-Suonio, H., Levi, M., Lilje, P., Meylan, G., Nichol, R., Pedersen, K., Popa, V., Lopez, R. R., Rix, H. W., Rottgering, H., Zeilinger, W., Grupp, F., Hudelot, P., Massey, R., Meneghetti, M., Miller, L., Paltani, S., Paulin-Henriksson, S., Pires, S., Saxton, C., Schrabback, T., Seidel, G., Walsh, J., Aghanim, N., Amendola, L., Bartlett, J., Baccigalupi, C., Beaulieu, J. P., Benabed, K., Cuby, J. G., Elbaz, D., Fosalba, P., Gavazzi, G., Helmi, A., Hook, I., Irwin, M., Kneib, J. P., Kunz, M., Mannucci, F., Moscardini, L., Tao, C., Teyssier, R., Weller, J., Zamorani, G., Osorio, M. R. Z., Boulade, O., Foumond, J. J., Giorgio, A. D., Guttridge, P., James, A., Kemp, M., Martignac, J., Spencer, A., Walton, D., Blümchen, T., Bonoli, C., Bortoletto, F., Cerna, C., Corcione, L., Fabron, C., Jahnke, K., Ligori, S., Madrid, F., Martin, L., Morgante, G., Pamplona, T., Prieto, E., Riva, M., Toledo, R., Trifoglio, M., Zerbi, F., Abdalla, F., Douspis, M., Grenet, C., Borgani, S., Bouwens, R., Courbin, F., Delouis, J. M., Dubath, P., Fontana, A., Frailis, M., Grazian, A., Koppenhöfer, J., Mansutti, O., Melchior, M., Mignoli, M., Mohr, J., Neissner, C., Noddle, K., Poncet, M., Scodeggio, M., Serrano, S., Shane, N., Starck, J. L., Surace, C., Taylor, A., Verdoes-Kleijn, G., Vuerli, C., Williams, O. R., Zacchei, A., Altieri, B., Sanz, I. E., Kohley, R., Oosterbroek, T., Astier, P., Bacon, D., Bardelli, S., Baugh, C., Bellagamba, F., Benoist, C., Bianchi, D., Biviano,

A., Branchini, E., Carbone, C., Cardone, V., Clements, D., Colombi, S., Conselice, C., Cresci, G., Deacon, N., Dunlop, J., Fedeli, C., Fontanot, F., Franzetti, P., Giocoli, C., Garcia-Bellido, J., Gow, J., Heavens, A., Hewett, P., Heymans, C., Holland, A., Huang, Z., Ilbert, O., Joachimi, B., Jennins, E., Kerins, E., Kiessling, A., Kirk, D., Kotak, R., Krause, O., Lahav, O., van Leeuwen, F., Lesgourgues, J., Lombardi, M., Magliocchetti, M., Maguire, K., Majerotto, E., Maoli, R., Marulli, F., Maurogordato, S., McCracken, H., McLure, R., Melchiorri, A., Merson, A., Moresco, M., Nonino, M., Norberg, P., Peacock, J., Pello, R., Penny, M., Pettorino, V., Porto, C. D., Pozzetti, L., Quercellini, C., Radovich, M., Rassat, A., Roche, N., Ronayette, S., Rossetti, E., Sartoris, B., Schneider, P., Semboloni, E., Serjeant, S., Simpson, F., Skordis, C., Smadja, G., Smartt, S., Spano, P., Spiro, S., Sullivan, M., Tilquin, A., Trotta, R., Verde, L., Wang, Y., Williger, G., Zhao, G., Zoubian, J., and Zucca, E. (2011). Euclid definition study report.

Lazeyras, T., Wagner, C., Baldauf, T., and Schmidt, F. (2016). Precision measurement of the local bias of dark matter halos. *Journal of Cosmology and Astroparticle Physics*, 2016(2):018–018.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, J., Shin, J., Snaith, O. N., Kim, Y., Few, C. G., Devriendt, J., Dubois, Y., Cox, L. M., Hong, S. E., Kwon, O.-K., Park, C., Pichon, C., Kim, J., Gibson, B. K., and Park, C. (2020). The Horizon Run 5 Cosmological Hydrodynamic Simulation: Probing Galaxy Formation from Kilo- to Giga-parsec Scales. *arXiv e-prints*, page arXiv:2006.01039.

Lesgourgues, J. and Pastor, S. (2006). Massive neutrinos and cosmology. *Physics Reports*, 429(6):307–379.

Li, Y., Hu, W., and Takada, M. (2014). Super-sample covariance in simulations. *Phys. Rev. D Physical Review D: Particles, Fields, Gravitation & Cosmology*, 89(8):083519.

Lin, Z., Wei, D., Petkova, M. D., Wu, Y., Ahmed, Z., K, K. S., Zou, S., Wendt, N., Boulanger-Weill, J., Wang, X., et al. (2021). Nucmm dataset: 3d neuronal nuclei instance segmentation at sub-cubic millimeter scale. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 164–174. Springer.

Linden, S. and Virey, J.-M. (2008). Test of the Chevallier-Polarski-Linder parametrization for rapid dark energy equation of state transitions. *Phys. Rev. D Physical Review D: Particles, Fields, Gravitation & Cosmology*, 78(2):023526.

Linder, E. V. (2003). Exploring the Expansion History of the Universe. *Phys. Rev. Lett. Physical Review Letters*, 90(9):091301.

Long, J., Shelhamer, E., and Darrell, T. (2014). Fully Convolutional Networks for Semantic Segmentation. *arXiv e-prints*, page arXiv:1411.4038.

López-Cano, D., Angulo, R. E., Ludlow, A. D., Zennaro, M., Contreras, S., Chaves-Montero, J., and Aricò, G. (2022). The cosmology dependence of the concentration-mass-redshift relation. *Monthly Notices of the Royal Astronomical Society*, 517(2):2000–2011.

López-Cano, D., Stücker, J., Pellejero Ibañez, M., Angulo, R. E., and Franco-Barranco, D. (2023). Characterizing Structure Formation through Instance Segmentation. *arXiv e-prints*, page arXiv:2311.12110.

Lucie-Smith, L., Adhikari, S., and Wechsler, R. H. (2022). Insights into the origin of halo mass profiles from machine learning. *arXiv e-prints*, page arXiv:2205.04474.

Lucie-Smith, L., Barreira, A., and Schmidt, F. (2023). Halo assembly bias from a deep learning model of halo formation. *Monthly Notices of the Royal Astronomical Society*, 524(2):1746–1756.

Lucie-Smith, L., Peiris, H. V., and Pontzen, A. (2019). An interpretable machine-learning framework for dark matter halo formation. *Monthly Notices of the Royal Astronomical Society*, 490(1):331–342.

Lucie-Smith, L., Peiris, H. V., Pontzen, A., and Lochner, M. (2018). Machine learning cosmological structure formation. *Monthly Notices of the Royal Astronomical Society*, 479(3):3405–3414.

Lucie-Smith, L., Peiris, H. V., Pontzen, A., Nord, B., and Thiyagalingam, J. (2020). Deep learning insights into cosmological structure formation. *arXiv e-prints*, page arXiv:2011.10577.

Ludlow, A. D. and Angulo, R. E. (2017). Einasto profiles and the dark matter power spectrum. *Monthly Notices of the Royal Astronomical Society*, 465(1):L84–L88.

Ludlow, A. D., Bose, S., Angulo, R. E., Wang, L., Hellwing, W. A., Navarro, J. F., Cole, S., and Frenk, C. S. (2016). The mass-concentration-redshift relation of cold and warm dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 460(2):1214–1232.

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

Ludlow, A. D., Navarro, J. F., Boylan-Kolchin, M., Bett, P. E., Angulo, R. E., Li, M., White, S. D. M., Frenk, C., and Springel, V. (2013). The mass profile and accretion history of cold dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 432(2):1103–1113.

Ludlow, A. D., Navarro, J. F., Li, M., Angulo, R. E., Boylan-Kolchin, M., and Bett, P. E. (2012). The dynamical state and mass-concentration relation of galaxy clusters. *Monthly Notices of the Royal Astronomical Society*, 427(2):1322–1328.

Ludlow, A. D., Navarro, J. F., White, S. D. M., Boylan-Kolchin, M., Springel, V., Jenkins, A., and Frenk, C. S. (2011). The density and pseudo-phase-space density profiles of cold dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 415(4):3895–3902.

Ludlow, A. D. and Porciani, C. (2011). The peaks formalism and the formation of cold dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 413(3):1961–1972.

Ludlow, A. D., Schaye, J., and Bower, R. (2019). Numerical convergence of simulations of galaxy formation: the abundance and internal structure of cold dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 488(3):3663–3684.

Ludlow, A. D., Schaye, J., Schaller, M., and Bower, R. (2020). Numerical convergence of hydrodynamical simulations of galaxy formation: the abundance and internal structure of galaxies and their cold dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 493(2):2926–2951.

Macciò, A. V., Dutton, A. A., and van den Bosch, F. C. (2008). Concentration, spin and shape of dark matter haloes as a function of the cosmological model: WMAP1, WMAP3 and WMAP5 results. *Monthly Notices of the Royal Astronomical Society*, 391(4):1940–1954.

Maiolino, R., Nagao, T., Grazian, A., Cocchia, F., Marconi, A., Mannucci, F., Cimatti, A., Pipino, A., Ballero, S., Calura, F., Chiappini, C., Fontana, A., Granato, G. L., Matteucci, F., Pastorini, G., Pentericci, L., Risaliti, G., Salvati, M., and Silva, L. (2008). AMAZE. I. The evolution of the mass-metallicity relation at z > 3. *Astronomy and Astrophysics*, 488(2):463–479.

Maion, F., Angulo, R. E., and Zennaro, M. (2022). Statistics of biased tracers in variance-suppressed simulations. *arXiv e-prints*, page arXiv:2204.03868.

Mandelbaum, R., Seljak, U., and Hirata, C. M. (2008). A halo mass—concentration relation from weak lensing. *Journal of Cosmology and Astroparticle Physics*, 2008(8):006.

Mandic, V., Bird, S., and Cholis, I. (2016). Stochastic Gravitational-Wave Background due to Primordial Binary Black Hole Mergers. *Phys. Rev. Lett. Physical Review Letters*, 117(20):201102.

Masaki, S., Nishimichi, T., and Takada, M. (2020). Anisotropic separate universe simulations. *Monthly Notices of the Royal Astronomical Society*, 496(1):483–496.

Merson, A., Smith, A., Benson, A., Wang, Y., and Baugh, C. (2019). Linear bias forecasts for emission line cosmological surveys. *Monthly Notices of the Royal Astronomical Society*, 486(4):5737–5765.

Meyer, F. (1994). Topographic distance and watershed lines. *Signal Processing*, 38(1):113–125.

Mohr, P. J. and Taylor, B. N. (2000). Codata recommended values of the fundamental physical constants: 1998. *Rev. Mod. Phys.*, 72:351–495.

Moon, J.-S. and Lee, J. (2023). Why Galaxies are Indeed Simpler than Expected. *arXiv e-prints*, page arXiv:2311.03632.

Musso, M. and Sheth, R. K. (2021a). Excursion set peaks in energy as a model for haloes. *Monthly Notices of the Royal Astronomical Society*, 508(3):3634–3648.

Musso, M. and Sheth, R. K. (2021b). Excursion set peaks in energy as a model for haloes. *Monthly Notices of the Royal Astronomical Society*, 508(3):3634–3648.

Musso, M. and Sheth, R. K. (2023a). Getting in shape with minimal energy: a variational principle for protohaloes. *Monthly Notices of the Royal Astronomical Society*, 523(1):L4–L8.

Musso, M. and Sheth, R. K. (2023b). Getting in shape with minimal energy: a variational principle for protohaloes. *Monthly Notices of the Royal Astronomical Society*, 523(1):L4–L8.

Navarro, J. F., Frenk, C. S., and White, S. D. M. (1996). The Structure of Cold Dark Matter Halos. *The Astrophysical Journal*, 462:563.

Navarro, J. F., Frenk, C. S., and White, S. D. M. (1997). A Universal Density Profile from Hierarchical Clustering. *The Astrophysical Journal*, 490(2):493–508.

Navarro, J. F., Hayashi, E., Power, C., Jenkins, A. R., Frenk, C. S., White, S. D. M., Springel, V., Stadel, J., and Quinn, T. R. (2004). The inner structure of ΛCDM haloes - III.

Universality and asymptotic slopes. *Monthly Notices of the Royal Astronomical Society*, 349(3):1039–1051.

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

Nguyen, T., Modi, C., Yung, L. Y. A., and Somerville, R. S. (2023). FLORAH: A generative model for halo assembly histories. *arXiv e-prints*, page arXiv:2308.05145.

Nuza, S. E., Sanchez, A. G., Prada, F., Klypin, A., Schlegel, D. J., Gottloeber, S., Montero-Dorta, A. D., and Manera, M. e. a. (2012). The clustering of galaxies at z˜0.5 in the SDSS-III Data Release 9 BOSS-CMASS sample: a test for the LCDM cosmology. *ArXiv e-prints*.

Okoli, C., Taylor, J. E., and Afshordi, N. (2018). Searching for dark matter annihilation from individual halos: uncertainties, scatter and signal-to-noise ratios. *Journal of Cosmology and Astroparticle Physics*, 2018(8):019.

Ondaro-Mallea, L., Angulo, R. E., Zennaro, M., Contreras, S., and Aricò, G. (2022). Non-universality of the mass function: dependence on the growth rate and power spectrum shape. *Monthly Notices of the Royal Astronomical Society*, 509(4):6077–6090.

Orsi, A., Padilla, N., Groves, B., Cora, S., Tecce, T., Gargiulo, I., and Ruiz, A. (2014). The nebular emission of star-forming galaxies in a hierarchical universe. *Monthly Notices of the Royal Astronomical Society*, 443(1):799–814.

Ortín, T. (2007). *Gravity and Strings*.

Osterbrock, D. E. (1989). *Astrophysics of gaseous nebulae and active galactic nuclei*.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66.

Paranjape, A., Sefusatti, E., Chan, K. C., Desjacques, V., Monaco, P., and Sheth, R. K. (2013). Bias deconstructed: unravelling the scale dependence of halo bias using real-space measurements. *Monthly Notices of the Royal Astronomical Society*, 436(1):449–459.

Parkinson, D., Riemer-Sørensen, S., Blake, C., Poole, G. B., Davis, T. M., Brough, S., Colless, M., Contreras, C., Couch, W., Croom, S., Croton, D., Drinkwater, M. J., Forster, K., Gilbank, D., Gladders, M., Glazebrook, K., Jelliffe, B., Jurek, R. J., Li, I. h., Madore, B., Martin, D. C., Pimbblet, K., Pracy, M., Sharp, R., Wisnioski, E., Woods, D., Wyder,

T. K., and Yee, H. K. C. (2012). The WiggleZ Dark Energy Survey: Final data release and cosmological results. *Phys. Rev. D Physical Review D: Particles, Fields, Gravitation & Cosmology*, 86(10):103518.

Peebles, P. J. E. (1980). *The large-scale structure of the universe*.

Peebles, P. J. E. (1982). Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations. *The Astrophysical Journal Letters*, 263:L1–L5.

Pellejero-Ibanez, M., Angulo, R. E., Zennaro, M., Stuecker, J., Contreras, S., Arico, G., and Maion, F. (2022). The Bacco Simulation Project: Bacco Hybrid Lagrangian Bias Expansion Model in Redshift Space. *arXiv e-prints*, page arXiv:2207.06437.

Penzias, A. A. and Wilson, R. W. (1965). A Measurement of Excess Antenna Temperature at 4080 Mc/s. *The Astrophysical Journal*, 142:419–421.

Perivolaropoulos, L. and Skara, F. (2022). Challenges for lcdm: An update. *New Astronomy Reviews*, 95:101659.

Perraudin, N., Srivastava, A., Lucchi, A., Kacprzak, T., Hofmann, T., and Réfrégier, A. (2019). Cosmological N-body simulations: a challenge for scalable generative models. *Computational Astrophysics and Cosmology*, 6(1):5.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., and et al. (2015). Planck 2015 results. XIII. Cosmological parameters. *ArXiv e-prints*.

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Carron, J., Challinor, A., Chiang, H. C., Chluba, J., Colombo, L. P. L., Combet, C., Contreras, D., Crill, B. P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J. M., Di Valentino, E., Diego, J. M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Fantaye, Y., Farhang, M., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Fraisse, A. A., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga, K., Génova-Santos, R. T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Handley, W., Hansen, F. K., Herranz, D., Hildebrandt, S. R., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T. S., Knox, L., Krachmalnicoff, N., Kunz, M.,

Kurki-Suonio, H., Lagache, G., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Le Jeune, M., Lemos, P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lilley, M., Lindholm, V., López-Caniego, M., Lubin, P. M., Ma, Y. Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Molinari, D., Montier, L., Morgante, G., Moss, A., Natoli, P., Nørgaard-Nielsen, H. U., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Peiris, H. V., Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J. L., Rachen, J. P., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E. P. S., Sirignano, C., Sirri, G., Spencer, L. D., Sunyaev, R., Suur-Uski, A. S., Tauber, J. A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valenziano, L., Valiviita, J., Van Tent, B., Vibert, L., Vielva, P., Villa, F., Vittorio, N., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Zacchei, A., and Zonca, A. (2020a). Planck 2018 results. VI. Cosmological parameters. *Astronomy and Astrophysics*, 641:A6.

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Carron, J., Challinor, A., Chiang, H. C., Chluba, J., Colombo, L. P. L., Combet, C., Contreras, D., Crill, B. P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J. M., Di Valentino, E., Diego, J. M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Fantaye, Y., Farhang, M., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Fraisse, A. A., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga, K., Génova-Santos, R. T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Handley, W., Hansen, F. K., Herranz, D., Hildebrandt, S. R., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T. S., Knox, L., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lagache, G., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Le Jeune, M., Lemos, P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lilley, M., Lindholm, V., López-Caniego, M., Lubin, P. M., Ma, Y. Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen,

J. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Molinari, D., Montier, L., Morgante, G., Moss, A., Natoli, P., Nørgaard-Nielsen, H. U., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Peiris, H. V., Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J. L., Rachen, J. P., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E. P. S., Sirignano, C., Sirri, G., Spencer, L. D., Sunyaev, R., Suur-Uski, A. S., Tauber, J. A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valenziano, L., Valiviita, J., Van Tent, B., Vibert, L., Vielva, P., Villa, F., Vittorio, N., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Zacchei, A., and Zonca, A. (2020b). Planck 2018 results. VI. Cosmological parameters. *Astronomy and Astrophysics*, 641:A6.

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

Pozzetti, L., Hirata, C. M., Geach, J. E., Cimatti, A., Baugh, C., Cucciati, O., Merson, A., Norberg, P., and Shi, D. (2016). Modelling the number density of Hα emitters for future spectroscopic near-IR space missions. *Astronomy and Astrophysics*, 590:A3.

Prada, F., Klypin, A., Yepes, G., Nuza, S. E., and Gottloeber, S. (2011). Measuring equality horizon with the zero-crossing of the galaxy correlation function. *arXiv e-prints*, page arXiv:1111.2889.

Prada, F., Klypin, A. A., Cuesta, A. J., Betancort-Rijo, J. E., and Primack, J. (2012). Halo concentrations in the standard Λ cold dark matter cosmology. *Monthly Notices of the Royal Astronomical Society*, 423(4):3018–3030.

Press, W. H. and Schechter, P. (1974a). Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *The Astrophysical Journal*, 187:425–438.

Press, W. H. and Schechter, P. (1974b). Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *The Astrophysical Journal*, 187:425–438.

Ragagnin, A., Saro, A., Singh, P., and Dolag, K. (2021). Cosmology dependence of halo masses and concentrations in hydrodynamic simulations. *Monthly Notices of the Royal Astronomical Society*, 500(4):5056–5071.

Richardson, T. R. G., Stücker, J., Angulo, R. E., and Hahn, O. (2022). Non-halo structures and

their effects on gravitational lensing. *Monthly Notices of the Royal Astronomical Society*, 511(4):6019–6032.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., Gilliland, R. L., Hogan, C. J., Jha, S., Kirshner, R. P., Leibundgut, B., Phillips, M. M., Reiss, D., Schmidt, B. P., Schommer, R. A., Smith, R. C., Spyromilio, J., Stubbs, C., Suntzeff, N. B., and Tonry, J. (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *Astronomical Journal*, 116(3):1009–1038.

Riess, A. G., Macri, L. M., Hoffmann, S. L., Scolnic, D., Casertano, S., Filippenko, A. V., Tucker, B. E., Reid, M. J., Jones, D. O., Silverman, J. M., Chornock, R., Challis, P., Yuan, W., Brown, P. J., and Foley, R. J. (2016). A 2.4% Determination of the Local Value of the Hubble Constant. *The Astrophysical Journal*, 826(1):56.

Robles, S., Gómez, J. S., Ramírez Rivera, A., Padilla, N. D., and Dujovne, D. (2022). A deep learning approach to halo merger tree construction. *Monthly Notices of the Royal Astronomical Society*, 514(3):3692–3708.

Rodríguez, A. C., Kacprzak, T., Lucchi, A., Amara, A., Sgier, R., Fluri, J., Hofmann, T., and Réfrégier, A. (2018). Fast cosmic web simulations with generative adversarial networks. *Computational Astrophysics and Cosmology*, 5(1):4.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints*, page arXiv:1505.04597.

Sánchez, A. G., Baugh, C. M., and Angulo, R. E. (2008). What is the best way to measure baryonic acoustic oscillations? *Monthly Notices of the Royal Astronomical Society*, 390(4):1470–1490.

Sánchez-Conde, M. A. and Prada, F. (2014). The flattening of the concentration-mass relation towards low halo masses and its implications for the annihilation signal boost. *Monthly Notices of the Royal Astronomical Society*, 442(3):2271–2277.

Schanz, A., List, F., and Hahn, O. (2023). Stochastic Super-resolution of Cosmological Simulations with Denoising Diffusion Models. *arXiv e-prints*, page arXiv:2310.06929.

Schaurecker, D., Li, Y., Tinker, J., Ho, S., and Refregier, A. (2021). Super-resolving Dark Matter Halos using Generative Deep Learning. *arXiv e-prints*, page arXiv:2111.06393.

Schmidt, A. S., White, S. D. M., Schmidt, F., and Stücker, J. (2018). Cosmological N-body simulations with a large-scale tidal field. *Monthly Notices of the Royal Astronomical Society*, 479(1):162–170.

Schuller, F. P. (2015). A thorough introduction to the theory of general relativity. https://www.youtube.com/watch?v=7G4SqIboeig&list=PLFeEvEPtX_0S6vxxiiNPrJbLu9aK1UVC_&index=1.

Scott, N., Graham, A. W., and Schombert, J. (2013). The Supermassive Black Hole Mass-Spheroid Stellar Mass Relation for Sérsic and Core-Sérsic Galaxies. *The Astrophysical Journal*, 768:76.

Sheth, R. K., Mo, H. J., and Tormen, G. (2001). Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 323(1):1–12.

Sheth, R. K. and Tormen, G. (2002). An excursion set model of hierarchical clustering: ellipsoidal collapse and the moving barrier. *Monthly Notices of the Royal Astronomical Society*, 329(1):61–75.

Smoot, G. F. (1999). COBE observations and results. In Maiani, L., Melchiorri, F., and Vittorio, N., editors, *3K cosmology*, volume 476 of *American Institute of Physics Conference Series*, pages 1–10.

Sobral, D., Stroe, A., Koyama, Y., Darvish, B., Calhau, J., Afonso, A., Kodama, T., and Nakata, F. (2016). The nature of Hα star-forming galaxies at z $\sim$ 0.4 in and around Cl 0939+4713: the environment matters. *Monthly Notices of the Royal Astronomical Society*, 458(4):3443–3454.

Somerville, R. S., Primack, J. R., and Faber, S. M. (2001). The nature of high-redshift galaxies. *Monthly Notices of the Royal Astronomical Society*, 320:504–528.

Sousbie, T. (2011). The persistent cosmic web and its filamentary structure - I. Theory and implementation. *Monthly Notices of the Royal Astronomical Society*, 414(1):350–383.

Spergel, D., Gehrels, N., Baltay, C., Bennett, D., Breckinridge, J., Donahue, M., Dressler, A., Gaudi, B. S., Greene, T., Guyon, O., Hirata, C., Kalirai, J., Kasdin, N. J., Macintosh, B., Moos, W., Perlmutter, S., Postman, M., Rauscher, B., Rhodes, J., Wang, Y., Weinberg, D., Benford, D., Hudson, M., Jeong, W. S., Mellier, Y., Traub, W., Yamada, T., Capak, P., Colbert, J., Masters, D., Penny, M., Savransky, D., Stern, D., Zimmerman, N., Barry, R., Bartusek, L., Carpenter, K., Cheng, E., Content, D., Dekens, F., Demers, R., Grady, K., Jackson, C., Kuan, G., Kruk, J., Melton, M., Nemati, B., Parvin, B., Poberezhskiy, I., Peddie, C., Ruffa, J., Wallace, J. K., Whipple, A., Wollack, E., and Zhao, F. (2015). Wide-field infrarred survey telescope-astrophysics focused telescope assets wfirst-afta 2015 report.

Spergel, D., Gehrels, N., Breckinridge, J., Donahue, M., Dressler, A., Gaudi, B. S., Greene, T., Guyon, O., Hirata, C., Kalirai, J., Kasdin, N. J., Moos, W., Perlmutter, S., Postman, M., Rauscher, B., Rhodes, J., Wang, Y., Weinberg, D., Centrella, J., Traub, W., Baltay, C., Colbert, J., Bennett, D., Kiessling, A., Macintosh, B., Merten, J., Mortonson, M., Penny, M., Rozo, E., Savransky, D., Stapelfeldt, K., Zu, Y., Baker, C., Cheng, E., Content, D., Dooley, J., Foote, M., Goullioud, R., Grady, K., Jackson, C., Kruk, J., Levine, M., Melton, M., Peddie, C., Ruffa, J., and Shaklan, S. (2013). Wide-field infrared survey telescope-astrophysics focused telescope assets wfirst-afta final report.

Springel, V. (2015). N-GenIC: Cosmological structure initial conditions.

Springel, V., Wang, J., Vogelsberger, M., Ludlow, A., Jenkins, A., Helmi, A., Navarro, J. F., Frenk, C. S., and White, S. D. M. (2008). The Aquarius Project: the subhaloes of galactic haloes. *Monthly Notices of the Royal Astronomical Society*, 391(4):1685–1711.

Springel, V., White, S. D. M., Tormen, G., and Kauffmann, G. (2001). Populating a cluster of galaxies - I. Results at z=0. *Monthly Notices of the Royal Astronomical Society*, 328(3):726–750.

Springel, V., White, S. D. M., Tormen, G., and Kauffmann, G. (2001a). Populating a cluster of galaxies – I. Results at z = 0. *Monthly Notices of the Royal Astronomical Society*, 328(3):726–750.

Springel, V., Yoshida, N., and White, S. D. (2001b). Gadget: a code for collisionless and gasdynamical cosmological simulations. *New Astronomy*, 6(2):79–117.

Stark, D. V., McGaugh, S. S., and Swaters, R. A. (2009). A First Attempt to Calibrate the Baryonic Tully-Fisher Relation with Gas-Dominated Galaxies. *Astronomical Journal*, 138:392–401.

Stein, G., Alvarez, M. A., and Bond, J. R. (2019). The mass-Peak Patch algorithm for fast generation of deep all-sky dark matter halo catalogues and its N-body validation. *Monthly Notices of the Royal Astronomical Society*, 483(2):2236–2250.

Stücker, J., Angulo, R. E., and Busch, P. (2021a). The boosted potential. *Monthly Notices of the Royal Astronomical Society*, 508(4):5196–5216.

Stücker, J., Schmidt, A. S., White, S. D. M., Schmidt, F., and Hahn, O. (2021b). Measuring the tidal response of structure formation: anisotropic separate universe simulations using TREEPM. *Monthly Notices of the Royal Astronomical Society*, 503(1):1473–1489.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv e-prints*, page arXiv:1409.4842.

Terasawa, R., Takahashi, R., Nishimichi, T., and Takada, M. (2022). Separate universe approach to evaluate nonlinear matter power spectrum for nonflat Λ CDM model. *Phys. Rev. D Physical Review D: Particles, Fields, Gravitation & Cosmology*, 106(8):083504.

The Dark Energy Survey Collaboration (2005). The Dark Energy Survey. *arXiv e-prints*, pages astro–ph/0510346.

Tierny, J., Favelier, G., Levine, J. A., Gueunet, C., and Michaux, M. (2017). The Topology ToolKit. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*. `https://topology-tool-kit.github.io/`.

Tosone, F., Neyrinck, M. C., Granett, B. R., Guzzo, L., and Vittorio, N. (2021). MUSCLE-UPS: improved approximations of the matter field with the extended Press-Schechter formalism and Lagrangian perturbation theory. *Monthly Notices of the Royal Astronomical Society*, 505(2):2999–3015.

Tremonti, C. A., Heckman, T. M., Kauffmann, G., Brinchmann, J., Charlot, S., White, S. D. M., Seibert, M., Peng, E. W., Schlegel, D. J., Uomoto, A., Fukugita, M., and Brinkmann, J. (2004). The Origin of the Mass-Metallicity Relation: Insights from 53,000 Star-forming Galaxies in the Sloan Digital Sky Survey. *The Astrophysical Journal*, 613:898–913.

Tutusaus, I., Martinelli, M., Cardone, V. F., Camera, S., Yahia-Cherif, S., Casas, S., Blanchard, A., Kilbinger, M., Lacasa, F., Sakr, Z., Ilić, S., Kunz, M., Carbone, C., Castander, F. J., Dournac, F., Fosalba, P., Kitching, T., Markovic, K., Mangilli, A., Pettorino, V., Sapone, D., Yankelevich, V., Auricchio, N., Bender, R., Bonino, D., Boucaud, A., Brescia, M., Capobianco, V., Carretero, J., Castellano, M., Cavuoti, S., Cledassou, R., Congedo, G., Conversi, L., Corcione, L., Costille, A., Cropper, M., Dubath, F., Dusini, S., Fabbian, G., Frailis, M., Franceschi, E., Garilli, B., Grupp, F., Guzzo, L., Hoekstra, H., Hormuth, F., Israel, H., Jahnke, K., Kermiche, S., Kubik, B., Laureijs, R., Ligori, S., Lilje, P. B., Lloro, I., Maiorano, E., Marggraf, O., Massey, R., Mei, S., Merlin, E., Meylan, G., Moscardini, L., Ntelis, P., Padilla, C., Paltani, S., Pasian, F., Percival, W. J., Pires, S., Poncet, M., Raison, F., Rhodes, J., Roncarelli, M., Rossetti, E., Saglia, R., Schneider, P., Secroun, A., Serrano, S., Sirignano, C., Sirri, G., Starck, J., Sureau, F., Taylor, A. N., Tereno, I., Toledo-Moreo, R., Valenziano, L., Wang, Y., Welikala, N., Weller, J., Zacchei, A., and Zoubian, J. (2020). Euclid: The importance of galaxy clustering and

weak lensing cross-correlations within the photometric Euclid survey. *arXiv e-prints*, page arXiv:2005.00055.

Vale, A. and Ostriker, J. P. (2004). Linking halo mass to galaxy luminosity. *Monthly Notices of the Royal Astronomical Society*, 353(1):189–200.

Verde, L., Treu, T., and Riess, A. G. (2019). Tensions between the early and late Universe. *Nature Astronomy*, 3:891–895.

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Spergel, D. N., Somerville, R. S., Dave, R., Pillepich, A., Hernquist, L., Nelson, D., Torrey, P., Narayanan, D., Li, Y., Philcox, O., La Torre, V., Maria Delgado, A., Ho, S., Hassan, S., Burkhart, B., Wadekar, D., Battaglia, N., Contardo, G., and Bryan, G. L. (2021). The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *The Astrophysical Journal*, 915(1):71.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Voivodic, R., Lima, M., and Abramo, L. R. (2019). Excursion Set Halos – ExSHalos: A New Parameter Free Method for Fast Generation of Halo Catalogues. *arXiv e-prints*, page arXiv:1906.06630.

Wagner, C., Schmidt, F., Chiang, C. T., and Komatsu, E. (2015a). Separate universe simulations. *Monthly Notices of the Royal Astronomical Society*, 448:L11–L15.

Wagner, C., Schmidt, F., Chiang, C. T., and Komatsu, E. (2015b). Separate universe simulations. *Monthly Notices of the Royal Astronomical Society*, 448:L11–L15.

Wang, J., Bose, S., Frenk, C. S., Gao, L., Jenkins, A., Springel, V., and White, S. D. M. (2020). Universal structure of dark matter haloes over a mass range of 20 orders of magnitude. *Nature*, 585(7823):39–42.

Wang, J. and White, S. D. M. (2009). Are mergers responsible for universal halo properties? *Monthly Notices of the Royal Astronomical Society*, 396(2):709–717.

Wechsler, R. H., Bullock, J. S., Primack, J. R., Kravtsov, A. V., and Dekel, A. (2002). Concentrations of Dark Halos from Their Assembly Histories. *The Astrophysical Journal*, 568(1):52–70.

Wei, D., Lin, Z., Franco-Barranco, D., Wendt, N., Liu, X., Yin, W., Huang, X., Gupta, A., Jang, W.-D., Wang, X., et al. (2020). MitoEM dataset: Large-scale 3D mitochondria instance segmentation from EM images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 66–76. Springer.

Weinberger, K. Q. and Saul, L. (2009). Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.*, 10:207–244.

White, S. D. M. (1984). Angular momentum growth in protogalaxies. *The Astrophysical Journal*, 286:38–41.

White, S. D. M. and Frenk, C. S. (1991). Galaxy formation through hierarchical clustering. *The Astrophysical Journal*, 379:52–79.

Wu, Z., Zhang, Z., Pan, S., Miao, H., Luo, X., Wang, X., Sabiu, C. G., Forero-Romero, J., Wang, Y., and Li, X.-D. (2021). Cosmic Velocity Field Reconstruction Using AI. *The Astrophysical Journal*, 913(1):2.

Xie, S. and Tu, Z. (2015). Holistically-Nested Edge Detection. *arXiv e-prints*, page arXiv:1504.06375.

Yang, L., Roberts-Borsani, G., Treu, T., Birrer, S., Morishita, T., and Bradač, M. (2021). The evolution of the size-mass relation at z = 1-3 derived from the complete Hubble Frontier Fields data set. *Monthly Notices of the Royal Astronomical Society*, 501(1):1028–1037.

Zee, A. (2013). *Einstein Gravity in a Nutshell*.

Zennaro, M., Angulo, R. E., Aricò, G., Contreras, S., and Pellejero-Ibáñez, M. (2019). How to add massive neutrinos to your ΛCDM simulation - extending cosmology rescaling algorithms. *Monthly Notices of the Royal Astronomical Society*, 489(4):5938–5951.

Zennaro, M., Angulo, R. E., Pellejero-Ibáñez, M., Stücker, J., Contreras, S., and Aricò, G. (2021). The BACCO simulation project: biased tracers in real space. *arXiv e-prints*, page arXiv:2101.12187.

Zennaro, M., Bel, J., Villaescusa-Navarro, F., Carbone, C., Sefusatti, E., and Guzzo, L. (2016). Initial conditions for accuraten-body simulations of massive neutrino cosmologies. *Monthly Notices of the Royal Astronomical Society*, 466(3):3244–3258.

Zennaro, M., Bel, J., Villaescusa-Navarro, F., Carbone, C., Sefusatti, E., and Guzzo, L. (2017). Initial conditions for accurate N-body simulations of massive neutrino cosmologies. *Monthly Notices of the Royal Astronomical Society*, 466(3):3244–3258.

Zentner, A. R. (2007). The Excursion Set Theory of Halo Mass Functions, Halo Clustering, and Halo Growth. *International Journal of Modern Physics D*, 16:763–815.

Zhai, Z., Benson, A., Wang, Y., Yepes, G., and Chuang, C.-H. (2019). Prediction of H α and [O III] emission line galaxy number counts for future galaxy redshift surveys. *Monthly Notices of the Royal Astronomical Society*, 490(3):3667–3678.

Zhai, Z., Chuang, C.-H., Wang, Y., Benson, A., and Yepes, G. (2021). Clustering in the simulated H α galaxy redshift survey from Nancy Grace Roman Space Telescope. *Monthly Notices of the Royal Astronomical Society*, 501(3):3490–3501.

Zhang, T., Liao, S., Li, M., and Gao, L. (2019). The optimal gravitational softening length for cosmological N-body simulations. *Monthly Notices of the Royal Astronomical Society*, 487(1):1227–1232.

Zhang, X., Lachance, P., Ni, Y., Li, Y., Croft, R. A. C., Di Matteo, T., Bird, S., and Feng, Y. (2023). AI-assisted super-resolution cosmological simulations III: Time evolution. *arXiv e-prints*, page arXiv:2305.12222.

Zhao, D. H., Jing, Y. P., Mo, H. J., and Börner, G. (2003). Mass and Redshift Dependence of Dark Halo Structure. *The Astrophysical Journal Letters*, 597(1):L9–L12.

Appendices

Appendix A

Definining smooth manifolds

This Appendix is devoted to defining what a "*four-dimensional topological manifold with a smooth atlas*" is. In Figure A.1 I provide a schematic representation contextualizing the different definitions I introduce encoded by color.

- A **set** $\mathcal{M}$ is a well-defined collection of elements m. For example, the set of real numbers $\mathcal{M} \doteq \mathbb{R}$ is formed by all real numbers $m \doteq \{..., -\pi, -2.34, -1, 0, 1/2, e, 10^{27}, ...\}$. I will also denote sets with $\mathcal{N}$ and elements as n.

- f is a **map** from $\mathcal{M}$ (domain) to $\mathcal{N}$ (target), denoted as $f : \mathcal{M} \to \mathcal{N}$, if $\forall m \in \mathcal{M}$, $\exists n \in \mathcal{N} : f(m) = n$. I will also employ g to denote maps.

- The **powerset** $\mathcal{P}(\mathcal{M})$ of a set $\mathcal{M}$ is the set of all possible subsets of $\mathcal{M}$.

- $\mathcal{O} \subseteq \mathcal{P}(\mathcal{M})$ is a **topology** on $\mathcal{M}$, denoted as $\mathcal{O}_{\mathcal{M}}$, if and only if:

  1. $\emptyset \in \mathcal{O}$ and $\mathcal{M} \in \mathcal{O}$
  2. $\forall U, V \in \mathcal{O} \to U \cap V \in \mathcal{O}$. I will use U, V to denote subsets of $\mathcal{O}$ (open sets).
  3. $\bigcup_{\alpha \in A} U \in \mathcal{O}$ where A denotes an arbitrary index set.

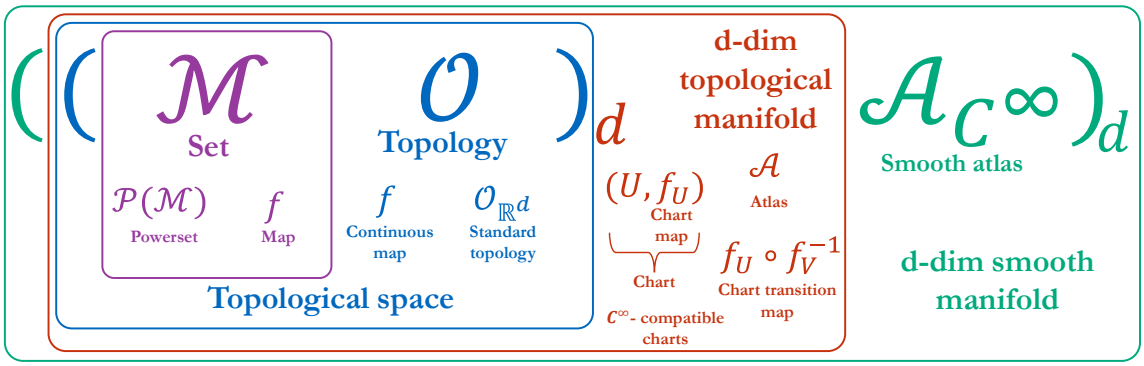

Figure A.1: Scheme for helping to relate the mathematical objects defined in this section.

- the doublet $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})$ is a **topological space**.
- The $\mathbb{R}^d$ **standard topology**, $\mathcal{O}_{\mathbb{R}^d}$, is defined as:

$$\mathcal{O}_{\mathbb{R}^d} \coloneqq \left\{U \in \mathcal{P}\left(\mathbb{R}^d\right) \middle| \forall p \in U \, \exists r \in \mathbb{R}^+ : B_r(p) \subseteq U \right\},$$

  where $B_r(p)$ is the soft-ball of radius r about p defined as:

$$B_r(p) \coloneqq \left\{ (q_1, \ldots, q_d) \in \mathbb{R}^d \mid \sum_{i=1}^{d} (q_i - p_i)^2 < r^2 \right\}.$$

- $f : \mathcal{O}_{\mathcal{M}} \to \mathcal{O}_{\mathcal{N}}$ is a **continuous map** with respect to $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})$ and $(\mathcal{N}, \mathcal{O}_{\mathcal{N}}) \iff \forall V \in \mathcal{O}_{\mathcal{N}} \to \text{preim}_f(V) \coloneqq \{m \in \mathcal{M} : f(m) \in V\} \in \mathcal{O}_{\mathcal{M}}$.
- A topological space $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})$ is a **d-dimensional topological manifold**, $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})_d$, if $\forall m \in \mathcal{M} \, \exists \{U \mid m \in U\} \in \mathcal{O}_{\mathcal{M}} : \exists \left\{ f_U : U \to f_U(U) \subseteq \mathbb{R}^d \right.$, where $\mathbb{R}^d$ implicitly belongs to $\left(\mathbb{R}^d, \mathcal{O}_{\mathbb{R}^d}\right)$, and the map f_U satisfies:
  1. f_U is invertible: $f_U^{-1} : f_U(U) \to U$,
  2. f_U is continuous with respect to $(U, \mathcal{O}_U|_{\mathcal{M}})$ and $\left(\mathbb{R}^d, \mathcal{O}_{\mathbb{R}^d}\right)$,
  3. f_U^{-1} is continuous with respect to $(U, \mathcal{O}_U|_{\mathcal{M}})$[1] and $\left(\mathbb{R}^d, \mathcal{O}_{\mathbb{R}^d}\right)$,
- The doublet (U, f_U) is a **chart** of $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})_d$ and $f_U : U \to f_U(U) \subseteq \mathbb{R}^d$ is known as a **chart map** defined by the coordinate maps $f_U(m) \coloneqq \left(f_U^{(1)}(m), \ldots, f_U^{(d)}(m)\right) \mid f_U^{(i)} : U \to \mathbb{R}$.
- Given some arbitrary index set A, the set $\mathcal{A} = \{(U_\alpha, f_U) \mid \alpha \in A\}$ is an **atlas** of $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})_d \iff \bigcup_{\alpha \in A} U_\alpha = \mathcal{M}$.
- The **chart transition map** between two chart maps, both from the same $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})_d$, (U, f_U), $(V, f_V) \mid U \cap V \neq \emptyset$, is the map $\left(f_U \circ f_V^{-1}\right) : f_V\left(U \cap V\right) \to f_U\left(U \cap V\right)$.
- Two chart maps (U, f_U) and (V, f_V) from $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})_d$, are C^∞**-compatible charts** if:
  1. $U \cap V = \emptyset$, or,
  2. $U \cap V \neq \emptyset$ and both chart transition maps $\left(f_U \circ f_V^{-1}\right) : f_V\left(U \cap V\right) \to f_U\left(U \cap V\right)$ and $\left(f_V \circ f_U^{-1}\right) : f_U\left(U \cap V\right) \to f_V\left(U \cap V\right)$ are C^∞ in the "oridinary multivariable calculus sense".
- An atlas is C^∞**-compatible**, $\mathcal{A}_{C^\infty}$, if all of its charts are C^∞-compatible.
- A **d-dimensional smooth-manifold** (or C^∞-manifold) is defined by the triplet $(\mathcal{M}, \mathcal{O}_{\mathcal{M}}, \mathcal{A}_{C^\infty})_d$

[1] $\mathcal{O}_U|_{\mathcal{M}}$ indicates the inherited topology on U from $\mathcal{O}_{\mathcal{M}}$

Appendix B

Additional validation plots

In this Appendix we provide supplementary plots that further show the validity and properties of the SAGE and ELG galaxies used throughout this study.

B.1 Halo Mass Function of flux-selected ELGs

Applying a SAM will eventually lead to selecting a sub-sample of the underlying dark matter haloes as galaxies, i.e. while every halo contains a galaxy, some might be too small to be detectable. To better understand which haloes host our ELGs, we show their halo mass functions for the two base models *RawELGs* and *DustELGs* for various redshifts in Fig. B.1. The dashed lines are without applying any flux cut, whereas the solid lines use the Euclid-inspired cut $F_{\rm cut} = 2 \times 10^{-16}$ erg s^{-1} cm^{-2}. We can see that the flux cut primarily affects low-mass haloes, i.e. the less luminous ELGs also live in lower mass host haloes. We further observe a shift of this 'cut-off' halo mass with redshift; while at $z \sim 0.5$ it is approximately $10^{11} M_{\odot}$, it increases to $\sim 10^{12} M_{\odot}$ at $z \sim 2$ for *RawELGs* and even $\sim 10^{13} M_{\odot}$ for *DustELGs*.

B.2 Baryonic properties of flux selected ELGs

In Section 1.3 we presented baryonic relations for the full set of SAGE galaxies, focusing on those properties that are relevant for the dust attenuation modelling. Here we now like to provide counterparts of those plots for the ELGs.

Stellar Mass Function

In order to view the effect of the flux selection and its relation to the stellar masses of the resulting sub-sample of ELGs, we show in Fig. B.2 both the SMF of all ELGs (i.e. no flux cut, dashed lines) and the flux-selected samples of ELGs (solid lines) for various redshifts. We restrict the results again to the two base models *RawELGs* and *DustELGs*. We appreciate that the majority of ELGs coincide with the most massive galaxies.

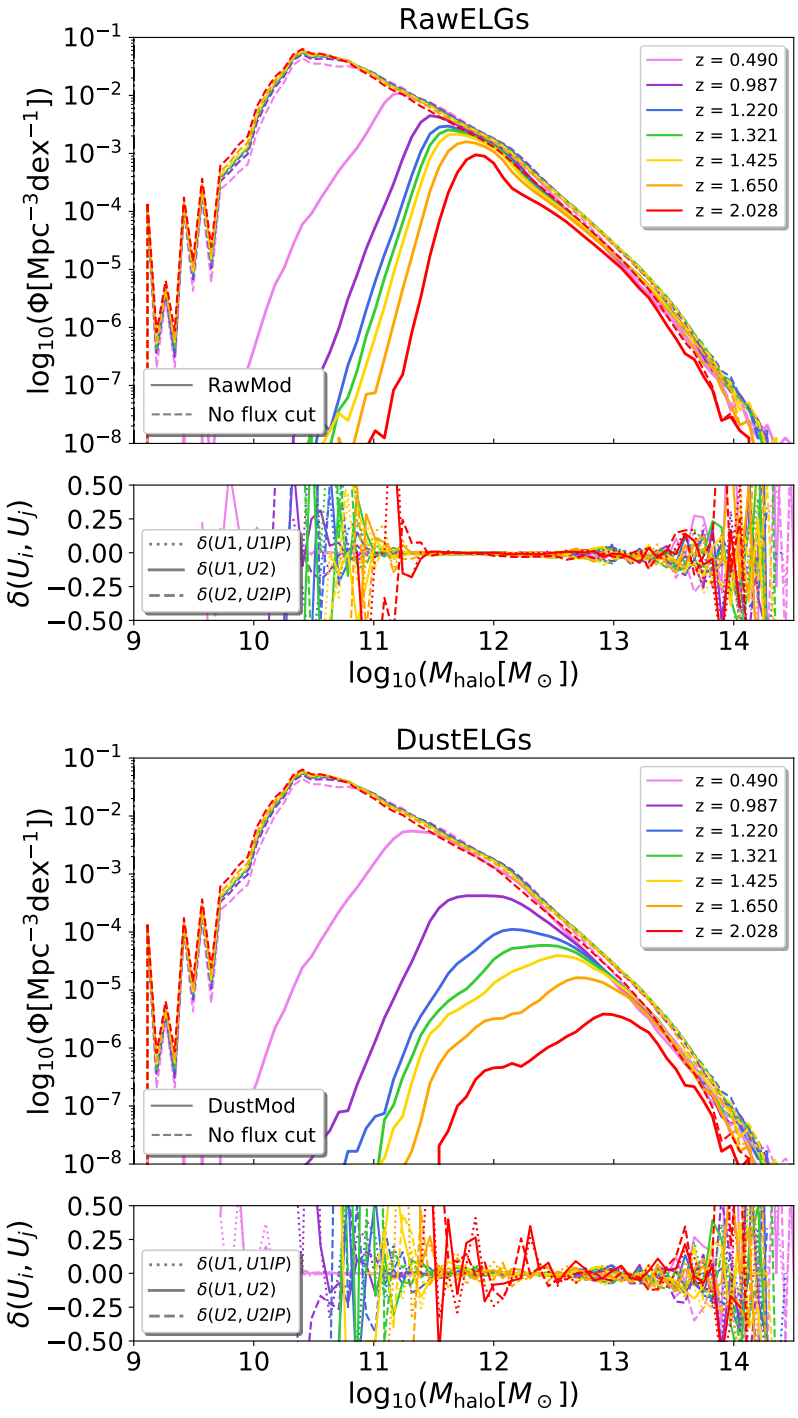

Figure B.1: Halo mass function of all ELGs (dashed lines) and the flux-selected samples (solid lines) for *RawELGs* (top) and *DustELGs* (bottom).

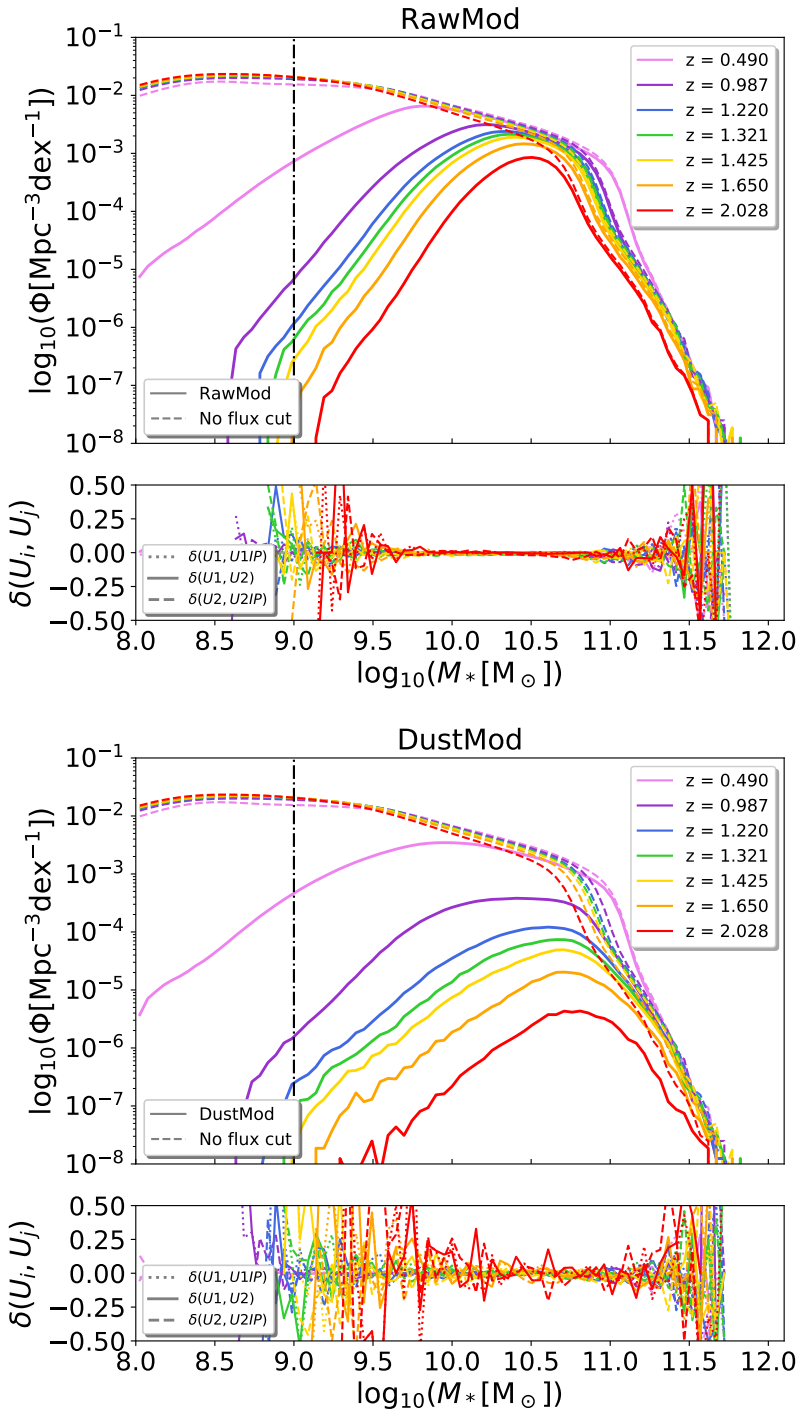

Figure B.2: Stellar mass function of all ELGs (dashed lines) and the flux-selected samples (solid lines) for *RawELGs* (top) and *DustELGs* (bottom). The vertical dot-dashed line shows our lower stellar mass limit.

Specific star formation rate

In Fig. 1.2 we show the specific star formation rate of all our SAGE galaxies in comparison to the observations of Daddi et al. (2007) at redshift $z \sim 2$. Here we now present in Fig. B.3 another version of that plot, this time using the (flux-cut) ELGs of the *RawELGs* and *DustELGs* catalogues. We further show results for $z \sim 1$ and add the best-fitting correlation for Hα emitting galaxies, as found by de los Reyes et al. (2015, eq. 3).[1]

The mass–metallicity relation

Here we reproduce Fig. 1.3 for the *RawELGs* and *DustELGs* catalogues, additionally adding the best-fit relation for Hα-emitting galaxies, as reported by de los Reyes et al. (2015, eq. 4). The results can be viewed in Fig. B.4, which shows that the SAGE-ELGs follow the observations sufficiently well.

The disc size–mass relation

At last we turn to the effective disc size of our *RawELGs* and *DustELGs* galaxies, shown in Fig. 1.4 for all SAGE galaxies. The results can be viewed in Fig. B.5, again in comparison to the general results of Yang et al. (2021).

[1] de los Reyes et al. (2015) studied 299 Hα-selected galaxies at redshift $z \sim 0.8$.

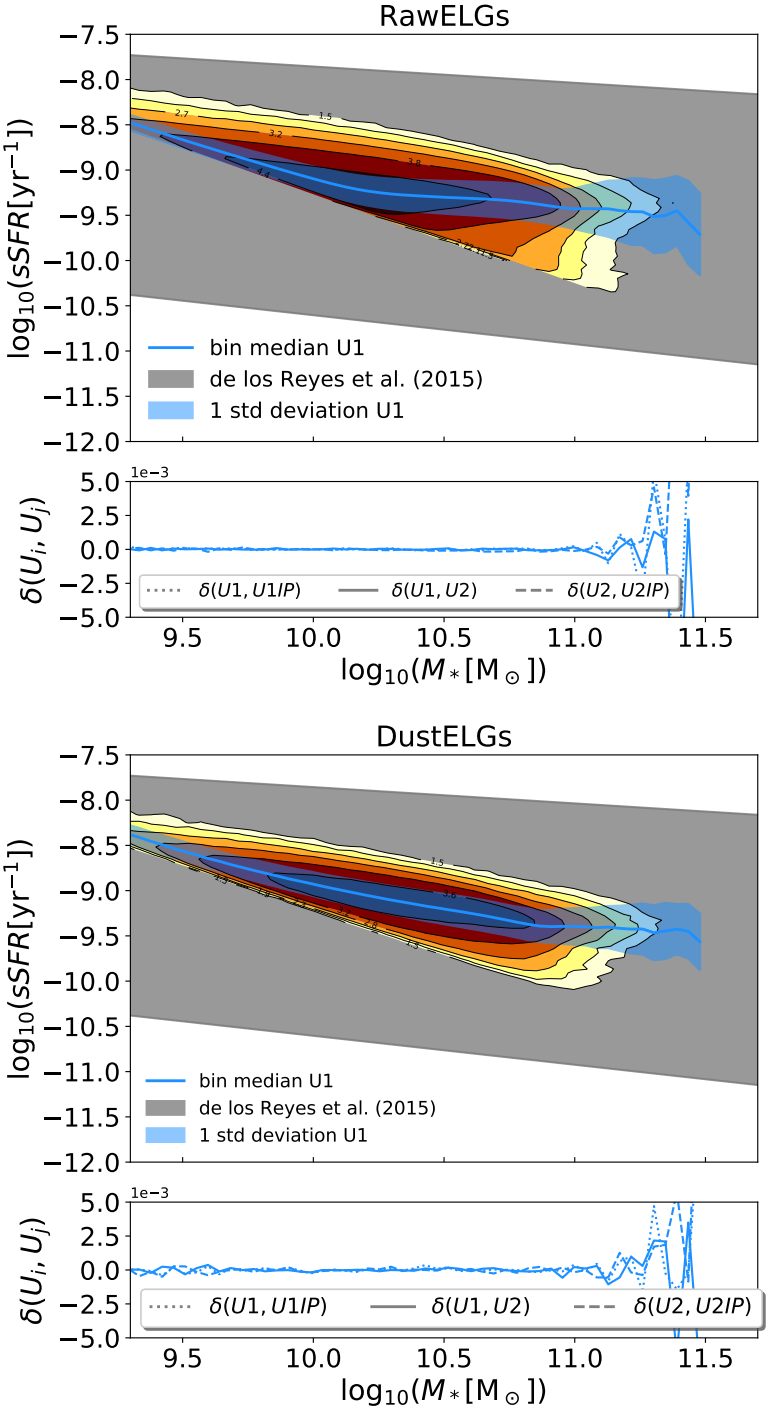

Figure B.3: Specific star formation rate of the *RawELGs* (top) and *DustELGs* (bottom) ELGs at redshift $z \sim 1$ in comparison to the best-fit relation as found by de los Reyes et al. (2015) at $z \sim 0.8$, shown as grey-shaded region. This figure is a reproduction of Fig. 1.2, but this time for our model ELGs.

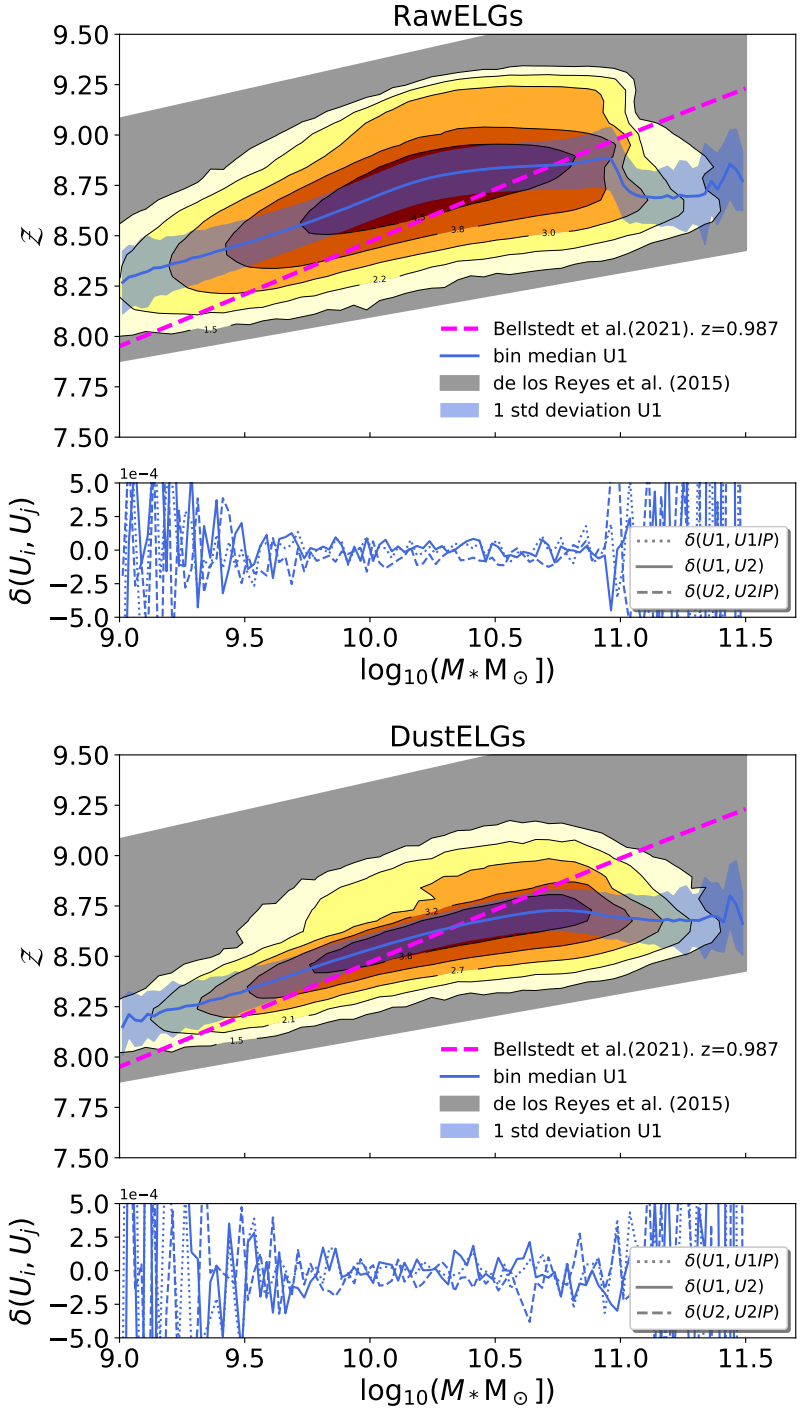

Figure B.4: Cold gas metallicity vs. stellar mass for all ELGs for *RawELGs* (top) and *DustELGs* (bottom) at redshift $z \sim 1$. This figure is a reproduction of Fig. 1.3, but this time for our model ELGs, but we also added the best-fitting relation as found by de los Reyes et al. (2015) at $z \sim 0.8$, shown as grey-shaded region.

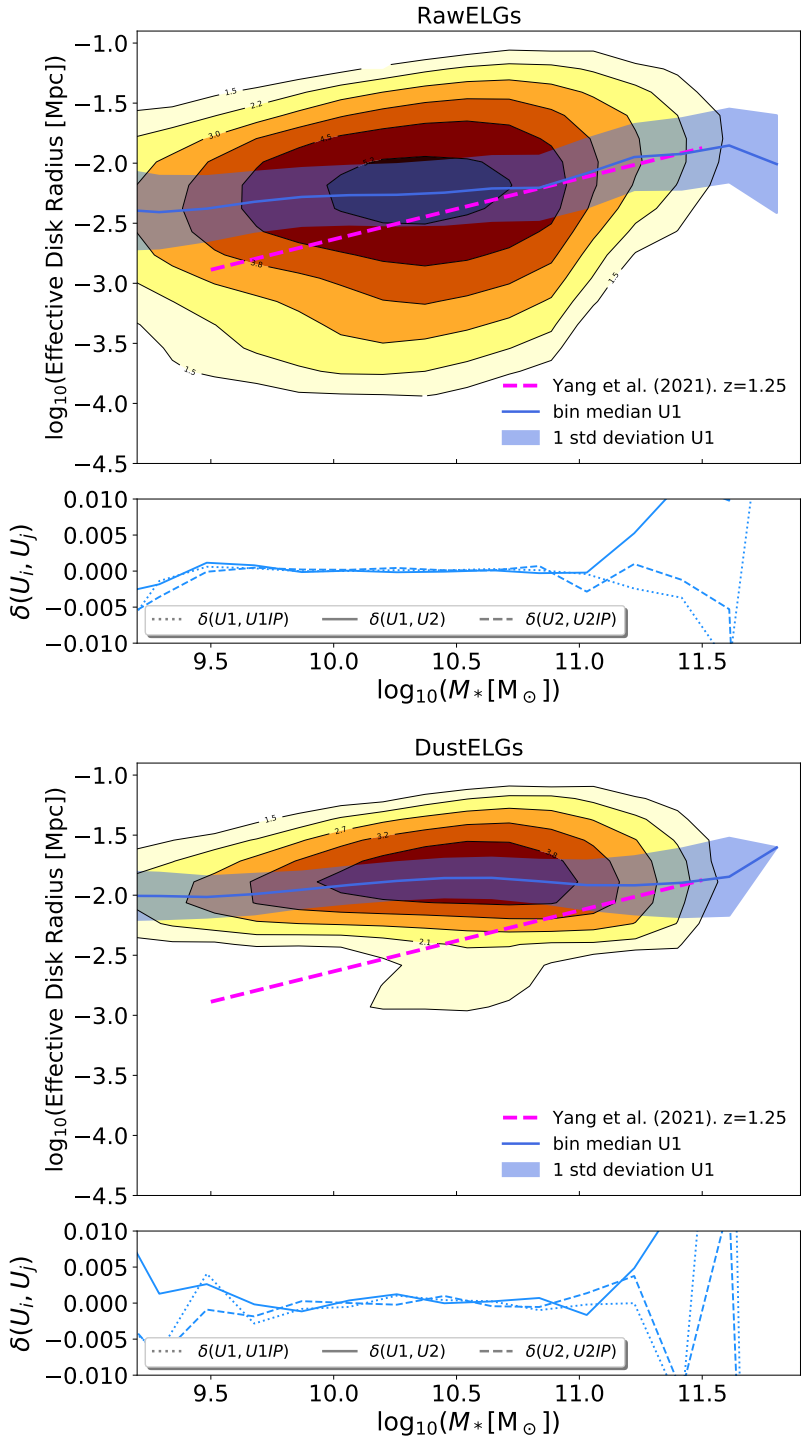

Figure B.5: Effective disc radius as a function of stellar mass at redshift $z = 1.25$ for all *RawELGs* (top) and *DustELGs* (bottom) galaxies. This figure is a reproduction of Fig. 1.4, but this time for our model ELGs.

Appendix C

Conversion of number densities

Here we show the steps necessary to go from volumetric number density

$$n = \frac{dN}{dV} \tag{C.1}$$

to the angular and redshift density

$$\eta = \frac{dN}{d\Omega\, dz}\,. \tag{C.2}$$

Taking into account

$$dV = d\Omega\, r^2 dr \tag{C.3}$$

where $d\Omega$ is the solid angle in stereoradians, we then get

$$\eta = n \cdot r^2 \frac{dr}{dz}\,. \tag{C.4}$$

Therefore, to go from number density $n = N/V$ of galaxies to number density of galaxies per square degree and redshift interval we find

$$\eta = n\, r^2(z) \frac{dr}{dz} \left(\frac{\pi}{180^\circ}\right)^2\,, \tag{C.5}$$

where $r(z)$ is the comoving distance

$$r(z) = \frac{c}{H_0} \int_0^z \frac{ds}{E(s)} \tag{C.6}$$

with

$$E^2(z) = \frac{1}{(\Omega_{r,0}(1+z)^4 + \Omega_{m,0}(1+z)^3 + \Omega_{k,0}(1+z)^2 + \Omega_{\Lambda,0})}\,, \tag{C.7}$$

where Ω_X are the usual density parameters of radiation ($X = r$), matter ($X = m$), curvature ($X = k$), and cosmological constant ($X = \Lambda$) at present time. We note that the derivative of $r(z)$ with respect to z as needed in Eq. (C.5) is simply

$$\frac{dr}{dz} = \frac{c}{H_0} \frac{1}{E(z)}. \tag{C.8}$$

Note that in the main body of the paper η is referred to as dN/dz, which is not fully consistent with the terminology used here, but compliant with how other workers in the field refer to this quantity. N as used in the main part is 'number of galaxies per unit area', whereas here it simply means 'number of galaxies'.

Appendix D

Description of L16 Model

Here we explain the main ideas behind the Ludlow et al. (2016) (L16) model for predicting the $c(M, z)$ relation. Throughout this work we carefully examined if the assumptions upon which L16 is founded are fulfilled or not for a variety of cosmologies, masses and redshifts. The original L16 paper shows that their model accurately predicts the $c(M, z)$ relation for relaxed haloes in different cosmologies, including both cold and warm dark matter scenarios; in this work we extend their analysis by considering a broader range of distinct cosmologies, including the effect of massive neutrinos and dynamical dark energy.

The L16 model is based on an empirical relation between $\rho_{-2}(z_0)$, i.e. the enclosed density of a halo within the scale radius, r_{-2} measured at redshift z_0, and $\rho_c(z_{-2})$, i.e. the critical density of the universe defined at a suitable formation redshift, z_{-2}, The relation can be written as:

$$\rho_{-2}(z_0) = A\rho_{\rm c}(z_{-2}), \tag{D.1}$$

where A is a proportionality constant. In L16 the halo formation redshift, z_{-2}, is defined as the redshift at which the collapsed-mass history (CMH) of a halo[1] first exceeds $M_{-2} \equiv M(r < r_{-2})$, i.e., the mass enclosed within a sphere of radius r_{-2}, at the z_0, centered around the potential minimum of the halo analyzed.

If indeed Eq. (D.1) is verified, we can predict the value of $\rho_{-2}(z_0)$ employing an analytical model capable of reproducing the CMH of a halo given its mass, which would allow us to infer $\rho_c(z_{-2})$. To obtain the synthetic CMHs, we make use of the extended Press-Schechter (EPS) formalism (Bond et al., 1991b; Lacey and Cole, 1993), according to which the mass contained in progenitors more massive than a certain fraction f of the final halo mass, M_0, at a given redshift, z, is given by:

[1]For a given halo identified at redshift z_0, the collapsed-mass history is defined as the sum of all the mass contained in progenitor haloes at redshift $z > z_0$ that end up being accreted by the halo of interest and whose mass exceeds a certain fraction f (in L16 $f \equiv 0.02$) of the halo's final mass. This can be calculated for simulated haloes using their merger trees, or predicted theoretically using the extended Press-Schechter formalism using Eq. (D.2).

$$\mathrm{M_{coll}}(z) = \mathrm{M_0}\,\mathrm{erfc}\left\{\frac{\delta_\mathrm{c}\left(\frac{D(z,k_{f\mathrm{M_0}})}{D(z_0,k_{f\mathrm{M_0}})}-1\right)}{\sqrt{2\left[\sigma^2\left(k_{f\mathrm{M_0}},z_0\right)-\sigma^2\left(k_{\mathrm{M_0}},z_0\right)\right]}}\right\} \tag{D.2}$$

where δ_c is the threshold for non-linear collapse extrapolated to $z = 0$ using linear theory. The value of δ_c can be calculated using the spherical collapse model, which predicts $\delta_\mathrm{sc} \approx 1.686$. However, we found that adopting a value of $\delta_\mathrm{c} = 1.46$ improves the agreement between the EPS-predicted collapsed mass histories and those obtained from our simulations, and therefore minimizes the error in the predicted redshifts of halo collapse (see Fig. 2.5). This is crucial for obtained accurate predictions for halo concentrations from the L16 model, since it relies on having accurate predictions for halo formation times.

Note $\sigma^2\left(k_\mathrm{M}, z\right)$ denotes the variance of the linear matter density field at redshift z and at scale k_M (associated with the mass $\mathrm{M} \propto k_\mathrm{M}^{-3}$). To compute $\sigma^2\left(k_\mathrm{M}, z\right)$ we use a sharp-k window function (which in real space can be written $W(x) = 3(\sin x - x\cos x)/x^3$, where $x \propto k^{-1}$) which exploits the fact that, for a Gaussian random field, the derivation of the EPS formula becomes simpler because overdensity "trajectories" in the smoothed density field follow Markovian random walks (Bond et al., 1991b; Lacey and Cole, 1993).

The scale dependent growth factor, $D(z, k_\mathrm{M})$, can be computed at redshift z and for scale k_M following Zennaro et al. (2017). The scale dependence of the growth factor introduces significant corrections when considering massive neutrinos, which impact the growth of structures differently at different scales in a manner that also depends on the neutrino mass.

Note that in Eq. (D.2) we evaluate the variance of the matter field and the growth factor at different scales. This is particularly important for calculating the CMHs in cosmologies with massive neutrinos, where the scale dependence of the growth factor can have a significant impact.

Appendix E

$c(M)$ relation at $z_0 = 0.5$

In Fig. E.1 we present the results for the concentration-mass relation (as measured in Fig. 2.2) at $z = 0.5$ (connected squares). We also show the predictions provided by the L16 model at that redshift employing the same calibration as the one used in Fig. 2.2.

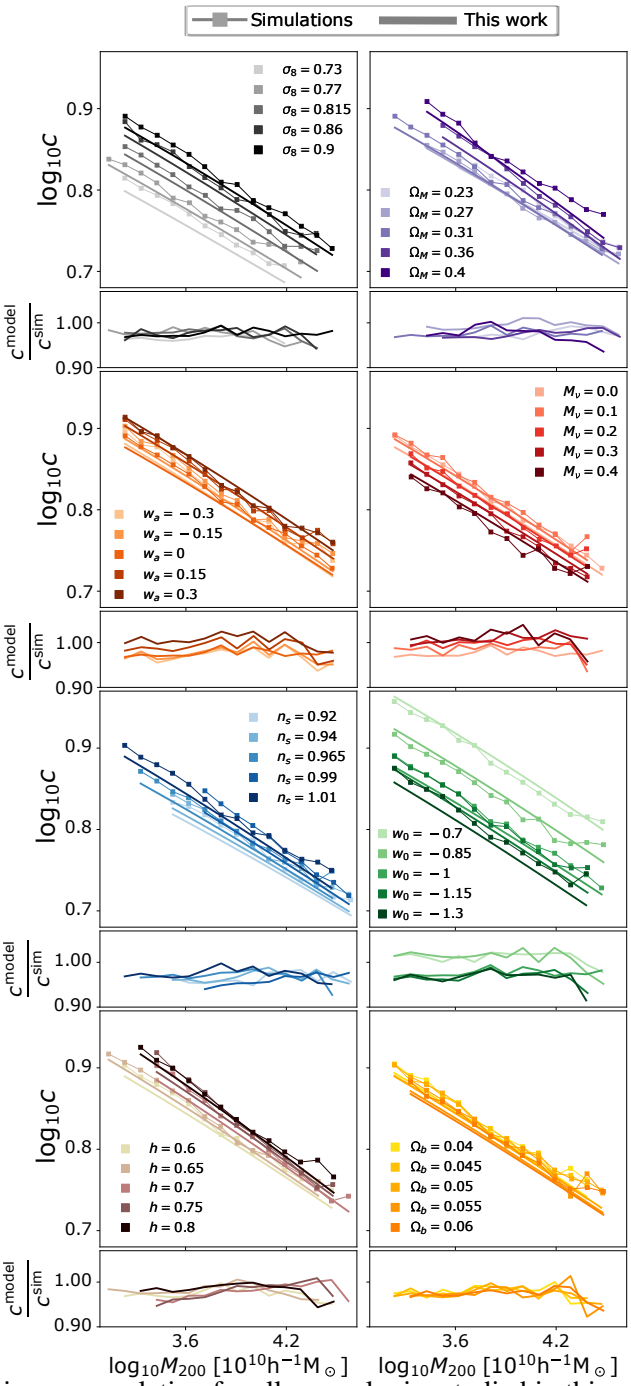

Figure E.1: Concentration-mass relation for all cosmologies studied in this paper at redshift $z = 0.5$ as a function of M_{200} analogous to the results presented in Fig. 2.2.

Appendix F

Watershed segmentation

In this appendix, we present an alternative approach to instance segmentation, based on the watershed approach. Originally we tried this technique to address the instance segmentation problem, but we finally decided to use the Weinberger approach presented in the main paper because of its theoretical advantages. These are that the loss function closer reflects the objective, that it is possible to predict disconnected regions, and that it is not necessary to define borders. However, during our exploration, we have gained some insights of how to make watershed-based instance segmentation techniques work for friends-of-friends proto-haloes. We will explain these here for the benefit of future studies.

Our watershed approach makes use of a U-Net-based architecture Ronneberger et al. (2015), specifically a 3D Residual U-Net based on previous work Franco-Barranco et al. (2021). The model's input consisting of $128 \times 128 \times 128 \times 2$ voxels for (x, y, z, $channels$) axes. The two input *channels* correspond to the initial density field and the potential.

The model is trained to predict two output channels: binary foreground segmentation masks and instance contours masks. Following the prediction, the two outputs are thresholded (automatically using Otsu's method Otsu (1979)) and combined. Next, a connected components operation is applied to generate distinct, non-touching halo instance seeds. Subsequently, a marker-controlled watershed algorithm Meyer (1994) is applied, using three key components: 1) the inverted foreground probabilities as the input image (representing the topography to be flooded), 2) the generated instance seeds as the marker image (defining starting points for the flooding process), and 3) a binarized version of the foreground probabilities as the mask image (constraining the extent of object expansion). To binarize the latter, we employed a threshold value of 0.372, which was determined through the application of the identical methodology outlined in Appendix H. The collective implementation of these components facilitates the creation of individual halo instances (see Fig. F.1 for a visual representation). This strategy has been extensively employed within the medical field with remarkable success Wei et al. (2020); Lin et al. (2021); Andres-San Roman et al. (2023).

In order to facilitate the generation of the two channels used to train the network, several

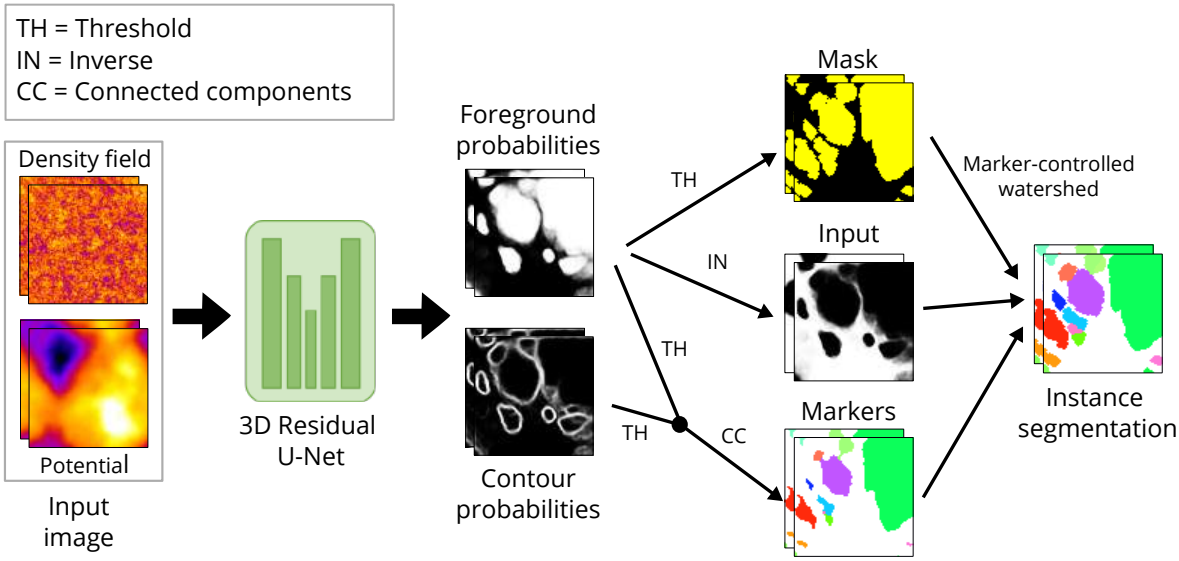

Figure F.1: Processing pipelines of our watershed segmentation approach. The input 3D image contains two channels: the density field and the potential. The model predicts foreground and contour probabilities that are fused to create three inputs for a marker-controlled watershed to produce individual instances.

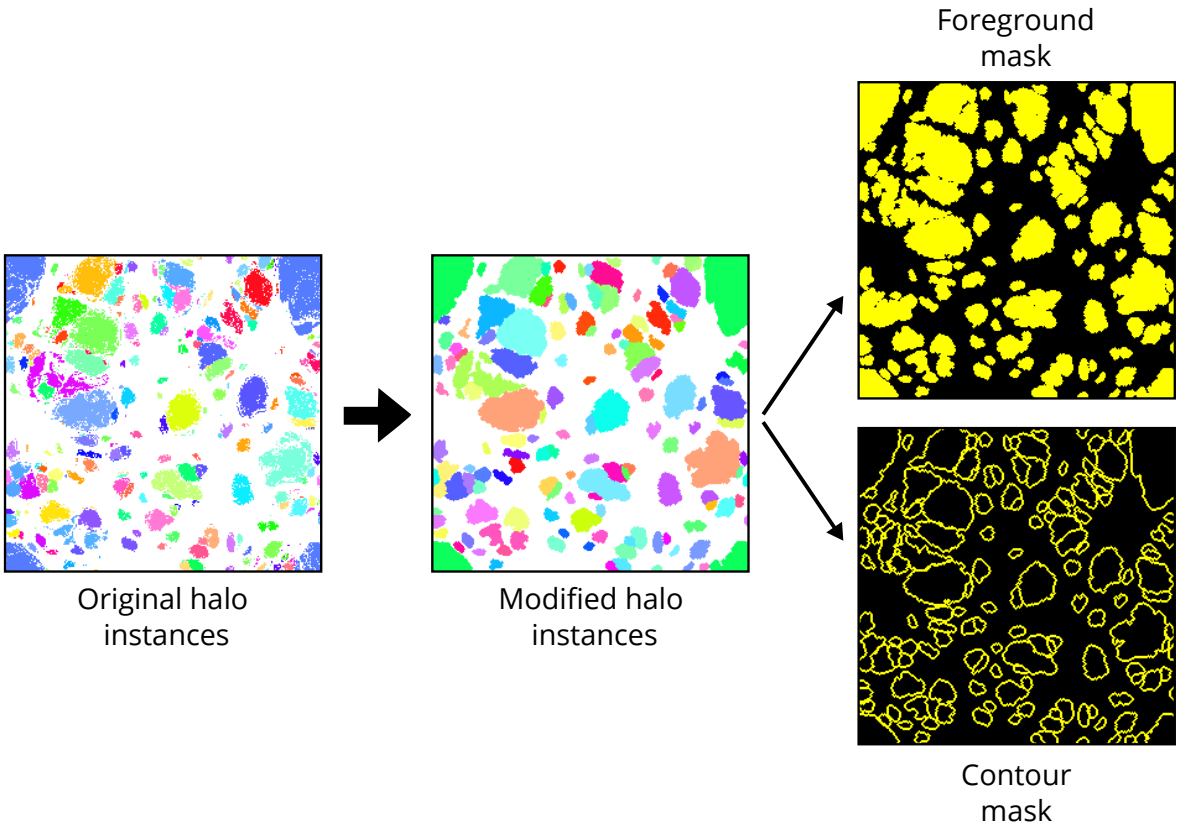

Figure F.2: Data preparation process of our watershed segmentation approach. From left to right: the original halo instances for the considered prediction problem, subsequent modifications involving the removal of small holes and spurious pixels and contour smoothing, and the presentation of both the foreground and contour masks utilized for model training. Pixels coloured in white do not belong to any halo. Pixels with the same colour belong to the same halo and different colours indicate different haloes.

transformations were applied to the labels. For each halo instance, small particles along the edges were removed, central holes were filled, and the labels were dilated by one pixel. This process results in instances with smoother boundaries, thereby aiding the network in training (see Fig. F.2).

The result of this method is depicted in Fig. F.3. The code is open source and readily available in BiaPy Franco-Barranco et al. (2023).

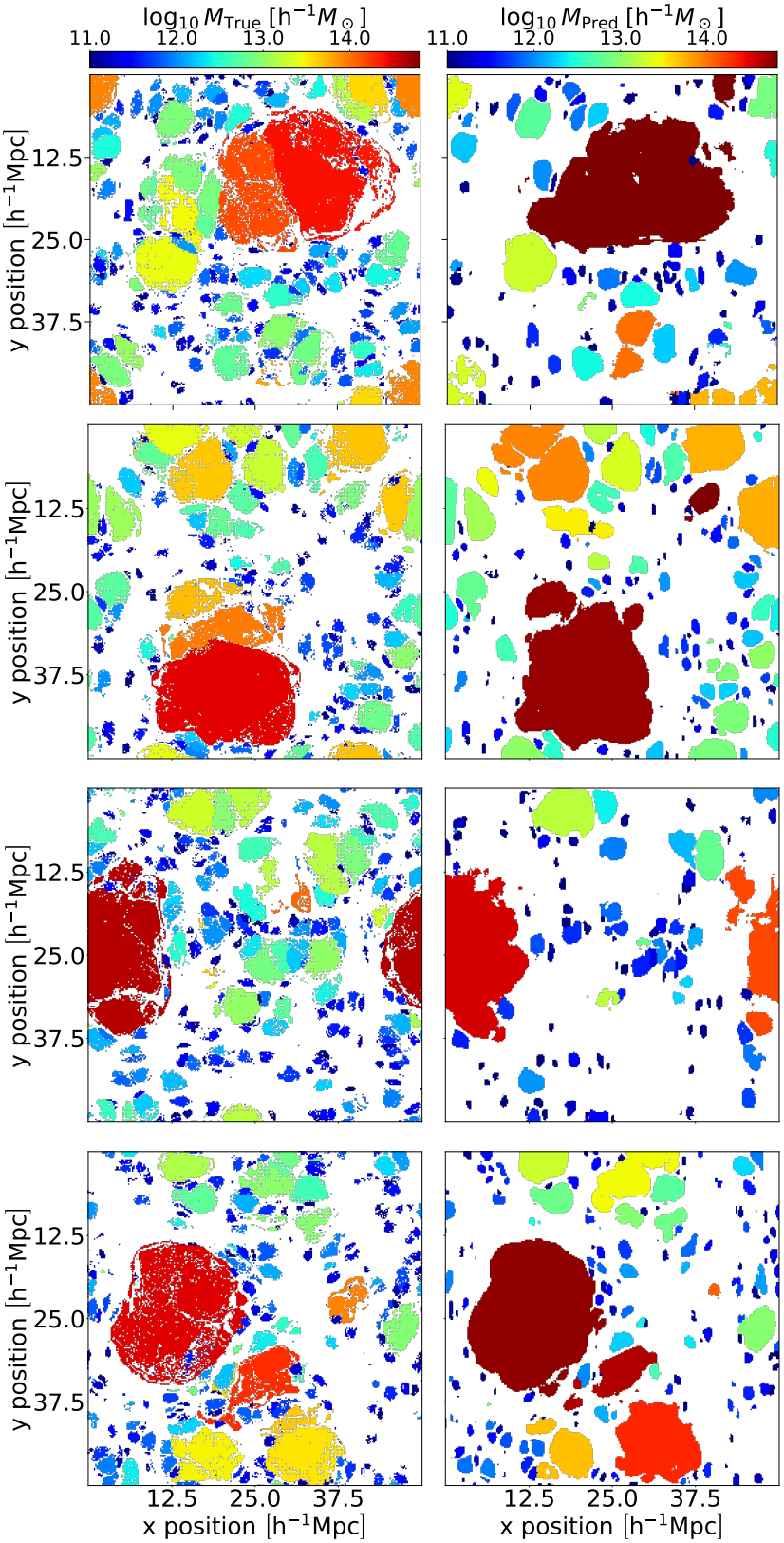

Figure F.3: Results of our watershed segmentation approach presented in an analogous way to results from Fig. 3.7.

Appendix G

Clustering algorithm

In this appendix, we describe the clustering algorithm that we have developed. This algorithm calculates instance predictions from the pseudo-space representations that are output by our instance segmentation network.

As described in §§3, the output of our instance network consists of a set of points that populate an abstract space (referred to as pseudo-space). Our instance network has been trained to minimize the Weinberger loss function 3.7, hence, we expect that the predicted mapping of points in the pseudo-space causes that points corresponding to the same instances to be close to each other, and separated to points that correspond to different instances. In the ideal case where $\mathcal{L}_{\text{Wein}}$=0, all points belonging to the same instance would be no farther apart from each other than a distance $2 \cdot \delta_{\text{Pull}}$, and the points corresponding to separate instances would be, as close as a distance $2 \cdot \delta_{\text{Push}} - \delta_{\text{Pull}}$ close to each other. However, we cannot expect that our network always separates perfectly the different instances. For example, if some Lagrangian voxel has a 60% chance to belong to halo A and a 40% chance to belong to halo B, then the optimal location in pseudo space (that statistically minimizes the loss) may be somewhere in between the centre of halo A and B in pseudo space and not inside the δ_{Pull} radius of neither. Therefore, we employ a clustering algorithm that can segment the pseudo-space distribution of points also when $\mathcal{L}_{\text{Wein}}$ is not exactly zero.

For this, we first estimate the local pseudo-space density ρ_i for each point i. For this we compute the distance $r_{k,i}$ to the kth-nearest neighbour of the point and assign

$$\rho_i = \frac{3k}{4\pi r_{k,i}^3} \tag{G.1}$$

where $k = N_{\text{dens}}$ is a hyper-parameter of the clustering algorithm. We accelerate this step with the CKD-TREE from the SCIPY package in PYTHON (Virtanen et al., 2020).

Then we determine groups as the descending manifold of the maxima that exceed a persistence ratio threshold $\rho_{\text{max}}/\rho_{\text{sad}} \geq p_{\text{thresh}}$ between maximum and saddle-point. The descending manifold corresponds to the set of particles from whose location following the local density gradient would end up in the same maximum (e.g. Sousbie, 2011; Tierny et al.,

2017). For this, we use a slightly modified version of the density segmentation algorithm used in SUBFIND (Springel et al., 2001):

We consider the particles from highest to lowest density. For each particle we consider from the $N_{\rm ngb}$ nearest particles the subset of particles that have a higher density than ρ_i (this set may be empty). Among these we select the set B_i of the (up to) two closest particles. This set can have zero, one or two particles.

- If the set B_i is empty, then there is a density maximum $\rho_{\rm max} = \rho_i$ and we start growing a new subgroup around it.
- If the set B_i contains a single particle or two particles that are of the same group, the particle i is attached to the corresponding group.
- If B_i contains two particles of different groups, then i is potentially a saddle-point. We check whether the group with the lower density maximum $\rho_{\rm max}$ has a sufficient persistence $\rho_{\rm max}/\rho_i \leq p_{\rm thresh}$. If not, then we merge the two groups (and keep the denser maximum). Otherwise, we keep both groups and we assign the particle to the group of the denser particle in B_i. (This step corresponds to following the local discrete density gradient.)

Note that unlike the SUBFIND algorithm, we merge groups not at every saddle-point, but only if they are below a persistence threshold. Therefore, sufficiently persistent groups are grown beyond their saddle point and ultimately correspond to the descending manifold of their maximum.

The clustering algorithm has three hyper-parameters $N_{\rm dens}$, $N_{\rm ngb}$ and $p_{\rm thresh}$. We have done a hyper-parameter optimization over these and found that $N_{\rm dens} = 20$, $N_{\rm ngb} = 15$ (quite close to the default parameters in the SUBFIND algorithm, 20 and 10 respectively) and $p_{\rm thresh} = 4.2$ give the best results, though our results are not very sensitive to moderate deviations from this. We can understand the quantitative value of the persistence ratio threshold by considering that the relative variance of our density estimate is

$$\sigma_{\log\rho} \approx \frac{\sigma_\rho}{\rho} = \frac{1}{\sqrt{N_{\rm dens}}} \approx 0.22 \tag{G.2}$$

so that at a fixed background density having a density contrast of $p_{\rm thresh} = 4.2$ due to Poisson noise corresponds to a

$$\Delta \log\rho = \log(p_{\rm thresh}) \approx 1.43 \approx 6.5\sigma_{\log\rho} \tag{G.3}$$

outlier. Therefore, the persistence ratio threshold $p_{\rm thresh}$ ensures that it is very unlikely that our algorithm mistakes a spurious overdensity in the pseudo space for a group.

Appendix H

Semantic threshold

In the bottom panel of Fig. H.1 we present how the predicted fraction of voxels that are members of a halo (that is $1-\beta$) evolves as we change the semantic threshold (black solid line). As it can be expected when the semantic threshold is close to zero, the majority of voxels are identified as members of haloes, and the contrary occurs when the semantic threshold approximates one. The horizontal dashed-dotted line corresponds to the ground truth value of $1-\beta = 0.418$, measured in the validation simulations. The semantic threshold value that we have selected is 0.589 (black dotted vertical line). This value corresponds to the intersection between the black solid line and the dashed-dotted line; it ensures that the total fraction of voxels that are members of haloes is correctly reproduced. Choosing this criterion to determine the semantic threshold also ensures more robust instance predictions since the number of FP cases is reduced, hence eliminating potentially uncertain pseudo-space particles that would complicate the clustering procedure.

In the top panel of Fig. H.1 we show the evolution of several metrics as a function of the semantic threshold value. These metrics allow us to asses the quality of our semantic predictions by comparing our results with values obtained using the baseline simulations. We study the behaviour of five different metrics: True Positive Rate TPR, True Negative Rate TNR, Positive Predictive Value PPV, Accuracy ACC and the F_1-score.

In the top panel of Fig. H.1 we also present the values obtained for the different metrics using the baseline simulations (horizontal dashed lines). We have obtained these results considering one of the baseline simulations as predicted maps and the other simulation as the ground truth. The values measured for the different metrics in the baseline simulations give us an expected ideal performance that we would like to reproduce with our model.

If we focus on the performance curves for the accuracy and the F_1-score (orange and yellow lines respectively) we can appreciate that they always remain under the baseline limit. The curve for the F_1-score peaks around the value for the semantic threshold of 0.5, which is a behaviour we expected since we considered the balanced cross-entropy loss to train our semantic model. The value for the F_1-score at its maximum is $\mathrm{F}_1(0.5) = 0.842$, which

is very similar to the value at the point in which we have fixed the semantic threshold, $\mathrm{F}_1(0.589) = 0.838$. The F_1-score obtained is only about 5% away from the optimal value obtained from the baseline simulations $\mathrm{F}_1^{\mathrm{Chaos}} = 0.884$. The accuracy reaches its maximum value around the semantic threshold of 0.58, where $\mathrm{ACC}(0.58) = 0.864$; the value for the model accuracy is even closer to the baseline limit $\mathrm{ACC}^{\mathrm{Chaos}} = 0.903$.

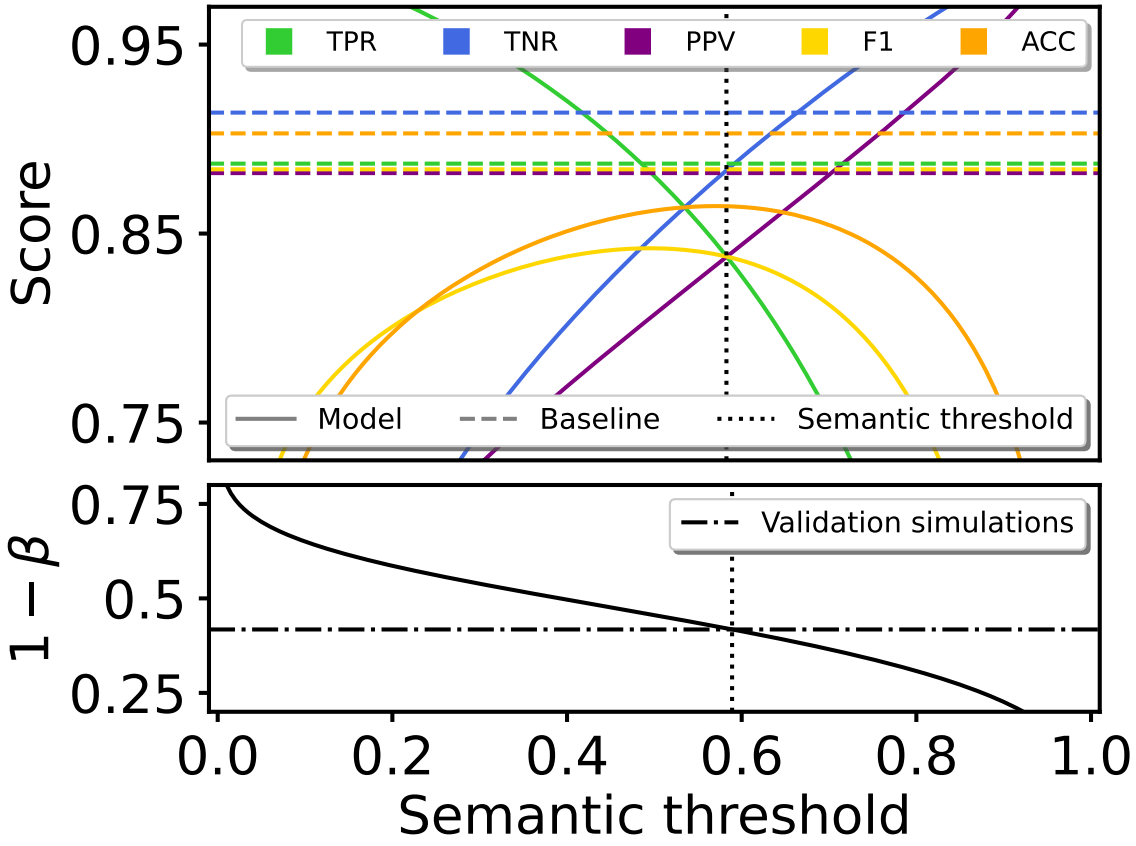

Figure H.1: Top panel: Evolution of different metrics (TPR - green, TNR - blue, PPV - purple, F_1-score - yellow & ACC - orange) measured employing the predictions of the semantic model as a function of the semantic threshold selected (solid lines); we also show the values measured for the corresponding metrics studying the differences between the baseline simulations (horizontal dashed lines). Bottom panel: Fraction of voxels predicted to be collapsed (equivalent to $1 - \beta$) as a function of the semantic threshold employed (solid black line); the horizontal black dashed line corresponds to the fraction of particles that end up in DM haloes measured in the validation simulations. In both panels, the vertical black dotted line shows the semantic threshold we employ; this threshold has been selected to match the fraction of collapsed voxels.

Appendix I

Generate full-box predictions from crops

In this appendix, we address the challenge of generating full-box predictions employing our instance segmentation model.

While our network architecture captures intricate features within simulation sub-volumes, the challenge arises when we aim to apply it to arbitrarily large input domains. Unlike some other ML approaches that rely on networks that are translational invariant, our model incorporates the Lagrangian positions of particles as input channels, making it dependent on the relative Lagrangian position. This design choice ensures that similar regions of the initial density field are mapped to distinct locations in the pseudo-space, allowing us to distinguish between separate structures, even if they are locally identical. However, this feature also presents a challenge when creating full-box predictions. Combining independent crop predictions straightforwardly may lead to inconsistencies due to the network's inherent non-translational invariance. To tackle this issue, we have developed a methodology for predicting sub-volumes independently and then merging these predictions to generate accurate full-box instance segmentation results.

To reduce the boundary effects that may result from such a method we employ the following strategy.

1. We evaluate the instance network centred several times, centred on locations $\vec{q}_{ijk}$ that are arranged on a grid

$$\vec{q}_{ijk} = \begin{pmatrix} i \cdot n_{\text{off}} \\ j \cdot n_{\text{off}} \\ k \cdot n_{\text{off}} \end{pmatrix}, \tag{I.1}$$

   where we choose an offset of $n_{\text{off}} = 64$ voxels and (i, j, k) run so far that the whole periodic volume is covered – e.g. from 0 to 4 each for a 256^3 simulation box. The network's input in each case corresponds to the 144^3 voxels (periodically) centred on $\vec{q}_{ijk}$ and the instance segmentation output will predict labels for the 128^3 central voxels.

2. From each prediction we only use the predicted labels of the central $n_{\text{off}}^3 = 64^3$ voxels, since we expect these to be relatively robust to field-of-view effects. We combine these

from all the predictions to a global grid that has the same dimensions as the input domain. In this step we add offsets to the labels so that the labels that originate from each predicted domain are unique in the global grid (this process will become relevant in step 4 where we define a graph used to link instances).

3. We repeat steps 1-2, but with an additional offset of $(n_{\rm off}/2, n_{\rm off}/2, n_{\rm off}/2)^T$. We additionally offset the labels in this second grid so that no label appears in both grids.

4. We use the two lattices and the intersections between instances to identify which labels should correspond to the same object. We do this by creating a graph[1] where each instance label is a node. Initially the graph has no edges, but we subsequently add edges if two labels should be identified (i.e. correspond to the same halo). Each connected component of the graph will then correspond to a single final label. To define the edges of the graph, we consider each quadrant Q of size $(n_{\rm off}/2)^3$ individually, since such quadrants are the maximal volumes over which two labels can intersect. We define the intersection $I_Q(l_1, l_2)$ of two labels l_1 and l_2 as the number of voxels that both carry label l_1 in grid one and label l_2 in grid two. We define as the union $U_Q(l_1, l_2)$ the number of voxels inside of quadrant Q that carry l_1 in grid 1 or l_2 in grid 2 (or both). We then add an edge between l_1 and l_2 into the graph if for any quadrant Q it is

$$\frac{I_Q(l_1, l_2)}{U_Q(l_1, l_2)} \geq IoU_{\rm thresh} \tag{I.2}$$

where we set $IoU_{\rm thresh} = 0.5$.

5. We summarize each connected component in the graph into a new label. After this operation for most voxels the new label in grid 1 and in grid 2 agree and we can choose that label as our final label. However, for a small fraction of voxels the labels still disagree, because the corresponding instances had too little overlap to be identified with each other. In this case, we assign to the corresponding voxel the label that contains the larger number of voxels in total.

We illustrate the different steps of this procedure in Fig. I.1. The top panel, labelled 'Lattice1', shows the individual instances predicted in the first lattice arrangement. Each colour represents a distinct label assigned to a group of voxels within the 64^3 central region of the sub-volumes. The middle panel, 'Lattice2', displays the second set of predictions using a shifted lattice by half the offset in each dimension. Here again, different colours represent unique instance labels. The bottom panel, 'Combined', presents the final merged full-box prediction. It is generated by synthesizing the labels from 'Lattice1' and 'Lattice2' using

[1] using the NETWORKX library (Hagberg et al., 2008)

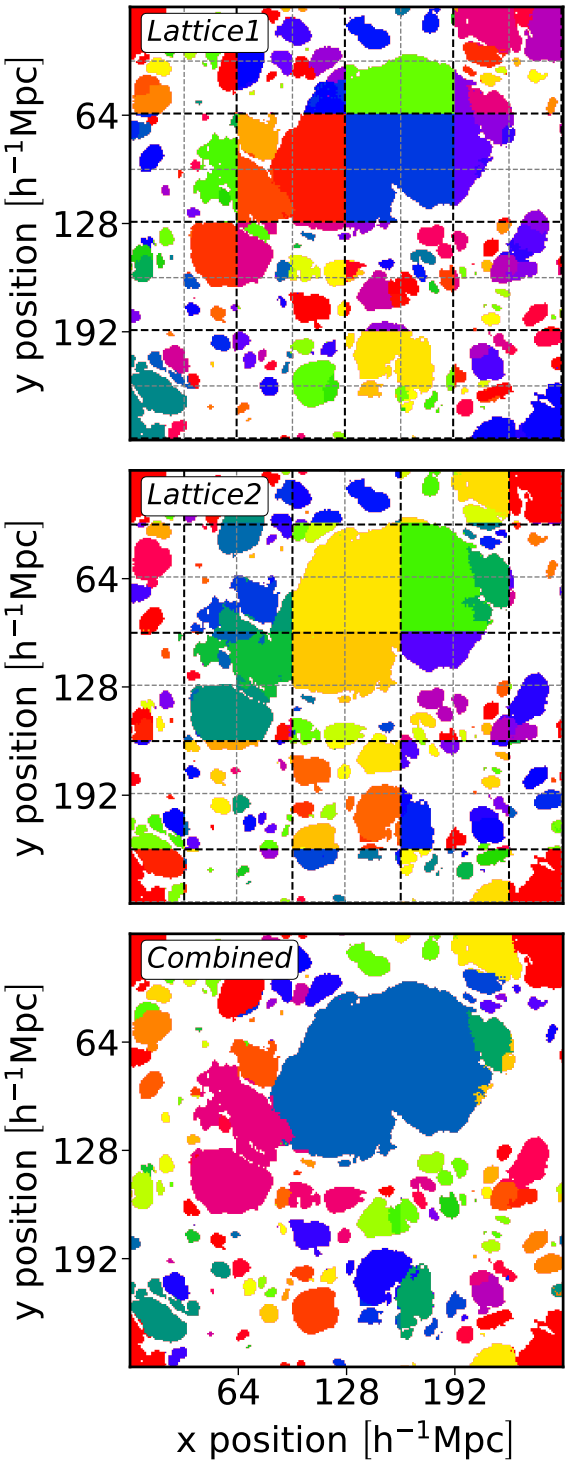

Figure I.1: Process of merging predictions from two overlapping lattice structures to produce a full-box instance segmentation map. 'Lattice1' (**top**) and 'Lattice2' (**middle**) represent predictions from initial and shifted lattice grids, respectively, with unique color-coded labels for instances. Black dashed lines indicate the lattice employed in each case, while thin dashed grey lines correspond to the lattice employed in the reciprocal scenario. 'Combined' (**bottom**) depicts the final synthesized full-box map, where instances have been merged based on their overlap, demonstrating the effectiveness of the methodology in generating contiguous and comprehensive halo segmentations from smaller, predicted sub-volumes.

the graph-based method to connect overlapping instances. The resulting image shows larger, coherent structures, indicative of the correct performance of combining both lattices.

Regarding the semantic segmentation network, we can merge the predictions corresponding to different crops independently since, in this case, we are truly working with a translation-invariant network. We employ the central 64^3 voxels (analogous to 'Lattice1') of separate predictions and merge them together to generate the final full-box predictions of the semantic segmentation network.

Appendix J

Comparison with ExSHalos

In this appendix, we explore how the results obtained with the ExSHalos code (Voivodic et al., 2019) compare against our semantic and instance predictions.

As mentioned in §§3, ExSHalos is an explicit implementation of the excursion set theory that identifies haloes in Lagrangian space by growing spheres around density peaks until the average density inside crosses a specified barrier for the first time. The barrier shape is motivated by the ellipsoidal collapse (Sheth et al., 2001; de Simone et al., 2011) and we have fitted the three free parameters in the model to reproduce the mean halo mass function of our simulations.

In Fig. J.1 we show a map-level comparison between the Lagrangian proto-haloes identified in one of our validation simulations with the friends-of-friends algorithm (left panel), and the ExSHalos detected employing the code presented in Voivodic et al. (2019) (central panel). The ExSHalos regions in Lagrangian space are spherical by construction (see the middle panel of Fig. J.1). The physical approach of the ExSHalos algorithm enables to identify, with a reasonable degree of accuracy, the location of proto-haloes in Lagrangian space, and their mass. However, the built-in assumption that proto-haloes are spherical gives only a crude approximation to the actual proto-halo shapes. In Table 3.2 we quantify the differences between ExSHalos and friends-of-friends employing several semantic metrics.

In Fig. J.2 we present a violin plot analogous to Fig. 3.8. This plot shows a comparison between the ground truth halo masses (friends-off-friends) and the predicted masses from our model associated with the particles/voxels in our validation set (black violin lines in the main panel). We also include the comparison between the masses of ExSHalos and of friends-off-friends haloes (purple violin lines). We have generated the violin lines of ExSHalos employing all our simulations (both training and validation) to achieve better statistics. Our model predictions are capable of achieving greater mass accuracy than ExSHalos throughout all mass bins considered here.

In the upper panel of Fig. J.2, we show the False Negative Rate (FNR) as solid lines against the ground truth halo mass, and the False Discovery Rate (FDR) as dashed lines against the

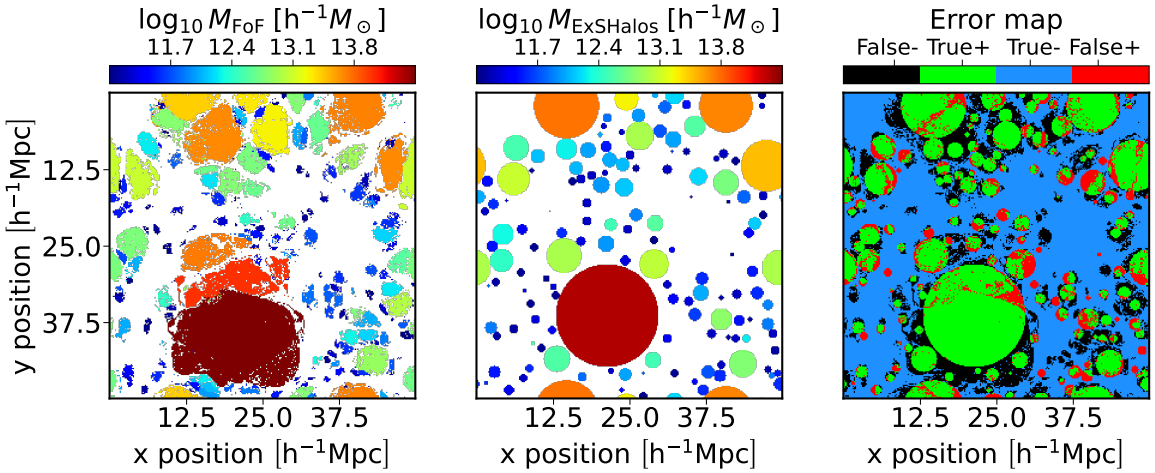

Figure J.1: Slices through the Lagrangian field of friends-of-friends proto-haloes, and the corresponding predictions using the ExSHalos algorithm. **Left panel**: ground truth masses obtained using N-body simulations (friends-of-friends proto-haloes). **Central panel**: predicted masses obtained using the ExSHalos algorithm. **Right panel** (analogous to left panel of Fig. 3.5): Semantic pixel-level error map between ExSHalos and friends-of-friends haloes indicating true positive (green), true negative (blue), false negative (black), and false positive (red) regions.

predicted mass. This plot is analogous to the top plot in Fig. 3.8 (See §§3 for details). We additionally include solid and dashed purple lines corresponding to the ExSHalos case. It's clear that ExSHalos predicts higher FNR and FDR values compared to the baseline case and our model predictions, indicating more semantically-misclassified particles.

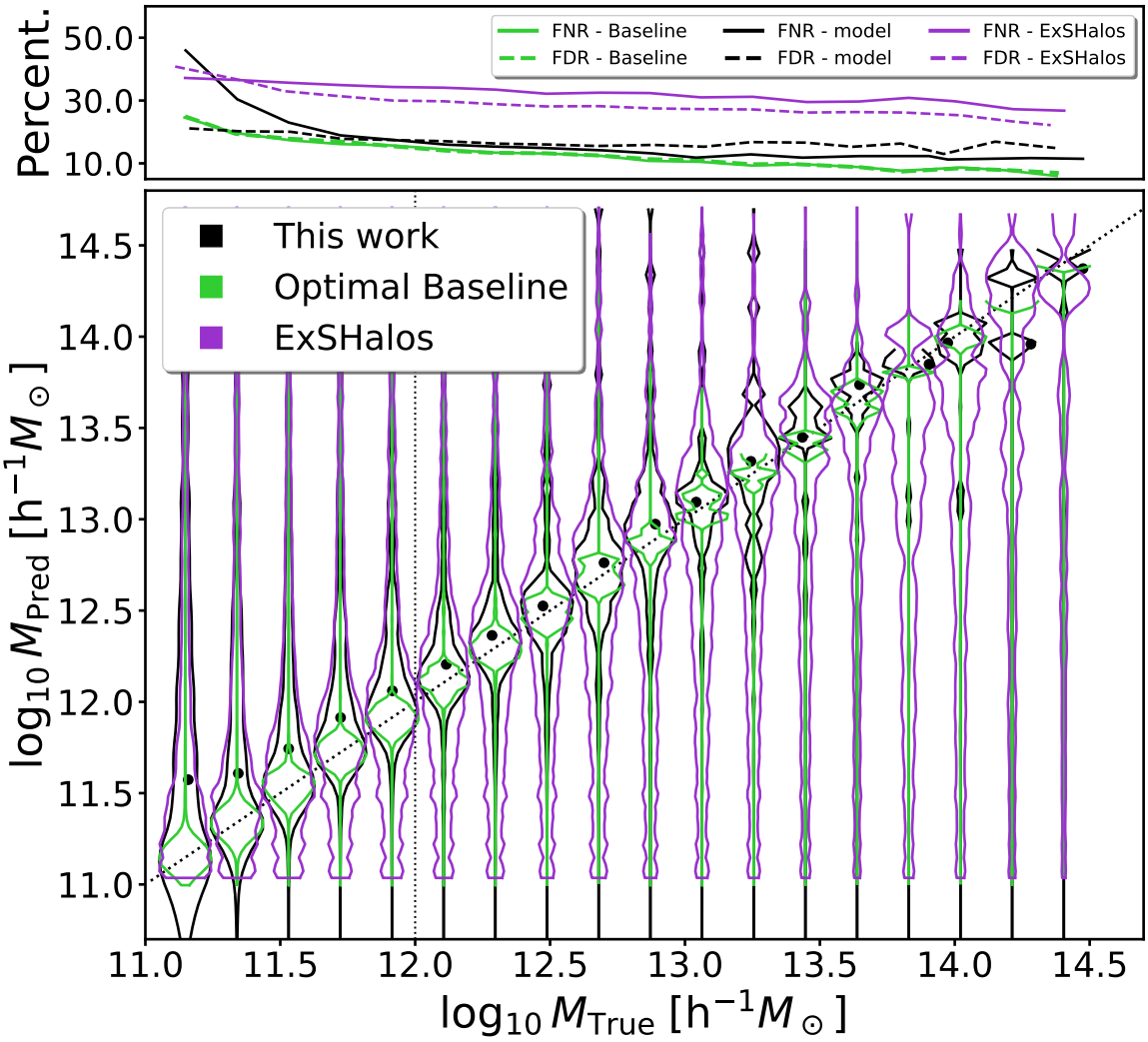

Figure J.2: "Violin plot", visualizing the distribution of predicted halo masses (at a voxel level) for different ground-truth mass bins. The black violin plots show the results obtained with our instance segmentation model. Green violin plots show the agreement between the two baseline simulations – representing an optimal target accuracy. The purple violin plots in the main panel correspond to the comparison with the ExSHalos predictions. The solid black line in the top panel shows the false negative rate, FNR, as a function of the ground truth halo mass. The dashed black line represents the fraction of predicted collapsed pixels that are not collapsed as a function of predicted halo mass (false discovery rate, FDR). The green and purple lines on the top panel correspond to the analogous results obtained from the baseline simulations and ExSHalos respectively.