UNIVERSITY COLLEGE LONDON

Faculty of Mathematics and Physical Sciences

Department of Physics & Astronomy

DEEP LEARNING INSIGHTS INTO DARK MATTER HALO FORMATION

NINGYUAN GUO

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

SUPERVISORS:

PROF. HIRANYA V. PEIRIS

PROF. ANDREW PONTZEN

EXAMINERS:

PROF. AMÉLIE SAINTONGE

PROF. ALAN HEAVENS

December 20, 2024

I, Ningyuan Guo, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Date: December 20, 2024

Name: Ningyuan Guo

Signature: _____

Zwei Dinge erfüllen das Gemüt mit immer neuer und zunehmender Bewunderung und Ehrfurcht, je öfter und anhaltender sich das Nachdenken damit beschäftigt: der bestirnte Himmel über mir, und das moralische Gesetz in mir.

— Immanuel Kant

Abstract

Unprecedented amounts of high-quality data will become available in the coming years from surveys such as *Euclid* and Rubin. They will allow stringent tests of the cosmological model and shed light on the nature of dark energy and dark matter. Understanding what determines properties of dark matter halos, the building blocks of cosmic structure, is crucial to developing robust and accurate theoretical models of quantities needed for cosmological inference, such as the halo mass function (HMF). This understanding is made difficult by the highly non-linear nature of halo formation. This thesis develops a novel approach to gaining physical insights on dark matter halo formation using deep learning. The approach takes advantage of deep learning models' ability to learn highly non-linear mappings, and combines this with techniques to interpret the model to extract new physical understanding. The approach is first developed using a synthetic image dataset: we investigate the use of an encoder-decoder architecture to learn an interpretable low-dimensional latent representation that captures underlying factors generating the dataset. We also build upon the novel approach of using mutual information to interpret the latent representation in a way that enables discovering physical factors not known *a priori*. The framework is then used to determine the information required to model the present-day HMF to the percent-level accuracy required by forthcoming surveys over a wCDM+$N_{\rm eff}$ parameter space. We find that three independent latent variables capture all relevant information. Interpreting these variables reveals that in addition to mass variance, the recent linear growth history and $N_{\rm eff}$ are also needed to accurately predict the HMF. The framework can be extended to additionally model the HMF's redshift evolution. The latent variables our framework discovers can be used to improve halo mass function modelling over a large cosmological parameter space and redshift range to meet future survey demands.

Impact statement

This thesis develops the use of deep learning for knowledge extraction and applies it to understand the factors that govern the cosmology dependence of the present-day dark matter halo mass function (HMF). The HMF is a key cosmological probe in forthcoming galaxy surveys, which aim to address important questions on the nature of dark energy, neutrino mass, and more. The approach combines representation learning, where information needed to produce the output is encoded in a low-dimensional latent representation, with mutual information (MI) that enables relating the latent representation to physical quantities.

The novel deep learning approach allows us to determine the inputs and the number of independent latent variables required to model the present-day HMF to a precision that meets the demand of current and forthcoming surveys. It further allows us to interpret the cosmological relevance of each latent variable, shedding light on the long-standing problem of what is needed to model the HMF beyond the theoretically known mass variance (i.e. beyond universality). Our approach has the advantage of being agnostic to prior assumptions, allowing us to discover a better parametrisation of the impact of growth history on non-universality compared to literature, and the relevance of the effective neutrino number $N_{\rm eff}$ to non-universality. The framework can be readily extended to model the HMF's redshift dependence, and to HMFs using different halo definitions. Furthermore, the latent space learnt by the model can be used both to inform the design of HMF emulator training sets, and to interpret information learnt by emulators, thereby offering valuable insights into emulator robustness. It could also lead to supplanting emulators with simpler, more interpretable models or mathematical expressions, whose robustness are easier to assess. Together, these could offer important insights that will help to develop robust and accurate HMF models over a wide cosmological parameter space and redshift range, which will be crucial for survey data analysis.

Methodology-wise, this thesis demonstrates the utility of interpretable deep learning in assisting with scientific discoveries, which is still a largely underexploited application of deep

learning. Interpretability is also crucial for scientists to trust results produced with the aid of machine learning; the latter is becoming indispensable in astrophysics and cosmology with the large amounts of data and high computational demands from surveys. The method of representation learning is especially promising for achieving interpretability, as all relevant information is compressed into a small number of independent variables, simplifying the problem of interpretation. The thesis also shows the important role that MI and conditional MI can play in interpreting the learned representation.

With the ever-increasing presence of artificial intelligence (AI) in every aspect of our life, being able to interpret how the AI model produces its outputs is not only important in the scientific community. It is crucial to gain trust on AI, especially in areas where accountability is key, e.g. in medicine, finance, and policy making. The methods developed in this thesis contribute towards this broader goal of achieving AI interpretability.

Acknowledgements

I express my deepest gratitude towards my supervisors Hiranya Peiris and Andrew Pontzen. Thank you for all the help, guidance, and insights you have offered me throughout this journey, both on the project and towards my growth as a scientist and person. Thank you also for your understanding and support during the pandemic, especially during the first nine months when I started my studies remotely in New Zealand. I also thank Hiranya for giving me the opportunity to visit Stockholm and for her help during my PhD application, without which I would not be here. I express my sincere gratitude towards my collaborator Luisa Lucie-Smith for the expertise and insights she has offered, which have been invaluable.

I have been lucky to receive the help from many along the way. Many thanks to Davide Piras for his help on GMM-MI, for the discussions on science and beyond, and for offering feedback on parts of this thesis. To Alessio Spurio Mancini, especially for his help in checkpointing subclassed TensorFlow models, which has been indispensable. To Justin Alsing for sharing his expertise on machine learning, especially for advices on model training. To Lurdes Ondaro-Mallea, Thomas McClintock, Eduardo Rozo, Joseph DeRose, Jeremy Tinker, Risa Wechsler, and Camila Correa for helping me in understanding their work. Thanks also to Claudia Muni for giving me feedback on parts of this thesis.

I am deeply grateful to Edd Edmondson for solving my innumerous problems with the cluster Hypatia, and for giving me priority time on the GPUs to complete the model tuning and training. I also thank John Deacon for help in setting up the GPU desktop and my laptop – and for posting it all the way to New Zealand when I started!

A big thank you to the graduate advisors Amélie Saintonge and Jason Sanders for all their care and support during my PhD, to my second supervisor Chamkaur Ghag, and to Nick Achilleos for his help during my PhD application. Thanks to Raphaële Raupp and Anita Maguire for making me feel welcome and supported in my daily life at the Cosmoparticle Hub, and to Nakum Kay, Nadia Waller, and Stan Zochowski for your support.

Acknowledgements

Many thanks to my Honours and Masters supervisor Richard Easther for his guidance and support with my PhD application, and to the cosmology group at the University of Auckland for making me feel connected while I worked on my PhD remotely in NZ.

My studies would not have been possible without the funding from the UCL Graduate Research Scholarship and Overseas Research Scholarship. I would also like to thank the Perren Fund for offering me the extra support I needed to wrap up the work in this thesis. Thanks to the Cosmoparticle Initiative for the great research environment and fun activities. I would also like to take the opportunity to acknowledge the use of open source software including `Python`, `Matplotlib`, `NumPy`, `SciPy`, `Scikit-Learn`, and `TensorFlow`, which I have used extensively.

I have been privileged to meet a group of extremely knowledgeable, friendly and kind people at the UCL Astrophysics Group and in the Cosmoparticle Hub. Thank you especially to Catarina Alves, Francesca Gerardi, Kiyam Lin, Jon Davies, Benjamin Stölzner, Max von Wietersheim-Kramsta, Corentin Cadiou, Claudia Muni, Davide Piras, Gandhali Joshi, Alex Jenkins, Alessio Spurio Mancini, Beatrice Crudele, Niall Jeffrey, Xie Cheng, Joana Teixeira, Grace Lawrence, Jacopo Siniscalco, Louis Hamaide, Arthur Loureiro, Pascal Förster, Vinooja Thurairethinam, Dirk Scholte, Mark Cunningham, and Luke Keyte for the company, discussions, and the joy you have brought to my PhD life.

Words cannot express my gratitude to my family for their ever unwavering support; I feel truly blessed to have you. Many thanks to my mother for her patience, care, and understanding, and for all the comfort and wise advice she has provided. To my father for advice on studies and life, and for making me feel loved. To my extended family for their concern and care for me. To my beloved cat Sherry, whose furry warm presence is deeply missed. A big thanks to all of my friends for their warm support. Thanks also to Jacob Maunders, Harry Gilbert, and Merve Gozum for mental guidance. Thank you to the great artists whose works have nourished and inspired me – to music by SID, Sodagreen, Beethoven, Rachmaninoff, Chopin, Debussy, and Lang Lang, to van Gogh's paintings, and to *GuJian3*, *Haikyu!!*, *Slam Dunk*, and *Bakuman*. Lastly, I would like to thank Yuzuru Hanyu for continuing to inspire me with your strive towards excellence, your determination and courage to challenge the unknown, and your strength and kindness.

Contents

List of Figures

Chapter 1

Introduction

Over the course of the last century, observations and theoretical developments have helped to establish the standard model of cosmology called Lambda Cold Dark Matter (ΛCDM; the name will become apparent in Sec. 1.1. Bocquet et al., 2019; Abdullah et al., 2020; Planck Collaboration et al., 2020a; Amon et al., 2023; Dark Energy Survey and Kilo-Degree Survey Collaboration et al., 2023; Li et al., 2023a; Sunayama et al., 2023; Ghirardini et al., 2024; Marques et al., 2024). In ΛCDM, the universe comprises of ordinary matter, radiation, *dark matter* which is a form of matter that only interacts gravitationally,[1] and *dark energy* which is an energy component driving the present-day accelerated expansion of the universe. While ΛCDM has successfully explained a wide range of observations to date, open questions remain. The nature of two important components of the standard model – dark matter and dark energy – remains elusive, and as the quality and amount of data improve, in recent years potential inconsistencies within ΛCDM have emerged (see Sec. 1.1.3 for more details). Very recent results from the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al., 2016) Data Release 1 (DESI Collaboration et al., 2024) provide tantalising hints that the dark energy density may be evolving with time. If this persists as more data become available, it would present a departure from ΛCDM and revolutionise our understanding of cosmology. Many other current and forthcoming observational campaigns such as *Euclid* (Laureijs et al., 2011), the Extended Roentgen Survey with an Imaging Telescope Array (eROSITA; Merloni et al., 2012), and the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al., 2009; Ivezić et al., 2019) also survey the late universe and will

[1] Some dark matter candidates are also proposed to interact via the weak force (Steigman and Turner, 1985, see Sec. 1.1.2.2).

provide a wealth of high-fidelity data to place the most stringent constraints on the cosmological model to date.

The analysis of most survey data currently relies on the *matter power spectrum*, which measures the amplitude of density contrasts, i.e. the fractional difference between the local density and the mean density of the universe, as a function of spatial scale (Reid et al., 2010; Lee et al., 2013; Troxel et al., 2018; Planck Collaboration et al., 2020b). At early times (or on large scales at late times) when the density contrast is small, the matter power spectrum can be calculated from linear perturbation theory, which describes the evolution of locally overdense regions with low density contrasts. The most accurate calculations rely on Einstein-Boltzmann solvers, though analytical approximations can be made accurate to a few percent (Eisenstein and Hu, 1998). The theory and solvers will be described in greater detail in Sec. 1.2. As the density contrasts grow and collapse under the influence of gravity, their evolution becomes non-linear, and analytical calculations become possible only by making many simplifying assumptions. Accurate models of the formation and evolution of small-scale structures like galaxy clusters (which contain hundreds to thousands of galaxies) rely on numerical simulations that follow the evolution of matter from early until late times.

The matter power spectrum only captures some of the cosmological information that is available in the high-fidelity data provided by current and forthcoming surveys (see Sec. 1.1.3.3). We therefore need to complement power spectrum analyses with other means of constraining cosmology. One way is through galaxy clusters, which historically provided one of the first evidences for dark matter (Zwicky, 1933); now it is believed that galaxy clusters reside in massive gravitationally-bound dark matter structures called halos. The abundance of dark matter halos as a function of halo mass is called the *halo mass function*, and depends sensitively on the cosmological model and cosmological parameters. This makes galaxy cluster number counts one of the major cosmological probes in surveys such as eROSITA (Merloni et al., 2012), *Euclid* (Laureijs et al., 2011), and LSST (LSST Science Collaboration et al., 2009; Ivezić et al., 2019). To fully exploit the constraining power provided by these surveys' data, the halo mass function must be modelled to 1% accuracy or better (Wu et al., 2010; Sartoris et al., 2016; McClintock et al., 2019b; Artis et al., 2021; Euclid Collaboration et al., 2023a; Ghirardini et al., 2024).

In Sec. 1.3, I describe how observations of galaxy clusters probe dark matter halos, and describe a commonly used analytical model of halo formation called spherical collapse (which makes several simplifying assumptions). Using ideas from spherical collapse, Press and Schechter (1974); Bond et al. (1991) derived an analytical halo mass function, which will be presented in Sec. 1.4. Comparisons with numerical simulations revealed that the analytical model only qualitatively describes the halo mass function, with up to $\sim 50\%$ deviations (Sheth and Tormen, 1999). As a result, semi-analytical models are developed by calibrating analytical formulae against numerical simulations; these are described in Sec. 1.5. While some semi-analytical models can be accurate to within 5% residuals (e.g. Tinker et al., 2008), with time, it became clear that semi-analytical fits are only accurate within a narrow cosmological parameter space and/or redshift range. In Sec. 1.6, I describe the state of the art on modelling the cosmology and redshift dependence of the halo mass function, and identify gaps in existing approaches. This motivates the work in this thesis, where we take a novel approach using deep learning to gain new insights on the physical quantities needed to accurately model the halo mass function over a wide cosmological parameter space. The aim and outline for this thesis are in Sec. 1.7.

1.1 The cosmological model

In the standard model of cosmology, the universe initially underwent a hypothetical period of rapid, near-exponential expansion called *inflation* (see e.g. Weinberg, 2008; Baumann, 2009). The end of inflation leaves the universe close to homogeneous and isotropic on large scales (Guth, 1981), while initially small quantum perturbations generated during inflation become stretched and seed the growth of structures into halos and galaxies today (Mukhanov and Chibisov, 1981, 1982; Guth and Pi, 1982; Hawking, 1982; Starobinsky, 1982; Bardeen et al., 1983; Mukhanov, 1985).

Inflation leaves a spectrum of small initial density perturbations, where the density contrast (or *overdensity*) at location $\mathbf{x}$ is defined as

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}}, \tag{1.1}$$

with $\rho(\mathbf{x})$ the local density and $\bar{\rho}$ the background density of the universe. Simple models of inflation (called single-field slow roll inflation, see e.g. Dodelson and Schmidt, 2020, for more details) predict the initial density perturbations to be very close to a Gaussian random field, where the values of $\delta(\mathbf{x})$ are randomly drawn and follow a Gaussian distribution. This is confirmed by measurements of the cosmic microwave background (Planck Collaboration et al., 2020c,d), which are the relic photons from the early universe and will be described in greater detail in Sec. 1.1.3.1. The information in a Gaussian random field can be completely characterised by the power spectrum $P(k)$, defined as

$$\langle \delta(\mathbf{k}) \delta^*(\mathbf{k}') \rangle = \delta_D^3(\mathbf{k} - \mathbf{k}') P(k)\,, \tag{1.2}$$

where $\delta(\mathbf{k})$ is the Fourier transform of the local overdensity $\delta(\mathbf{x})$, and δ_D^3 is the Dirac delta function in three dimensions that arises because different Fourier modes of a statistically homogeneous density field are uncoupled. Isotropy means the power spectrum is only a function of the wavenumber $k = |\mathbf{k}| = 2\pi/\lambda$, and not of $\mathbf{k}$. As mentioned before, the power spectrum describes the amplitude of density perturbations at different length scales. Inflation predicts a primordial power spectrum of the form (see e.g. Dodelson and Schmidt, 2020)

$$P(k) \propto A_s k^{n_s}\,. \tag{1.3}$$

The amplitude of the primordial power spectrum is measured to be $A_s \approx 2 \times 10^{-9}$, i.e. the initial overdensities are very small (Planck Collaboration et al., 2020c). The exponent n_s is the scalar spectral index characterising the tilt of the power spectrum. If $n_s = 1$, the variance of density fluctuations on scales entering the horizon is independent of time (see eg. Padmanabhan, 2003; Jaffe, 2012), and the power spectrum is referred to as *scale-free*. A value of $n_s < 1$ gives a 'red-tilted' spectrum with higher power at larger scales, and $n_s > 1$ gives a 'blue-tilted' spectrum. Current measurements of $n_s = 0.9665 \pm 0.0038$ (Planck Collaboration et al., 2020c) show the power spectrum is close to scale-free with a slight red-tilt. This agrees with the predictions of simple models of inflation (called slow-roll inflation, see e.g. Dodelson and Schmidt, 2020, for more details), giving credit to the theory of inflation (Planck Collaboration et al., 2020e).

After inflation, the universe continues to expand under gravitational influence described by general relativity; the main equations are outlined in Sec. 1.1.1. The different constituents

of the universe will each be described in Sec. 1.1.2, while Sec. 1.1.3 introduces some of the observational probes used to constrain the cosmological model.

1.1.1 General relativity and the Friedmann-Lemaître-Robertson-Walker metric

General relativity provides the foundation of the standard model of cosmology. It generalises Newtonian gravity to regimes of strong gravity and high velocity, and accurately describes a wide range of observations including planet orbits (Clemence, 1947; Biswas and Mani, 2008), the gravitational lensing of light (Dyson et al., 1920), gravitational time delay (Bertotti et al., 2003), black hole shadows from direct imaging (The Event Horizon Telescope Collaboration et al., 2019), and the existence and propagation of gravitational waves (Abbott et al., 2016).

In general relativity, gravity is not viewed as a force like in Newtonian theory, but rather as a geometric property of spacetime. The geometry of spacetime can be described by a *metric* tensor $g_{\mu\nu}$ which defines distances on a manifold. On large scales, the universe is approximately homogeneous and isotropic (this is confirmed by measurements of the cosmic microwave background and large-scale galaxy surveys; see Sec. 1.1.3 for further details), so it can be described by the Friedmann-Lemaître-Robertson-Walker (FLRW) metric, where a small proper distance ds is described by (in units of $c = 1$, Weinberg, 2008)

$$ds^2 = dt^2 - a^2(t)\left(\frac{dr^2}{1-Kr^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right). \tag{1.4}$$

Here, t is cosmic time and $a(t)$ is the time-dependent *scale factor* that quantifies the expansion (or contraction) of the universe and is normalised to $a(t_0) = 1$ today. We have expressed the spatial line element in terms of reduced-circumference polar coordinates (r, θ, ϕ), and the spatial curvature of the universe is described by K, with $K = 0$ for a flat universe. Note that as the scale factor has been factored out of the spatial coordinates, the coordinate distance (also called the *comoving distance*) between two points which move only along the t-direction remains constant, but the physical distance (also called the *proper distance*) between the two points increases proportional to the scale factor $a(t)$.

As a consequence, the physical wavelength of photons travelling in an FLRW spacetime scales proportional to $a(t)$. This allows us to use the observed photon wavelength $\lambda_{\rm obs}$ as a

measure of the scale factor at photon emission if the wavelength at emission $\lambda_{\rm em}$ is known (e.g. through known atomic transitions; Weinberg, 2008). The two are related via

$$\frac{a(t_{\rm obs})}{a(t_{\rm em})} = \frac{\lambda_{\rm obs}}{\lambda_{\rm em}} = 1 + z_{\rm em}\,, \tag{1.5}$$

where we have defined the cosmological *redshift* z. It is called redshift as observations of distant light sources such as galaxies show that $\lambda_{\rm em} < \lambda_{\rm obs}$, which also provided the first evidence that the universe is expanding (Slipher, 1915, 1917; Hubble, 1929). Redshift is often used instead of the scale factor a or time coordinate t as a measure of time in cosmology. Since observations are made in the present day, $a(t_{\rm obs}) = 1$, and the relation between redshift and the scale factor becomes

$$a = \frac{1}{1+z}\,, \tag{1.6}$$

where we have dropped the subscripts for simplicity of notation.

Another way of measuring time uses the *conformal time* η defined as

$$\mathrm{d}\eta \equiv \frac{\mathrm{d}t}{a(t)}\,. \tag{1.7}$$

The conformal time provides the *comoving distance* that a photon has travelled since $t = 0$ (in absence of any interactions):

$$\eta = \int_0^t \frac{\mathrm{d}t'}{a(t')}\,. \tag{1.8}$$

This is the *causal horizon*, or the *comoving horizon*, and presents the maximum comoving separation between two objects that could have been in causal contact since $t = 0$ (Pontzen and Peiris, 2020).

The expansion rate of the universe is given by the *Hubble parameter*

$$H(t) = \frac{\dot{a}}{a}\,, \tag{1.9}$$

where the overdot indicates a derivative with respect to t. The value of the Hubble parameter today is the Hubble constant H_0, which is often expressed through the dimensionless Hubble constant h defined as $H_0 \equiv 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. Its value is approximately $h \approx 0.7$, although

different observational probes currently disagree on its exact value; this will be further discussed in Sec. 1.1.3.

To describe the expansion of the universe and how this relates to the universe's energy contents, we use the Einstein field equations (Pontzen and Peiris, 2020)

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} R\, \mathrm{g}_{\mu\nu} = 8\pi G T_{\mu\nu} + \Lambda\, \mathrm{g}_{\mu\nu} \,. \tag{1.10}$$

The left hand side of the equation is the Einstein tensor $G_{\mu\nu}$, which is a combination of the Ricci tensor $R_{\mu\nu}$ and the Ricci scalar R (multiplied by the metric tensor $\mathrm{g}_{\mu\nu}$). These describe the geometry of spacetime: the Ricci tensor describes how a volume moving along geodesics changes and depends only on the metric and its derivatives, and the Ricci scalar is the contraction of the Ricci tensor $R = R^{\mu}_{\ \mu}$ (see e.g. Weinberg, 2008, for more details). On the right hand side, G is Newton's gravitational constant, $T_{\mu\nu}$ is the energy-momentum tensor that describes the energy distribution and flux, and Λ is the *cosmological constant* which has the same value at all points in spacetime (we will return to it in Sec. 1.1.2.4). The Einstein field equations describe how the geometry of spacetime interacts with the energy contents of the universe.

Energy contents in a homogeneous and isotropic universe behave as perfect fluids, which are completely characterised by their density ρ and pressure P. The energy-momentum tensor on the right hand side of Eq. (1.10) is therefore given by that of perfect fluids,

$$T_{\mu\nu} = (\rho + P) U_\mu U_\nu - P\, \mathrm{g}_{\mu\nu} \,, \tag{1.11}$$

where U_μ is the velocity 4-vector (Pontzen and Peiris, 2020). Calculating the left hand side of the Einstein field equation using the FLRW metric in Eq. (1.4), we find a set of equations known as the Friedmann equations (Friedmann, 1922, 1924):

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda}{3} - \frac{K}{a^2} \,, \tag{1.12}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3P) + \frac{\Lambda}{3} \,. \tag{1.13}$$

Since observations to date are consistent with a flat universe (Planck Collaboration et al., 2020c; DESI Collaboration et al., 2024), we will use $K = 0$ from now on. The Friedmann equations

describe the expansion of a homogeneous and isotropic universe as a function of time. The right hand side of the Friedmann equations show that expansion is dictated by the density and pressure of the energy contents of the universe, which I describe in the next section.

1.1.2 Energy contents of the Universe

As we just saw, the expansion of the universe depends on its energy contents, which are well-described by perfect fluids completely characterised by their density ρ and pressure P. The two are related via the *equation of state parameter*

$$w \equiv \frac{P}{\rho}. \tag{1.14}$$

The density evolution can be found by considering the divergence-free property of the energy-momentum tensor $\nabla_\mu T^\mu_\nu = 0$ due to local conservation of energy-momentum (Weinberg, 2008). Considering the $\nu = 0$ component gives the equation (which can also be derived by combining the two Friedmann equations Eq. 1.12 and 1.13):

$$\frac{\partial \rho}{\partial t} + 3\frac{\dot{a}}{a}(\rho + P) = 0. \tag{1.15}$$

Solving this equation shows the density of the fluid evolves according to

$$\rho \propto a^{-3(1+w)} \tag{1.16}$$

for fluids with a constant equation of state parameter w (Weinberg, 2008). For a flat universe, combining Eq. (1.16) with Eq. (1.12) gives us the time evolution of the scale factor

$$a(t) \propto \begin{cases} t^{2/(3(1+w))} & , \quad w \neq -1 \\ e^{H_0 t} & , \quad w = -1 \end{cases}. \tag{1.17}$$

The density of fluids can be expressed as a fraction of the *critical density* ρ_c, which is defined as the total density of the contents of a flat universe (see e.g. Huterer and Shafer, 2017)

$$\rho_c \equiv \sum_i \rho_i \tag{1.18}$$

for ρ_i the density of any species such as matter, radiation, etc. The critical density is then given by

$$\rho_c(t) = \frac{3H^2(t)}{8\pi G} \tag{1.19}$$

at a given time t (or equivalently at a given scale factor a or redshift z). The density of fluids can be expressed using the density parameter

$$\Omega_i = \frac{\rho_i}{\rho_c}\,. \tag{1.20}$$

From the above, we see that different species with different equations of state affect the universe's expansion differently. To describe the expansion history of the universe, we only need to know the equations of state w for the different species and their density parameter at a given time, which we take to be the present day $\Omega_{i,0}$. In the standard model of cosmology ΛCDM, the universe consists of the four species below.

1.1.2.1 Radiation

Radiation refers to relativistic components of the universe such as photons γ. In the standard model of particle physics, neutrinos ν are also massless (Navas et al., 2024) and hence count as radiation. While measurements of neutrino oscillation reveal at least two of the three active neutrino mass eigenstates are non-relativistic today (Fukuda et al., 1998; Ahmad et al., 2001; Esteban et al., 2020; de Salas et al., 2021; Capozzi et al., 2021; see also a review in Navas et al., 2024), they were still relativistic in the past and count as radiation at high redshifts, including when the universe transitioned from radiation to matter domination (see Sec. 1.1.2.2) and when cosmic structures started forming (Lesgourgues et al., 2013). The simplifying assumption that all neutrinos are massless is still sometimes used in numerical simulations of structure formation (DeRose et al., 2019).

Most of the radiation density today is in the form of the cosmic microwave background and the cosmic neutrino background. These are relics of the early hot universe filled with an electromagnetic plasma of baryons and leptons interacting with photons. The radiation energy density $\rho_r = \rho_\gamma + \rho_\nu$ can be expressed solely in terms of the photon energy density as (see e.g. Lesgourgues et al., 2013)

$$\rho_r = \rho_\gamma \left(1 + \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} N_{\rm eff}\right)\,, \tag{1.21}$$

where $N_{\rm eff}$ is the effective number of neutrino species we will return to shortly. The present-day photon energy density $\rho_{\gamma,0}$ can be precisely determined from the temperature of the cosmic microwave background $T = 2.7255\,\mathrm{K}$ (Fixsen, 2009; Planck Collaboration et al., 2020c; see e.g. Weinberg, 2008 for the derivation). Combined with Eq. (1.21), this gives a present-day radiation density $\Omega_{r,0} = \Omega_{\gamma,0} + \Omega_{\nu,0} \approx 0.009\%$ (assuming massless neutrinos). Radiation has an equation of state parameter $w = 1/3$, so Eq. (1.16) shows that $\rho_r \propto a^{-4}$, and from Eq. (1.17) $a \propto t^{1/2}$. This means that while the present-day radiation density is negligible, early on, radiation dominated the energy density in the universe and drove the initial period of cosmic expansion after the end of inflation.

We now return to $N_{\rm eff}$. This parameter was historically used because the number of active light neutrino generations was unclear. Laboratory measurements have now shown there are three generations (Beringer et al., 2012), but $N_{\rm eff} = 3.046$ is often assumed (Mangano et al., 2005, though a more recent calculation gives $N_{\rm eff} = 3.044$ Gariazzo et al., 2019). The slight increase comes from higher-order corrections due to neutrino mixing, and due to the decoupling of neutrinos from the early electromagnetic plasma not being instantaneous (Mangano et al., 2002; Mangano et al., 2005; de Salas and Pastor, 2016; Akita and Yamaguchi, 2020).

Cosmology also allows for constraining $N_{\rm eff}$. Current cosmological constraints give $N_{\rm eff} = 3.10 \pm 0.17$ (DESI Collaboration et al., 2024), in agreement with the prediction by the standard model of particle physics. However, the cosmological constraint can be significantly tightened with forthcoming survey data (Dodelson et al., 2016). This is of significant interest despite laboratory results showing there are three generations of neutrino species, as measuring deviations in $N_{\rm eff}$ would point to unknown particle physics such as unidentified light particles contributing to the radiation density, or that neutrino phase space distributions (distributions in the 6D space of positions and momenta) deviate from a naive expectation of the Fermi-Dirac distribution. Because of this, constraining $N_{\rm eff}$ and other neutrino properties such as their mass is a key interest in current and forthcoming observational campaigns (LSST Science Collaboration et al., 2009; Laureijs et al., 2011; Amendola et al., 2018; DESI Collaboration et al., 2016, 2024).

1.1.2.2 Cold dark matter

Cold dark matter refers to any form of non-relativistic matter that only interacts gravitationally (and possibly via the weak force). While laboratory experiments are yet to detect dark

matter directly, observational evidence for the existence of dark matter has been gathered from many astrophysical and cosmological probes (see Bertone and Hooper, 2018, for a review). Dark matter explains the discrepancy between the amount of luminous matter compared to the gravitational mass required to hold a galaxy cluster together (Zwicky, 1933), and why galaxy rotation curves flatten as the distance from galaxy centre continues to increase, rather than decrease at large radii (Babcock, 1939; Rubin and Ford, 1970). The presence of dark matter can also be inferred from gravitational lensing and the cosmic microwave background, which will be discussed in greater detail in Sec. 1.1.3.

Modern cosmological measurements consistently measure dark matter to have $\Omega_c \approx 25\%$ (Bocquet et al., 2019; Abdullah et al., 2020; Planck Collaboration et al., 2020c; Amon et al., 2023; Dark Energy Survey and Kilo-Degree Survey Collaboration et al., 2023; Li et al., 2023b; Sunayama et al., 2023; Marques et al., 2024; DESI Collaboration et al., 2024; Ghirardini et al., 2024). Dark matter is the dominant form of matter in the universe and therefore drives structure formation, providing the gravitational potential wells that baryonic matter (see Sec. 1.1.2.3) gathers in to form galaxies. Numerical simulations show that to reproduce the observed large-scale structures like in Fig. 1.1, dark matter must be cold, i.e. it should be non-relativistic at the time of structure formation (Schramm and Steigman, 1981; Peebles, 1982). This gives cold dark matter (CDM) the equation of state parameter $w = 0$, so $\rho_{CDM} \propto a^{-3}$. As the universe expanded, the energy density of matter decreased slower than radiation, such that at $z \sim 3400$ (Planck Collaboration et al., 2020c), the universe transited from radiation domination to matter domination, and then expanded approximately as $a \propto t^{2/3}$.

The particle nature of dark matter remains an active area of research. At the turn of the century, weakly interacting massive particle (WIMPs) were the strongly favoured dark matter candidate (Steigman and Turner, 1985; see e.g. Arcadi et al., 2018, for a review). This hypothetical type of particle can interact with baryons via the weak force in addition to gravity, and sparked many laboratory efforts to search for their presence. As dark matter detector sensitivity and analysis methods improved, increasingly large portions of the possible parameter space have been ruled out, and there is yet to be a detection. At the same time, potential discrepancies between small-scale predictions of CDM and observations spark studies in dark matter candidates with light masses, such as ultralight dark matter (e.g. Hui et al., 2017; Ferreira, 2021). Potential candidates for dark matter span some 50 orders of magnitude in mass, from ultralight

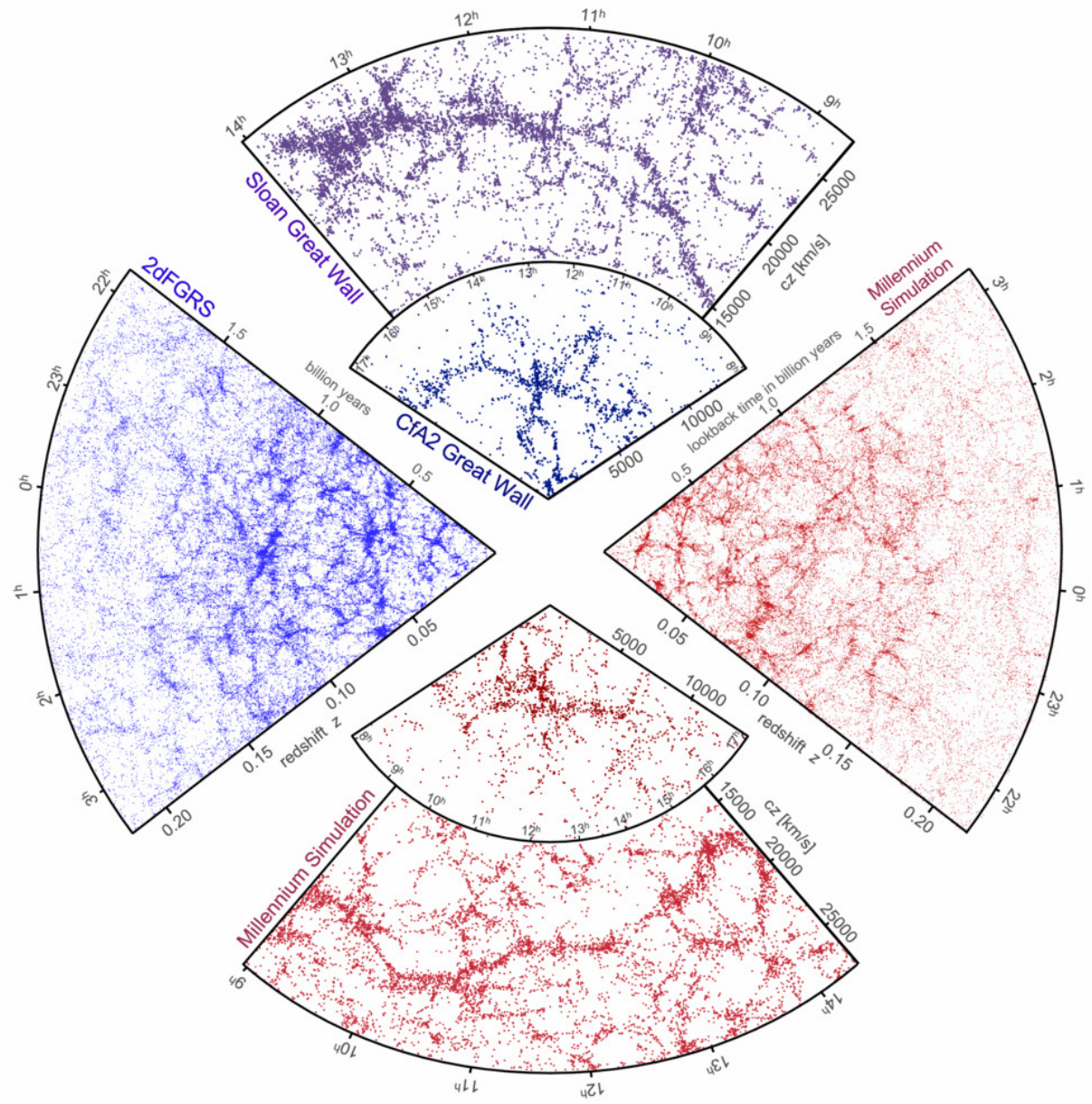


Figure 1.1: The large-scale structure from galaxy redshift surveys, compared to numerical simulations. The top slices show subregions of the Center for Astrophysics (CfA) survey (Davis et al., 1982) and the Sloan Digital Sky Survey (Gott et al., 2005); the Coma galaxy cluster Zwicky studied (Zwicky, 1933) is at the centre of the "CfA Great Wall". The left cone shows half of the Two-degree-Field Galaxy Redshift Survey (2dFGRS) (Colless et al., 2001). The lower and right cones are mock galaxy surveys produced using data from the Millennium Simulation (Springel et al., 2005). Figure is figure 1 in Springel et al. (2006).

scalar fields to primordial black holes or massive compact halo objects, though different regions of the parameter space have been ruled out by a combination of laboratory experiments, astrophysical and cosmological probes (for a review, see Bertone and Hooper, 2018).

1.1.2.3 Baryonic matter

In cosmology, *baryonic matter* refers to all massive non-relativistic particles that couple to the electromagnetic force, including electrically-charged leptons which are in a strict particle-physics sense non-baryonic. The abundance of baryonic matter can be constrained from Big Bang nucleosynthesis (Reeves et al., 1973; Burles and Tytler, 1998a,b; Burles et al., 2000; O'Meara et al., 2001) and the cosmic microwave background (Planck Collaboration et al., 2020c, see Sec. 1.1.3.1 for more details), giving the present-day density parameter $\Omega_{b,0} \approx 5\%$. Although baryons only constitute a small fraction of the universe's energy contents, because they are directly observable, they play a crucial role in cosmology. As described in the last subsection, baryons gather in gravitational potential wells provided by dark matter, so they trace dark matter structures which are otherwise invisible. Baryons especially impact the formation and evolution of structures on small scales relevant to galaxy formation, and affect the overall background evolution of the universe by contributing to the total matter density of the universe $\Omega_m = \Omega_b + \Omega_c$.

1.1.2.4 Dark energy

Observations of distant light sources such as supernovae (Riess et al., 1998; Perlmutter et al., 1999; see Sec. 1.1.3.2) reveal the universe is currently expanding at an accelerated rate, contrary to what would be expected if the universe were dominated by matter today. This accelerated expansion is hypothesised to be driven by an unknown form of energy called dark energy.

In the standard model of cosmology, dark energy is a cosmological constant Λ which we saw in the Einstein field equations Eq. (1.10). Comparing the $\Lambda g_{\mu\nu}$ term in Eq. (1.10) with the energy-momentum tensor of a perfect fluid in Eq. (1.11), we see the cosmological constant term can be absorbed into the energy-momentum tensor if we view it as a perfect fluid with an effective pressure $P_\Lambda = -\rho_\Lambda$, so its equation of state parameter is $w = -1$. Using Eq. (1.16), its energy density

$$\rho_\Lambda = \frac{\Lambda}{8\pi G} \tag{1.22}$$

remains constant with time as the universe expands. We can also assign to it a density parameter

$$\Omega_\Lambda = \frac{\Lambda}{3H^2(t)}\,. \tag{1.23}$$

As the universe continues to expand, the energy density of matter decreases such that at $z \sim 0.3$, the universe transited from matter domination to dark energy domination. The scale factor then started to increase exponentially with time after dark energy domination (from Eq. 1.17); the universe has re-entered a period of accelerated expansion, which continues to today.

The cosmological constant is often interpreted as vacuum energy (Sakharov, 1968), however its value poses a problem. The present-day value is $\rho_\Lambda \approx 2.6 \times 10^{-47}\,\mathrm{GeV}^4$ (Planck Collaboration et al., 2020c). On the other hand, calculating the zero-point energy densities of fields in the standard model of particle physics using quantum field theory leads to an estimate of $\rho_{\mathrm{vacuum}} \sim \mathcal{O}(10^8)\,\mathrm{GeV}^4$ (see Martin, 2012, for further details). The two values are discrepant by about 55 orders of magnitude, a problem that is known as the cosmological constant problem.

Beyond the cosmological constant, there is also a rich set of alternative dark energy models. For example, dark energy could be a minimally-coupled scalar field that evolves slowly towards a minimum of its potential. Solving the scalar field's equation of motion in an FLRW universe shows that a universe dominated by the scalar field's energy density would expand close to exponentially (such theories are called *quintessence*; see e.g. Peebles and Ratra, 2003; Weinberg, 2008; Tsujikawa, 2013, for more details). Or, instead of an energy density, the accelerated expansion could be due to a modification of gravity on cosmological scales, with additional terms on the left hand side of the Einstein field equations Eq. (1.10).

The equation of state parameter w defined in Eq. (1.14) provides a useful phenomenological characterisation of the possible dark energy models. For $-1 < w < -\frac{1}{3}$,[2] the dark energy density decreases as the universe expands; if $w < -1$, its energy density increases as the universe expands, often referred to as *phantom dark energy*. The dark energy equation of state parameter w can either be a constant that is allowed to deviate from -1, or a function of time $w(z)$. Excluding the case of a cosmological constant, there are no special reasons why w should be constant; both quintessence and modified gravity have time-varying w. However, many observational analyses still consider the case of a constant w (such a universe is called wCDM where dark matter is still cold), as any deviations of a constant w from -1 indicates physics beyond ΛCDM. Note the converse is not necessarily true; if w is measured to be -1 when

[2] We require $w < -\frac{1}{3}$ to drive accelerated expansion.

assuming w is a constant, the equation of state parameter may still be evolving with time in a way that the average w is consistent with -1 (Linder, 2007).

A time-varying w is usually observationally probed by parametrising the equation of state parameter as

$$w(a) = w_0 + w_a(1-a) = w_0 + w_a \frac{z}{1+z}, \tag{1.24}$$

where w_0 and w_a are free parameters. The energy density is then (Huterer and Shafer, 2017)

$$\frac{\rho_{DE}}{\rho_{c,0}} = \Omega_{DE,0} a^{-3(1+w_0+w_a)} e^{-3w_a(1-a)}. \tag{1.25}$$

This simple $w_0 - w_a$ parametrisation has the advantage of involving only two free parameters to constrain, and it avoids unphysical behaviours at large redshifts which alternative parametrisations often have (Chevallier and Polarski, 2001; Linder, 2003). In Fig. 1.2, the blue shaded region shows the $\pm 1\sigma$ dark energy density allowed by data constraints assuming a $w_0 - w_a$ parametrisation, produced by Huterer and Shafer (2017).

Observations have been in line with dark energy being a cosmological constant. Very recently, DESI Collaboration et al. (2024) report that assuming $w_0 w_a$CDM, they measure $w_0 > -1$ and $w_a < 0$ at up to 3.9σ significance when combining their measurement of the baryonic acoustic oscillation together with measurements of the cosmic microwave background and Type Ia supernovae (these probes will be described in Sec. 1.1.3). While the current results are still consistent with statistical fluke, it is expected that further data both from DESI and from other current and forthcoming surveys (DES, The Dark Energy Survey Collaboration, 2005; eROSITA, Merloni et al., 2012; Hofmann et al., 2017; *Euclid*, Laureijs et al., 2011; Amendola et al., 2018; and LSST, LSST Science Collaboration et al., 2009; Ivezić et al., 2019) will tighten the constraint and test this tantalising hint for the breakdown of ΛCDM in the near future.

Having looked at the contents of the universe and their equations of state, we can now summarise the background expansion of the universe by writing the Friedmann equation Eq. (1.12)

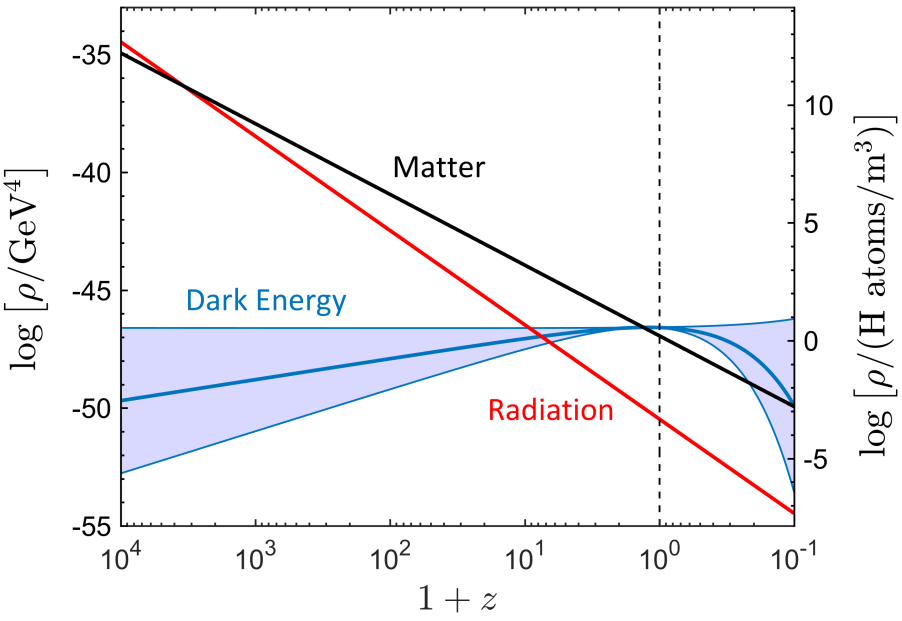


Figure 1.2: Energy density of different species described in Sec. 1.1.2 as a function of $(1+z=a^{-1})$ for redshift z. The dashed vertical line indicates the present day ($z=0$); past is to the left and future is to the right. Dark energy density is a constant for the case of cosmological constant Λ; the shaded region indicates $\pm 1\sigma$ allowed by data constraints assuming a dynamical dark energy equation of state parameter Eq. (1.24). Radiation dominates the early energy density until $z\sim 3400$, when matter takes over. At $z\sim 0.3$, matter energy density diluted enough such that dark energy is now the dominant energy component. The figure is taken from figure 1 in Huterer and Shafer (2017).

as

$$\frac{H^2}{H_0^2} = \left(\Omega_{b,0}+\Omega_{c,0}\right)a^{-3} + \left(\Omega_{\gamma,0}+\Omega_{\nu,0}\right)a^{-4} + \Omega_{DE,0}a^{-3(1+w_0+w_a)}e^{-3w_a(1-a)}\,, \tag{1.26}$$

or $$\frac{H^2}{H_0^2} = \left(\Omega_{b,0}+\Omega_{c,0}\right)a^{-3} + \left(\Omega_{\gamma,0}+\Omega_{\nu,0}\right)a^{-4} + \Omega_{\Lambda,0} \quad \text{for } \Lambda\text{CDM with massless neutrinos} \tag{1.27}$$

in a flat universe where $\sum_i \Omega_i = 1$. Fig. 1.2 summarises the evolution of different energy components as a function of redshift.

1.1.3 Observational constraints on the cosmological model

The evolution history of the universe and the energy density of its different components described in the last few subsections are constrained by a variety of observations. The standard model of cosmology ΛCDM describes the universe's expansion (following Eq. 1.27) and the growth of structures using just six free parameters: A_s and n_s from Eq. (1.3) describing the

primordial density fluctuations, the physical density of baryons $\Omega_b h^2$ and cold dark matter $\Omega_c h^2$, the present-day expansion rate H_0, and the reionisation optical depth τ which provides a measure of when the first stars formed and began to reionise the neutral hydrogen gas in the universe.[3] Dark energy density is determined in ΛCDM by assuming spatial flatness, and the photon and neutrino densities are accurately measured from the cosmic microwave background temperature (see Sec. 1.1.2.1) assuming $N_{\rm eff} = 3.046$. The values of the six free parameters in ΛCDM, as well as additional parameters like the spatial curvature, the dark energy equation of state parameters, and parameters of neutrino properties such as the sum of neutrino masses and $N_{\rm eff}$ are constrained by several observational probes covering a wide range of spatial scales and epochs in the universe's history. ΛCDM has been largely in agreement with observations to date, though tensions in some parameter values have recently emerged, and very recent results suggest possible hints for a non-constant dark energy. These will be described in this section while introducing some of the main observational probes.

1.1.3.1 Cosmic microwave background

The discovery of the universe's expansion (Slipher, 1915, 1917; Eddington, 1923; see a review in Peacock, 2013) led to the Big Bang theory and the idea that the universe expanded from an initially hot, dense state filled with radiation (Gamow, 1946; Alpher et al., 1948; Alpher and Herman, 1948, 1949). This radiation is initially tightly coupled to the baryons and electrons, so photons only had a very low mean free path compared to the horizon scale. As the universe expanded, the temperature cooled enough for *recombination*: baryons and electrons combined into electrically neutral atoms, and the mean free path of photons increased to greater than H. Free-streaming photons emitted from recombination permeate the universe. The central photon wavelength of the black body radiation has since redshifted into the low-energy microwave band due to the expansion of the universe. Discovering the cosmic microwave background (CMB; Penzias and Wilson, 1965) provided strong evidence for the Big Bang theory (Dicke et al., 1965).

Since the physics that describe the CMB power spectrum is well-known, the CMB provides us with robust constraints on the cosmological model. Measuring its mean temperature $T = 2.7255\,\mathrm{K}$ (Fixsen, 2009) accurately allows us to determine the present-day photon density

[3] A larger τ means the CMB photons are more likely to be scattered, resulting in a suppression of temperature anisotropies at scales below the horizon at the time of reionisation (see Dodelson and Schmidt, 2020, for more details).

$\Omega_{\gamma,0}$ (see Sec. 1.1.2.1), while its *temperature anisotropies* (small variations in the CMB temperature at $\frac{\delta T}{T} \sim 10^{-5}$) reflect the initially small density perturbations which have since grown to form all structure in the universe. Before recombination, the baryons and photons are tightly coupled together as a baryon-photon plasma against a backdrop of gravitational potential wells mostly provided by dark matter. The (proper) Jeans length, the scale below which pressure provides support against gravitational collapse, is $\lambda_J = c_s\sqrt{\frac{\pi}{G\bar{\rho}}} \sim 6ct$, using the speed of sound $c_s \sim c/\sqrt{3}$, $\bar{\rho} \simeq \rho_{\rm crit}$ and $H(t) = (2t)^{-1}$ (Mo et al., 2010). Since it is comparable to the horizon size $\sim 2ct$ (Eq. 1.8 times a), sub-horizon perturbations do not grow (Mo et al., 2010). Instead, as gravity compresses the fluid, its pressure (and temperature) increases, eventually causing the fluid to expand again (and cool). This interplay between gravity and radiation pressure results in acoustic oscillations in the fluid. The oscillations stop at recombination since photons decouple from the baryons and start to free-stream, while the Jeans length drops significantly so perturbations start to grow due to gravitational instability (Eisenstein and Hu, 1999; Mo et al., 2010; more on this in Sec. 1.2). Modes of the oscillation reaching their maxima at recombination have enhanced temperature fluctuations, leaving an imprint on the CMB referred to as *baryon acoustic oscillations* (BAO; Hu, 1995).

The amplitude of temperature fluctuations on different angular scales is measured by the CMB temperature-temperature (*TT*) power spectrum, shown in Fig. 1.3. In particular, there is a 'fundamental frequency' at the first peak in the CMB *TT* power spectrum with $\ell \sim 200$, where the fluid just had enough time to compress once before recombination (Hu and Dodelson, 2002). Half the wavelength of this mode is equal to the *sound horizon* r_s, the distance that sound in the baryon-photon plasma can travel by recombination, which depends on quantities such as the ratio of baryon-to-photon density (Eisenstein and Hu, 1998). The angular size of the first peak on the sky, $\sim 1^\circ$ (Planck Collaboration et al., 2020a), depends on the distance photons can travel since recombination d_s. This depends on the present-day $\Omega_{m,0}$, $\Omega_{\Lambda,0}$, H_0, and the universe's curvature. The location of the peak provides strong evidence that the universe is flat when an additional probe constraining either Ω_m or H_0 is combined with the *TT* power spectrum, e.g. CMB lensing, which we will return to shortly (otherwise, because the *TT* power spectrum is most sensitive to $\Omega_m h^2$ rather than Ω_m and H_0 individually, the same d_s can be obtained by adjusting Ω_m together with curvature; see e.g. Page et al., 2003; Planck Collaboration et al., 2020c). Successive peaks in the *TT* power spectrum can be viewed as 'overtones' of the

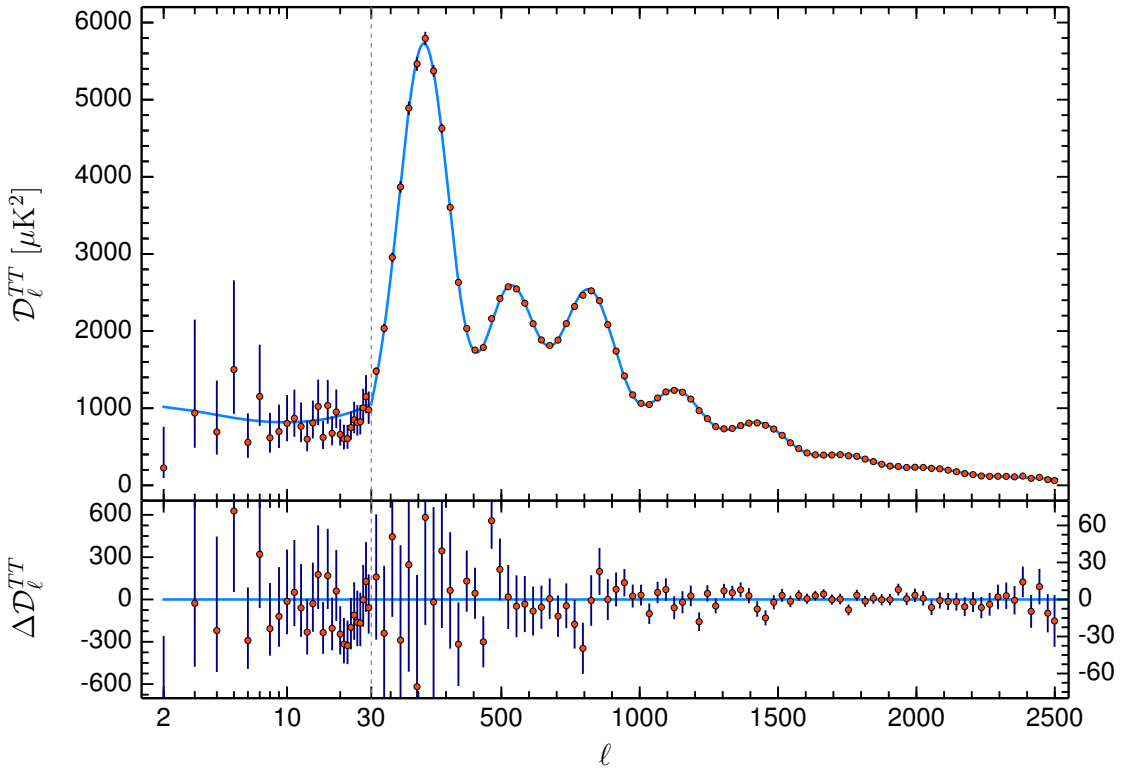


Figure 1.3: The *Planck* 2018 cosmic microwave background temperature (TT) power spectrum. The x-axis shows the multipole ℓ, with higher multipoles corresponding to smaller angular scales. The blue line in the top panel shows the theoretically calculated ΛCDM power spectrum using best-fit parameters to the *Planck TT, TE, EE* + low$E$ + lensing likelihoods. The lower panel shows residuals with respect to this model. More details can be found in the original caption of this figure 1 in Planck Collaboration et al. (2020c).

fundamental frequency, where the second peak corresponds to the scale on which the fluid just had time to compress and rarefy, and so on. Measuring the amplitude of the successive peaks provides tight constraints on the baryon density $\Omega_{b,0}h^2$ and dark matter density $\Omega_{c,0}h^2$. In fact, the locations and amplitudes of the CMB peaks provide tight constraints on all six parameters of flat ΛCDM (see e.g. Hu, 1995, 2001).[4]

In addition to the temperature anisotropies, cosmological information can be extracted from CMB polarisation and lensing (see e.g. Hu and Dodelson, 2002, for a review). Polarisation occurs due to the Thomson scattering between CMB photons and free electrons, and can be decomposed into an E-mode and a B-mode. In particular, theories of inflation predict the generation of tensor perturbations (gravitational waves) which induce B-mode polarisation, un-

[4]Note though that CMB directly constrains the angular scale of sound horizon at recombination $\theta_\star$, and H_0 is a derived parameter.

like scalar (density) perturbations which only induce E-mode polarisation. Therefore, detecting B-mode polarisation in the primary CMB (removing foreground effects)[5] would provide strong evidence for inflation (see Planck Collaboration et al., 2020e; BICEP/Keck Collaboration et al., 2022, for recent results and constraints on inflation). As the CMB photons travel towards us, they are gravitationally lensed by the matter distribution between us and when photons decoupled. This lensing smooths out peaks and troughs in the TT power spectrum, converts some of the E-mode polarisation into B-mode, and causes correlations between different Fourier modes so the CMB signal becomes non-Gaussian. Measuring this last correlation allows the lensing potential to be reconstructed, which can be used to constrain a combination of Ω_m, H_0, and σ_8, which measures the variance of matter fluctuations on scales of $R = 8h\,\mathrm{Mpc}^{-1}$ (see Eq. 1.59; Planck Collaboration et al., 2016a, 2020b). Combining CMB lensing potential measurements with the temperature and polarisation power spectra can constrain the spatial curvature of the universe (Planck Collaboration et al., 2020c). It also allows internal consistency checks for the cosmological parameters derived from the TT power spectrum. For example, combining CMB lensing potential measurements with measurements of the baryon acoustic oscillation (see Sec. 1.1.3.3) gives $H_0 = 67.9^{+1.2}_{-1.3}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ assuming flat ΛCDM, which agrees with $H_0 = (66.88 \pm 0.92)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ inferred from the TT power spectrum alone under the same modelling. On the other hand, the lensing amplitude expected from the smoothing effect on the TT power spectrum is larger than that predicted by the lensing potential reconstruction (Planck Collaboration et al., 2020c). The relative amplitudes between the two are described by a phenomenological parameter A_{lens} (or A_L, Calabrese et al., 2008); the nature of this so-called "A_{lens} anomaly" is expected to become clearer with current and future CMB experiments (Renzi et al., 2018). Nevertheless, combining the information from the CMB temperature, polarisation, and lensing maps gives some of the tightest constraints we have on cosmological parameters to date, with the results so far supporting ΛCDM (Planck Collaboration et al., 2020a,c; Costanzi et al., 2021; Madhavacheril et al., 2024).

1.1.3.2 Type Ia supernovae

Another independent test of the cosmological model comes from supernovae, which probe the expansion history of the universe. Type Ia supernovae (SNIa) are thought to arise from explod-

[5] Galactic foregrounds and lensing are the main sources of B-mode polarisations measured (BICEP/Keck Collaboration et al., 2022).

ing white dwarfs when their masses reach the Chandrasekhar limit, and they are characterised by a lack of hydrogen lines and the presence of a silicon absorption line in their spectra. They are *standardisable candles*: the peak luminosity of Type Ia supernovae can be standardised so that the scatter in their absolute magnitudes (the apparent magnitude if the object were at a distance of 10 parsecs) becomes only ~ 0.1 magnitude (Brout and Scolnic, 2021). The difference between their apparent and absolute magnitudes can then be attributed to the distance between us and the supernovae, which is the luminosity distance given by

$$d_L = (1+z)\int_0^z \frac{\mathrm{d}z'}{H(z')} \tag{1.28}$$

in a flat universe (Pontzen and Peiris, 2020). This probes the present-day Hubble parameter H_0, as well as $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ (or more generally $\Omega_{DE,0}$) through Eq. (1.27). While supernovae measurements also agree with a ΛCDM cosmology, as measurement precision improved, there now appears to be a discrepancy between the value of H_0 inferred from the CMB, $H_0 = (67.4 \pm 0.5)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (combining temperature, polarisation and lensing; Planck Collaboration et al., 2020c), compared to that from the supernovae, $H_0 = (73.04 \pm 1.04)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Riess et al., 2022). The statistical significance of the H_0 *tension* is in excess of 5σ.

Measuring H_0 using SNIa requires calibrating their absolute magnitudes using known distances. A common method, also used by Riess et al. (2022), is to find galaxies that contain both SNIa and classical Cepheids, a type of variable star whose pulsation period is strongly correlated with its absolute magnitude. The distance to the galaxy, and hence the supernova, can be determined from the Cepheids' apparent magnitudes (see e.g. Shah et al., 2021).[6] Although currently, Cepheids enable the most precise local-universe H_0 measurement, alternative methods such as the Tip of the Red Giant branch (TRGB) are becoming competitive. TRGB consists of post-main-sequence stars just before the onset of helium fusion. Helium fusion occurs in a narrow temperature range, allowing TRGB to be used as a standard(isable) candle (Freedman et al., 2024). Recent measurements by the James Webb Space Telescope (JWST; Freedman et al., 2024) using TRGB yielded a $H_0 = (69.85 \pm 1.75)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, which agrees with results from both the CMB and the Cepheid-calibrated SNIa. Interestingly, in the same study, Freedman et al. (2024) also used the high-resolution JWST data to mea-

[6] The absolute magnitudes of Cepheids are themselves calibrated using e.g. parallax to Milky Way Cepheids from *Gaia* (Riess et al., 2021); see Shah et al. (2021); Verde et al. (2023) for more details.

sure distances using two other methods: Cepheids, as well as a relatively new method utilising J-region asymptotic giant branch stars (JAGB). JAGB stars serve as standard candles due to their near-constant luminosity in the near-infrared J band and a low intrinsic dispersion (Freedman et al., 2024; Lee et al., 2024; see also Verde et al., 2023). Using Cepheid distances gave $H_0 = (72.05 \pm 1.86)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ in good agreement with Riess et al. (2022), whereas JAGB distances gave $H_0 = (67.96 \pm 1.57)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ in excellent agreement with *Planck*. Because the three methods each have their own sources of systematic errors, comparing between the three could lead to a better understanding of whether the H_0 tension is due to systematic errors. Freedman et al. (2024) found the distances determined using JAGB agree with TRGB to better than 1%, whereas both methods agree with Cepheid distances to better than 4%. They note that interestingly, both JAGB and TRGB methods are less prone to effects of crowding or blending (due to nearby stars) and reddening (by dust along line of sight) compared to Cepheids, since the latter is found in star-forming disks whereas JAGB and TRGB stars can be found in low surface brightness regions with less dust. This could point to systematics in Cepheids affecting SNIa distances and hence H_0 (see however Riess et al., 2024a,b). Alternatively, the H_0 tension could be due to systematic errors possibly associated with supernovae measurements, such as the potential evolution in SNIa at higher redshifts, or due to new physics beyond the ΛCDM standard model (for a review on the H_0 tension, see e.g. Shah et al., 2021; Verde et al., 2023).

1.1.3.3 Large-scale structure

In the standard model of cosmology, the small anisotropies seen in the cosmic microwave background (CMB) grow by gravitational instability into the large-scale structures observed today. Hence, probes of the large-scale structure test whether the present-day inhomogeneities are consistent with those expected from the CMB under ΛCDM. Furthermore, the large-scale structure probes density fluctuations on scales smaller than those accessible from the CMB, and does so as a function of redshift, providing a three-dimensional view of the universe rather than the two-dimensional view offered by the CMB (see e.g. the top and left wedges of Fig. 1.1).

Structural growth via gravitational instability causes coupling between different Fourier modes, so the density field becomes non-Gaussian. This means the matter power spectrum defined in Eq. (1.2) cannot capture all the information in the density field at late times. Nevertheless, as mentioned at the beginning of the chapter, most existing analyses of large-scale structure data primarily constrain cosmology via the matter power spectrum (it will be dis-

cussed in greater detail in Sec. 1.2), as capturing information beyond the power spectrum in a robust and computationally-efficient way is challenging.

Although the matter power spectrum cannot be directly measured, as the bulk of matter fluctuations is in dark matter, galaxies are expected to trace the underlying matter distribution. Hence, measuring the clustering of galaxies indirectly probes the matter power spectrum. For example, the Two-degree-Field Galaxy Redshift Survey (2dFGRS; Colless et al., 2001), the Sloan Digital Sky Survey (SDSS; Gott et al., 2005; Percival et al., 2007; Alam et al., 2017, 2021), and most recently DESI (DESI Collaboration et al., 2016, 2024) measure the baryon acoustic oscillation feature (BAO, introduced in Sec. 1.1.3.1) in the matter power spectrum, which results from baryons being frozen at the scale of the sound horizon at the *drag epoch* $r_{\rm drag}$, when baryons decoupled from photons (Eisenstein and Hu, 1998). Since $r_{\rm drag}$ is well-estimated from theory, the angular scale of the BAO feature as a function of redshift provides a standard ruler to probe the universe's expansion history (Dodelson and Schmidt, 2020). There are, however, sources of uncertainty in connecting between galaxy distributions and matter distributions (often referred to as *bias*), as it is known that galaxies cluster differently to dark matter (see Wechsler and Tinker, 2018, for a review). Accurate predictions of bias are complicated by the challenge in accurately modelling galaxy formation, which requires accounting for complex astrophysical phenomena (*baryonic effects*) not well-constrained (see Vogelsberger et al., 2020; Crain and van de Voort, 2023, for reviews).

A probe which avoids the problems of bias is weak gravitational lensing (a.k.a. cosmic shear). It measures the correlation in galaxy shapes due to the distortion of light by matter between us and the source galaxies, and therefore directly probes the matter density along the line of sight like CMB lensing, though at smaller scales. Weak lensing measurements are especially sensitive to the parameter combination $S_8 \equiv \sqrt{\Omega_{m,0}/0.3}\sigma_8$. Recently, a tension has emerged in the value of S_8 measured by weak lensing compared to the CMB, though the statistical significance at $2-3\sigma$ is lower than the H_0 tension (e.g. Li et al., 2023b; see Abdalla et al., 2022, for a review). The situation on this may become clearer with data from forthcoming surveys like *Euclid*, and with improved data analyses addressing the main sources of uncertainties in weak lensing analyses, which include the estimation of the lensed sources' redshift distribution, and the modelling of intrinsic alignments in galaxies (Joachimi et al., 2015; Mandelbaum, 2018). There is still much untapped constraining power in weak lensing data: weak lensing can probe

scales much smaller than those probed by the CMB, but data at such small scales are often not used in existing analyses because baryonic effects have significant impacts on such small scales (Abbott et al., 2022).

Matter fluctuations on smaller scales is also probed by the Lyman-alpha forest. This is a series of absorption lines in spectra of bright distant quasars and galaxies caused by the Lyman-alpha transition when light emitted from these distant sources passes through neutral hydrogen at different redshifts. It therefore provides a one-dimensional map of the matter distribution between us and the distant source. It can probe the universe at higher redshifts and smaller scales than those accessible from galaxy clustering measurements, though relating the absorption lines to the underlying matter density requires modelling more complex physics which are less well-constrained (Semelin et al., 2023).

Galaxy clusters also provide a means of constraining the cosmological model. In particular, the abundance of galaxy clusters is a sensitive probe of cosmology, and will be the focus of this thesis. I describe this in greater detail in Sec. 1.3.

Finally, small scale probes of the dark matter halo density profiles (e.g. galaxy rotation curves) and halo abundance (e.g. dwarf galaxy counts) also provide constraints on the cosmological model. These are especially interesting as potential problems of the CDM scenario mostly arise on small scales. These potential problems inspire alternative dark matter candidates such as warm dark matter (see e.g. Bode et al., 2001; Viel et al., 2013), self-interacting dark matter (see review in Tulin and Yu, 2018), and ultralight dark matter (see reviews in Hui et al., 2017; Ferreira, 2021). These are designed to agree with CDM predictions on large scales but resolve small-scale issues, such as the excess of predicted low-mass halos (Moore et al., 1999; Klypin et al., 1999), and overly-cuspy halo profiles (Flores and Primack, 1994; Oh et al., 2015). However, discrepancies on small scales may also be resolved by accounting for baryonic effects (see a review in Bullock and Boylan-Kolchin, 2017).

In summary, observations ranging from the cosmic microwave background to galaxy rotation curves have provided strong constraints on the cosmological model, which appears mostly consistent with flat ΛCDM. However, there are still exciting open questions left to answer, from understanding the source of tensions in the values of H_0 and S_8, to searching for evidence of inflation, probing neutrino properties, and understanding the nature of dark matter and whether dark energy is a cosmological constant. Current and forthcoming experiments measuring the

cosmic microwave background (Abazajian et al., 2016; Ade et al., 2019), mapping galaxies and weak lensing using both photometric and spectroscopic surveys (e.g. The Dark Energy Survey Collaboration, 2005; Laureijs et al., 2011; Takada et al., 2014; DESI Collaboration et al., 2016; Hofmann et al., 2017; Aihara et al., 2018; Ivezić et al., 2019), as well as gravitational waves (Punturo et al., 2010; Amaro-Seoane et al., 2017; Reitze et al., 2019) will provide high-quality data to probe physics beyond the standard model of cosmology.

1.2 Linear structure growth

In the standard model, structures grow from initially small quantum perturbations generated during inflation. As introduced at the beginning of Sec. 1.1, inflation leaves us with initial small overdensities which can be described by a power spectrum

$$P(k) \propto A_s k^{n_s} . \qquad \text{(1.3 reproduced)}$$

Over time, matter accumulates in slightly overdense regions, and the overdensities grow in amplitude under the influence of gravity, eventually forming the highly clustered structures in the universe today. The initial growth of the overdensities can be accurately described by linear perturbation theory, leading to accurate theoretical predictions for quantities like the linear matter power spectrum and the linear growth function, which can be exploited for cosmological constraints.

Since the total matter abundance is dominated by cold dark matter (CDM), baryons play a minor role compared to dark matter for linear structure growth. To obtain a qualitative understanding of the shape of the linear matter power spectrum, I therefore only consider perturbations in CDM. Sec. 1.2.1 summarises the evolution of different modes under linear theory, leading to the basic shape of the linear matter power spectrum in Sec. 1.2.2. I then briefly describe the effect of including baryons and changing $N_{\rm eff}$ on the power spectrum shape. Sec. 1.2.3 gives a short overview of how the linear functions are calculated, and Sec. 1.2.4 briefly summarises observational constraints on the matter power spectrum.

1.2.1 Mode evolution

The evolution of small perturbations generated from inflation is described by considering a small perturbation $\delta g_{\mu\nu} \ll 1$ to the background metric tensor $\bar{g}_{\mu\nu}$:

$$ds^2 = (\bar{g}_{\mu\nu} + \delta g_{\mu\nu}) dx^\mu dx^\nu , \tag{1.29}$$

where $\bar{g}_{\mu\nu}$ solves the Einstein field equations Eq. (1.10) in a homogeneous and isotropic universe (Pontzen and Peiris, 2020). For convenience, let us continue to use $c = \hbar = 1$ in this subsection.

General relativity leaves us with gauge freedom, i.e. a freedom to choose the coordinate system in which to express the metric perturbation. For structure formation, we only need to consider the scalar mode of metric perturbations; a commonly used gauge for this is the conformal Newtonian gauge,[7] wherein Eq. (1.29) becomes

$$ds^2 = a^2(\eta) \left((1 + 2\Phi) d\eta^2 - (1 - 2\Psi) \delta_{ij} dx^i dx^j \right) ; \tag{1.30}$$

η is the conformal time defined in Eq. (1.8), and Φ and Ψ are two potentials. The advantage of this gauge choice is that in the Newtonian limit, the potential Ψ plays the role of the Newtonian gravitational potential, facilitating its physical interpretation (Ma and Bertschinger, 1995). Furthermore, as the contents of the universe are well-described by perfect fluids (see Sec. 1.1.2), solving the Einstein field equations Eq. (1.10) using the metric in Eq. (1.30) and considering perturbations in the energy-momentum tensor of perfect fluids, we find (Pontzen and Peiris, 2020)

$$\Phi = \Psi , \tag{1.31}$$

i.e. the evolution of the scalar perturbations are governed by a single Newtonian-like potential.[8]

As described at the beginning of Sec. 1.1, the initial density field is approximately homogeneous, so different Fourier modes of the field are uncoupled and evolve independently. Therefore, let us consider the evolution of modes with wavenumber k in Fourier space. This subsection will outline the main results that allow us to obtain a qualitative description of the

[7] While the conformal Newtonian gauge is only applicable to the scalar mode of metric perturbations, it can be generalised to include vector and tensor modes (Bertschinger, 1995).

[8] This holds if there is no anisotropic stress, which is true for CDM.

shape of the linear matter power spectrum; further details can be found in e.g. Weinberg (2008); Dodelson and Schmidt (2020); Pontzen and Peiris (2020).

The evolution of the potential Φ can be found by considering the Einstein equations using the perturbed metric Eq. (1.30) in Fourier space. For a fluid with an equation of state parameter w as defined in Eq. (1.14), the evolution is given by

$$\Phi'' + 3(1+w)\mathcal{H}\Phi' + wk^2\Phi = 0\,, \tag{1.32}$$

where $\mathcal{H} \equiv aH$.

For the overdensity defined in Eq. (1.1), the evolution equation can be found by considering the conservation of energy momentum, $\nabla_\mu T^\mu_\nu = 0$ (see Ma and Bertschinger, 1995). For CDM with $w = 0$ and no anisotropic stress, it is

$$\delta_C'' + \mathcal{H}\delta_C' = -k^2\Phi + 3\mathcal{H}\Phi' + 3\Phi''\,. \tag{1.33}$$

Let us first consider the super-horizon modes that remain outside the horizon at all times. Such modes have a very large wavelength so that $k \ll \mathcal{H}$. This means Eq. (1.32) reduces to $\Phi'' + 3(1+w)\mathcal{H}\Phi' = 0$, which admits a constant solution $\Phi(\mathbf{k},\eta) = \Phi(\mathbf{k},\eta_0)$. So, super-horizon modes are frozen outside the horizon. Since the potential Φ is constant, using Eq. (1.33), we find that the density contrast δ_C remains constant too.[9]

For the modes that entered the horizon during radiation domination, we have $w = 1/3$, and $\mathcal{H} = 1/\eta$. Therefore, Eq. (1.32) becomes

$$\Phi'' + \frac{4}{\eta}\Phi' + \frac{k^2}{3}\Phi = 0\,. \tag{1.34}$$

Its solution (see e.g. Dodelson and Schmidt, 2020) tends to a constant for $\eta \to 0$, but as soon as the mode enters the horizon, the potential Φ starts oscillating due to the interaction between gravity and radiation pressure (this is the acoustic oscillations described in Sec. 1.1.3.1). The amplitude of the potential oscillations also decays due to the universe's expansion. While baryons are coupled to radiation so further growth in baryon overdensities is prevented by the

[9] Note that the behaviour of modes on super-horizon scales is gauge-dependent, so while they appear constant in the conformal Newtonian gauge, this is not necessarily the case in other gauges (Ma and Bertschinger, 1995).

radiation pressure, CDM only interacts gravitationally, so the CDM perturbations continue to grow. The overdensity in CDM evolves according to Eq. (1.33), with a solution of the form

$$\delta_C \propto \ln(Bk\eta) + \text{constant} \tag{1.35}$$

for B a constant. This shows that during the radiation dominated era, CDM perturbations only grow logarithmically. The growth is suppressed compared to during matter domination (which we will come to later) because the universe expands more rapidly during radiation domination.

As the universe transitions into matter domination, the CDM perturbations which entered during radiation domination have grown to become larger than radiation perturbations. The evolution of the CDM perturbations is then described by the Mészáros equation (Mészáros, 1974), which has a general solution consisting of a growing mode and a decaying mode. The general solution tends to the solution in Eq. (1.35) at early times deep in radiation domination. Deep inside matter domination, the decaying mode has long decayed away, so the CDM perturbation is governed by the growing mode, which scales as $\delta \propto a/a_{\rm eq}$ for late times when $a \gg a_{\rm eq}$, the scale factor at matter-radiation equality. The solution of the Mészáros equation shows that a mode entering during radiation domination initially experiences suppressed growth (compared to if it were growing $\propto a$), but after the universe transitions to matter domination, at late times all modes tend towards growing proportional to the scale factor. The suppressed growth during radiation leaves an imprint on the amplitude of CDM perturbations depending on when the modes entered the horizon; smaller modes that entered earlier experience more suppression. This leads to the following expression for very small-scale CDM perturbations deep inside matter domination (see e.g. Dodelson and Schmidt, 2020):

$$\delta_C \propto \frac{a}{a_{\rm eq}} \ln\left(C\frac{k}{k_{\rm eq}}\right), \qquad a \gg a_{\rm eq}\,, \quad k \gg k_{\rm eq}\,, \tag{1.36}$$

where C is a constant, and $k_{\rm eq}$ is the wavenumber entering the horizon at matter-radiation equality:

$$k_{\rm eq} \equiv a_{\rm eq}H_{\rm eq} = \sqrt{2\Omega_m}H_0 a_{\rm eq}^{-1/2} = H_0\sqrt{2\Omega_m(1+z_{\rm eq})} \tag{1.37}$$

for k measured in units without factors of h.

On the other hand, for modes entering the horizon during matter domination, $w = 0$ and $\mathcal{H} = 2/\eta$ such that Eq. (1.32) and (1.33) give $\Phi =$ constant, $\delta_C \propto \eta^2 \propto a$, i.e. during matter domination, all modes grow as $\delta_C \propto a$.

From this analysis, we see there is a significant difference in the growth of perturbations entering the horizon before and after matter-radiation equality, which occurs at a redshift of

$$z_{\rm eq} = \frac{\Omega_m h^2}{\Omega_\gamma h^2 \left(1 + \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} N_{\rm eff}\right)} - 1\,, \tag{1.38}$$

where I have accounted for the energy density of neutrinos in the radiation energy density. Note that from here on the density parameters are the values evaluated at the present day $z = 0$, unless they are written as explicit functions of redshift or scale factor. As we shall see, the difference of growth before versus after matter-radiation equality leaves a clear imprint on the shape of the linear matter power spectrum.

1.2.2 Linear matter power spectrum

As introduced earlier, the density perturbations can be described by the power spectrum $P(k,a)$ defined (for a given a) in Eq. (1.2). While the matter power spectrum depends on the density perturbation of all matter contents, as CDM is the dominant component, for now we will assume $\delta_m = \delta_C$.

We predominantly observe the distribution of matter in the universe at late times, after all modes evolve identically again. This means the shape of the linear matter power spectrum will remain the same, while its amplitude increase can be described using a *growth factor* $D_+(a)$, with $D_+(a) = a$ deep in matter domination. The shape of the power spectrum at late times depends on the initial density perturbations given in Eq. (1.3), as well as the scale-dependent growth of perturbations discussed in Sec. 1.2.1. The latter is described by the *transfer function* $T(k)$, which is defined such that $T(k) = 1$ on large scales (small k; see e.g. Dodelson and Schmidt, 2020, for details). The linear matter power spectrum is then

$$P(k,a) \propto k^{n_s} D_+^2(a) T^2(k)\,, \tag{1.39}$$

keeping only the terms dependent on k and a. On small scales, using Eq. (1.36), we can find that $T(k) \propto \frac{k_{\rm eq}^2}{k^2} \ln\left(C\frac{k}{k_{\rm eq}}\right)$ (Dodelson and Schmidt, 2020).

To summarise, the shape of the linear matter power spectrum is given by

$$P(k) \propto \begin{cases} k^{n_s} & , \quad k \ll k_{\rm eq} \\ k^{n_s-4} \ln^2\left(C\frac{k}{k_{\rm eq}}\right) & , \quad k \gg k_{\rm eq} \end{cases} . \tag{1.40}$$

Since inflation left us with a near scale-invariant power spectrum with $n_s \sim 1$, we find on large scales approximately $P(k) \propto k$: power increases with increasing wavenumber, as those modes entered the horizon earlier and have had more time to grow. On the other hand, for the modes which entered the horizon well before matter domination, their growth was suppressed due to the rapid expansion of the universe during radiation domination. The smaller the mode, the earlier it entered the horizon during radiation domination, and the more suppression it experienced, such that for $k \gg k_{\rm eq}$, the power falls off approximately as $P(k) \propto k^{-3} \ln^2\left(\frac{k}{k_{\rm eq}}\right)$. Hence, the power spectrum has a distinct turnaround at $k \sim k_{\rm eq}$ given by Eq. (1.37). This can be seen in Fig. 1.4 around $k_{\rm eq} \sim 10^{-2}\,h\text{Mpc}^{-1}$.

We arrived at the overall qualitative shape of the linear matter power spectrum by assuming $\delta_m = \delta_C$. The presence of baryons has mainly two additional effects, which are illustrated in Fig. 1.4. The first is that during radiation domination, baryons are tightly coupled to photons and oscillate instead of cluster. Because a fraction of matter now does not participate in gravitational collapse (compared to in the CDM-only scenario), the balance between gravitational attraction and background expansion shifts compared to what we considered in Sec. 1.2.1. This causes a further reduction in the growth rate of CDM perturbations for modes which entered the horizon before matter-radiation equality. Hence, at the end of the baryon drag epoch when baryons decoupled from photons, the CDM perturbations have lower amplitudes compared to if no baryons were present. After the drag epoch, eventually the growth rate of baryon fluctuations catches up with that of CDM again. This results in suppressed power at small scales (Eisenstein and Hu, 1998; Lesgourgues et al., 2013). Secondly, baryon acoustic oscillations (BAO, see Sec. 1.1.3.3) leave an imprint in the matter power spectrum. At the baryon drag epoch, the baryons are deposited at the scale of the sound horizon $r_{\rm drag}$ (much like how CMB anisotropies were 'frozen in' at the scale of the sound horizon at photon decoupling). Then, over time, baryons aggregate in the gravitational potentials of CDM, while CDM are also attracted by the gravitational potentials of baryons, resulting in a series of small oscillations in

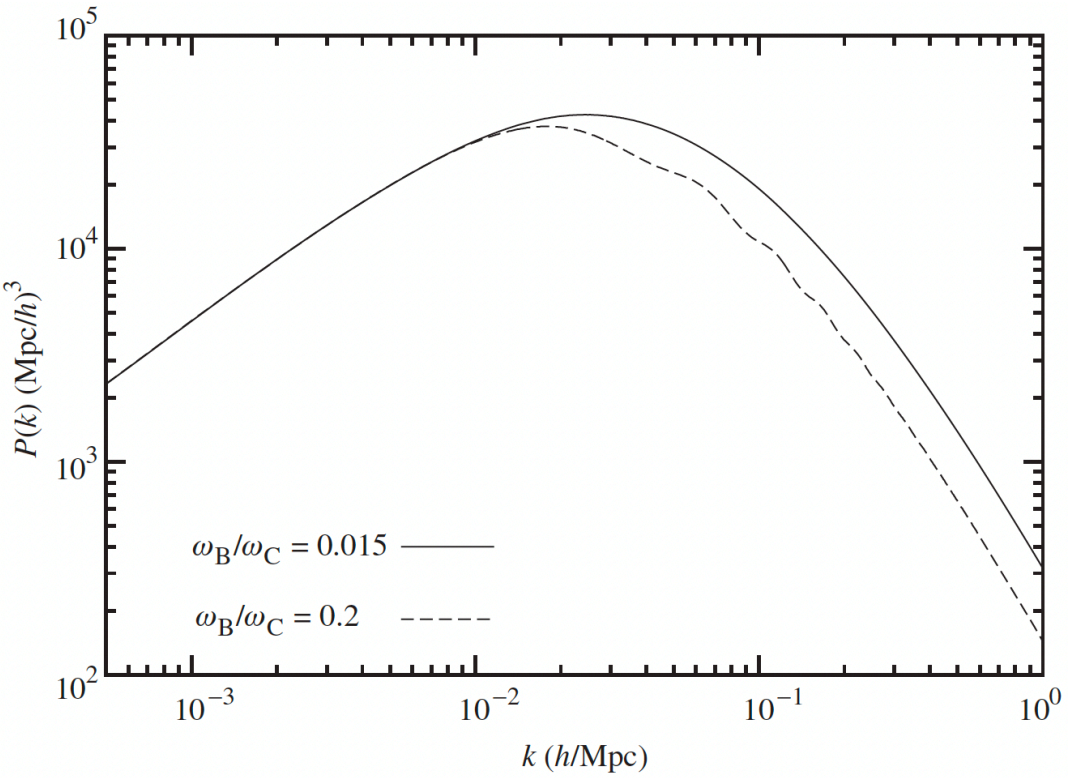


Figure 1.4: The linear matter power spectrum for two neutrinoless ΛCDM models with identical primordial power spectra and the same Ω_m, but different baryon-to-CDM ratios ω_B/ω_C, where e.g. $\omega_B = \Omega_b h^2$. An increase in baryon fraction suppresses the power at scales $k > k_{\rm eq}$, and the imprint of baryon acoustic oscillations becomes significant. Figure taken from Lesgourgues et al. (2013) figure 6.1.

the matter power spectrum spaced $k_s = \pi/r_{\rm drag}$ apart (see Eisenstein and Hu, 1998; Dodelson and Schmidt, 2020). The amplitude of BAOs decreases with increasing k due to Silk damping (Silk, 1968), the erasing of fluctuations caused by photon diffusion with respect to the baryons in between photon decoupling and the baryon drag epoch.

Finally, let us briefly touch on the effects of additionally including massless neutrinos (or other relativistic species). We saw in Sec. 1.1.2.1 that these effects can be accounted for by the single parameter $N_{\rm eff}$. Varying $N_{\rm eff}$ can affect the shape of the linear matter power spectrum by changing $k_{\rm eq}$ (combine Eq. 1.37 with 1.38), and by affecting the perturbations primarily via reducing the amplitude and shifting the phase of BAOs. These effects are illustrated in Fig. 1.5. More detailed discussions of neutrino effects, including the effect of massive neutrinos, can be found in e.g. Lesgourgues et al. (2013).

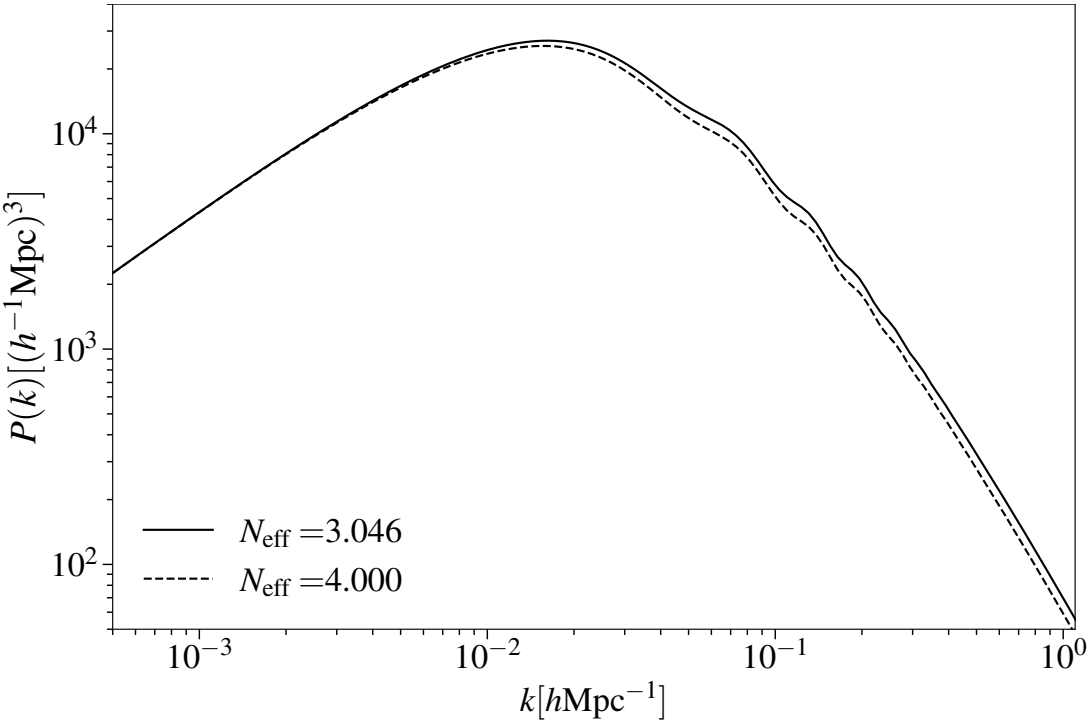


Figure 1.5: The linear matter power spectrum for two ΛCDM models with identical cosmological parameters, except different N_{eff}. Changing N_{eff} changes both k_{eq} and affects the baryon acoustic oscillations. The power spectra are calculated using CAMB (Lewis et al., 2000).

1.2.3 Calculating the linear functions

In the previous sections, we obtained approximate expressions for the evolution of overdensities, leading to an understanding of the shape of the matter power spectrum. Highly accurate calculations of the mode evolutions for different energy components of the universe (cold dark matter, baryons, radiation, etc.) that take into account the interactions and couplings between species is done nowadays using Einstein-Boltzmann equation solvers such as the Code for Anisotropies in the Microwave Background (CAMB, Lewis et al., 2000) and the Cosmic Linear Anisotropy Solving System (CLASS, Lesgourgues, 2011; Blas et al., 2011). These codes numerically solve the linearised Einstein and Boltzmann equations in an expanding universe, where the Boltzmann equations describe how the phase space distributions of different species evolve in time, and the Einstein equations describe how these interact with spacetime. These codes can evaluate the transfer function $T(k)$ as a function of redshift, thereby providing very accurate predictions of the linear matter power spectrum at different redshifts.

The growth factor $D_+(a)$ encodes information on the late-time growth of structure, from deep in matter domination, through matter-dark-energy equality occurring at

$$z_{\mathrm{m=DE}} = \left(\frac{\Omega_m}{\Omega_{DE}}\right)^{\frac{1}{3w}} - 1 \tag{1.41}$$

for dark energy with a constant equation of state parameter w, until the present day when dark energy dominates and the accelerated expansion slows down the growth of structures. In a ΛCDM universe, the growth factor can be found via an analytic integral. However, for $w \neq -1$, it needs to be found by solving for the sub-horizon late time evolution of matter perturbations δ_m (see e.g. Dodelson and Schmidt, 2020; Linder and Jenkins, 2003). On the other hand, the growth rate is well-approximated by

$$\frac{\mathrm{d}\ln D}{\mathrm{d}\ln a} \approx \Omega_m^{\gamma}(z)\,, \qquad \gamma = 0.55 + 0.05\,(1 + w(z=1)) \tag{1.42}$$

to within percent-level accuracy even for some dynamical dark energy models (Linder, 2005).

Alternatively, since it represents the scale-independent growth at late times, the growth factor can be found by evaluating the transfer function at a given k as a function of redshift using CAMB or CLASS (for redshifts from deep in matter domination to today).[10] When introducing the growth factor in Eq. (1.39), we defined it to be $D_+(a) = a$ deep in matter domination. However, it can be normalised arbitrarily (as long as it combines with the transfer function to give the correct power spectrum at redshift z). In this thesis, I will be using

$$D(z) = \frac{D_+(z)}{D_+(z_{\mathrm{norm}})} \tag{1.43}$$

to refer to the growth factor at redshift z normalised to unity at z_{norm}. In most cases I use $z_{\mathrm{norm}} = 0$, i.e. the growth factor is normalised to unity today, which is commonly used in literature, though later in Ch. 4 I will be using $z_{\mathrm{norm}} = 50$ to produce inputs for the deep learning model.

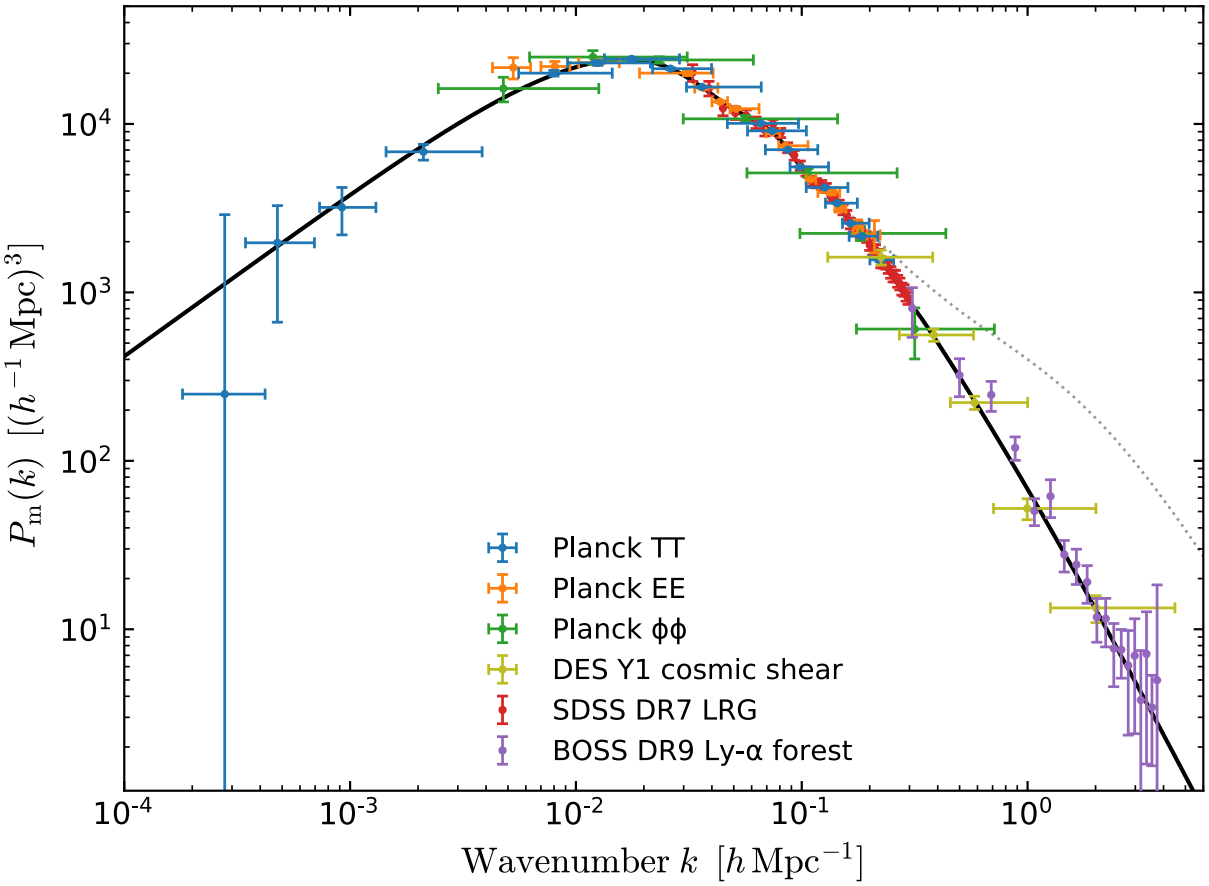


Figure 1.6: Figure 19 from Planck Collaboration et al. (2020a) showing the linear matter power spectrum $P(k)$ at $z = 0$ predicted from theory using CMB-inferred cosmological parameters as the solid line. It is overplotted with that inferred from different cosmological probes at different scales: CMB temperature (TT), polarisation (EE) and lensing ($\phi\phi$) reconstruction power spectra measured by *Planck* (Planck Collaboration et al., 2020a,b), Dark Energy Survey Year 1 (DES Y1) weak lensing cosmic shear measurements (Troxel et al., 2018), galaxy clustering measured from the Sloan Digital Sky Survey Data Release 7 luminous red galaxy sample (SDSS DR7 LRG, Reid et al., 2010), and Lyman-alpha measurements from the Baryon Oscillation Spectroscopic Survey Data Release 9 (BOSS DR9 Ly-α forest, Lee et al., 2013). The dotted line straying from the linear theory prediction at small scales shows the impact that non-linear clustering has at $z = 0$.

1.2.4 Observational constraints on the matter power spectrum

As we have seen, the matter power spectrum at redshift z summarises how fluctuations on different scales evolved until redshift z, and encodes rich information on the underlying cosmological model. It is a key function constrained by many probes such as galaxy clustering, weak lensing and the Lyman-alpha forest (see Sec. 1.1.3). Fig. 1.6, taken from Planck Collaboration et al. (2020a), shows the linear matter power spectrum inferred from the different cosmological

[10] If dark energy has $w \neq -1$, the growth for very large modes that entered the horizon when dark energy density becomes significant can be scale-dependent.

probes compared to the ΛCDM prediction. On large scales, the matter power spectrum is constrained by measurements from the CMB (Planck Collaboration et al., 2020a,b). Intermediate scales featuring the BAOs are probed by galaxy clustering measured from SDSS (Reid et al., 2010); smaller scales are probed by measurements of the Lyman-alpha (Ly-α) forest from the Baryon Oscillation Spectroscopic Survey (BOSS; Lee et al., 2013), and weak lensing cosmic shear measurements from the Dark Energy Survey (DES; Troxel et al., 2018). The constraints from different late-time probes on the linear matter power spectrum are broadly consistent with the ΛCDM prediction made using cosmological parameters constrained from the CMB at $z \sim 1100$. We have seen though in Sec. 1.1.3 that there are tensions within the ΛCDM paradigm regarding the value of H_0 and, to a lesser extent, S_8 (see Planck Collaboration et al., 2020a, for a more detailed discussion of discord between *Planck* and other probes).

Also shown in Fig. 1.6 as the dotted line is the non-linear matter power spectrum. As overdensities continue to grow, $\delta_m \sim \mathcal{O}(1)$, and linear perturbation theory no longer accurately describes the growth of structures. Overdensities $\delta \gtrsim 1$ collapse and cluster under the influence of gravity, resulting in enhanced power at small scales for modes which have had sufficient time to grow since they have entered the horizon. The scale $k_{\rm NL}$ above which non-linearities cannot be ignored can be estimated by considering the variance of linear density perturbations

$$\Delta_L^2(k,a) = \frac{k^3 P(k,a)}{2\pi^2}\,. \tag{1.44}$$

It becomes $\Delta_l^2 \simeq 1$ for $k_{\rm NL}(z=0) \simeq 0.25h\,{\rm Mpc}^{-1}$ today (Dodelson and Schmidt, 2020), which can be seen in Fig. 1.6. The next section describes the evolution of structures in the non-linear regime.

1.3 Beyond linear growth and halos

As overdensities continue to grow, different modes start coupling with each other, and linear perturbation theory can no longer accurately describe their growth. The density contrast continues to increase, and at some stage it decouples from the expanding background and collapses under the influence of gravity. Numerical simulations show collapse occurs firstly along one dimension, forming 'pancakes', then along the second dimension, forming 'filaments'. Fi-

nally, dark matter halos form, which are three-dimensionally collapsed, gravitationally bound dark matter structures. Dark matter halos form the building blocks of cosmic structures within which galaxies are believed to form. Constraining halo properties, such as their mass, abundance, and density profiles provides great tests of the cosmological model, as they are sensitive to cosmological parameters, and/or the nature of dark matter. Sec. 1.3.1 describes how halo properties are probed observationally. I then describe the analytical model of halo formation via spherical collapse in Sec. 1.3.2.

1.3.1 Observational probes of dark matter halos

Halo properties such as their density profiles and abundance as a function of halo mass are highly interesting. They can offer a means of testing the nature of dark matter. Cold-dark-matter-only (CDM-only) simulations[11] suggest that halos should have cuspy central density profiles which scale as $\rho \propto r^{-1}$. This appears discrepant with observations, where the galaxy rotation curves suggest a relatively flat central density profile (e.g. Flores and Primack, 1994; Oh et al., 2015). CDM-only simulations also suggest that halos should form down to very small halo masses,[12] but the number of observed Milky Way satellites appeared lower than expected (e.g. Klypin et al., 1999; see however recent advances in e.g. Bose et al., 2020; Homma et al., 2024). These "small-scale problems" spurred active investigations into alternative dark matter models such as ultralight dark matter and warm dark matter, although baryonic feedback may be sufficient to resolve such discrepancies (see e.g. recent reviews in Weinberg et al., 2015; Bullock and Boylan-Kolchin, 2017; Del Popolo and Le Delliou, 2017).

Halo abundance also probes cosmological parameters and the nature of dark energy. As introduced earlier, the abundance of especially high-mass halos hosting galaxy groups and galaxy clusters is a sensitive probe of cosmological parameters, in particular Ω_m and σ_8. This makes galaxy cluster counts a competitive cosmological probe (Dodelson et al., 2016; Abbott et al., 2020; Abdullah et al., 2020; Costanzi et al., 2021) that current and forthcoming surveys utilise (e.g. DES, The Dark Energy Survey Collaboration, 2005; eROSITA, Hofmann et al., 2017; *Euclid*, Sartoris et al., 2016; and LSST, Ivezić et al., 2019). Measuring cluster mass from

[11] Technically these should be called gravity-only simulations, as the effect of baryons on the background evolution is included in the simulations, but baryonic effects such as feedback and pressure are not modelled; matter interact only via gravity in these simulations.

[12] This supports the halo model, a framework widely used to analyse cosmological data, in which all matter is seen as residing in halos (Asgari et al., 2023).

observations is currently the main source of error in cluster count experiments; I summarise in this subsection how halo masses are estimated especially from galaxy cluster observations.

As some of the most massive collapsed structures in the universe, galaxy clusters host a large number of galaxies, mostly early-type red galaxies. The space between them is filled by the *intracluster medium*, which is hot, diffuse ionised gas not directly associated with any galaxy. For massive clusters, the bulk of the baryonic mass is in the intracluster medium, with the ratio of stellar mass M_* to gas mass M_g at $M_*/M_g \sim 0.2-0.05$, decreasing for larger clusters (Lin et al., 2012). The deep gravitational potentials inside massive galaxy clusters compress the diffuse plasma, heating it to high temperatures and prompting X-ray emissions due to collisional processes such as bremsstrahlung, recombination, and line radiation (Allen et al., 2011). At the same time, cosmic microwave background (CMB) photons travelling through the cluster scatter off the thermal electrons in the intracluster medium, causing a spectral distortion in the CMB, an effect known as the thermal Sunyaev-Zel'dovich (tSZ) effect (Sunyaev and Zeldovich, 1970). Galaxy clusters can therefore be observed in the optical and near-infrared band via observing galaxies and starlight, in the X-ray band to probe the intracluster medium, and in the microwave band through the tSZ effect. The mass of clusters also causes gravitational lensing of background galaxies, so cosmic shear measurements can also be used to infer galaxy cluster mass.

Cluster mass can be measured in X-ray by assuming that the intracluster medium is in hydrostatic equilibrium, so that the mass M within radius r is related to the temperature profile $T(r)$ and gas density profile $\rho(r)$. Similarly, in the optical band the mass profile $M(r)$ can be estimated by assuming dynamical equilibrium, with galaxies acting as test particles in the cluster. Compared to X-ray mass measurements, optical mass measurements have the advantage of being insensitive to effects such as magnetic fields and turbulence. However, the assumption of dynamical equilibrium is less well-justified, as galaxies are largely collisionless and have longer relaxation timescales. Optical mass measurements are also more sensitive to triaxiality than X-ray measurements, and identification of the cluster centre is more difficult due to the finite number of galaxies (see reviews in Allen et al., 2011; Kravtsov and Borgani, 2012).

Combining the assumptions of hydrostatic equilibrium and spherical symmetry, together with additional assumptions that galaxy clusters form in a self-similar way, cluster scaling relations can be derived using the Kaiser model (Kaiser, 1986, see also a review in Kravtsov and

Borgani, 2012). These relate the halo mass to observables such as gas temperature T, gas mass $M_{\rm gas}$, and the product of the gas mass with X-ray temperature $Y_X = TM_{\rm gas}$ using simple power laws. In practice, the scaling relations between observables ("mass proxies") and halo mass are calibrated observationally and using simulations. Earlier calibrations of the scaling relations often employ X-ray mass measured assuming hydrostatic equilibrium (Vikhlinin et al., 2009; Mantz et al., 2010; Planck Collaboration et al., 2016b). However, numerical simulations have shown that such assumptions can lead to significant biases in the cluster mass measured (Lau et al., 2009; Rasia et al., 2012). In the recent years, calibration of observable-mass scaling relations often utilise weak lensing mass measurements in order to mitigate bias. For example, to produce the recent cosmological constraints from cluster abundances measured by the eROSITA first All-Sky Survey, Ghirardini et al. (2024) used the X-ray count rate in the soft X-ray band as the observable, and calibrated its scaling relation with the total cluster mass utilising weak lensing measurements from DES (The Dark Energy Survey Collaboration, 2005; Gatti et al., 2021; Sevilla-Noarbe et al., 2021), the Kilo Degree Survey (KiDS, de Jong et al., 2013; Giblin et al., 2021; Hildebrandt et al., 2021), and the Hyper Suprime-Cam (HSC) survey (Aihara et al., 2018; Kleinebreil et al., 2024). For optical surveys, *richness* is a commonly used mass proxy measuring the number of galaxies associated with the cluster photometrically. The mass-richness relation is also calibrated using weak lensing data by McClintock et al. (2019a) for redMaPPer clusters identified in DES Year 1 data. *Euclid* also plans to calibrate the mass-richness relation using its weak lensing and spectroscopic measurements, in addition to cross-correlating its measurements with X-ray samples from eROSITA and Sunyaev-Zel'dovich (SZ) samples from the South Pole Telescope (SPT; Carlstrom et al., 2011), Atacama Cosmology Telescope (ACT; Marriage et al., 2011), and *Planck* (Planck Collaboration et al., 2016b; Sartoris et al., 2016). The SZ observations are sensitive to the line-of-sight intracluster medium pressure (the "Compton y-parameter", see e.g. Bleem et al., 2015; Planck Collaboration et al., 2016b). For SZ observations, commonly used mass proxies include the volume-integrated intracluster medium pressure (integrating y out to a given radius, e.g. in Planck Collaboration et al., 2016b; Zubeldia and Challinor, 2019; Salvati et al., 2022) and the SZ signal significance (e.g. Bocquet et al., 2019); the SZ observable-mass relations are now also commonly calibrated using weak lensing mass measurements.

As mentioned before, the dominant source of uncertainty in cluster cosmology currently lies in mass-observable relations, both due to the scatter in the scaling relation and its calibration. Many efforts have been made to define mass observables which have small scatter and are relatively insensitive to details of cluster dynamics and astrophysics (e.g. Kravtsov et al., 2006). Recently, efforts are also being made to find low-scatter cluster mass proxies using machine learning (Ntampaka et al., 2015; Armitage et al., 2019; Ho et al., 2019; Kodi Ramanah et al., 2020; Ho et al., 2021). Applied to dynamical mass measurements, deep learning methods (introduced in greater detail in Ch. 2) can achieve better than a factor of two reduction in scatter compared to traditional power-law relations between halo mass and line-of-sight velocity dispersion (Ntampaka et al., 2015; Ho et al., 2019). Machine learning is also being used to directly predict halo mass (Calderon and Berlind, 2019; Kodi Ramanah et al., 2021; Villanueva-Domingo et al., 2022; Bowden et al., 2023).

With these efforts, the calibration accuracy has improved, e.g. McClintock et al. (2019a) quote $\sim 5\%$ calibration in the mass-richness relation of redMaPPer clusters. As calibration of the mass-observable relations continues to improve, tightly constraining cosmological parameters requires percent-level accurate halo mass function models (Sartoris et al., 2016; McClintock et al., 2019b; Euclid Collaboration et al., 2023a; Ghirardini et al., 2024). The rest of this chapter is dedicated to introducing models of the halo mass function, starting with the analytic theories of halo formation via spherical collapse.

1.3.2 Spherical collapse

Numerical simulation is currently the most accurate and robust way to model the deeply non-linear regime, where structures collapse and form. Nevertheless, analytical approximations using simplifying assumptions allow us to develop physical insights into the process of halo formation, and provide us with a language we can use to describe halo properties.

Although simulations show that gravitational collapse occurs at different rates along different dimensions, it is useful to consider the simplest possible scenario of halo formation, *spherical collapse* (Gunn and Gott, 1972). In this scenario, halos form from the collapse of an isolated, uniform spherical region that has a slightly higher mean density than its surrounding. It also assumes different shells of matter do not cross each other. This subsection summarises

some of the main results from spherical collapse; more details can be found in e.g. Dodelson and Schmidt (2020); Mo et al. (2010).

The comoving size of a region enclosing mass M is given by its Lagrangian radius R_L, such that

$$M = \frac{4\pi}{3}\Omega_m \rho_c R_L^3 = \frac{4\pi}{3}\rho_b R_L^3 , \tag{1.45}$$

for ρ_c the critical density of the universe in Eq. (1.19), and we have defined (as is common in literature) $\rho_b \equiv \Omega_m \rho_c$ to be the mean background density (note the subscript b stands for background and not for baryons!). The evolution of this isolated spherical region can be described by the second Friedmann equation Eq. (1.13) (Dodelson and Schmidt, 2020)

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho - 2\rho_{\rm DE}) , \tag{1.46}$$

where $\rho > \rho_b$ is the mean density within the spherical region, and we are considering structure growth since the matter dominated era so only matter and dark energy are relevant. We have also used dark energy density $\rho_{\rm DE}$, which is given in Eq. (1.22) for the cosmological constant. To find an analytical solution, let us consider from now on an Einstein-de Sitter (EdS) cosmology where the universe is flat and dominated by cold dark matter so $\Omega_m = 1$. This is also a good approximation for structure formation at high redshifts, deep in the matter dominated era.

The physical radius r of the overdensity is proportional to the scale factor a, so it evolves as (Dodelson and Schmidt, 2020)

$$\ddot{r} = -\frac{GM}{r^2(t)} . \tag{1.47}$$

This is just a Newtonian equation of motion for a spherical shell enclosing mass M. By integrating, we can find that for a gravitationally bound system, Eq. (1.47) has parametric solutions of the form (Mo et al., 2010)

$$r = A(1-\cos\theta) , \qquad t = B(\theta - \sin\theta) , \qquad A^3 = GMB^2 , \tag{1.48}$$

with $\theta \in [0, 2\pi)$ a parametric time variable. The solutions imply that at early times, the shell continues to expand, but expansion slows down. At $\theta = \pi$, the shell reaches its maximum size, turns around, and starts to collapse. The turnaround time $t_{\rm ta} = \pi B$, and the radius at turnaround is $r_{\rm ta} = 2A$. At $\theta = 2\pi$, the shell collapses back to $r = 0$ at the time of collapse $t = t_{\rm coll} = 2t_{\rm ta}$.

At early times, $\theta \ll 1$, so we can write r and t as power series in θ. Keeping the first two non-zero terms, we can combine them to show that r evolves according to linear theory as (Hu, 2017)

$$r = \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left(1 - \frac{1}{20}\left(\frac{6t}{B}\right)^{2/3}\right) = r_{\rm EdS} + \delta r\,, \tag{1.49}$$

where $r_{\rm EdS} = \left(\frac{9}{2}GM\right)^{1/3} t^{2/3}$ is the radius of a region that expands along with the background universe (if the density within the shell $\rho = \rho_b$), and δr is the slight decrease in radius due to the overdensity. Using the fact that the mass in the spherical region is conserved also when the overdensity δ becomes non-trivial, we can relate δ to δr and find that to first order, the overdensity evolves as (Hu, 2017)

$$\delta \approx \frac{3}{20}\left(\frac{6t}{B}\right)^{2/3} \tag{1.50}$$

in linear theory. Extrapolating the linear theory to the turnaround time $t_{\rm ta} = \pi B$ and the time of collapse $t_{\rm coll} = 2t_a$, we find that it predicts a linear overdensity of (Mo et al., 2010)

$$\delta_{\rm ta} = \frac{3}{20}(6\pi)^{2/3} \approx 1.062\,, \tag{1.51}$$

$$\delta_{\rm coll} = \frac{3}{20}(12\pi)^{2/3} \approx 1.686\,. \tag{1.52}$$

Therefore, shortly after the linear overdensity exceeds unity, it turns around and collapses.

Let us compare the linear predictions with the evolution of the overdensity according to the spherical collapse model (which is in general non-linear). According to spherical collapse, the overdensity of the spherical region is (van den Bosch, 2022a)

$$1 + \delta = \frac{\rho}{\rho_b} = \frac{9}{2}\frac{(\theta - \sin\theta)^2}{(1-\cos\theta)^3}\,, \tag{1.53}$$

where the density within the spherical region can be found using Eq. (1.48), and the background density of the universe as a function of time is found for an EdS cosmology using the first Friedmann equation Eq. (1.12) together with Eq. (1.17). Using our non-linear spherical collapse

model, we have (van den Bosch, 2022a)

$$1+\delta_{\rm ta} = \frac{9\pi^2}{16} \approx 5.552\,, \tag{1.54}$$

$$\delta_{\rm coll} = \infty\,. \tag{1.55}$$

Comparing this with the linear theory predictions Eq. (1.51) and (1.52), we find at turnaround time the spherical collapse model predicts an overdensity that is four times greater, and at $t_{\rm coll}$ the spherical collapse model predicts the overdensity should collapse to infinite density. This is, however, a highly idealised situation; in reality the collapse is never perfectly spherical, shell crossing happens, and eventually the halo formed virialises into a state of virial equilibrium (van den Bosch, 2022a).

The virial equilibrium is characterised by the final potential energy W_f and kinetic energy K_f satisfying the condition

$$2K_f + W_f = 0\,. \tag{1.56}$$

Using conservation of energy, we can relate the final energy E_f to the energy at turnaround $E_{\rm ta}$, when there is no kinetic energy. From this, we can find that $r_{\rm vir} = r_{\rm ta}/2 = A$ (see e.g. Hu, 2017); the radius at virialisation $r_{\rm vir}$ has decreased by a factor of two compared to at turnaround, and the mean density within the virialised halo is given by $\rho_h = 3M/(4\pi A^3)$. We can view virialisation as happening at $t_{\rm coll} = 2\pi B$, so the background density of the universe at $t_{\rm coll}$ is $\rho_b = (24\pi^3 GB^2)^{-1}$ (Kravtsov and Borgani, 2012). This means the overdensity at virialisation, which we denote using $\Delta_{\rm vir}$ to be consistent with most of the literature, is

$$\Delta_{\rm vir} = \frac{\rho_h}{\rho_b} = 18\pi^2 \approx 178\,. \tag{1.57}$$

The virial overdensity $\Delta_{\rm vir}$ is often used to define halos as regions enclosing a mean overdensity equal to $\Delta_{\rm vir}$. Since $\Delta_{\rm vir} \approx 200$, it also motivates defining halos as regions enclosing a mean overdensity 200 times the background matter density, Δ_{200b} (Dodelson and Schmidt, 2020).

To recap, under spherical collapse, we expect an overdense region to have collapsed and formed a halo with a density contrast of $\Delta_{\rm vir} \approx 200$ by the time its linear overdensity reaches $\delta_{\rm coll} = 1.686$. The quantity $\delta_{\rm coll}$, often denoted as δ_c, is therefore useful for predicting whether a small density perturbation δ_i at an initial redshift z_i will have collapsed to form a dark matter

halo at a later redshift z. According to linear theory, the collapse condition is simply

$$D(z)\delta_i > \delta_c \tag{1.58}$$

for $D(z)$ the growth factor in Sec. 1.2 normalised at redshift $z_{\rm norm} = z_i$ (see e.g. Kravtsov and Borgani, 2012).[13] For this reason, δ_c is usually referred to as the *collapse threshold*.

So far we have only considered spherical collapse in an EdS universe. For a universe with lower matter density $\Omega_m < 1$, a fluctuation enclosing mass M would have a lower initial mean density and a larger Lagrangian radius (Eq. 1.45), thus taking longer to collapse. The density contrast of the collapsed object is also expected to be larger, as the mean background density at the time of collapse is lower (Kravtsov and Borgani, 2012). The collapse threshold, however, only has a weak dependence on Ω_m and Ω_Λ; for ΛCDM with parameters similar to those in Sec. 1.1.2, $\delta_c \approx 1.675$ (Kravtsov and Borgani, 2012), so often a fixed $\delta_c = 1.686$ is used in practice. Approximations for $\Delta_{\rm vir}$ in an open universe with $\Omega_\Lambda = 0$, or a flat ΛCDM universe, were found by Bryan and Norman (1998); Mo et al. (2010) also summarise the spherical collapse model in EdS and flat ΛCDM.

1.4 Extended Press-Schechter and the universality of the halo mass function

One of the most important insights we have obtained from considering spherical collapse in Sec. 1.3.2 is that whether a region will collapse to form a halo depends only on the initial mean overdensity within the spherical region in our simplified scenario (given by Eq. 1.58). This led to the *extended Press-Schechter* formalism, also called the *excursion set* formalism (Press and Schechter, 1974; Bond et al., 1991), which provides a theoretical framework that is widely-used for modelling structure formation such as halo abundance, halo bias, and halo growth. We focus on the modelling of the halo mass function, which describes the abundance of halos as a function of halo mass M at a given redshift z. The halo mass function is sensitive to cosmological parameters such as Ω_m and the matter fluctuation amplitude σ_8 especially at

[13] Alternatively, the collapse condition is often written as $\delta_0 > \delta_c/D(z)$ for $D(z)$ normalised at $z_{\rm norm} = 0$, where δ_0 is the overdensity extrapolated to $z = 0$.

the high-halo-mass tail, allowing cluster number count experiments (see Sec. 1.3.1) to be a sensitive cosmological probe. I present the extended Press-Schechter formalism for modelling the halo mass function in Sec. 1.4.1. The Press-Schechter halo mass function agrees with numerical simulations only qualitatively; this is already impressive given its many simplifications. However, its lack of quantitative agreement with simulations spurs further efforts to improve theoretical modelling, including the Sheth-Mo-Tormen equation, commonly referred to as "ellipsoidal collapse" (Sheth and Tormen, 1999), which I summarise in Sec. 1.4.2. Sec. 1.4.3 introduces the concept of universality, which plays a central role in many halo mass function models.

1.4.1 Extended Press-Schechter formalism

In their seminal paper, Press and Schechter (1974) propose a model where the halo mass function at late times is estimated using statistical properties of the initial density fluctuations. As mentioned at the beginning of Sec. 1.1, measurements of the cosmic microwave background confirm that the initial density fluctuations δ (Eq. 1.1) are small and highly Gaussian (Planck Collaboration et al., 2020a). Using spherical collapse, halos with mass M form from spheres enclosing such mass in the initial overdensity field. The number of collapsed regions at redshift z should then be determined by how the amplitude of overdensities within such spheres compares to the collapse threshold δ_c at z.

The typical amplitude of overdensities within spheres enclosing mass M can be found by smoothing the Gaussian random overdensity field with a spherically-symmetric smoothing filter. The variance of the smoothed overdensity field δ_M is then (Kravtsov and Borgani, 2012)

$$\langle \delta_M^2 \rangle = \sigma^2(M,z) = \frac{1}{(2\pi)^3} \int P(k,z) |\tilde{W}(\mathbf{k},R)|^2 \mathrm{d}^3 k \,, \tag{1.59}$$

with $\tilde{W}$ the Fourier transform of a spherically-symmetric smoothing filter, and R is the radius of the sphere such that $M = \frac{4}{3}\pi\rho_b(1+\delta_M)R^3$ (recall $\rho_b = \Omega_m \rho_c$ is the background matter density). The linear matter power spectrum $P(k,z)$ described in Sec. 1.2.2 fully characterises the Gaussian random overdensity field at redshift z. Note that Eq. (1.59) evaluated at $R = 8\,h^{-1}\mathrm{Mpc}$ and $z = 0$ gives the parameter σ_8.

Press and Schechter (1974) postulate that the probability for the smoothed overdensity at mass scale M to be above δ_c is equal to the fraction of volume (i.e. the fraction of mass)

that has collapsed into halos with mass $> M$, assuming that the overdensity grows linearly as $\delta_M(z) = D(z)\delta_M(z_{\rm init})$. The probability for $\delta_M > \delta_c$ is given by (Kravtsov and Borgani, 2012)

$$F(M,z) = \int_{-1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma(M,z)} \exp\left[-\frac{\delta_M^2}{2\sigma^2(M,z)}\right] \Theta(\delta_M - \delta_c) \mathrm{d}\delta_M \,, \tag{1.60}$$

$$= \frac{1}{2} \mathrm{erfc}\left[\frac{\delta_c}{\sqrt{2}\sigma(M,z)}\right] , \tag{1.61}$$

where $\Theta(\delta_M - \delta_c)$ is the Heaviside step function, and erfc is the complementary error function that has a high value for small peak height $\nu(M,z) = \delta_c/\sigma(M,z)$, and a low value for high peak height. The halo abundance is then the fraction of volume collapsing into halos with mass $(M, M + \mathrm{d}\ln M)$ divided by the comoving volume that each collapsing halo occupies in the initial density field M/ρ_b, giving the halo mass function (Kravtsov and Borgani, 2012)

$$\frac{\mathrm{d}n(M)}{\mathrm{d}\ln M} = \frac{\rho_b}{M}\left|\frac{\mathrm{d}F}{\mathrm{d}\ln M}\right| = \frac{\rho_b}{M}\left|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\frac{\partial F}{\partial \ln\sigma}\right| \equiv \frac{\rho_b}{M}\left|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\right| f(\sigma) \,. \tag{1.62}$$

The multiplicity function $f(\sigma)$ determines the shape of the halo mass function. Eq. (1.62) can also be written as a function of peak height $\nu(M,z)$, with the multiplicity function $f(\nu)$.

Assuming spherical collapse has allowed Press and Schechter (1974) to derive an analytical formula for the halo mass function using Eq. (1.61). However, their postulate fails to account for small underdensities that can still collapse into halos if they reside in large overdensities. This "cloud-in-cloud" problem of miscounting the number of small-scale halos results in their formula accounting for only half of the matter in the universe. Press and Schechter (1974) corrected for this by inserting a "fudge factor" of 2, so that the halo mass function has

$$f_{\rm PS}(\sigma) = \sqrt{\frac{2}{\pi}}\frac{\delta_c}{\sigma}\exp\left[-\frac{\delta_c^2}{2\sigma^2}\right] . \tag{1.63}$$

The Press-Schechter halo mass function complete with the factor of 2 is derived fully analytically by Bond et al. (1991), whose approach is known as the excursion set formalism or the extended Press-Schechter (EPS) theory. Unlike the Press-Schechter postulate which only considers the field smoothed at a single mass scale, Bond et al. (1991) consider the field smoothed on a range of scales. They equate the mass fraction in halos of mass $> M$ with the fraction of space in the initial density field that is above the collapse threshold δ_c when

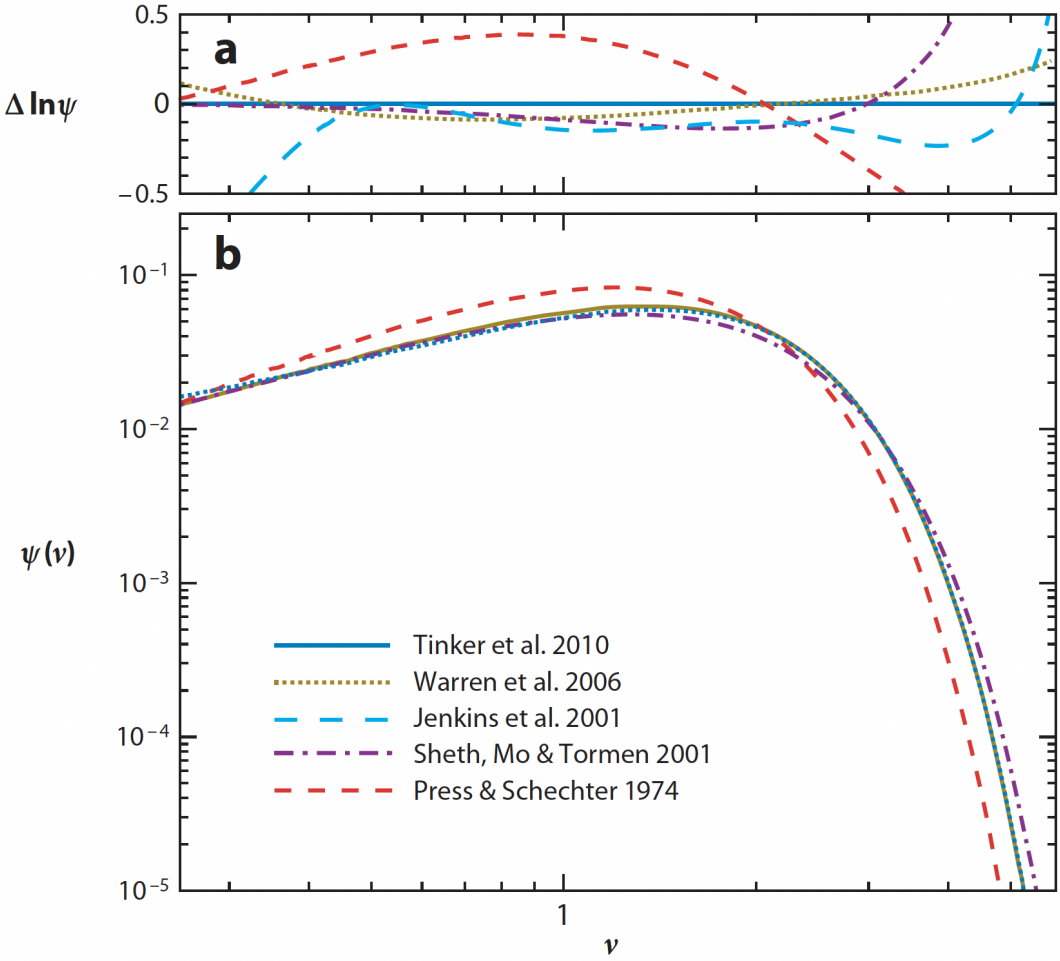


Figure 1.7: Figure 7 from Kravtsov and Borgani (2012) showing the difference between the (extended) Press-Schechter, Sheth-Mo-Tormen, and other simulation-based calibrations of the halo mass function. The figure plots $\psi(\nu) = \left|\frac{\mathrm{d}\ln\nu}{\mathrm{d}\ln M}\right| f(\nu)$ from Eq. (1.62). The upper panel shows deviations of each function from the Tinker mass function (see Sec. 1.6.2.1) at $z = 0$.

smoothed with any filter of scale $R \geq R(M)$. Assuming spherical collapse in an EdS universe for which the collapse threshold is a constant results in Eq. (1.63).

While the extended Press-Schechter formalism predicts a halo mass function that qualitatively agrees with simulations (Efstathiou et al., 1988), as the resolution of N-body simulations increased, Eq. (1.63) is found to be inaccurate and especially underpredicts the number density of massive halos. This can be seen in Fig. 1.7, where the Press-Schechter mass function in red can deviate from functions fitted on numerical simulations by 50%.

1.4.2 Sheth-Mo-Tormen equation

To improve the agreement with simulations, Sheth and Tormen (1999) subsequently modified the Press-Schechter halo mass function. The authors argued that under the excursion set for-

malism, their modified halo mass function must arise from a collapse threshold that is not constant as assumed in Bond et al. (1991) from spherical collapse, but is instead a function of mass: $\delta_c = \delta_c(M)$. Sheth et al. (2001) motivate such a mass-dependent collapse threshold by considering the evolution and collapse of ellipsoidal density perturbations that are described by the initial ellipticity, prolateness, as well as the overdensity δ.

The halo mass function is obtained using the mass-dependent collapse threshold following the excursion set formalism, giving the Sheth-Mo-Tormen (SMT) halo mass function

$$f(\nu) = 2A\left(1+\frac{1}{\nu^{2q}}\right)\left(\frac{\nu^2}{2\pi}\right)^{1/2}\exp\left(-\frac{\nu^2}{2}\right), \qquad \nu = \frac{\delta_c}{\sigma} \tag{1.64}$$

with $q = 0.3$ and $A \approx 0.3222$. Eq. (1.64) is very similar to the function Sheth and Tormen (1999) found by fitting their simulations, except that to agree, ν in Eq. (1.64) needs to be replaced by $\sqrt{a}\nu$ where $a = 0.707$ is a fitted constant.

While initially the SMT halo mass function agreed with numerical simulations, Reed et al. (2003) soon reported its tendency to overpredict high-mass halos. Analytical attempts to derive the halo mass function persist (e.g. Paranjape and Sheth, 2012; Castorina et al., 2016), however there is currently no fully satisfactory first-principles approach that predicts the halo mass function to the percent-level accuracy required by current and future surveys.

For this reason, numerical simulations have become a standard tool for determining the halo mass function; the function f in Eq. (1.62) is fitted to the halo mass distribution found in simulations. These semi-analytical halo mass function models will be introduced in greater detail in Sec. 1.5.2.

1.4.3 Universality of the halo mass function

While the number density of halos depends both on redshift and cosmology, in the extended Press-Schechter formalism, the multiplicity function f in Eq. (1.62) is a sole function of the mass variance $\sigma(M,z)$ (or as sometimes used in literature, e.g. SMT, it is only a function of the peak height $\nu = \delta_c/\sigma$; recall that δ_c depends very weakly on cosmology). All the cosmology and redshift dependence is embodied in the mass variance (or peak height) such that f is not explicitly dependent on cosmological parameters or redshift.

The ability to use a single function f to accurately model the halo mass function for different cosmologies and redshifts is referred to as the *universality* of the halo mass function. This

is an important property underlying the approach of semi-analytical halo mass function models: because of universality, halo mass functions calibrated against simulations ran using one set of cosmological parameters can be used for alternative cosmologies.[14] The semi-analytical halo mass function model can then be used in inference pipelines for constraining cosmological parameters instead of running simulations at each step in the Markov chain Monte Carlo (MCMC). Sec. 1.5 introduces numerical simulations and semi-analytical halo mass functions in greater detail.

Universality was initially confirmed by Jenkins et al. (2001). They found the same fitted $f(\sigma)$ described simulation results from cosmologies with matter density parameter $\Omega_m = 0.3 - 1$ at redshifts $z \leq 5$, but their halo mass function fit was only accurate to $\sim 20\%$. As simulations are conducted at higher resolutions and more cosmologies were tested, studies revealed that universality does not hold to below $\sim 10\%$ (Reed et al., 2003; Reed et al., 2007; Tinker et al., 2008; Courtin et al., 2010; Bhattacharya et al., 2011; Watson et al., 2013; Diemer, 2020; Ondaro-Mallea et al., 2021; Euclid Collaboration et al., 2023a). These findings were again confirmed with simulation suites ran to train state-of-the-art halo mass function emulators (McClintock et al., 2019b; Bocquet et al., 2020, see Sec. 1.6.3.3).

As the calibration of mass-observable relations (see Sec. 1.3.1) is expected to reach uncertainties at a few-percent level (Euclid Collaboration et al., 2023a), this requires the halo mass function to be modelled to 1% accuracy or better to keep uncertainties in theoretical modelling negligible (McClintock et al., 2019b; Wu et al., 2010; Sartoris et al., 2016; Euclid Collaboration et al., 2023a). Not accounting for non-universality (the redshift and cosmology dependence of $f(\sigma)$) could therefore be a significant source of error in the analysis of forthcoming surveys. Sec. 1.6 will introduce in greater detail the current state of understanding and modelling non-universality.

1.5 Numerical Simulations

While we have gained physical intuitions and a qualitative halo mass function through analytical models in the previous section, the simplifications that had to be made result in discrepancies with numerical simulations by up to $\sim 50\%$ (see e.g. a review in Kravtsov and Borgani,

[14] By cosmology I refer to both the cosmological model and cosmological parameter values.

2012). On the other hand, forthcoming surveys require the halo mass function to be modelled to 1% accuracy or better. As announced earlier, such modelling relies on numerical simulations, which are the only way to accurately and robustly model structure formation deep into the non-linear regime, where the overdensity $\delta \gg 1$.

Cosmological simulations follow the evolution of matter from early times, when perturbations are still small and can be described by perturbation theory, until the present day when matter is highly clustered. A common approach is to use N-body simulations, where only gravitational interactions are considered. This is justified since on large scales, the evolution of matter is primarily driven by gravitational interactions. Baryons are subject to hydrodynamical forces and astrophysical processes (the baryonic effects mentioned in Sec. 1.1.3), however they only constitute a small fraction of matter, and baryonic effects only become important on galactic scales. I introduce N-body simulations in Sec. 1.5.1. A summary of semi-analytical halo mass function models fit to numerical simulations is in Sec. 1.5.2, and the rest of this section outlines some of the numerical aspects that affect the halo mass function modelling.

1.5.1 N-body simulations

Simulations of structure formation on large scales are typically done using N-body simulations, or gravity-only simulations. They are comparatively simple, since the more complicated and uncertain baryonic physics do not need to be modelled. A further simplification is that on the scales of interest for structure formation, matter evolution is well-described by Newtonian gravity, avoiding the need for computationally expensive full general relativistic treatments.

Given a cosmological model and initial conditions, N-body simulations follow the evolution of matter particles under gravity through cosmic time (Davis et al., 1985; Efstathiou et al., 1985; Springel et al., 2021). In this context, particles are tracers representing an extended piece of phase space (the 6D space of possible positions and momenta for the matter fluid). In simulations ran to calibrate the halo mass function, particles often have a mass $M_p \sim \mathcal{O}(10^8)\, h^{-1}\mathrm{M}_\odot$ or more. Particles are evolved inside boxes of comoving length L, with periodic boundary conditions so that particles near the edge of the box would not be affected by the lack of matter outside the box. The box size limits the largest scales on which density fluctuations can be represented, and therefore limits the maximum mass of simulated halos. The mass of tracer particles limits the minimum halo mass that can be resolved by the simulation, as well as the

resolution of internal halo structures. Larger volume and lower particle masses increase the accuracy of simulations at the cost of increasing the computational resources required.

Since evolution is well-described by Newtonian gravity, a tracer particle i feels the sum of all forces exerted on it by other particles. This can be calculated as (see e.g. van den Bosch, 2022b)

$$F_i = \sum_{i \neq j} \frac{GM_p^2 \left(\mathbf{r}_j - \mathbf{r}_i\right)}{\left(\left|\mathbf{r}_j - \mathbf{r}_i\right|^2 + \varepsilon^2\right)^{3/2}}, \tag{1.65}$$

where j denotes all other particles such that $r_j - r_i > \varepsilon$, with ε being the *softening scale* or *force resolution*. The softening scale is introduced to modify the force on small scales to avoid strong particle-particle interactions, which would be unphysical as the particles do not actually represent physical particles. Since gravity is modified below the softening scale, simulation results below this scale cannot be trusted.

While direct summation is the most accurate method for force calculation, it scales as $O(N^2)$ with the N number of particles, quickly becoming computationally unfeasible. Instead, algorithms for sufficiently accurate force approximations are used. The simplest is the particle-mesh (PM) algorithm that solves for the gravitational potential on a regular grid via fast Fourier transforms (FFT). The force resolution is determined by the grid spacing, and a finer grid is required for higher resolution. But, because the matter distribution in the late universe is highly non-homogeneous (see Fig. 1.1), most of the evaluation will be at locations of little interest and not where structures form, making a high-resolution PM evaluation inefficient. Another method, the tree algorithm (Barnes and Hut, 1986), hierarchically groups particles together and computes their joint effect on distant other particles if the distance is above some criterion. Hybrid tree-particle-mesh (TreePM) algorithms are also used by e.g. the popular N-body code GADGET (Springel et al., 2001), now at GADGET-4 (Springel et al., 2021)[15]. TreePM takes advantage of both algorithms: for short range force calculation it uses the tree algorithm, which does not have intrinsic resolution limits of meshes, whereas for long range force calculation it uses a mesh to avoid the approximations used by the tree algorithm (see e.g. Angulo and Hahn, 2022 for a more detailed review of algorithms).

[15]GADGET-4 has both TreePM and an optional alternative method called fast multipole method, which will not be discussed here (Springel et al., 2021).

The initial conditions of N-body simulations are prescribed by the cosmological model and a set of cosmological parameters. As the initial conditions are highly Gaussian, they are characterised by the power spectrum $P(k,z)$. The power spectrum is typically obtained by numerically integrating the Einstein-Boltzmann equations at linear order using state-of-the-art codes CAMB (Lewis et al., 2000) and CLASS (Lesgourgues, 2011; Blas et al., 2011) introduced in Sec. 1.2.3.

Given the power spectrum and a grid, the amplitude of the Fourier modes of density perturbations $\delta_{\mathbf{k}}$ can be calculated. Particles initialised on a regular grid are then displaced to match the Fourier amplitudes. Especially for earlier simulations, particles with position $\mathbf{q}$ are often displaced to $\mathbf{x}$ according to first order approximation, the *Zel'dovich approximation* (Zel'Dovich, 1970), as

$$\mathbf{x}(\mathbf{q},z) = \mathbf{q} + D(z)\mathbf{f}(\mathbf{q}), \qquad \mathbf{f}_{\mathbf{k}} = -i\frac{\delta_k}{k^2}\mathbf{k}, \tag{1.66}$$

where $\mathbf{f}(\mathbf{q})$ is the displacement field whose Fourier modes are $\mathbf{f}_{\mathbf{k}}$ for mode δ_k of the density fluctuations, and $D(z)$ is the growth factor in Eq. (1.43). Simulations initialised using the Zel'dovich approximation need to be started at early redshifts (often $z \sim 100-200$) to avoid significant transients caused by the difference between linear input initial conditions and the true non-linear evolution (Crocce et al., 2006). However, this makes the simulation more prone to numerical errors since the small density perturbations at early times are easily overshadowed by numerical noise (Michaux et al., 2020). With increasing accuracy requirements, higher order perturbation theory is sometimes used to initialise simulations at lower starting redshifts. For example, the AEMULUS suite of numerical simulations (DeRose et al., 2019, we will return to them in Sec. 1.6.3.3) uses second-order Lagrangian perturbation theory (2LPT) initial conditions (described in e.g. Angulo and Hahn, 2022). In a recent paper, Michaux et al. (2020) even advocate for using third order Lagrangian perturbation theory at a starting redshift as low as $z = 12$.

Given initial conditions, the N-body simulation integrates equations of motion over time, and outputs *snapshots* which are catalogues of particle positions and velocities at given points in time. These are then post-processed, for example to identify halos.

For modelling the halo mass function, studies of numerical errors and their correction methods (most extensively in Lukić et al., 2007; Bhattacharya et al., 2011) highlight the importance of starting simulations at early redshifts when generating initial conditions with first order perturbation theory. This is required to avoid suppression of the halo mass function at high redshifts (Lukić et al., 2007). They also highlight the importance of accounting for finite volume effects due to the lack of fluctuations on scales larger than the simulation box size, and accounting for run-to-run variations due to finite sampling of density fluctuation modes especially at scales near the simulated box size (Reed et al., 2007). It is also important to require a high minimum number of particles in the halos used to fit the halo mass function:[16] particularly for low-mass halos with a small number of particles (~ 20), the halo mass can be systematically too high because the halo mass resolution is limited by the number of particles (Warren et al., 2006). A larger number of particles is needed if identifying the halo requires resolving its density profile (Tinker et al., 2008). While earlier studies have used a minimum of 20 particles in their halos (Sheth and Tormen, 1999; Jenkins et al., 2001), from Warren et al. (2006) onwards it is common for studies to use a minimum of 400 particles; e.g. Warren et al. (2006); Tinker et al. (2008) use a minimum of 400 particles, and Watson et al. (2013) uses a minimum of 1000 particles.

1.5.2 Semi-analytical models of the halo mass function

Although in the Sheth-Mo-Tormen (SMT) halo mass function (Sec. 1.4.2) the modified collapse threshold is motivated considering ellipsoidal collapse, the SMT halo mass function involves parameters which are calibrated against numerical simulations. This heralds subsequent efforts in the last two decades dedicated to finding accurate models of the halo mass function via calibration against numerical simulations.

Jenkins et al. (2001) initially found the SMT halo mass function in Eq. (1.64) agrees well with their fit produced using a large suite of simulations, a subset of which were used by Sheth and Tormen (1999). However, as mentioned in Sec. 1.4.2, Reed et al. (2003) soon after reported a tendency of the SMT function to overpredict high mass objects. Warren et al. (2006) found that both SMT and Jenkins et al. (2001) are inconsistent with ΛCDM mass functions at $\sim 10\%$ level for intermediate halo masses and at $\sim 30\%$ level for high halo masses (see Fig. 1.7). Since

[16] Particles in these simulations can have masses from $\sim 10^3$ to $\sim 10^{10}\, h^{-1}\mathrm{M}_\odot$.

then, attempts to fit halo mass functions to simulations have worked to obtain accurate fits by increasing the box size and resolution of simulations and improving the modelling of numerical errors.

The halo mass function has been fitted to N-body simulations with box lengths up to $L \sim 3\,h^{-1}$Gpc, e.g. with the Millennium-XXL simulation, a very large ($L = 3\,h^{-1}$Gpc) and high resolution simulation (6720^3 simulation particles of masses $\sim 9 \times 10^9\,h^{-1}\mathrm{M}_\odot$ each, with force resolution of $13.7\,h^{-1}$kpc; Angulo et al., 2012), and the more recent *Euclid* Flagship 2 simulation (Euclid Collaboration et al., 2024). Different studies focussed on different halo mass ranges, from $\sim 10^5\,h^{-1}\mathrm{M}_\odot$ at high redshifts ($z \lesssim 30$) to $\sim 10^{15}\,h^{-1}\mathrm{M}_\odot$ at $z = 0$. As awareness for the need of better statistics and corrections to systematic errors increased (Warren et al., 2006; Lukić et al., 2007; Bhattacharya et al., 2011), studies generally use an increasing number of simulations with repeated realisations of the same cosmology. The precision of the fitting function has increased from $\sim 20\%$ by Jenkins et al. (2001) to a few-percent-level accuracy for the most accurate fitting functions to date: Tinker et al. (2008) to $\sim 5\%$, Bhattacharya et al. (2011) to $\sim 2\%$, Angulo et al. (2012) to $\sim 5\%$, and Euclid Collaboration et al. (2023a) to $\lesssim 1\%$.

However, efforts to determine the halo mass function have been complicated by differences in the approaches used, and the literature has not converged towards a consensus fitting function form or fitting parameters. The studies mainly differ on two respects. The first is whether a universal fitting function is assumed, where a single fitting function parametrised as $f(\sigma)$ or $f(\nu)$ is used to fit simulations of different cosmologies and redshifts. As discussed at the end of Sec. 1.4.3, the assumption of universality was found to not hold below $\sim 10\%$ accuracy (Reed et al., 2003; Reed et al., 2007; Tinker et al., 2008; Courtin et al., 2010; Bhattacharya et al., 2011; Watson et al., 2013; Diemer, 2020), hence the fitted functions also differ depending on the cosmologies and redshift ranges used for calibration. The second difference is the definition of halos used (summarised in Sec. 1.5.3), which changes the halo mass function by $\sim 20\%$ or higher (Tinker et al., 2008). Different halo finding algorithms, summarised in Sec. 1.5.4, are also found to impact the halo mass function (Euclid Collaboration et al., 2023a).

1.5.3 Halo definition

In the highly idealised spherical collapse scenario in Sec. 1.3.2, we considered halo formation from an isolated region, which provides a natural halo boundary. However, in simulations,

halos smoothly blend into the surrounding matter, so there is no clear boundary. This causes a lack of consensus in literature on which halo definition to use.

The two most commonly used classes of definitions are spherical overdensity (SO, Warren et al., 1992; Lacey and Cole, 1994; Bond and Myers, 1996) and friends-of-friends (FOF, Davis et al., 1985; Barnes and Efstathiou, 1987), and depending on the halo definition used, the halo mass function differs significantly (Jenkins et al., 2001).

Spherical overdensity (SO) is motivated by spherical collapse. The mass of a halo is measured by growing a sphere around the centre of the halo until the mean density in the sphere is equal to the overdensity criterion Δ. SO definitions differ mainly on the overdensity criterion chosen (e.g. 200 or 500 times), and whether the overdensity is defined in reference to the critical density of the universe, the mean background matter density, or the virial overdensity $\Delta_{\rm vir}$ (see Eq. 1.57). A common choice of SO mass definition is $\Delta_{200b} = 200\rho_b$, where ρ_b is the mean matter density of the universe (see Sec. 1.3.2). Different overdensity criteria result in different halo mass functions (e.g. fig. 5 in Tinker et al., 2008).

However, halos in simulations are never perfectly spherical, and virialisation at higher masses and redshifts is often incomplete so that objects identified with SO may not closely follow regions of high matter density (e.g. in Fig. 1.8, the SO sphere includes regions with low matter density). For this reason, especially earlier studies often use friends-of-friends (FOF) halo definition, which allows halos to be non-spherical. In FOF, particles belong to the same halo if they lie within the linking distance (b times the mean inter-particle spacing) of one another, with b the linking parameter commonly set to $b = 0.2$.[17] However, while SO masses can be measured in X-ray or optical analyses (see Sec. 1.3.1 and Allen et al., 2011) or via gravitational lensing analyses, the physical interpretation of FOF masses is unclear. Furthermore, FOF suffers from the problem of overlinking, where two visually distinct halos are bridged together and identified as one halo by the FOF algorithm. Lukić et al. (2007) show that approximately $15-20\%$ of all FOF halos with $b = 0.2$ can be bridged.

Tinker et al. (2008) show that there is a large scatter between SO and FOF mass (SO mass can range between 50% to 110% of FOF mass), and the median of the ratio of SO to FOF

[17] This is because in an Einstein de-Sitter universe, if the halo is a perfectly spherical and singular isothermal sphere, then the matter enclosed when $b = 0.2$ has roughly $\Delta = \Delta_{\rm vir}$ (White, 2002).

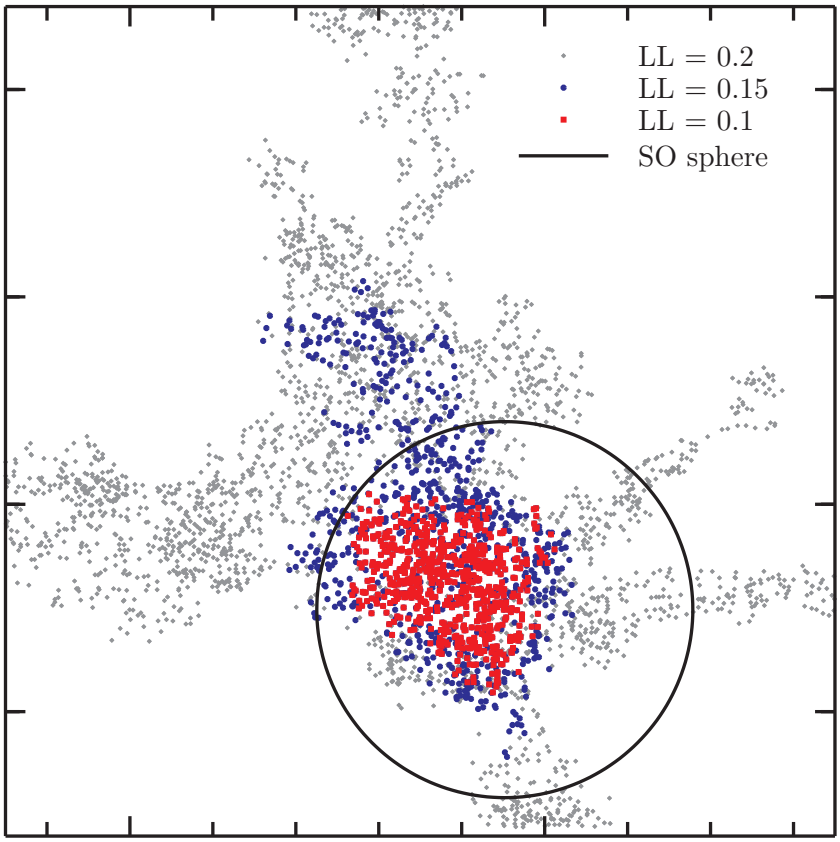


Figure 1.8: Halos identified with different halo definitions show large discrepancies. The SO sphere is identified with $\Delta = 178\rho_b$, while the grey, blue and red show FOF halos identified with a linking length ("LL") of 0.2, 0.15 and 0.1. Figure taken from figure 15 in Watson et al. (2013).

masses decreases with redshift but scatter increases. The clear difference between SO and FOF halos, and FOF halos defined by different linking parameters, is illlustrated in Fig. 1.8.

Since neither SO nor FOF halos are entirely satisfactory, some recent literature has started investigating the use of a physically motivated splashback radius to define halos. This is located at the first apocentre radius of the most recently infalling matter, with the idea that it should separate infalling from orbiting material and hence define a physical halo boundary (Diemer and Kravtsov, 2014; Adhikari et al., 2014; More et al., 2015; Diemer, 2020). However, this is currently difficult to calculate in practice, and calculation can fail especially if there is insufficient particle splashback events (Diemer, 2020). Hence, SO and FOF remain the most widely used halo definitions to date.

1.5.4 Halo finders

Because of their clearer connections to observables, currently spherical overdensity (SO) halo mass functions are more commonly used in surveys and analyses (for example, the Tinker halo mass function, Tinker et al., 2008, described later in Sec. 1.6.2.1 is widely used, e.g. in Bocquet et al., 2019; Costanzi et al., 2021; Bocquet et al., 2024; Ghirardini et al., 2024).

SO halo identification is based on identifying a halo centre and growing a sphere around it, but different SO halo finders vary in how the centre is identified. For example, a widely used halo finder ROCKSTAR (Behroozi et al., 2013) makes use of the FOF algorithm to identify the halo centre. It first runs a 3D FOF finder on a simulation volume to identify overdensities. Then, for each overdensity, it builds a hierarchy of FOF subgroups in phase space by repeatedly running a 6D FOF finder on the parent group, adaptively choosing the linking length so that 70% of the particles are linked together in a subgroup. The halo centre is then computed by averaging the positions of the particles in the lowest-level subgroup, i.e. the lowest-level substructure identified. Once the sphere is grown around the halo centre, there is also a choice of whether to keep all particles within the sphere, or only the particles that are gravitationally bound to the halo.

Recently, Euclid Collaboration et al. (2023a) studied the effect of using different SO halo finders on the halo mass function, for halos defined using $\Delta_{\rm vir}$. They compared a variety of algorithms, and find that algorithms utilising 3D FOF or non-adaptive 6D FOF in their halo centre definitions tend to find fewer halos with mass $< 10^{14}\,h^{-1}\mathrm{M}_{\odot}$ compared to algorithms such as ROCKSTAR (which will be used this thesis; Behroozi et al., 2013). This appears to be associated with the known tendency of pure FOF methods to bridge smaller structures, which also impacts the halo centre location. The choice of only including gravitationally bound particles in the halo mass calculation, as opposed to using all particles within the sphere, also leads to a $\sim 5\%$ decrease in the number of halos (Euclid Collaboration et al., 2023a). For an *Euclid*-like survey whose mass-observable relation may be calibrated to $< 3\%$ precision, Euclid Collaboration et al. (2023a) conclude that significant bias can come from using a different halo mass definition and halo finder when calibrating the mass-observable relation compared to when modelling the halo mass function.

A source of uncertainty in halo mass function calibration is baryons. Calibrating the halo mass function for cluster cosmology requires simulations with large volumes, while resolving

SO halos accurately puts a lower bound on the mass resolution. Therefore, calibrating the halo mass function on hydrodynamical simulations accounting for baryonic effects is very expensive, and an overwhelming majority of halo mass functions are calibrated on N-body simulations. Bocquet et al. (2016) used a suite of hydrodynamical simulations with dark-matter-only counterparts to study the effect of baryons on the halo mass function. The Magneticum simulations they used incorporate radiative cooling, and feedback associated with supernova-driven winds and active galactic nuclei. They find that baryons tend to decrease the mass of clusters, leading to a suppression of the halo mass function. This can have a $\lesssim 15\%$ effect on the halo mass function for $10^{12}\,h^{-1}\mathrm{M}_\odot \lesssim M_{200b} < 10^{14}\,h^{-1}\mathrm{M}_\odot$ halos defined using Δ_{200b}, where the lower bound in mass is due to their mass resolution. They find baryonic effects above this mass scale is negligible, confirmed again by Schaye et al. (2023). Instead of fitting hydrodynamical simulations, an alternative way of accounting for baryonic effects is to apply a baryonic correction model to a halo mass function fit on N-body simulations, as in e.g. Euclid Collaboration et al. (2023b). Nevertheless, efforts of including baryons in the halo mass function modelling are fundamentally limited by baryonic physics still being poorly constrained.

1.6 Non-universal halo mass functions

In Sec. 1.4 we saw that in the extended Press-Schechter formalism, the abundance of halos depends only on the initial density field. All cosmology and redshift dependence of the halo mass function is captured by the mass variance $\sigma(M,z)$ (or the peak height $\nu = \delta_c/\sigma$, as in the Sheth-Mo-Tormen equation Eq. 1.64). The multiplicity function f in Eq. (1.62) therefore remains unchanged, i.e. it is universal. Although the Press-Schechter halo mass function turned out to be inaccurate, universality enables f calibrated against a single set of numerical simulations to be used across a cosmological parameter space. However, we saw at the end of Sec. 1.4.3 that universality was found to not hold below $\sim 10\%$. Not accounting for non-universality (the redshift and cosmology dependence of f) could therefore be a significant source of error in the analysis of forthcoming surveys, which require the halo mass function to be modelled to 1% accuracy or better (McClintock et al., 2019b; Wu et al., 2010; Sartoris et al., 2016; Euclid Collaboration et al., 2023a).

While N-body simulations give accurate halo mass function predictions, it becomes computationally prohibitive to run simulations for the large number of cosmological parameter samples that would be needed for parameter inference. Hence, it is important to find a way to model the halo mass function's cosmology and redshift dependence accurately. At the same time, the presence of non-universality also indicates there is physics in the formation and evolution of halos that the extended Press-Schechter formalism has not accounted for. Understanding the source of non-universality and how to model it is therefore both practically and theoretically interesting.

Sec. 1.6.1 summarises the main results of explorations aimed at understanding which quantities affect non-universality. Sec. 1.6.2 introduces the state of the art on accounting for non-universality in semi-analytical halo mass function models. Sec. 1.6.3 introduces emulators, which are essentially flexible fitting functions trained on simulations, that produce percent-level-accurate halo mass functions over a cosmological parameter space.

1.6.1 Explorations of non-universality

Studies that investigate and quantify the degree of non-universality generally take two approaches. The first investigates how the degree of universality depends on the halo definition, and searches for halo definitions that lead to more universal halo mass functions; this is summarised in Sec. 1.6.1.1. The second studies which quantities the halo mass function depends on apart from the mass variance σ (or peak height $\nu = \delta_c/\sigma$) using specifically designed simulations, summarised in Sec. 1.6.1.2.

1.6.1.1 Dependence on halo definition

Literature on the degree of non-universality given a halo definition is quite heterogeneous: different studies explore different cosmologies and redshift ranges, and the conclusions drawn on whether the halo mass function is universal depend on the accuracy criterion used in the study. These make it difficult to obtain a coherent overview of the results; the details of a given study must be taken into account.

White (2002) already found that the degree of universality depends on halo definition. The halo mass function at $z = 0$ was found to be universal at $\sim 20\%$ level (i.e. the fractional change in f is $\lesssim 20\%$) between cosmologies with $\Omega_m = 0.3$ and $\Omega_m = 1$ if halos are identified using friends-of-friends (FOF) with a fixed linking parameter, or using spherical overdensity

(SO) with an overdensity criterion of $\Delta = 180\rho_b$ (see Sec. 1.5.3); these findings agree with Jenkins et al. (2001). However, for SO halos with a higher overdensity threshold ($\Delta = 1000\rho_b$) or overdensities defined with respect to the critical density ρ_c given in Eq. (1.19), White (2002) finds the halo mass function shows non-universality $> 20\%$.

Further studies find that neither the FOF nor SO halo mass functions are universal at the percent-level. For ΛCDM cosmologies with $\Omega_m \sim 0.25$, the FOF halo mass function is found to evolve by $\sim 10\%$ within $z = 0-2$ (Tinker et al., 2008; Crocce et al., 2010; Bhattacharya et al., 2011; Watson et al., 2013). For the same redshift range and cosmology, the SO halo mass function defined with $\Delta = 200\rho_b$ evolve by $\sim 20\%$ (Tinker et al., 2008). FOF halos are generally found to be more universal than SO definitions, but there is no convincing explanation for why this is the case, or if such universality is an artefact of the halo definition and bears no physical significance (see e.g. Tinker et al., 2008; Watson et al., 2013).

Studies also confirm that the non-universality of SO halos depends on the overdensity definition. Despali et al. (2016); Diemer (2020) confirmed the overall decrease in the amplitude of f with redshift that Tinker et al. (2008) observed for halos defined with respect to the background matter density (including Δ_{200b}). Despali et al. (2016) further showed that this decrease with redshift is not observed if halos are defined using the virial overdensity $\Delta_{\rm vir}$; instead, the amplitude remains approximately unchanged at lower halo masses but increases at the high-mass end. They claimed that defining halos using the virial overdensity accounts for the redshift and cosmology dependence of the halo virialisation process, so that the halo mass function becomes universal for ΛCDM cosmologies with $\Omega_m = 0.2-0.4$ and $z \leq 5$. The universality they found is at $\sim 10-20\%$ fractional residuals.[18] Diemer (2020) also finds for $\Delta_{\rm vir}$ an approximately $\sim 20\%$ change in the multiplicity function between $z = 0$ and $z = 4$ for ΛCDM cosmology, which the author viewed as a sign of non-universality.

Recently, Diemer (2020) compared the extent of non-universality for SO halo mass functions defined using $\Delta_{\rm vir}$, Δ_{200b}, and 200 and 500 times the critical density of the universe (Δ_{200c} and Δ_{500c}). The study considered ΛCDM cosmologies and Einstein-de Sitter (EdS) cosmologies for the redshift range $z = 0-6$. The results confirmed that the halo mass function is non-universal for all SO definitions; for most SO definitions except Δ_{200b}, the multiplicity function

[18] Despali et al., 2016 quote logarithmic residuals within $5-8\%$, which I convert to fractional residuals to facilitate comparison with other studies.

increases with redshift especially for high masses, and the increase is particularly strong between $z = 0.5$ and $z = 0$. Diemer (2020) also studied universality for halos defined using the splashback radius, and found that while the halo mass function is more universal for splashback halos than SO halos, there is still significant non-universality up to $\sim 40\%$ when comparing between ΛCDM and Einstein-de Sitter (EdS) cosmologies; for comparison, between the same two cosmologies, the $\Delta_{\rm vir}$ definition leads to over 100% changes in f. Interestingly, halo definitions probing only the inner regions of haloes (e.g. Δ_{500c}) tend to exhibit more non-universality than halo definitions that probe out to the halo outskirts (splashback mass).

Despite the literature being confusing to navigate, the results are broadly consistent. For our purpose, it suffices to draw two broad conclusions: the degree of non-universality depends on the halo definition used, and there appears to be no halo definition that gives a halo mass function which is universal to the percent-level required by forthcoming surveys.

1.6.1.2 Dependence on quantities beyond mass variance

Because no commonly-used halo definition gives a universal halo mass function at the $\leq 1\%$ level required by forthcoming surveys, another approach is to take a given halo definition, and explore what quantities the halo mass function depends on apart from σ or ν.

Courtin et al. (2010) investigated the prospect that non-universality is related to different growth histories characterised by different linear growth functions $D(z)$ (see Eq. 1.43), via running simulations of cosmologies with different dark energy models. To isolate the effect of growth history, they used a ΛCDM power spectrum to initialise their simulations such that the present-day linear matter power spectrum and hence $\sigma(R)$ are identical across simulations. Courtin et al. (2010) found the present day halo multiplicity function $f(\sigma)$ (for FOF halos) shows clear non-universality between different cosmologies, so growth history likely affects the halo mass function. They also hypothesise that $f(\sigma)$ at $z = 0$ shows greater difference between two cosmologies with greater deviation between their growth histories, though they suggest no method for quantifying deviation between growth histories. Courtin et al. (2010) also suggest that accounting for the redshift and cosmology dependence of the linear collapse threshold δ_c (see Sec. 1.3.2) reduces non-universality for high-mass halos: the fractional difference in f between an EdS cosmology and a quintessence dark energy cosmology decreases from $\sim 50\%$ to $\sim 7\%$ for high-mass halos if δ_c is taken into account.

Bhattacharya et al. (2011) also found a relation between growth history and non-universality. They found that FOF halo mass functions fitted to 2% accuracy on ΛCDM cosmologies are only accurate to 10% for wCDM, the simplest extension of ΛCDM in which dark energy has a constant equation of state parameter $w \neq -1$ (see Sec. 1.1.2.4). They give no hypotheses to explain this non-universality. They also found that accounting for the cosmology dependence of the collapse threshold δ_c reduces non-universality from $\sim 10\%$ to $\sim 4\%$, but it does not fully account for the degree of non-universality with different cosmologies.

Recently, Ondaro-Mallea et al. (2021) and Euclid Collaboration et al. (2023a) also ran simulations to investigate the individual effects that growth history and the linear matter power spectrum shape have on non-universality. For example, Ondaro-Mallea et al. (2021) initialised their simulations using $P(k)$ calculated with a fixed $\Omega_m = 0.307$, while the growth history evolves according to $\Omega_m = [0.148, 1]$. They also ran simulations with the same growth history using $\Omega_m = 0.307$, while the spectral index n_s is varied from 0.75 to 1.25 to change the shape of the power spectrum. They found that both growth history and the shape of the linear matter power spectrum affect non-universality of $f(\sigma)$ (they assume a constant δ_c, and consider FOF and SO halos with Δ_{200b}, Δ_{200c} and $\Delta_{\rm vir}$). The degree of non-universality and the trend depend on the halo definition used, in agreement with studies summarised in Sec. 1.6.1.1. They also remark that changing growth history appears to have a stronger effect on non-universality than the power spectrum shape. Euclid Collaboration et al. (2023a) also find that non-universality depends on both the shape of the linear matter power spectrum and the background evolution, even if the cosmology dependence of δ_c is taken into account. They consider SO halos defined with $\Delta_{\rm vir}$, and find that in an Einstein-de Sitter (EdS) universe, varying n_s changes fitting parameters of $f(\nu)$. For simulations with the same power spectra, the fitting parameters of $f(\nu)$ change between an EdS universe and a ΛCDM universe with *Planck* 2018 parameters (Planck Collaboration et al., 2020c).

1.6.2 Semi-analytical models of non-universal halo mass functions

To improve the accuracy of halo mass function modelling across different redshifts and/or cosmologies, semi-analytical models often use a Press-Schechter-like (Eq. 1.63) or Sheth-Mo-Tormen-like (Eq. 1.64) halo mass function and include additional parameters. Sec. 1.6.2.1 introduces the widely used Tinker halo mass function which includes explicit redshift depen-

dence in its fitting parameters. Other semi-analytical models accounting for redshift evolution are outlined in Sec. 1.6.2.2. The state-of-the-art semi-analytical halo mass functions accounting for both redshift and cosmology dependence are summarised in Sec. 1.6.2.3.

1.6.2.1 Tinker halo mass function

One of the most accurate fitting functions (accurate to within 5% for the range it is calibrated on) was produced by Tinker et al. (2008). It is commonly used in literature both as a reference for other halo mass function models and in analyses (e.g. Watson et al., 2013; Bocquet et al., 2016; Ondaro-Mallea et al., 2021; Euclid Collaboration et al., 2023a; Costanzi et al., 2021; Bocquet et al., 2024; Ghirardini et al., 2024). It is the first accurate fitting function for SO halos, calibrated for a $\sigma(M)$ range corresponding to $10^{10.5}h^{-1}M_\odot \leq M \leq 10^{15.5}h^{-1}M_\odot$ at $z = 0$,[19] over redshifts $z \in [0, 2.5]$. This makes it useful for cluster cosmology. The halo mass function was fit to ΛCDM cosmologies with $\Omega_m = 0.23 - 0.3$ using simulations with box lengths $L = 80 - 1280\,h^{-1}\,\mathrm{Mpc}$ and particle masses from $6.96 \times 10^7 - 5.99 \times 10^{11}\,h^{-1}\mathrm{M}_\odot$, ran with different N-body algorithms to avoid possible bias from using a particular algorithm.

Halos used for fitting the halo mass function are found by estimating the local density around each particle and growing a sphere around the highest density particle. The fitting function is separately fitted for a range of overdensity criteria from $\Delta = 200\rho_b$ up to $\Delta = 3200\rho_b$. To avoid systematic errors from insufficient resolution, only halos with a minimum of 400 particles were considered for the calibration of the $\Delta = 200\rho_b$ mass function; even higher thresholds were used for higher overdensity criteria.

The halo mass function is found by fitting the equation

$$f_\mathrm{T}(\sigma, z) = A\left(\left(\frac{b}{\sigma}\right)^a + 1\right)\exp\left[-\frac{c}{\sigma^2}\right] \tag{1.67}$$

after binning the halos in bins of $\log_{10}(M/h^{-1}\mathrm{M}_\odot) = 0.1$, using only data points with $< 25\%$ error after accounting for both cosmic variance and Poisson errors.

As mentioned in Sec. 1.6.1.1, Tinker et al. (2008) observe a $\sim 20\%$ decrease in $f(\sigma)$ for Δ_{200b} halos at $z = 1.25$ compared to at $z = 0$, as well as evolution in the shape of $f(\sigma)$. They suggest that the evolution may be related to the decrease in $\Omega_m(z)$ as the universe transitioned from matter to dark energy domination. However, because their simulations only covered a

[19] The calibrated mass range is $\log_{10}\sigma^{-1} \in [-0.6, 0.4]$.

narrow range of Ω_m, they did not test this hypothesis further. Instead, they turned the fitting parameters in Eq. (1.67) into explicit functions of redshift. For the overdensity criterion of Δ_{200b}, they find $A = 0.186(1+z)^{-0.14}$, $a = 1.47(1+z)^{-0.06}$, $b = 2.57(1+z)^{-\alpha}$, and $c = 1.19$, where $\alpha = \exp\left(-\left(\frac{0.75}{\ln(\Delta_{200b}/75)}\right)^{1.2}\right)$. The authors do note that residuals for the snapshot with the highest redshift $z = 2.5$ appear larger than at lower redshifts, and the numerical results from simulations at $z = 2.5$ are comparatively noisy and cover a small mass range. Hence, the fitting function at $z \geq 2.5$ requires further checks. Explicitly incorporating redshift dependence into the fitting function was also done by Bhattacharya et al. (2011).

Tinker et al. (2008) also produce an alternative fitting function such that $\int \tilde{f}(\sigma) \mathrm{d}\ln\sigma^{-1} = 1$ at $z = 0$, i.e. all mass lies in dark matter halos. This avoids the infinite mass density implied by integrating Eq. (1.67), and the alternative fit agrees well with their main fit. The alternative fit is used for halo bias modelling in Tinker et al. (2010).

Subsequent fits by Watson et al. (2013); Bocquet et al. (2016) for similar cosmologies agree well with the Tinker mass function for Δ_{200b} over similar redshift and mass ranges as that probed by Tinker et al. (2008).

1.6.2.2 Other models accounting for redshift evolution

To fit the redshift evolution of the halo mass function in their simulations, Watson et al. (2013) included the matter density parameter $\Omega_m(z)$ in their SO halo mass function's fitting parameters (using Δ_{178b}). The fitting function they proposed remains accurate to $< 20\%$ up to a high redshift $z \sim 20$ for their simulation of a single ΛCDM cosmology with $\Omega_m = 0.27$. Watson et al. (2013) did not test how well the model fits other cosmologies.

Rather than directly including redshift or cosmological parameters, Reed et al. (2007) used the effective local spectral index n_{eff}, defined as the effective spectral slope at the scale of the halo:

$$n_{\mathrm{eff}} = -3 - 2\frac{\mathrm{d}\log\sigma(R)}{\mathrm{d}\log R}\bigg|_{R_L}, \tag{1.68}$$

where R_L is the Lagrangian radius of the halo in Eq. (1.45), to improve their halo mass function fit in the redshift range $z = 10 - 30$ they examined for a single ΛCDM cosmology with $\Omega_m = 0.25$. It accounted for the suppression of the halo mass function at higher redshifts, seen in their simulations especially for the high-mass halos. The statistical significance of this observed suppression is, however, unclear, since Lukić et al. (2007) did not observe a statistically

significant suppression at $z \leq 20$ by using an alternative method of correcting for finite volume effects.

1.6.2.3 Models of both cosmology and redshift dependence

As described in Sec. 1.6.1.2, Ondaro-Mallea et al. (2021) and Euclid Collaboration et al. (2023a) found through running simulations varying only growth history, or varying only the power spectrum shape, that both the growth history and the shape of the power spectrum affect non-universality.

Ondaro-Mallea et al. (2021) parametrise the information needed from the growth function and the linear matter power spectrum using two variables in addition to peak height $\nu = \delta_c/\sigma$ (essentially in addition to σ, since the authors assume constant $\delta_c = 1.686$). To parametrise the effect of growth history, Ondaro-Mallea et al. (2021) propose the recent growth rate parameter

$$\alpha_{\rm eff}(a) \equiv \left.\frac{{\rm d}\log(D)}{{\rm d}\log a}\right|_{a=a_{\rm ev}}, \tag{1.69}$$

where a is the scale factor. They adopted a trial and error approach to find that evaluating the growth rate at an earlier scale factor $a_{\rm ev}$, when the growth factor $D(a_{\rm ev}) = \frac{4}{5}D(a)$, allows them to better fit an additional suite of test simulations varying w_0 (see Eq. 1.24 in Sec. 1.1.2.4). They parametrise additional dependence on the power spectrum using the effective spectral slope $n_{\rm eff}$ given in Eq. (1.68).

The final fitting function they proposed, $f(\nu, \alpha_{\rm eff}, n_{\rm eff})$, has a form where a Sheth-Mo-Tormen-like $f(\nu)$ (Eq. 1.64) is multiplied by a linear function of $\alpha_{\rm eff}$ and a quadratic function of $n_{\rm eff}$. The fitting function achieves residuals of $\lesssim 10\%$ over redshifts $z = 0 - 1$ on their nine calibration simulations, ran by combining three different linear power spectra with three different growth histories. Compared to a fit using only $f(\nu)$, the residuals decreased by $\gtrsim 10\%$. As many of their calibration simulations have growth histories evolved using a different value of Ω_m to the power spectrum used to initialise the same simulation, they assess their fit on an additional suite of test simulations with $w_0 - w_a$ cosmology and massive neutrinos, initialised and evolved consistently. While the improvement over a universal fitting function is not evident when varying other cosmological parameters, Ondaro-Mallea et al. (2021) find the scatter in residuals decreases when using their fitting function compared to a universal fitting

function when varying w_0 while holding other parameters fixed. This shows their additional fitting parameter $\alpha_{\rm eff}$ captures information related to non-universality.

Unlike Ondaro-Mallea et al. (2021), in the more recent halo mass function calibration by Euclid Collaboration et al. (2023a), the authors chose instead to make the fitting parameters dependent on power spectrum shape and growth history. They parametrise growth history using $\Omega_m(z)$, a good proxy for the linear growth rate (see Eq. 1.42), and evaluate it at the redshift of interest rather than at an earlier redshift as was done by Ondaro-Mallea et al. (2021). This may be because Euclid Collaboration et al. (2023a) only calibrated the halo mass function on ΛCDM cosmologies, whereas Ondaro-Mallea et al. (2021) adjusted their effective growth rate parameter on wCDM cosmologies. To parametrise the power spectrum, Euclid Collaboration et al. (2023a) uses $\mathrm{d}\ln\sigma/\mathrm{d}\ln R$.[20] Their fitted halo mass function describes a suite of ΛCDM cosmologies with $\Omega_m \simeq 0.17 - 0.37$ to percent-level accuracy, but it is yet to be tested on cosmologies beyond ΛCDM that a *Euclid*-like survey aims to probe.

Interestingly, Euclid Collaboration et al. (2023a) compare their fitting function to Ondaro-Mallea et al. (2021) on a simulation ran with *Planck*-like cosmological parameters, and find that the Ondaro-Mallea et al. (2021) fitting function underpredicts the abundance of lower mass halos by $> 5\%$ while overpredicting the abundance of high-mass halos by $\sim 10\%$ at $M_{\rm vir} = 10^{15}\,h^{-1}\mathrm{M}_\odot$. While the difference at lower halo masses can be attributed to Ondaro-Mallea et al. (2021) using a non-adaptive FOF halo finder (see Sec. 1.5.4), the cause for the deviation at high masses is unclear.

In summary, recent semi-analytical models accounting for non-universality have incorporated growth history and power spectrum shape into the fitting function. While this achieved some success, the literature is yet to reach a consensus on what information is most relevant from growth history, and how to parametrise growth history or power spectrum shape information. The two formulae proposed by Ondaro-Mallea et al. (2021) and Euclid Collaboration et al. (2023a) also do not agree to percent-level, and there is yet to be further literature that can confirm the accuracies of either formula.

[20] Euclid Collaboration et al. (2023a) do not specify the radius at which $\mathrm{d}\ln\sigma/\mathrm{d}\ln R$ is evaluated, though it is reasonable to assume that it is evaluated at the Lagrangian radius of the halo given in Eq. 1.45.

1.6.3 Emulating the halo mass function

Given the difficulties in finding semi-analytical halo mass functions that are percent-level accurate over a wide cosmological parameter space, another approach is to use halo mass function emulators, which bypasses the challenge of explicitly modelling non-universality. An emulator is a surrogate model that learns, from a small number of numerical simulations forming the *training set*, how to map from cosmological parameters to the corresponding halo mass function. After the emulator is trained, it predicts the halo mass function for any point within a target cosmological parameter space without having to run additional numerical simulations to make the prediction.

Emulators learn to produce their outputs by interpolating between the halo mass functions from the training simulations, which should cover the target cosmological parameter space well. Usually, the target cosmological parameter space is high-dimensional to encapsulate physics beyond ΛCDM, so running a grid of simulations becomes highly computationally expensive. Instead, the cosmological parameter space is sampled via space-filling designs, e.g. via Latin hypercube sampling (DeRose et al., 2019; Nishimichi et al., 2019), which divides each dimension of the parameter space into M even intervals, and places M samples in the parameter space such that only one sample falls in each interval when projected into any one dimension. Existing emulators have $\sim 40-100$ parameter samples uniformly distributed in a prior cosmological parameter space (Nishimichi et al., 2019; McClintock et al., 2019b; Bocquet et al., 2020). N-body simulations for each sample are run and halo mass functions fitted. Interpolation within the prior cosmological parameter space is done by training Gaussian process (GP) emulators on these $40-100$ simulations that form the training set. I give an overview of Gaussian process emulators in Sec. 1.6.3.1, and describe how emulator training sets are designed in Sec. 1.6.3.2. Sec. 1.6.3.3 introduces the state-of-the-art halo mass function emulators. These are often trained on cosmologies varying the dark energy equation of state parameter and neutrino-related parameters, and reach high accuracy within their domain of validity. Outside the parameter space they are trained on, however, emulators are not reliable.

1.6.3.1 Gaussian process emulators

A Gaussian process (GP) emulator interpolates between training samples to predict the mean and variance of the output, using an assumed form for the correlation across the target parameter

space. It can model complex trends in data and naturally provides uncertainty estimates through the predicted variance.

To emulate a function, a GP emulator interpolates a given set of function values f_i evaluated at points x_i in a parameter space. It assumes the function values are drawn from a multivariate Gaussian distribution $\mathcal{N}(0, K(x_i, x_j))$, where K models the covariance between two points x_i, x_j in the parameter space. The mean of f can always be subtracted to meet the zero mean condition. The function value $f_\star$ at a test point $x_\star$ is then predicted from the posterior predictive distribution,

$$f_\star \sim \mathcal{N}(K_\star K^{-1} f_i, K_{\star\star} - K_\star K^{-1} K_\star^T) \tag{1.70}$$

where $K_\star = K(x_\star, x_i)$ and $K_{\star\star} = K(x_\star, x_\star)$ (Rasmussen, 2003).

Eq. (1.70) shows that constructing a GP emulator requires two key ingredients: the covariance function $K(x_i, x_j)$ which models the correlation between function values, and the training data. These two ingredients jointly determine the accuracy of the emulator prediction.

The covariance function describes the correlation between function values at two points in the parameter space. Its functional form is generally chosen given some prior knowledge on the data structure. A commonly used covariance function, also for halo mass function emulation (McClintock et al., 2019b), is the radial basis function (RBF) kernel

$$K\left(x_i, x_j\right) = \sigma_0^2 \exp\left(-\frac{\left|x_i - x_j\right|^2}{2l^2}\right), \tag{1.71}$$

which assumes the function value at point x_j is locally informed by neighbouring points x_i. The smoothness of the emulated function is controlled by the hyperparameter l, while the hyperparameter σ_0^2 gives the variance of the function value. Training the GP emulator amounts to tuning the covariance function hyperparameters to model the covariance of the data, by maximising the likelihood of the training data given the hyperparameters.

The training data consist of points x_i sampled from the (N-dimensional) parameter space, and the corresponding function values $f_i = f(x_i)$. For example, existing halo mass function emulators have as x_i vectors of cosmological parameters sampled from a prior volume. Running N-body simulations using the sampled cosmological parameter values gives the function values

$f(x_i)$, which are e.g. the halo mass function fitting function parameters (McClintock et al., 2019b; Nishimichi et al., 2019).

1.6.3.2 Design of emulator training sets

The performance of a GP emulator critically depends on the design of the training dataset. Generally, the GP emulator error scales inversely with the distance to the nearest training points in the parameter space. Hence, the emulator uncertainty is minimised if: a) there are more training points, and b) if training points cover the region of interest well. The training sets of existing halo mass function emulators are designed to be space-filling, i.e. they are evenly spaced throughout the prior volume. This is achieved by e.g. Latin-hypercube sampling the prior volume, and adjusting the samples such that the distance of any sample to its closest neighbour is maximised. The emulator then has approximately uniform uncertainty throughout the cosmological parameter space (Heitmann et al., 2016; DeRose et al., 2019; McClintock et al., 2019b; Nishimichi et al., 2019; Bocquet et al., 2020).

Such design is not specifically optimised for accurately emulating the halo mass function (Nishimichi et al., 2019). Further improving the emulator accuracy requires additional training samples, whose locations can be optimised so the emulator accuracy improves using a small number of additional samples.

An example of emulator training set optimisation is Bayesian optimisation, explained as follows. For applications such as cosmological parameter inference, only regions of high posterior probability require high emulator accuracy. Rogers et al. (2019) and Rogers and Peiris (2021) showed that accurate emulators (for Lyman-alpha forest flux power spectrum) can be built using training sets starting with a good space-filling design, such as a Latin hypercube, and then adding training points where both posterior probability (computed using a specified likelihood function) and emulator uncertainty are high. Such Bayesian optimisation of the training set often achieves tighter posterior constraints using emulators trained on fewer simulations than if the training set was only designed using a Latin hypercube, as emulator uncertainty is reduced in the region of interest. This approach however requires specifying a likelihood for a given dataset, which could impact the generalizability of the emulator to other datasets.

1.6.3.3 State-of-the-art halo mass function emulators

State-of-the-art halo mass function emulators achieve high precision over their target cosmological parameter space. The most commonly referenced ones are the AEMULUS emulator

(McClintock et al., 2019b), the DARK EMULATOR (Nishimichi et al., 2019), and the Mira-Titan emulator (Bocquet et al., 2020).

The AEMULUS emulator (McClintock et al., 2019b) can achieve to within 1% precision over its seven-dimensional cosmological parameter space varying $\Omega_b h^2$, $\Omega_c h^2$, w_0, n_s, $\ln 10^{10} A_s$, H_0, and $N_{\rm eff}$ (McClintock et al., 2019b).[21] Its cosmological parameter space spans approximately the $\pm 4\sigma$ region allowed by CMB+BAO+SNIa (Anderson et al., 2014). Training samples were chosen by projecting a 7D Latin hypercube into the cosmological parameter space such that the resulting samples follow degeneracies among cosmological parameters. Each of the 40 training simulations were run with a version of the GADGET-2 N-body solver (Springel, 2005, see Sec. 1.5.1), with box size $(1.05\,h^{-1}\,{\rm Gpc})^3$ and 1400^3 particles (DeRose et al., 2019). Halos are identified using ROCKSTAR (Behroozi et al., 2013, see Sec. 1.5.4) using an overdensity threshold Δ_{200b} on snapshots ranging from $z = 0$ to $z = 3$.[22] The binned halo data at $z = 0$ reliably cover the mass range $\log(M/h^{-1}{\rm M}_\odot) = [13.2, 15.0]$, where I always use $\log(x) = \log_{10}(x)$. Below $\log(M/h^{-1}{\rm M}_\odot) = 13.2$, the simulations that AEMULUS was trained on did not converge to within 1% with respect to simulations of higher mass resolution (DeRose et al., 2019). Above $\log(M/h^{-1}{\rm M}_\odot) = 15.0$, the Poisson noise in the binned halo counts data used to train AEMULUS is $\gtrsim 10\%$ (McClintock et al., 2019b).[23]

Given cosmological parameters, AEMULUS outputs the halo mass function by first emulating the fitting parameters d, e, f, g to the multiplicity function (Tinker et al., 2008; Tinker et al., 2010)

$$f(\sigma) = B\left[\left(\frac{\sigma}{e}\right)^{-d} + \sigma^{-f}\right]\exp\left(-g/\sigma^2\right), \tag{1.72}$$

where B is a normalisation constant such that all dark matter resides in halos. This is the alternative fitting function by Tinker et al. (2008) mentioned in Sec. 1.6.2.1. Then, AEMULUS converts Eq. (1.72) into the halo mass function by calculating $\sigma(M,z)$ from the linear matter power spectrum $P(k,z)$ computed by CLASS (Lesgourgues, 2011; Blas et al., 2011) using Eq. (1.59). As the fitting parameters are explicit functions of cosmology and redshift, AEMU-

[21] Note the quoted emulator accuracies do not fully account for numerical uncertainties associated with fitting halo mass functions to N-body simulations such as the effect of halo finders discussed in Sec. 1.5.

[22] The snapshots are at 10 redshifts z={0.0, 0.1, 0.25, 0.4, 0.55, 0.7, 0.85, 1.0, 2.0, 3.0}.

[23] The binned halo data were accessed from https://github.com/tmcclintock/Aemulus_data/tree/master/aemulus_data/mass_functions.

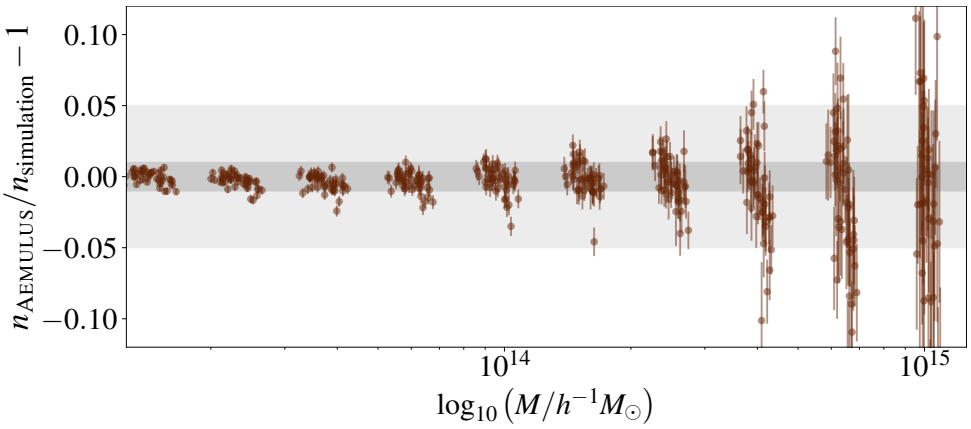


Figure 1.9: Residuals of the AEMULUS halo mass function emulator (McClintock et al., 2019b) on its training and test sets totalling 47 cosmologies, which span $\pm 4\sigma$ of the region allowed by CMB+BAO+SNIa in a wCDM+N_{eff} parameter space. The y-axis shows the halo abundance per unit volume $n = N/V$ predicted by AEMULUS vs. simulation data used to train the emulator (see footnote 23). Error bars show Poisson error. Dark shaded region shows $\pm 1\%$ and light shaded region shows $\pm 5\%$.

LUS learns a non-universal halo mass function. The emulator accuracy on the training and test sets is shown in Fig. 1.9.

The DARK EMULATOR (Nishimichi et al., 2019) is also trained for M_{200b} halos (i.e. halos identified with Δ_{200b}), and covers a six-dimensional wCDM cosmological parameter space for a smaller redshift range of $z = [0, 1.48]$. Similar to AEMULUS, it emulates fitting parameters to the Tinker halo mass function, although it uses Eq. (1.67) instead, and it only emulates two of the four fitting parameters while keeping the others to the original Tinker expressions. Nevertheless, it achieves $\sim 1-2\%$ accuracy for $10^{13}\,h^{-1}\mathrm{M}_\odot \lesssim M \lesssim 10^{14}\,h^{-1}\mathrm{M}_\odot$, while errors at the high mass end of $M \sim 10^{15}\,h^{-1}\mathrm{M}_\odot$ become $> 5\%$.

The Mira-Titan emulator (Bocquet et al., 2020) is trained on halos identified using Δ_{200c} covering $M_{200c} \sim 10^{13} - 10^{15}\,h^{-1}\mathrm{M}_\odot$ for cosmologies including dynamical dark energy via the $w_0 - w_a$ parametrisation (Eq. 1.24) and massive neutrinos. Its 111 training simulations are chosen to sample an eight-dimensional cosmological parameter space uniformly (Heitmann et al., 2016). Heitmann et al. (2016) do find however that sampling w_0 and w_a as independent parameters leads to poorer prediction accuracy where $w_0 + w_a$ is close to zero. To achieve a more consistent prediction accuracy across the entire parameter space, they instead sample over a $w_0 - w_b$ parameter space for $w_b \equiv (-w_0 - w_a)^{1/4}$, which places more samples where $w_0 + w_a$ is close to zero. Mira-Titan can reach percent-level precision at $z = 0$ for lower mass halos,

though the precision at the high mass end can be at $\sim 10\%$. The error increases with redshift between the range covered from $z = 0$ to $z = 2$.

1.7 Thesis overview

The standard model of cosmology – ΛCDM – has successfully described many phenomena, but it still has unresolved tensions and leaves questions, such as the nature of dark energy, unanswered. Therefore, major current and forthcoming surveys aim to probe cosmology beyond the standard model.

A sensitive cosmological probe is galaxy cluster counts, as it traces the halo mass function where it is exponentially sensitive to a change in cosmology. Until now, the strength of cosmological constraints from cluster counts is limited by the accuracy of the mass-observable relation. As its accuracy improves to a few-percent level (Sartoris et al., 2016; McClintock et al., 2019a; Euclid Collaboration et al., 2023a), the uncertainties associated with modelling the halo mass function start to become significant. Modelling error of the halo mass function can degrade the constraining power of forthcoming surveys if it is not kept at a percent-level or lower (McClintock et al., 2019b; Euclid Collaboration et al., 2023a).

As we have seen in Sec. 1.4 to 1.6, modelling the halo mass functions traditionally assumes universality, but universality does not hold at the percent-level. To find a halo mass function this accurate for (beyond-)ΛCDM cosmologies, literature either tries to improve semi-analytical fits by including additional parameters to $f(\sigma)$, or it approaches this problem by training halo mass function emulators. The state-of-the-art semi-analytical halo mass functions incorporate parametrisations of the linear growth history and the power spectrum shape (Ondaro-Mallea et al., 2021; Euclid Collaboration et al., 2023a). While physically motivated, such semi-analytical approaches require assuming the relevant quantities and their parametrisation, which literature has not converged on. On the other hand, emulators bypass the challenges associated with modelling non-universality and can reach percent-level precision in their halo mass function predictions. But, due to the lack of theoretical underpinning, emulators cannot generalise well outside the parameter space they are trained on. It also leaves rooms of improvement on designing emulator training sets to efficiently achieve high accuracy; current designs uniformly covering the target cosmological parameter space require an increasingly large number of addi-

tional simulations to achieve higher emulator accuracy (Heitmann et al., 2016), making further improvements to the accuracy of existing emulators expensive.

Given the current situation, there is a need to find more generalisable accurate halo mass functions, which requires a physical understanding of the factors that affect non-universality of the halo mass function. Such knowledge can, in turn, be used to inform the design of halo mass function emulators: additional simulations would only need to be placed such that the relevant factors are more uniformly sampled.

Obtaining such physical understanding using existing approaches is challenging. This thesis therefore develops a novel approach using deep learning, which are machine learning algorithms that excel at learning complex non-linear mappings. Recent work by Iten et al. (2020); Lucie-Smith et al. (2022b, 2024a) have shown that it is possible to extract physical knowledge from deep learning models. We aim to train a deep learning model to predict the non-universal halo mass function, and then interpret what the model learnt to discover knowledge on the factors required to model the halo mass function. This approach has the advantage of requiring minimal prior assumptions on the relevant quantities and parametrisations, while enabling physical insights that emulators alone do not offer.

The following chapters start with an introduction to deep learning models and methods of knowledge extraction in Ch. 2. Preparatory work done to explore deep learning models that enable knowledge extraction, and to develop tools for interpreting deep learning models, is presented in Ch. 3. Using the knowledge and tools gained, Ch. 4 applies this to discover the factors required to model the non-universal halo mass function at $z = 0$. Ch. 5 outlines future directions of the work: the framework developed in this thesis can be extended to model the redshift dependence of the halo mass function in addition to its cosmology dependence, and the factors identified can be used to improve emulator training sets. Concluding remarks are presented in Ch. 6.

Chapter 2

Methodology Background

We aim to gain novel insights on a problem in physics through deep learning, a subclass of *machine learning*. Machine learning algorithms (e.g. Gaussian process emulators discussed in Sec. 1.6.3.1) learn how to map from some input to output without explicit instructions. They learn the mapping through seeing examples of data, which form the *training set*; generally, the larger the training set, the better the algorithm learns the mapping. How well the the mapping is learnt depends also on what information is used from the data. The information can be 'manually' extracted from the data by performing *feature extraction*, where the high-dimensional raw data is projected into a lower-dimensional feature space, followed by *feature selection*, where the most relevant and informative features are selected to form inputs of the machine learning algorithm (Li et al., 2016). Alternatively, if the machine learning algorithm is complex and flexible enough, it can learn to extract and select the relevant features internally; such algorithms are often *deep learning* algorithms and will be introduced in greater detail in Sec. 2.1.

Because deep learning algorithms do not require identifying the features beforehand, they are well-suited for high-dimensional, highly non-linear problems where the relevant features are not necessarily known *a priori*. Interpreting the features used by a deep learning model therefore offers an avenue towards extracting knowledge about the underlying mapping, which can lead to new scientific insights. Extracting such knowledge is non-trivial; the expressiveness and flexibility of deep learning algorithms that make them so powerful also make them difficult to interpret, i.e. understand how the algorithm produces its output. This earns deep learning models the reputation of being "black-boxes". Because of this, knowledge extraction is still a small and developing field of deep learning applications in astrophysics and cosmology. The difficulty in interpreting deep learning models can further make it harder to assess the robust-

ness of the model and the scientific results produced. Efforts to interpret machine learning models in astrophysics and cosmology therefore mainly focus on determining if the model is using physically sensible quantities, rather than on extracting new knowledge. An overview of approaches towards interpreting machine learning models is presented in Sec. 2.2.

More broadly, the popularity of machine learning (especially deep learning) in astrophysics has soared in recent years as the rapid increase in data availability and computing power make the algorithms increasingly powerful (see Fluke and Jacobs, 2019; Huertas-Company and Lanusse, 2023, for recent reviews). Machine learning algorithms' ability to learn high-dimensional non-linear mappings and theirs fast deployment once the model is trained make machine learning especially well-suited to complex problems involving many variables, and to speeding up time-consuming calculations that need to be repeated a large number of times.

A prominent use of machine learning is assisting with astronomical data analysis. Object identification and classification is one of the earliest applications of machine learning in astrophysics (e.g. Whitmore, 1984; Storrie-Lombardi et al., 1992; Odewahn et al., 1992; Bertin, 1994; Lahav et al., 1995), and remains prevalent today (e.g. Domínguez-Sánchez et al., 2018; Huertas-Company et al., 2018; Walmsley et al., 2020). Machine learning enables automatic classification of a large number – $\mathcal{O}(10^7 - 10^8)$ – of sources from surveys like *Gaia* (Gaia Collaboration et al., 2016) and SDSS (Gott et al., 2005; Clarke et al., 2020; Rimoldini et al., 2023). It is also used for anomaly detection, the identification of rare or unknown astrophysical phenomena (e.g. Giles and Walkowicz, 2019; Liang et al., 2023), which is especially useful in the present era when surveys produce datasets too large to be examined by humans. Such use of machine learning also has lower requirements on the model's interpretability, as the outputs (anomalies detected) are then further analysed by humans. Interestingly, large language models which have recently surged in popularity in the wider society, such as Chat Generative Pre-Trained Transformer (ChatGPT; OpenAI et al., 2023), are also finding their way into astrophysics. For example, they are being used to access *Gaia* data and produce plots from the data.[1] With time, these may revolutionise the way we access astronomical databases.

In cosmology, a prominent use of machine learning is building emulators that accelerate time-consuming tasks such as simulations and inference. An example is emulating N-body simulations. Doeser et al. (2024) trained an emulator that maps from first-order Lagrangian

[1] ChatGaia, https://www.whatplugin.ai/gpts/chatgaia

perturbation theory predictions to N-body simulations, which significantly accelerates evaluations of the physics model during inference compared to running N-body simulations; similar N-body simulation emulators are also built by e.g. Jamieson et al. (2023); Piras et al. (2023a). Machine learning is also used to emulate hydrodynamical simulations, which are much more computationally expensive, by mapping from dark matter to baryonic properties (e.g. stellar mass, thermal Sunyaev-Zel'dovich maps, etc.; Tröster et al., 2019; Dai and Seljak, 2021; Dai et al., 2023), or by emulating the calculation of baryonic subgrid physics in each time step of the simulation (Hirashima et al., 2023). As described in Sec. 1.6.3.3, machine learning is also used to emulate quantities derived from simulations like the halo mass function, as well as the non-linear matter power spectrum (Angulo et al., 2021; Moran et al., 2022) and other quantities (e.g. Zhai et al., 2019; McClintock et al., 2019c; Aricò et al., 2021).

Because deep learning algorithms are typically automatically differentiable, neural network emulators can be combined with gradient-based inference methods such as Hamiltonian Monte Carlo (HMC, Duane et al., 1987; Neal, 1996) to significantly speed up inference even when individual evaluations of the emulated calculation are less time-consuming than e.g. running N-body simulations. Neural network emulators further take advantage of graphics processing unit (GPU) acceleration, leading to even more significant speed up if the rest of the inference pipeline can also be run on GPUs. For example, Piras and Spurio Mancini (2023) show that compared to a more traditional inference approach using Einstein-Boltzmann equation solvers (e.g. CAMB Lewis et al., 2000 introduced in Sec. 1.2.3) coupled with nested samplers (Handley et al., 2015a,b), cosmological inference can be accelerated by orders of magnitude when using a neural network emulator of Einstein-Boltzmann solvers (Spurio Mancini et al., 2022) combined with HMC and ran on several GPUs. This would be particularly useful for meeting the large computational demands of forthcoming surveys.

Another major application of machine learning in cosmology is simulation-based (a.k.a. likelihood-free) inference (SBI), which allows Bayesian inference when the likelihood of observed data given some parameters is either not available or too costly to evaluate. Machine learning allows simulation-based inference using smaller numbers of simulations than traditional SBI approaches such as Approximate Bayesian Computation (ABC; see e.g. Alsing et al., 2019), leading to orders of magnitude speed-ups. Major surveys such as the Kilo Degree Survey (KiDS; Lin et al., 2023; von Wietersheim-Kramsta et al., 2024) and the Dark Energy

Survey (DES; Jeffrey et al., 2024) explore using SBI for parameter inference, and SBI allows fast inference also when using information beyond two-point statistics such as the power spectrum, leading to tighter parameter constraints (Jeffrey et al., 2024).

Machine learning is being widely adopted in assisting with the analysis of current and forthcoming surveys. Areas where its use is most well-established are where there are lower requirements on the interpretability of the machine learning model. For example, an exact understanding of how a model learns the likelihood in SBI is not required as long as tests ensure the reliability of the learnt likelihood (Lemos et al., 2023a). At the same time, the lack of interpretability inevitably makes it difficult to understand the machine learning model's limitations and impacts the model's robustness. Therefore, developing methods to interpret machine learning, and in particular deep learning models, is of great importance in science. Furthermore, as discussed at the beginning of this chapter, there is significant potential in using deep learning to gain new scientific insights if the model can be interpreted; this use of deep learning is the focus of this thesis. In Sec. 2.3, I introduce representation learning, which is an approach that is especially promising for knowledge extraction. The representation learnt can be interpreted using an information-theoretic metric call *mutual information*, which will be introduced in Sec. 2.4 and used in this thesis.

2.1 Neural networks

Deep learning models are made of artificial neural networks (or simply *neural networks*); the basic components of a neural network are introduced in Sec. 2.1.1 together with the simplest type of neural networks, *fully-connected neural networks*. These are capable of accurately approximating a wide range of multi-input and multi-output functions (Nielsen, 2018). Convolutional neural network (CNN), a type of neural network often used for image data, is described in Sec. 2.1.2, and Sec. 2.1.3 describes the training and tuning of neural networks.

2.1.1 Fully-connected neural network

The elementary building blocks of a neural network are known as *neurons*. They are usually organised into *layers*, as illustrated in Fig. 2.1; each neuron is linked to neurons in the layer before it from which it receives data, and is linked to neurons in the layer after it to which it

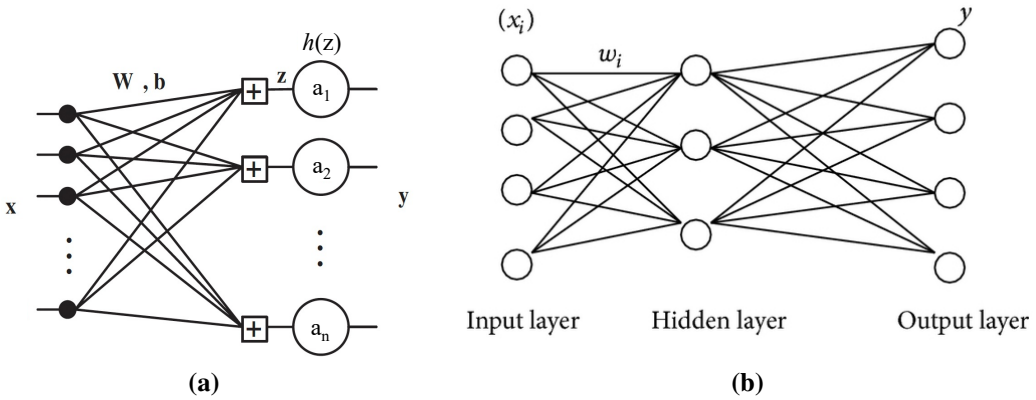


Figure 2.1: *Left:* The value of each neuron a_i (open circle) is a weighted sum of its inputs (filled circles) plus a bias b, passed through a non-linear activation function h. Figure adapted from figure 1 in Fontenla-Romero et al. (2003). *Right:* An illustration of a neural network with an input layer, one hidden layer, and an output layer. Circles represent neurons and lines show links between neurons. This is a fully-connected neural network since all neurons between two layers are interconnected. Figure taken from figure 5 in Sengottuvelan and Arulmurugan, 2014.

sends data. Each of the links between two neurons carries a parameter w_i known as the *weight*. A neuron takes as inputs some vector $\mathbf{x}$ weighted by $\mathbf{w}$, then applies a linear transformation $z = \mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b$, where b is another parameter known as *bias*. An *activation function*, h, which is typically non-linear to enable a neural network to model non-linear mappings, is applied to z to give the neuron output $a = h(z)$. Earlier neural networks often used the sigmoid function as the activation function (see e.g. Odewahn et al., 1992; Bertin, 1994 for examples in astronomy). However, its outputs are not symmetric around zero, which can lead to slower training, and the hyperbolic tangent (tanh) function

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{2.1}$$

became a more preferable choice (see e.g. a review in Szandała, 2021). It is smooth and continuously differentiable. However, the gradient becomes vanishingly small for values lying close to -1 or 1, which could hinder gradient-based neural network training (described in Sec. 2.1.3). In recent years, rectified linear units (ReLU, Fukushima, 1969; Nair and Hinton, 2010) became more popular. These activation functions output 0 if $z < 0$ and is otherwise a linear identity function. Although they are not smooth and differentiable at zero, they avoid the vanishing gradient problem of sigmoid and tanh functions for positive values, and perform

well for a range of tasks (Nwankpa et al., 2018). For negative values, however, ReLUs still suffer the vanishing gradient problem (they do not utilise negative values). This prompted the development of the leaky ReLU activation (Maas et al., 2013), defined as

$$\text{leakyReLU} = \begin{cases} \alpha z & , \quad z < 0 \\ z & , \quad z \geq 0 \end{cases} \tag{2.2}$$

for α a small value that is typically set manually. In this way, the negative values are also utilised, avoiding the vanishing gradient problem. A review of activation functions can be found in Nwankpa et al. (2018); Dubey et al. (2021).

The simplest form of neural networks is a feed-forward neural network made of dense layers. *Feed forward* refers to data only moving through layers of neurons in one direction, and *dense* (a.k.a. *fully-connected*) means each neuron within a layer is connected to all neurons in the layer before it, and to all neurons in the next layer. Fig. 2.1b shows a shallow fully-connected neural network with just one layer between the input and output layers, i.e. one *hidden layer*. Deep neural networks have many hidden layers with many neurons per layer. Generally, a wider network allows extracting more features, while a deeper network can combine the features to form higher level features, thus achieving greater levels of abstraction.

2.1.2 Convolutional neural networks

While fully-connected neural networks in theory can approximate a wide range of functions (Cybenko, 1989; see also Nielsen, 2018), in practice, alternative designs of neural network architectures can exploit symmetries and data properties to achieve better performance. For data with spatial structure such as images, a fully-connected neural network would require compressing the image down to a one-dimensional vector, losing information on spatial relations between pixels. To preserve such information, *convolutional layers* are often used which can take two- (or higher-)dimensional inputs. A convolution is applied by taking the dot product between inputs and a set of weights plus a bias term. However, in this case the weights are organised in a two- (or higher-)dimensional matrix called a *convolution kernel* or *filter*. The filter is often much smaller than the input and is applied to a patch of the input at a time, outputting a single scalar for each patch. The filter with the same set of weights is shifted systematically across the input so the output forms a matrix known as a *feature map*. Typically several filters

are used each to detect different features; the feature maps then undergo activation before the next layer.

2.1.3 Training and hyperparameters

Training a neural network amounts to adjusting the weights and biases so the network gives 'correct' outputs. The 'correctness' of outputs is specified by the *loss function* $\mathscr{L}$, which measures the difference between the network outputs and the desired outputs. For *supervised learning* where *ground truths* (the true value that outputs should assume) are known and are provided as labels, a common choice for the loss function is the mean squared error between the ground truths and model predictions. For *unsupervised learning* where the data are not labelled, the loss function could be e.g. the mean squared error between the input and the model reconstruction for a neural network trained to reconstruct the inputs from a compressed representation (see more details in Sec. 2.3.1).

Neural networks are trained iteratively. In one *epoch* of training, the data are first passed once through the network to obtain outputs. Gradients of the loss function with respect to the weights and biases are then calculated via the chain rule for partial differentiation, starting from the output layer back through each layer of the network. This procedure is called *backpropagation*. All weights and biases in the network are updated in proportion to the calculated gradients according to (Nielsen, 2018)

$$w_i \to w_i - \eta \frac{\partial \mathscr{L}}{\partial w_i}, \qquad b_i \to b_i - \eta \frac{\partial \mathscr{L}}{\partial b_i}, \tag{2.3}$$

where η is a *hyperparameter* that is manually set (as opposed to learned during training) called the learning rate, which we will return to shortly.

Gradient descent, i.e. updating weights and biases in proportion to the gradients like in Eq. (2.3), can be done in three ways: the weights and biases can be updated per training sample, every N samples, or once every time the entire training set passes through. Updating the trainable parameters every training sample is called *stochastic gradient descent*. The gradient estimate given just one training sample is noisy, although the noise can help to overcome local minima in the loss landscape, and updating training parameters every sample can lead to faster training (Ruder, 2016). On the other end, *batch gradient descent* updates weights and biases using the mean gradient of the entire training set. It leads to accurate estimates of the gradient

for the training set, but is slower in training and intractable if the dataset does not fit in memory. Moreover, it does not necessarily lead to finding minima that generalise well beyond the training set (Shirish Keskar et al., 2016), which is ultimately the goal of most machine learning models. Hence, batch gradient descent is rarely used. The third way is known as *mini-batch gradient descent* (though it is sometimes also referred to as stochastic gradient descent or batch gradient descent); it combines the best of two worlds, reducing the variance in the estimated gradients by calculating the mean gradient on small batches of data, while preserving some noise and allowing more frequent parameter updates that lead to more efficient training. The number of samples per batch is known as *batch size*, and is another hyperparameter that must be manually tuned (Goodfellow et al., 2016).

The *learning rate* η in Eq. 2.3 determines the amount by which the network parameters are adjusted for each epoch. A higher learning rate can cause drastic changes to the weights, resulting in divergent behaviour in the loss function, whereas a very small learning rate causes slow convergence towards the minimum of the loss function. Choosing a suitable learning rate often requires experimentation. Gradient calculation and weights and biases updates are done using *optimisers*, such as the Adam optimiser (Kingma and Ba, 2017) or AMSGrad optimiser (Reddi et al., 2018). Such optimisers take in a single user-defined value for the learning rate, but dynamically adapt the learning rate for each individual weight and bias during training. This can lead to faster convergence than using the same learning rate for all parameters.

At the beginning of the training, the weights are initialised using *initialisation schemes*. This is usually done so that the initial weights are small and randomly scatter around zero. The small zero-centred values prevent the activations from combining into excessively large (or vanishing) values that then cause very large (or negligible) updates to the trainable parameters, resulting in unstable (or stalled) training. For this same reason, the inputs to a neural network need to be scaled so the values are $\mathcal{O}(1)$. A commonly used initialisation scheme for layers using ReLU activation is the He initialisation scheme (He et al., 2015), where weights are drawn from a normal distribution with zero mean and a standard deviation of $\sqrt{2/n}$ for n the number of neurons in the layer. For layers using tanh activation (which we will use in Ch. 4), the Glorot uniform (a.k.a uniform Xavier) initialisation (Glorot and Bengio, 2010) is commonly used, where weights are sampled from a uniform distribution between $\pm\sqrt{6}/\sqrt{n_{l-1}+n_l}$, where n_{l-1} is the number of inputs to layer l, and n_l is the number of outputs from layer l.

A neural network is generally trained for multiple epochs until the loss function no longer decreases, at which point the network has reached convergence. Because neural networks are so flexible, a danger lies in *overfitting*, where the model starts to learn specificities of the training data such as noise that are unrelated to the mapping that the network aims to learn. This hinders the ability of the neural network to generalise to unseen data. To prevent this, especially for supervised learning, the performance of a neural network is often monitored during training on a *validation set* comprising of data not used for parameter updates. The loss function is regularly evaluated on the validation set (usually at the end of every epoch), and if the validation loss stops decreasing over a number of epochs (this number is referred to as *patience*), training is terminated (called *early stopping*), and the model is deemed converged. Beyond this point, the training loss will continue to decrease, but the model's performance on unseen data is not expected to improve. After a model is converged, its generalisability is often assessed on a *test set* consisting of data samples that the model has never seen during training nor validation. If the model performance on the test set is comparable to its performance on the training set, the model generalises well (Nielsen, 2018). Note that generally, one can only expect the neural network to perform well within the domain that it is trained on, and the test set is also drawn from the same domain as the training set.

2.2 Approaches to interpretability and explainability

Deep neural networks typically have hundreds of thousands to billions of parameters that need to be trained. While this gives a neural network the flexibility to well-approximate any continuous function (Nielsen, 2018), it makes deep learning models extremely difficult to interpret, giving them the reputation of being "black boxes". This is less of a concern when the neural network is used as emulators for a known mapping, as long as the neural network performance is well-tested and the neural network is used within the domain it is trained on. Even then, one would like to ensure the model is robust and basing its decisions on relevant quantities. As discussed at the beginning of this chapter, understanding how deep learning models produce their outputs is important for science results obtained using deep learning to gain trust of the science community. The domain of machine learning concerned with interpreting and explaining what

a machine learning model learnt and why it reaches certain decisions is known as *explainable artificial intelligence* (XAI).

The community has not yet converged on a set of clear definitions for the notions of *interpretability* and *explainability*. Sometimes interpretability is used to refer to models that are simple enough for humans to understand their workings while explainability refers to the ability of explaining a model post-hoc (e.g. in Pasquato et al., 2023), though most of the times the terms are used interchangeably. I provide here the definitions used in this thesis:

- *Interpretability* refers to the ability to account for why a machine learning model reaches a certain decision or prediction. It is important both for ensuring the robustness of the model, and for extracting new knowledge from a neural network.

- *Explainability* denotes the ability to map the interpretation of the machine learning model onto existing knowledge in the relevant science domain and formulating physical explanations and hypotheses accordingly. Presently this step can only be done by a human.

The rest of this section gives an overview of methods used to achieve interpretability in astrophysics and cosmology. These methods are applied to neural networks as well as other complex machine learning algorithms that are difficult to understand (e.g. gradient boosted trees, Friedman, 2001).

2.2.1 Saliency maps

In astrophysics and cosmology, methods for interpreting machine learning models are often used for confirming whether model outputs are based on physically relevant quantities. For example, Huertas-Company et al. (2018) use a saliency maps method (Simonyan et al., 2013) called integrated gradients (Sundararajan et al., 2017) to identify which pixels of an image contribute the most to a galaxy being identified as being in a compact star-forming phase. Saliency maps highlight pixels of an input image that contribute the most to the model's output, and is a widely used method for interpreting features learnt by deep learning models (Zhang and Zhu, 2018). Using saliency maps, Huertas-Company et al. (2018) confirm that model predictions are only sensitive to pixels of the galaxy, and not to spurious information located elsewhere. Saliency maps have also been used by Bhambra et al. (2022) to assist with measuring galaxy bar length. However, the disadvantage of saliency maps is that a human needs to assemble the

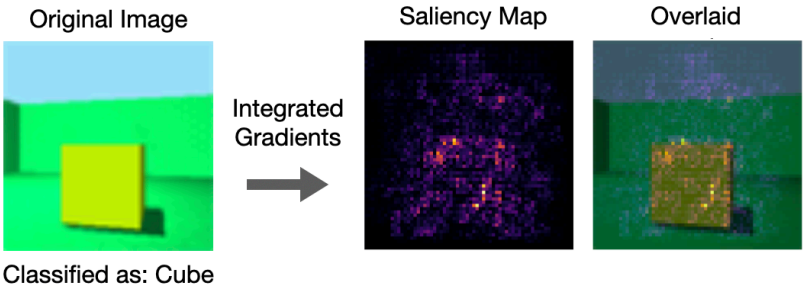


Figure 2.2: Saliency maps (calculated with integrated gradients; Sundararajan et al., 2017) is applied to an image that is correctly classified as a cube. It highlights some pixels that lie on the edge of the cube, but also pixels around the cube. It is difficult to understand what feature (e.g. object edge or face) is being highlighted.

pixels into features, which is often not straightforward. The features highlighted by the saliency map can also vary from image to image.

The difficulty of interpreting saliency maps can be illustrated on a simple toy problem. I train a neural network to classify images based on object shape, and interpret features learnt by the network using integrated gradients (Sundararajan et al., 2017). Integrated gradients scores each pixel by accumulating the gradient of the probability for classifying the image as e.g. a cube with respect to varied pixel intensities. Fig. 2.2 shows an example saliency map that was produced for the case where the image is classified (correctly) as a cube. In this example, it is not clear how the highlighted pixels can be interpreted as features (such as object edge). Furthermore, because saliency maps are computed given specific images, they only show what the model used to produce its prediction for the particular image. For some other example images, the edge of the cube or its shadow becomes highlighted instead. Hence, saliency maps offer limited insights on whether the same identified features generalise to the entire dataset, but such insight is crucial for extracting knowledge by interpreting a neural network.

For non-image data, a similar idea is used in *sensitivity analysis*, where one measures the change in model outputs as a set of features in inputs are locally perturbed. Yip et al. (2021) uses this method to confirm that their neural networks infer atmospheric parameters using features of exoplanet spectra that agree with physical expectations.

2.2.2 Change inputs and analyse corresponding change in model outputs

Lucie-Smith et al. (2018, 2019, 2024b) determine the change in the machine learning model outputs when provided with different inputs to understand the relevance of each input. The pa-

pers test the interpretation that the Sheth-Mo-Tormen (SMT) halo mass function (see Sec. 1.4.2) fits simulations better than Press-Schechter (PS; Sec. 1.4.1) because SMT incorporates ellipsoidal collapse into excursion set theory (Sheth et al., 2001; Jenkins et al., 2001; Reed et al., 2003; Warren et al., 2006) and accounts for deformation of spherical regions by tidal shear effects (Lukić et al., 2007). Lucie-Smith et al. (2018, 2019) train random forests (Ho, 1995) to predict the friends-of-friends halo mass (see Sec. 1.5.3). They compare the model's performance when it is given a large sample of spherically smoothed densities δ as a function of smoothing scale (centred around N-body particles), to when it is also given the ellipticity and prolateness (and hence information on the local tidal shear field). Contrary to the existing interpretations of why SMT is more accurate than PS, they find that providing the algorithm with local tidal shear field information does not significantly improve the algorithm performance, suggesting that the local tidal shear field plays no significant role in determining halo masses.

While the higher accuracy of SMT compared to PS is often interpreted as its incorporation of tidal shear, the SMT function only incorporates this through parameters fitted to simulations which are motivated by considering ellipsoidal collapse (see Sec. 1.4.2). Hence, it does not directly demonstrate the significance of tidal shear. Lucie-Smith et al. (2024b) expand upon their previous work and test whether any anisotropic information (including tidal shear) is relevant to predicting the halo mass. They do so by removing information in the density field: they compare the prediction accuracy of a CNN given the raw local initial density field around each simulation particle to a CNN given just spherically averaged densities centred on each particle. All anisotropic features are removed in the latter input. Compared to their previous work, the use of a deep learning algorithm removes the need to manually construct features, and therefore tests whether their previous finding could be because ellipticity and prolateness do not adequately describe the anisotropic information. They find that while there is a small statistically significant improvement in the predictions when the input also contains anisotropic information, the scatter in the predicted halo masses does not reduce qualitatively. Therefore, it appears that anisotropic information in the initial density field does not play a significant role in determining the final halo mass. Such result demonstrates that interpreting deep learning models through removing information from the inputs can establish their relevance and lead to new physical insights.

2.2.3 Other methods of feature attribution

Saliency maps in Sec. 2.2.1 belong to a class of methods called *feature attribution*. In addition to assessing the robustness of the model, feature attribution is also used to extract knowledge on the important or relevant features of a problem. This class of techniques quantifies the amount that each feature contributes to the model outputs. In Lucie-Smith et al. (2019), feature importance score is used to confirm that the machine learning model barely uses ellipticity or prolateness features to predict halo mass. Feature importance is also used by Lucie-Smith et al. (2022a) to identify that the spherically-averaged halo mass profile is mostly affected by the initial density field smoothed at two scales (one close to the halo's Lagrangian radius, and another in the halo's large-scale environment). They also found the halo mass profile is significantly affected by the halo's mass accretion history at three timescales: the halo's half-mass formation time, its dynamical time, and a recent timescale related to recent massive mergers (which mostly affects the halo profile at the outskirts).

Machado Poletti Valle et al. (2021) used Shapley Additive exPlanations (SHAP, Lundberg and Lee, 2017), another feature attribution method, to identify that the shape of gas in halos is best predicted by baryonic properties in the halo centre, while dark matter shape is the best predictor at halo outskirts, agreeing with physical intuition. Wu et al. (2024) predict galaxy stellar mass from halo and large-scale environment properties, and use SHAP to find that the most relevant feature apart from the maximum circular velocity and halo mass is the linear overdensity on scales of 3 Mpc.

Shao et al. (2022) perform feature attribution by calculating the average gradient of the model predictions with respect to input features, and identify a set of halo and galaxy properties that predict subhalo mass. They further apply *symbolic regression* (Koza, 1994; Schmidt and Lipson, 2009) to find an analytical formula for predicting the subhalo mass given the most important features identified.

2.2.4 Symbolic regression

The use of symbolic regression as an approach towards interpretable machine learning and achieving physical insights is gaining momentum in the recent years. Symbolic regression is a class of algorithms that predicts the output by combining input features through mathematical operators and constants; the set of mathematical operators are often pre-defined by the user. It is

often applied to trained deep learning algorithms to gain interpretability, as it returns mathematical expressions which give an account of how the output is produced. However, mathematical expressions are not necessarily explainable, and the algorithm can often produce expressions that cannot be made sense of (e.g. Piras and Lombriser, 2024). This issue limits the current applications of symbolic regression to situations where there is already ample prior knowledge. For example, Lemos et al. (2023b) use symbolic regression to interpret a graph neural network (Battaglia et al., 2018) trained to simulate dynamics of the solar system, and rediscover Newton's law of gravity. However, this required incorporating Newton's second and third laws into the graph neural network. Encouraging symbolic regression to return meaningful expressions that enable new physical insights is an active area of research. Current efforts try to achieve this by incorporating physical priors on the neural network that symbolic regression is applied to, e.g. through choosing neural networks well-suited to the problem and encouraging sparse representations (Cranmer et al., 2020), or by incorporating priors into the symbolic regression algorithm through e.g. imposing unit constraints (Tenachi et al., 2023).

The problem of interpreting machine learning models becomes easier if there is only a small number of relevant parameters that need to be interpreted. Such small set of parameters can be found through representation learning, which next section introduces.

2.3 Representation learning

Representation learning is a class of machine learning algorithms that compresses relevant information from the raw input into a low-dimensional representation. A widely used representation learning algorithm is principle component analysis (PCA; see Greenacre et al., 2022), which performs linear dimensionality reduction. Non-linear dimensionality reduction can be done using neural networks such as the *autoencoder* (Kramer, 1991; Kramer, 1992), which learns to reconstruct inputs after first compressing the input into a low-dimensional *latent representation*.

In astrophysics, representation learning is often used for morphological classification (e.g. Whitmore, 1984; Francis et al., 1992; Spindler et al., 2021) and outlier detection (e.g. Melchior et al., 2023; Liang et al., 2023), as it reduces data into a low-dimensional space where similar data are clustered together. As recent works demonstrate (Iten et al., 2020; Ntampaka

and Vikhlinin, 2022; Lucie-Smith et al., 2022b, 2024a), representation learning also offers a promising route towards gaining physical insight. It is particularly suited for knowledge extraction because it can significantly reduce the number of parameters that require interpretation.

For example, Ntampaka and Vikhlinin (2022) used an autoencoder-inspired architecture to predict σ_8 after first compressing galaxy cluster observables down into a one-dimensional latent representation for each cluster. They then only need to interpret the 1D latent representation to understand what X-ray observables are being used to predict σ_8. They do so using a combination of methods, including saliency maps (between the *latent representation* and the input), and directly plotting the correlation between the latent representation and cluster parameters such as σ_8, gas mass, or temperature. Through this, they find a previously unknown self-calibration mode which allows constraining fundamental cluster properties together with cosmological parameters: they find that the total number of clusters in a survey together with Y_X (see Sec. 1.3.1) contains enough information to calibrate mass and constrain σ_8. Inspecting the correlation between the latent representation and cluster parameters allows them to check that the neural network learnt physically reasonable information.

Ntampaka and Vikhlinin (2022) only needed to interpret a one-dimensional latent representation. For a higher-dimensional latent representation, interpreting the latent space is facilitated if the latent representation is *disentangled*, i.e. if each component of the latent representation captures independent information. This is rarely the case for latent representations learnt by autoencoders, as there is no incentive in the loss function for the autoencoder to disentangle. However, it is possible using variational autoencoders (VAEs; Kingma and Welling, 2013; Higgins et al., 2017); Sec. 2.3.1 introduces them. As Iten et al. (2020); Lucie-Smith et al. (2022b, 2024a) demonstrate, a framework inspired by variational autoencoders allows the discovery of physically relevant factors; Sec. 2.3.2 presents this in greater detail.

2.3.1 Variational autoencoder

A variational autoencoder (VAE; Kingma and Welling, 2013) is a deep learning model comprising of two parts: an *encoder* and a *decoder*. It is similar to the autoencoder; both learn the identity mapping with an information bottleneck in the middle. The VAE's encoder network compresses an input $\mathbf{x}$ into a low-dimensional representation described by a small number of variables called *latents*. For each latent, unlike autoencoders which only learn a single latent

value given each sample, the VAE learns a Gaussian probability distribution called the *latent distribution* $q(\boldsymbol{\kappa} \mid \mathbf{x})$.[2] The decoder takes in, as input, one realisation of the latent value randomly sampled from each latent distribution, and maps the sampled latent vector $\boldsymbol{\kappa}$ to the output (a reconstruction of the input).

VAEs are trained with the objective of approximating the true posterior $p(\boldsymbol{\kappa}|\mathbf{x})$ as closely as possible using the learnt latent distribution $q(\boldsymbol{\kappa} \mid \mathbf{x})$, while maximising the marginal likelihood of the data, $p(\mathbf{x})$ (Kingma and Welling, 2019). The former can be expressed through minimising the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the latent distribution and the true posterior,

$$\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{\kappa} \mid \mathbf{x}) \| p(\boldsymbol{\kappa}|\mathbf{x})) = \int q(\boldsymbol{\kappa} \mid \mathbf{x}) \ln \frac{q(\boldsymbol{\kappa} \mid \mathbf{x})}{p(\boldsymbol{\kappa}|\mathbf{x})} \mathrm{d}\boldsymbol{\kappa}, \tag{2.4}$$

which is non-negative, and is zero if and only if the two distributions are equal. Eq. (2.4) can be manipulated into a sum of three terms (see e.g. Odaibo, 2019), one of which is the marginal log-likelihood that should be maximised. Hence, the VAE is trained to minimise the sum of the other two terms, which are collectively known as the (negative of) the evidence lower bound (ELBO; Kingma and Welling, 2019). The two terms consist of a reconstruction term and a regularisation term. The *reconstruction term* measures the difference between VAE outputs and the desired outputs; it gets its name because VAEs are trained to reconstruct inputs using the information compressed in a small number of latents. The *regularisation term* promotes structure in the *latent space* (spanned by the latents) by separately encouraging the Gaussian distribution learnt for each latent to be close to a standard Gaussian $\mathcal{N}(0,1)$. This encourages a continuous latent space where all data samples are mapped to a localised region in the latent space (see Burgess et al., 2018), and similar latent values represent data with similar features, so that latents can reflect underlying structure in the data. This is because regularisation pushes the latent distributions given different inputs as close together as possible, while sampled latent vectors drawn from the partially overlapping latent distributions should still decode to correct outputs. This drives the encoder to map similar inputs to overlapping latent distributions (Burgess et al., 2018).

[2] Note that literature often uses z for latents; I use $\boldsymbol{\kappa}$ to avoid possible confusion with redshift.

A VAE as proposed by Kingma and Welling (2013) does not necessarily achieve disentanglement. To encourage latent disentanglement, Higgins et al. (2017) propose β-VAE, which weights the regularisation term with a hyperparameter $\beta \geq 0$. The loss function (for a single sample) becomes

$$\mathscr{L} = -\mathbb{E}_{q(\kappa|\mathbf{x})}\left[\ln p(\mathbf{x} \mid \kappa)\right] + \beta \mathcal{D}_{\mathrm{KL}}\left(q(\kappa|\mathbf{x}) \| p(\kappa)\right), \tag{2.5}$$

$$= \text{reconstruction term} + \beta \cdot \text{regularisation loss}. \tag{2.6}$$

In the reconstruction term, $p(\mathbf{x} \mid \kappa)$ is the probability density of the input given a sampled latent vector, parametrised by the decoder. The regularisation loss is the Kullback-Leibler divergence (Kullback and Leibler, 1951) between the latent distribution returned by the encoder $q(\kappa|\mathbf{x})$ and the prior $p(\kappa)$. As mentioned, the latent distribution of the i-th latent is parametrised as a Gaussian $\mathcal{N}(\mu_i, \sigma_i)$, and the prior is usually a diagonal standard Gaussian. The diagonal standard Gaussian prior encourages continuity in the latent space and promotes independence between latent variables, since the marginal latent distribution $q(\kappa) = \int q(\kappa|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$ is encouraged to be diagonal. These together encourage each latent variable to learn a different, independent factor governing variation in the data.

Mean squared error is often used for the reconstruction term.[3] This gives a closed form for the loss function, which for a batch of N samples reads

$$\mathscr{L} = \frac{1}{N}\sum_{j}^{N}\left(\left\|\mathbf{x}_j - \hat{\mathbf{x}}_j\right\|^2 - \beta\frac{1}{2}\sum_{i}^{L}\left(1 + \ln\left(\sigma_{i,j}^2\right) - \mu_{i,j}^2 - \sigma_{i,j}^2\right)\right), \tag{2.7}$$

where $\hat{\mathbf{x}}$ is the VAE output and L is the number of latents. Increasing $\beta > 0$ pressures each latent distribution closer to a standard Gaussian, so the latents become statistically independent and can be disentangled if factors are also statistically independent. This however increases the variance of each latent distribution, which impacts reconstruction accuracy (Higgins et al., 2017; Burgess et al., 2018), so β must be tuned to achieve a balance between accuracy and disentanglement. Other models such as FactorVAE (Kim and Mnih, 2018) have been proposed in an attempt to avoid the decrease in reconstruction accuracy when achieving disentanglement, though this leads to more complex architectures and/or more hyperparameters to tune.

[3] This can be justified if we assume the decoder outputs a Gaussian distribution with mean $\hat{\mathbf{x}}$ and identity covariance, in which case $-\mathbb{E}_{q(\kappa|\mathbf{x})}\left[\log p(\mathbf{x} \mid \kappa)\right]$ resembles the mean squared error (plus constants).

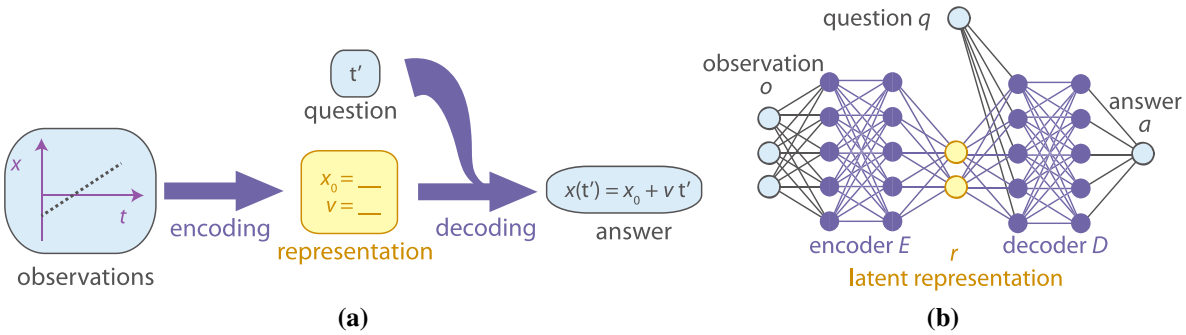


Figure 2.3: *Left:* The human process of formulating physical theories involves forming a low-dimensional representation of observations such that when given a question, one uses just the representation to produce answers. *Right:* the model translated into neural network architecture. Figures are from figure 1 in Iten et al. (2020).

2.3.2 *SciNet* and the interpretable variational encoder

Using the idea of β-VAEs, Iten et al. (2020) propose *SciNet*, a neural network architecture modelled after the human reasoning process designed for extracting physical concepts with minimal prior assumptions. The architecture is illustrated in Fig. 2.3; the left shows how physical theories are formulated by humans. Given observations (particle trajectory in Fig. 2.3a), a human compresses these down into a compact representation (the initial position and velocity). When later asked a question about the physical system ("where is the particle at a later time t'?"), one should be able to produce an accurate answer using only information in the compact representation. The right plot shows this idea translated to a neural network. Given inputs, an encoder compresses information into a low-dimensional latent representation. The decoder uses the learnt latent representation to produce an answer given a question about the system. Unlike a VAE, in *SciNet* the latent representation does not need to contain all the information needed to reconstruct the input, but only the relevant information needed to produce the answer.

SciNet is trained with the β-VAE loss function in Eq. 2.7. Iten et al. (2020) has applied *SciNet* to numerous toy problems and showed that it can learn disentangled latent representations where each latent corresponds to a physically-relevant factor. For example, to predict the position of a damped pendulum, two latent variables each learn the frequency and damping factors respectively, while additional latent variables carry no information as no further factors were varied.

Lucie-Smith et al. (2022b, 2024a) extend the *SciNet* framework so it can be applied to more complex problems where the underlying physically relevant factors are not known *a pri-*

ori: they investigate what factors determine the spherically averaged dark matter halo density profiles. To interpret the latent representation without knowing the true factors, they utilise mutual information, an information-theoretic metric for non-linear dependency between variables (see Sec. 2.4). They calculate the mutual information between each latent variable and the halo density profile to identify that predicting the spherically averaged halo density profile out to the halo's virial radius requires two latent variables; this agrees with the widely used Navarro-Frenk-White (NFW) profile requiring two degrees of freedom (Navarro et al., 1996). Beyond the virial radius, they find three latent variables are required to predict the halo density profile. Two latent variables each learn information on the central and the outer profile respectively, while the third latent variable learns information around the splashback radius. Lucie-Smith et al. (2024a) calculate the mutual information between the latent variables and the halo mass accretion history, and find that the latent variable learning the outer profile captures information on the recent mass accretion rate. This highlights the power of representation learning combined with mutual information to obtain novel physical insights.

2.4 Mutual Information

Literature on learning disentangled representations uses a variety of metrics to measure disentanglement, which can give discrepant results (see Carbonneau et al., 2022 for a review). At the same time, how to best interpret the information encoded in latent representations remains an open question. Latent representations are mostly interpreted using *latent traversal*, where a series of outputs (usually images) are generated by varying one latent at a time while holding other latent variables fixed. The changing feature is identified visually (see e.g. Higgins et al., 2017; Kim and Mnih, 2018), so it only gives a qualitative description of the information in the latent, and the results could depend on the values that other latents are fixed to. These problems can be addressed by mutual information.

Mutual information quantifies the amount of information we obtain about one variable by observing another. Unlike measures of correlation such as the Pearson correlation, mutual information is able to quantify non-linear dependences between two variables (Kinney and

Atwal, 2014). Mutual information between two variables x and y is given by

$$\mathrm{MI}(x,y) = \iint p(x,y) \ln \frac{p(x,y)}{p(x)\,p(y)} \mathrm{d}x\,\mathrm{d}y\,, \tag{2.8}$$

where $p(x,y)$ is the joint distribution and $p(x)$ and $p(y)$ are the marginal distributions of x and y respectively. Mutual information is zero if and only if x and y are mutually independent (see e.g. a review in Vergara and Estévez, 2013).

A related quantity is the conditional mutual information (Wyner, 1978), which measures the mutual information between two variables x and y given knowledge about additional variables. For three variables, it is defined as

$$\mathrm{MI}(x,y \mid z) = \iiint p(x,y,z) \ln \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,, \tag{2.9}$$

where $p(x,y,z)$ is the joint probability density function of x, y and z. The integral over z reflects that conditional mutual information measures the expected mutual information between x and y given knowledge of the variable z, rather than the mutual information given a particular value of z (MacKay, 2003). Conditioning on more than one variable follows by promoting z to a vector of quantities. Generally, an increase in the mutual information after conditioning on z means that z and x provide synergistic information on y, whereas a decrease in mutual information after conditioning on z means that z provides redundant information to x about y (see e.g. use in Bhattacharjee et al., 2020).

Estimating mutual information requires approximating the joint and marginal distributions in Eq. (2.8) and (2.9). Accurately estimating the joint probability distributions remains a challenge. Early attempts at MI estimation use a binning scheme, where the data is binned into joint histograms, and Eq. (2.8) is transformed into a discrete sum over all the bins. This approach, however, is heavily reliant on the bin width chosen (Fraser and Swinney, 1986; Moon et al., 1995; Darbellay and Vajda, 1999; Kwak and Choi, 2002; Kraskov et al., 2004; Suzuki et al., 2008; Saxe et al., 2019; Holmes and Nemenman, 2019; Pichler et al., 2022). More recent alternatives use estimators based on k-nearest neighbours (Kraskov et al., 2004; Frenzel and Pompe, 2007), kernel density estimation (Rosenblatt, 1956; Moon et al., 1995), or neural networks (Belghazi et al., 2018). Unfortunately, such estimates are still heavily dependent on the hyperparameter choice and do not provide error estimates; this will be discussed in greater

detail in Ch. 3, and leads to the development of a robust mutual information estimator based on Gaussian mixture models (Piras et al., 2023b) which I describe in Sec. 3.2.

Despite the difficulty in accurately measuring mutual information, because it is a well-established information-theoretic metric, it is used both to assess the level of disentanglement, and to interpret neural networks. For example, Chen et al. (2018) propose a metric named the *mutual information gap* to measure the extent to which each latent encodes a unique factor when the factors varied to generate the dataset are known. Sedaghat et al. (2021) interpret the latent representation of a VAE trained to reconstruct stellar spectra though calculating mutual information between latent variables and known stellar physics parameters.

Representation learning coupled with mutual information appears a promising way of gaining novel physical insights on a complex non-linear problem. Before delving into the problem of finding factors governing the halo mass function, however, we need to first gain an understanding of β-VAE-based models, and develop ways of using mutual information to interpret the latent space. In the next chapter, we investigate the use of β-VAE for identifying physically relevant factors on a synthetic dataset where the factors are known, and explore ways of using mutual information to achieve interpretability.

Chapter 3

Preparatory Work

As we saw in the last chapter, to gain physical insights on a complex non-linear problem, a promising approach is to interpret a low-dimensional latent representation learnt by a neural network. For the latent representation to be interpretable, it should have independent components each capturing a different, physically relevant factor that is needed to produce the output. Such disentangled latent representations can be found by models based on the β-variational autoencoder (β-VAE; Higgins et al., 2017), as we saw in Sec. 2.3. This chapter presents preparatory work done to explore the use of β-VAEs for finding disentangled latent representations, and presents the tools we develop for interpreting the latents. These help build the framework we will use in the next chapter to gain physical insights into what determines the cosmology dependence of the halo mass function.

In the first part of this chapter, Sec. 3.1, we use a synthetic dataset generated from a known set of factors to explore how to train β-VAEs so that by interpreting the latents, we identify the relevant factors. We also develop a novel approach to interpret latents by calculating the mutual information between latents and the quantities which we believe are relevant.[1] While we find mutual information is an important tool for achieving interpretability, as mentioned in Sec. 2.4, existing methods to estimate mutual information are highly dependent on hyperparameter choices and lack robustness. This prompted my collaborators to work on a new robust estimator of mutual information, leading to GMM-MI (pronounced 'Jimmie'), the mutual information estimator with Gaussian mixture models presented in Piras et al. (2023b), to which I also contributed. Both the estimator and my contribution are described in Sec. 3.2.

[1] While we developed and tested this approach independently, as noted in Sec. 2.4, using mutual information to interpret latents has been done in Sedaghat et al. (2021); it has since also been used in Lucie-Smith et al. (2022b, 2024a).

3.1 β-VAE on 3D Shapes

β-VAE has been shown to find disentangled latent representations of image datasets in which different latents encode different factors controlling variation (Higgins et al., 2017; Kim and Mnih, 2018). Furthermore, as described in Sec. 2.3.2, encoder-decoder networks trained using β-VAE loss have been shown to learn physically meaningful latent representations (Iten et al., 2020; Lucie-Smith et al., 2022b). Because of these successes, we would like to train an encoder-decoder architecture with β-VAE loss to predict the halo mass function using a disentangled latent representation, and interpret the latents to discover the factors needed to model the halo mass function. To this end, we require an understanding of how to train β-VAE-like models to find disentangled latent representations. We develop this understanding by training β-VAEs on a dataset of synthetic images generated by varying a known set of factors. The synthetic dataset we use is described in Sec. 3.1.1. The architecture of the β-VAE model and the training procedure are in Sec. 3.1.2. Sec. 3.1.3 describes how we calculate mutual information and presents the metric we use to assess the interpretability of the latents. Using this metric, in Sec. 3.1.4 we explore strategies to determine the main hyperparameters impacting interpretability: the number of latent variables, and β. In Sec. 3.1.5, we use mutual information to inspect the latent representation of the best model found, and assess whether the latent representation allows easy identification of the factors. Sec. 3.1.6 draws the conclusions.

3.1.1 Dataset

For this investigation, we choose to use 3D Shapes (Burgess and Kim, 2018), a simple synthetic dataset consisting of images of three-dimensional shapes that Kim and Mnih (2018) use to

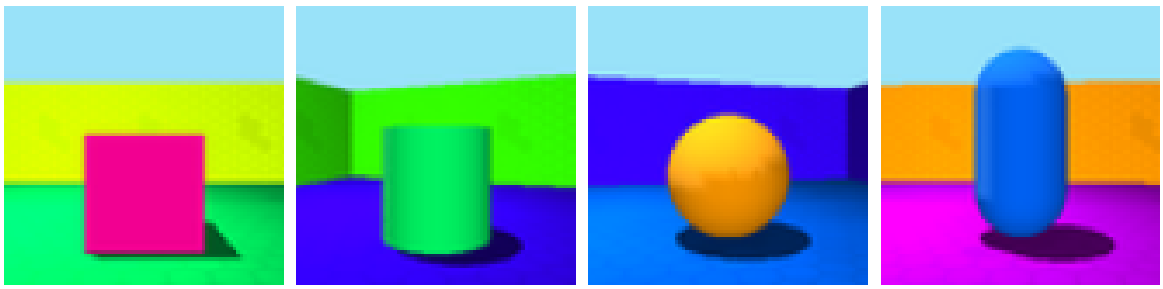


Figure 3.1: Example images from the 3D Shapes dataset. One image is shown for each type of shape. The images shown are chosen to have the same factor value for scale, but otherwise have randomly selected factor values for each of the remaining four factors.

assess disentanglement. It consists of images generated by varying a known set of six factors: shape (4 values), scale (8 values), object hue (10 values), floor hue (10 values), wall hue (10 values), and orientation (15 values). Each combination of factor values is used to generate one RGB image of 64×64 pixels, producing 480000 images in total. The images also have additional features like sky and lighting that are not varied to generate the dataset. Example images from the dataset are shown in Fig. 3.1.

We use this dataset to explore how to train the β-VAE and interpret the latent representation so that we can identify the six factors varied to generate the dataset. This dataset was chosen since it was readily available, generated using known factors, and it allows easy visual identification of which factors change as each latent variable is systematically varied. The last point facilitates assessing whether the β-VAE learnt a disentangled latent representation, and serves as a cross-check of our results obtained using mutual information (more details in the following sections.) As the dataset contains other features that were not varied, it also lets us test if we only recover the six relevant factors when interpreting the latent representation.

3.1.2 β-VAE architecture and training

The 3D Shapes dataset was used in Kim and Mnih (2018) to compare disentanglement of different VAE-variants including β-VAE, and we adopt the same β-VAE architecture they used for our investigation to avoid the need to tune the model architecture. The model takes in RGB images of 64×64 pixels $\times 3$ channels, and the encoder applies four convolutional layers (introduced in Sec. 2.1.2) with 32, 32, 64, and 64 filters respectively. Each filter has size 4×4, and we use a stride of two, i.e. the filter slides across the input in steps of two pixels. The output of the fourth convolutional layer is then flattened and passed to a fully-connected layer of 256 neurons, which outputs the means and variances of L latent distributions. We let the latent space dimensionality L vary. A value is sampled from each latent distribution, forming an input of size L to the decoder. The decoder consists of two fully-connected layers (see Sec. 2.1.1) with 256 and 1024 neurons each, followed by four transposed convolutional layers with 64, 32, 32, and 3 filters respectively. The transposed convolutional layers can be thought of as applying an 'inverse operation' to convolutional layers (more details can be found in e.g. Dumoulin and Visin, 2016); they upsample their inputs, and the final transposed convolutional layer outputs a reconstructed $64 \times 64 \times 3$ RGB image. The decoder has the same filter size and stride as in

the encoder. Outputs of all (transposed) convolutional layers are followed by ReLU activation (Fukushima, 1969; Nair and Hinton, 2010).

We use the β-VAE loss as in Eq. (2.7), and train the neural networks using the `Adam` optimiser (Kingma and Ba, 2017) as implemented in TensorFlow (Martín Abadi et al., 2015) with a learning rate of 10^{-4} and a batch size of 32, following Kim and Mnih (2018).

This investigation has two aims: i) determine how the hyperparameters β and the latent space dimensionality L impact the interpretability of the latent representation, and ii) assess the latent representation of the most interpretable model to determine how well β-VAE allows us to recover the factors. This requires us to tune β and L using a metric of interpretability based on mutual information (see Sec. 2.4) that we develop in the next subsection. The tuning and testing are done by dividing the dataset so that 90% forms the training set used for tuning, and 10% is used to evaluate the final model. The hyperparameters are tuned using *k-fold cross-validation* with $k = 5$. This means the training set is evenly split into five sets; for each combination of L and β, for each fold we take three sets for training and use one set as the validation set to determine early stopping (see Sec. 2.1.3). The trained model's performance is evaluated on the last set. The model is converged if validation loss fails to decrease over five consecutive epochs, or if the training reaches 60 epochs, since by then the rate of training loss decrease is small. Using cross-validation, we step through $L = 6, 8, 10$ and $\beta = 1, 2, 3, 4$ (and 5 for $L = 10$). The latent space dimensionalities are chosen based on the number of known factors: we expect perfectly disentangling all six factors to require at least six latents, while the latent space is already significantly larger than the number of factors when $L = 10$ (empirically we also find a higher L unnecessary). The values of β are chosen as Higgins et al. (2017) and Kim and Mnih (2018) find $\beta > 1$ is required to disentangle the latent space, while by the time β reaches its maximum value in the list, the interpretability of the latent representation (see the next subsection) shows a clear decreasing trend with increasing β.

3.1.3 Interpreting latents using mutual information

As introduced in Sec. 2.4, mutual information is an information-theoretic metric which can be useful for measuring disentanglement and interpreting latents. We use it for both purposes in this investigation, which require calculating the mutual information between factors and latents. For the 3D Shapes dataset where each factor f (e.g. scale) can take on M discrete

values $f_i = f_1, ...f_M$ but the latents are continuous, the mutual information between one latent κ and one factor f is calculated as

$$\mathrm{MI}(f,\kappa) = \frac{1}{M}\sum_{i=1}^{M}\int_{-\infty}^{\infty} p(\kappa \mid f_i)\left[\ln p(\kappa \mid f_i) - \ln \frac{1}{M}\sum_{j=1}^{M} p(\kappa \mid f_j)\right] \mathrm{d}\kappa \tag{3.1}$$

(the derivation can be found in Appendix A.1).

We calculate the mutual information as per Eq. (3.1) in three steps: we first fix the factor value f_i and estimate the probability density $p(\kappa \mid f_i)$, then we evaluate the integral over κ, and finally we sum over the integrals for different factor values. Note that the calculations described in this subsection are not the GMM-MI algorithm, which is developed later in response to the need of a quick and robust mutual information estimator partly uncovered in this investigation.

To estimate $p(\kappa \mid f_i)$, we filter the dataset so only images with a certain factor value, e.g. only images of cubes, are input to the trained β-VAE. For each image, the encoder outputs the mean and variance of the latent distribution, and one sample is drawn from the latent distribution given each image. We turn the resulting discrete distribution into a continuous probability density using *kernel density estimation* (KDE, Rosenblatt, 1956). Kernel density estimation is a non-parametric method of estimating probability densities from a set of discrete samples. Each sample x_i is approximated by a *kernel function*, K, for which we use a Gaussian distribution with a given *bandwidth*, h. The probability density at y is estimated as

$$p_K(y) = \sum_i K(y - x_i; h) \propto \sum_i \exp\left(-\frac{(y-x_i)^2}{2h^2}\right). \tag{3.2}$$

The bandwidth is a hyperparameter that controls the smoothness of the returned probability density function; a higher bandwidth results in a smoother distribution. The bandwidth for KDE can be chosen by k-fold cross-validation on the data. This makes it advantageous compared to other commonly used methods for estimating mutual information, such as the k-nearest-neighbours-based KSG estimator (Kraskov et al., 2004), for which it is unclear how to efficiently optimise the number of nearest neighbours to use (Holmes and Nemenman, 2019). We selected the bandwidth by first taking one discrete distribution and performing a grid search on h; for each value of h we performed five-fold cross-validation. We then visually inspected a comparison between the best fit, which was found to be $h = 0.06$, and a histogram (or scatter

plot) of the discrete distribution for every distribution fitted. Kernel density estimates using $h = 0.06$ fitted other discrete distributions too, so we used a fixed bandwidth for all mutual information calculations in this section.[2] Sometimes, however, it can be difficult to tell if a fit is overly smooth, or not smooth enough. As the bandwidth is the crucial hyperparameter that affects the fitted distribution, this potentially affects the robustness of the mutual information estimation. One could perform grid search combined with cross-validation for each KDE fit, however this becomes computationally inefficient, especially because the bandwidth is continuous so that searching over a fine grid may be needed. These caveats have led to the development of GMM-MI, which will be introduced in Sec. 3.2, and we will revisit some of the calculations in this section using GMM-MI then.

After obtaining $p(\kappa \mid f_i)$ using KDE, we integrate over κ using *Monte Carlo integration*. This involves drawing samples of κ from the distribution $p(\kappa \mid f_i)$, and using the sampled κ to evaluate the terms in the square brackets in Eq. (3.1). The sample-and-evaluate process is repeated a large number of times (10000 times) and the results averaged to obtain the integral. We perform Monte Carlo integration for each of the M factor values $f_i = f_1, \ldots f_M$, and find the average (more on this in Sec. 3.1.4).

3.1.3.1 Metric for disentanglement

As mentioned in Sec. 2.4, literature uses a variety of metrics to assess whether each latent encodes a different factor of disentanglement. Such metrics can involve the training of additional classifiers and therefore introduce additional hyperparameters (e.g. Higgins et al., 2017; Kim and Mnih, 2018), and different metrics can give discrepant results (Carbonneau et al., 2022). We choose to use mutual information, which is well-established in information theory, and derive a metric for interpretability to help us explore the impact that the hyperparameters β and the latent space dimensionality L have on the latent representation.

For the latent space to be interpretable, the latents should have high mutual information with the factors that we know or believe are relevant for the problem, and ideally different factors should be encoded in different latents. Combining both criteria in the end amounts to identifying, for each factor, the highest mutual information it shares with a single latent,

$$\mathrm{MI}(f, \kappa_t)\,, \tag{3.3}$$

[2] We used KDE implemented in `scikit-learn` (Pedregosa et al., 2011).

where κ_t is the latent that has the highest mutual information with f. The total score for the entire model is the sum of the scores for all known factors; we call this the *interpretability score*.

Other commonly used disentanglement metrics based on mutual information between factors and latents, such as the mutual information gap (MIG; Chen et al., 2018), focus solely on quantifying if each factor is encoded in a different latent, but do not penalise models for encoding little information on the factors. Unlike these metrics, our interpretability score penalises models both for encoding little information on a factor, and for encoding a factor in multiple latent variables. We use this metric to find the β-VAE model with the optimal β and number of latent variables for extracting knowledge on the 3D Shapes dataset factors.

3.1.4 Effects of β and latent space dimensionality on interpretability

For the first part of our investigation, we investigate how β and the number of latents impact the latent space's interpretability. As outlined in Sec. 3.1.3.1, for the latent space to be interpretable, we would like the latents to encode high information on the relevant factors, and different latents should encode different factors. The latter is encouraged by finding disentangled latents that are statistically independent of each other.

In Sec. 2.3.1, we saw that disentanglement in β-VAEs is achieved via tuning the single hyperparameter β that changes the weight of the regularisation term. Increasing $\beta > 0$ encourages disentanglement. However, the range of values a latent can take becomes restricted, so latents may encode less information (Burgess et al., 2018).

In addition to β, another important hyperparameter is the number of latents, L. A smaller number of latents makes the task of interpreting the latents more tractable to humans. However, L also determines the amount of total information the latent space can contain and sets an upper limit on the number of factors that can be disentangled. To disentangle all six factors of 3D Shapes, $L \geq 6$.

We investigate whether there are guidelines for choosing the number of latents L and the value of β to achieve high latent space interpretability. As described in Sec. 3.1.2, we train models on a grid of L and β values using five-fold cross-validation, and evaluate the model trained in each fold using the metric described in Sec. 3.1.3.1.

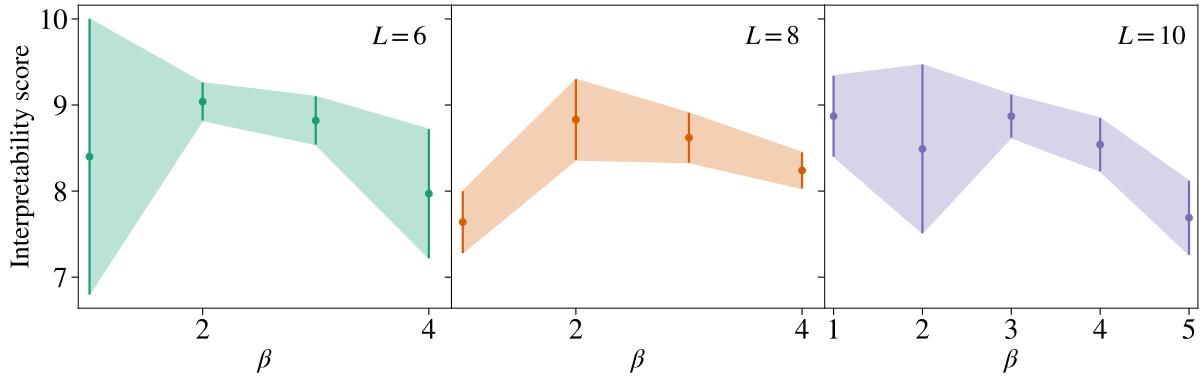


Figure 3.2: Interpretability score (Eq. 3.3 summed over all six factors) as a function of β for β-VAE models trained on 3D Shapes dataset with different numbers of latents L. Dots indicate the mean and error bars indicate one standard deviation of the interpretability score evaluated over five-fold cross-validation (described in Sec. 3.1.2).

The results are presented in Fig. 3.2. The combination of $L = 6$ and $\beta = 2$ has the highest mean interpretability score. Because its error bars overlap with scores of models with $(L, \beta) = (6, 3)$ and $(10, 3)$,[3] to decide the best combination, we train models with all three combinations $(L, \beta) = (6, 2), (6, 3)$ and $(10, 3)$ respectively on the full training set, and evaluate the interpretability score using sample images from the test set (see Sec. 3.1.2). The model with $(L, \beta) = (6, 2)$ scores the highest, so we analyse its latent representation in Sec. 3.1.5.

While it is interesting that the best model we found has the latent space dimensionality equal to the number of factors, Fig. 3.2 shows that $L = 6$ does not generally lead to more interpretable representations compared to other dimensionalities; similar metric scores can be achieved by combining L with different β. There appears to be no clear optimal number of latents that leads to better interpretability. Even if we use L greater than the number of factors, it is still possible to find a disentangled and interpretable latent representation, because regularisation pushes extra latents to resemble the prior and encode no information. The β value giving the highest interpretability score varies with different numbers of latents L, with no clear relationship between β and L. Therefore, there are no general guidelines to finding β and L for achieving high interpretability.

To provide a starting point for tuning β and L, we explore further and find empirically that when training models with $\beta = 0$, the mean square error (MSE) increases if the latent space

[3] Several other β-L combinations also have scores with error bars that overlap with the best performing model of $(L, \beta) = (6, 2)$. However, their error bars are more than twice the error bars of the best model, indicating that the quality of their representation varies considerably depending on the training data used. These parameters therefore give less robust models and are not chosen.

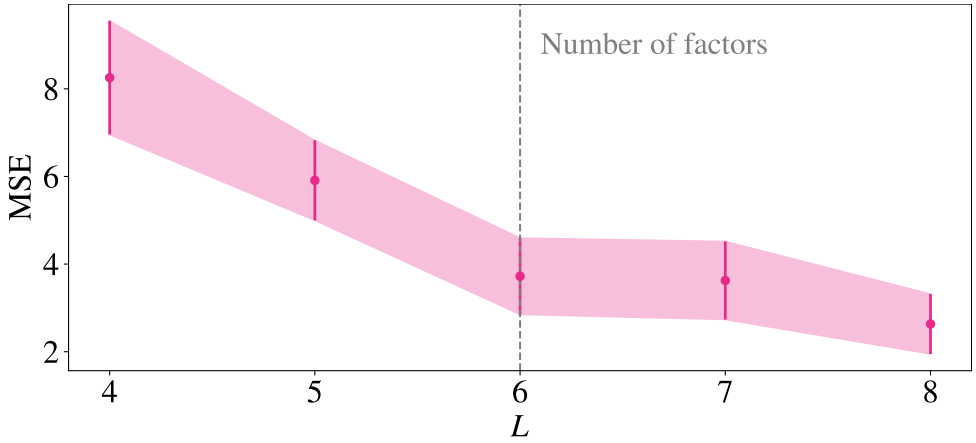


Figure 3.3: Mean squared error (MSE) as a function of the number of latents L for β-VAE models trained on 3D Shapes with $\beta = 0$. Dots indicate the mean and error bars indicate one standard deviation of the MSE evaluated over five-fold cross-validation (described in Sec. 3.1.2). Grey dashed line indicates the number of factors varied to generate 3D Shapes, and therefore the minimum latent space dimensionality required for different latents to encode different factors.

dimensionality becomes lower than the number of factors varied to generate the dataset, as shown in Fig. 3.3. While this is not guaranteed to happen,[4] this empirical observation suggests the following approach when we do not know the exact number of factors. We could start by training $\beta = 0$ models with a large L greater than the expected number of factors, and decrease L until we observe a significant decrease in the model's accuracy. Using this latent space dimensionality, we can then tune β to disentangle the latents. Any further redundant latents can be identified as those for which the encoder essentially returns the prior (a standard unit Gaussian) regardless of input.[5] The approach of fixing L to a number larger than the number of factors and tuning β is taken by existing literature investigating disentanglement on the 3D Shapes dataset; both Kim and Mnih (2018) and Locatello et al. (2019) fix $L = 10$ and vary β to find the most disentangled representation (although their disentanglement metric does not award models for encoding more information as ours does). After identifying the total number of latents needed, β and L can be tuned again to achieve both high accuracy and disentanglement.

[4] Theoretically, in the absence of any regularisation, any data can be encoded on a real line if the encoder and decoder are powerful enough. This rarely happens in practice, however, as there is always some form of regularisation (e.g. model architecture or even gradient descent can be a form of regularisation, Barrett and Dherin, 2020).

[5] On the other hand, if the initial L is very small and we find it practically impossible to disentangle the latents without significantly losing model accuracy, we can consider increasing L.

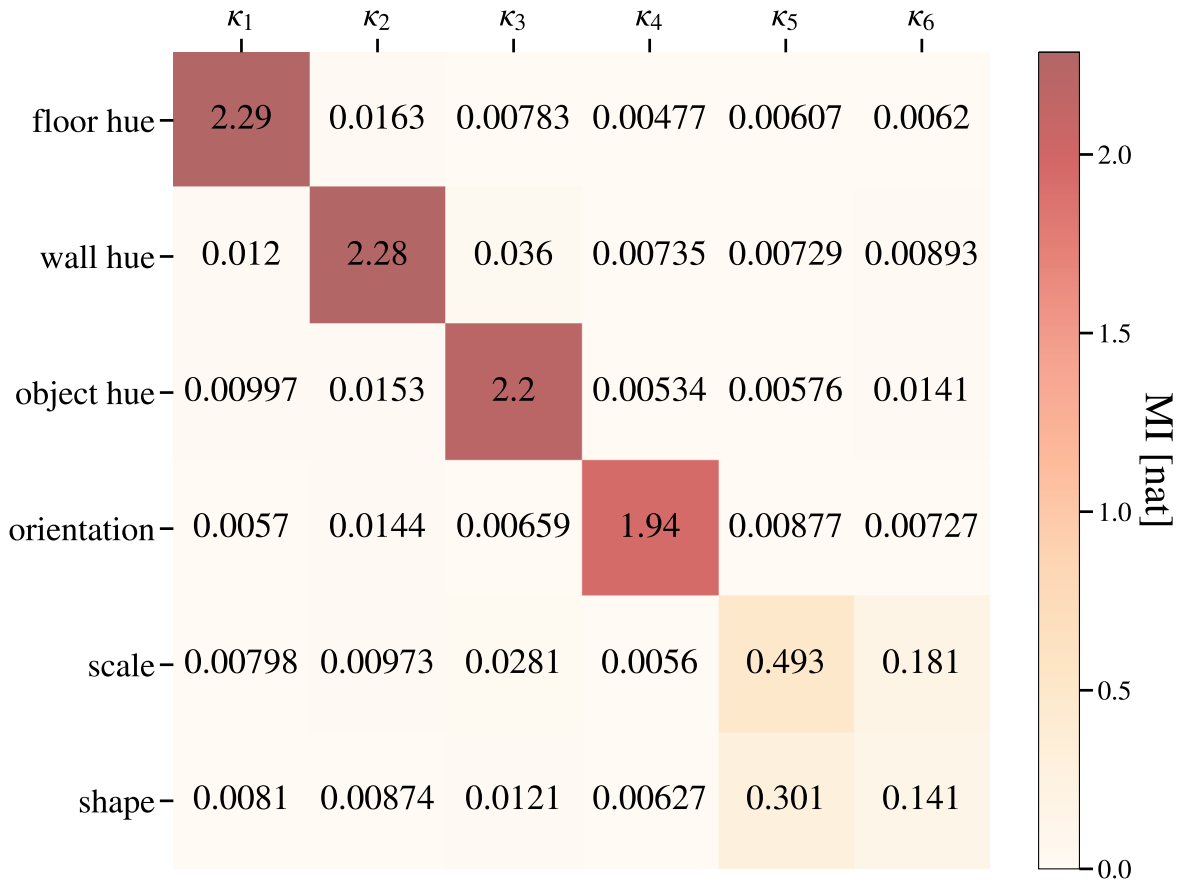


Figure 3.4: Mutual information between each factor varied to generate the 3D Shapes dataset and latents of the trained β-VAE with $L = 6$ and $\beta = 2$. Note that latents have been re-labelled so that the first latent κ_1 has the highest mutual information with a factor, and so on. The interpretability score metric is the sum of the entries with the deepest colour in each row.

3.1.5 The latent representation

From the previous subsection, we find the model trained with six latents and $\beta = 2$ has the most informative and disentangled latent representation. We now assess if the latents indeed learn the relevant factors, if each factor is encoded in a different latent, and if the factors can be identified with mutual information.

Fig. 3.4 shows the mutual information between each factor and each latent κ_i evaluated on the test set, as described in Sec. 3.1.3. We find that the latent space has indeed encoded information on all six data generating factors. The model has disentangled four out of the six factors: floor hue, wall hue, object hue, and orientation. The mutual information between these disentangled factors and their corresponding latents are high so the latent space is informative on these factors. This means even if we only guessed that those factors played a role in

generating the dataset, by interpreting latents κ_1 to κ_4 separately via calculating their mutual information with the guessed factors, we can easily confirm the relevance of these factors.

We find information on scale and shape are entangled together in two latent variables. This agrees with the findings of Kim and Mnih (2018), who used β-VAE as well as FactorVAE (mentioned at the end of Sec. 2.3.1) to disentangle the 3D Shapes dataset. Even though Kim and Mnih (2018) found a best β-VAE model that was able to disentangle all data generation factors, this was highly dependent on the random seed used; across many random seeds, they found that both β-VAE and FactorVAE struggled to disentangle the factors for shape and scale.

Although not shown, we have also cross-checked the factors we have identified in each latent from Fig. 3.4 with a visual inspection of the latent traversals, where we generate images by systematically varying one latent at a time while holding other latents fixed. The two methods agree, confirming that mutual information allows us to identify factors encoded in the latents.

3.1.6 Conclusions

Our investigation of training β-VAEs on the synthetic 3D Shapes dataset shows that β-VAEs can learn meaningful latent representations that capture information on all six factors varied to generate the dataset. By interpreting the latent representation using mutual information, we are able to identify these factors, and therefore extract knowledge on which factors are needed to produce the model's outputs. While there appears to be no general guideline for finding β and the number of latents L, we may tentatively start by determining L as the minimum number of latents below which model accuracy significantly worsens, and then proceed to tuning β and fine-tuning both parameters.

Using these insights, we will employ an encode-decoder architecture trained with β-VAE loss to find a low-dimensional latent representation of the information required to model the halo mass function. We will then interpret the latent representation using mutual information.

Before this, as mentioned in Sec. 3.1.3, we find a need for a robust mutual information estimator in order to achieve interpretability during our investigation on 3D Shapes. In our investigation we used kernel density estimation, but the bandwidth controlling the smoothness of the probability density function was chosen by combining grid search with cross-validation on one fit, then assessed by eye for fitting other distributions. Performing a grid search combined with cross-validation for every single fit can be computationally expensive, especially if we

plan to use mutual information to inspect the information that latents encode on functions (we will do this in Ch. 4), which can involve tens to hundreds of mutual information evaluations.

As described in Sec. 2.4, most existing methods for calculating mutual information rely heavily on hyperparameter choice, and they also do not quantify the uncertainty in the mutual information estimate. Because we will be relying on mutual information to interpret the latents, a lack of robustness in the mutual information calculation could impact our latent interpretation. To address this, my collaborators developed a robust mutual information estimator, which we will use for interpreting factors governing the halo mass function in Ch. 4. The estimator is described in the next section along with my contributions towards its development.

3.2 Towards a robust estimator of mutual information – GMM-MI

From our investigation on the 3D Shapes dataset in Sec. 3.1, we find that mutual information provides a powerful tool for identifying the information encoded in the latent representations. Because it is an important quantity for achieving interpretability, we require a method of robustly calculating mutual information that provides error estimates and is also computationally efficient; as mentioned in Sec. 2.4, existing methods do not satisfy all these criteria. To address this need, Piras et al. (2023b) develop the GMM-MI algorithm. It estimates the joint distribution $p(x,y)$ in Eq. (2.8) by fitting Gaussian mixture models, i.e. the data are modelled as being drawn from a finite sum of Gaussian distributions. GMM-MI also estimates the uncertainty due to finite sample size via bootstrapping, which enables us to assess the statistical significance of any trends in mutual information. Sec. 3.2.1 describes the GMM-MI algorithm in greater detail. Sec. 3.2.2 describes my contribution towards the development of GMM-MI, and Sec. 3.2.3 shows GMM-MI applied to the best model found on 3D Shapes in Sec. 3.1 to check and refine the results in Fig. 3.4.

3.2.1 Description of the GMM-MI algorithm

The mutual information (MI) between two continuous variables x and y is given in Eq. (2.8), which I reproduce here for convenience:

$$\mathrm{MI}(x,y) = \iint p(x,y) \ln \frac{p(x,y)}{p(x)\,p(y)} \mathrm{d}x\,\mathrm{d}y \qquad \text{(2.8 reproduced.)}$$

As described earlier, GMM-MI fits $p(x,y)$ using a Gaussian mixture model of c components, i.e. a weighted sum of c Gaussian distributions $\mathcal{N}$ each having a mean of μ_i and a covariance of Σ_i,

$$p(x,y \mid \theta) = \sum_{i=1}^{c} w_i\, \mathcal{N}(x,y \mid \mu_i, \Sigma_i)\,, \tag{3.4}$$

where θ refers to the set of weights w_i along with the means μ_i and covariances Σ_i for all c components of the Gaussian mixture model. The marginals $p(x)$ and $p(y)$ are then also Gaussian mixture models with parameters given by θ; by fitting the joint distribution we obtain all distributions needed in Eq. (2.8). To estimate the conditional mutual information in Eq. (2.9), GMM-MI uses an alternative form, which for three variables reads

$$\mathrm{MI}(x,y \mid z) = \iiint p(x,y,z) \ln \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,. \tag{3.5}$$

GMM-MI fits the three-dimensional $p(x,y,z)$ with a Gaussian mixture model, which then provides all the distributions in Eq. (3.5). When conditioned on more than one variable, z in Eq. (3.5) is promoted to a vector, and GMM-MI similarly fits the multi-dimensional joint distribution.

Fig. 3.5 summarises the algorithm GMM-MI uses to obtain mutual information estimates. The algorithm essentially comprises two parts. The first part fits the joint distribution by optimising the Gaussian mixture model parameters θ using an expectation-maximisation algorithm and determining the number of components. The second part uses the optimal Gaussian mixture model parameters and performs bootstrapping to evaluate the mutual information and obtain an error estimate. I describe each of the two parts in greater detail below, using evaluation of the mutual information in Eq. (2.8) as an example (conditional mutual information is evaluated in the same way).

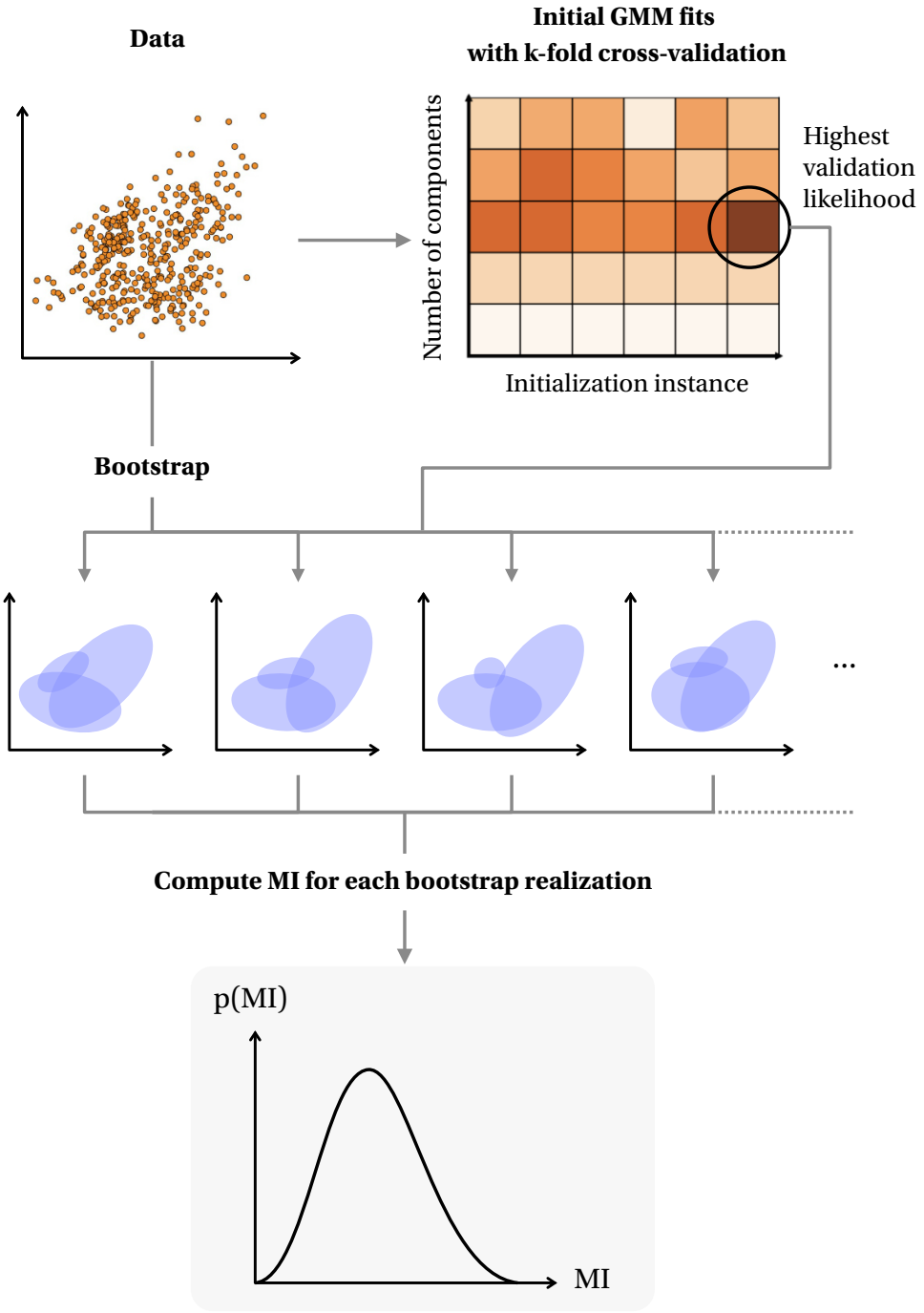


Figure 3.5: Flowchart summarising GMM-MI from figure 1 in Piras et al. (2023b); details of the steps are given in Sec. 3.2.1.

Fitting the joint distribution: For a given number of components c, GMM-MI randomly initialises n_{init} different Gaussian mixture models. For each initialisation, the initial values of θ can be obtained through a variety of different initialisation schemes. The default scheme used

in this thesis determines μ_i, Σ_i, and w_i for the i-th component by first randomly assigning each sample in the dataset with a probability that it belongs to the i-th component. The mean μ_i and covariance Σ_i are then calculated as the sample mean and covariance matrix of all samples in the dataset, weighted by the assigned probabilities. The initial weight w_i is the average of the assigned probabilities. This initialisation scheme (called "random-sklearn" in GMM-MI) is quick, though sometimes it can produce overlapping components that then cause training to become stuck in local optima (discussed further in Sec. 3.2.2). When this happens, alternative initialisation schemes such as k-means clustering (Lloyd, 1982) can be used. Because the optimisation of Gaussian mixture models can become stuck in local minima, having multiple initialisations ($n_{\text{init}} > 1$) is essential to mitigate this issue (see e.g. Baudry and Celeux, 2015).

For each initialisation, the Gaussian mixture model is fit to data using a k-fold cross-validation process, where $k-1$ folds of data are used for training and the remaining fold is used for evaluation. GMM-MI defaults to $k = 2$ to ensure there are enough samples in each fold so the distribution in each fold is representative of the data; we tested that using a higher $k = 3$ does not affect the mutual information estimates beyond 2σ, where σ here is the GMM-MI estimated error (more details below). The model is fit using an expectation-maximisation algorithm, which iteratively adjusts the parameters θ in Eq. (3.4) to maximise the expected log-likelihood on the data (Dempster et al., 1977). The fitting terminates once the change in log-likelihood becomes smaller than a threshold value. The threshold is chosen such that the GMM-MI fit matches the data visually, and is usually 10^{-6}; decreasing this by an order of magnitude does not significantly improve the result in most cases (see however Sec. 3.2.2). To avoid singular matrices, a small regularisation constant (for which I use the default value 10^{-15}) is added to the diagonals of the covariance matrix. The model with the highest mean validation log-likelihood across all folds generalises the best, so it is chosen as the best model given c components.

The number of Gaussian mixture model components plays a similar role to the bandwidth for KDEs, described in Sec. 3.1.3. However, it is discrete (unlike bandwidth), so we determine it by iteratively increasing c and performing the initialise-and-fit process, until further increasing the number of components does not improve the mean validation log-likelihood above a threshold (I use the default component-threshold of 10^{-5}). This avoids overfitting the data with too many components and also avoids increased runtime when performing the bootstrap

step (discussed below) due to an overly complex model. Other criteria for choosing c are also implemented in GMM-MI; Piras et al. (2023b) show that although these can lead to different numbers of components, the mutual information estimates agree to within the errors GMM-MI estimate by bootstrapping (see below).

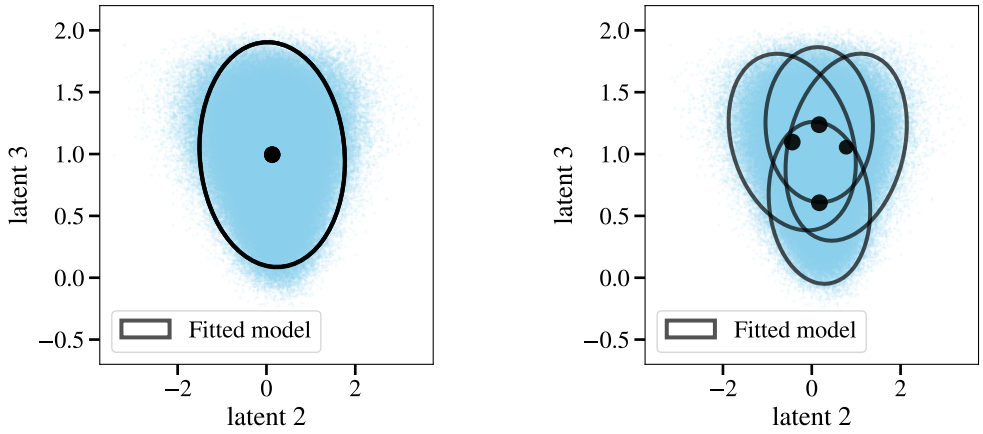
The best overall Gaussian mixture model has the chosen number of components, and the parameters $\hat{\theta}$ are found from the fold with the highest validation log-likelihood.

Estimating mutual information and its uncertainty: Once the Gaussian mixture model is determined, the data is bootstrapped n times and the Gaussian mixture model is fit to each bootstrapped realisation of the data. Each fit is initialised with parameters $\hat{\theta}$. For each fit, mutual information is calculated by Monte Carlo integrating Eq. (2.8): the integral is the mean over M evaluations of $\ln \frac{p(x,y)}{p(x)p(y)}$, where each evaluation uses samples of x and y drawn from the fitted Gaussian mixture model. This gives a distribution of the mutual information values, and GMM-MI returns the sample mean and standard deviation of the distribution. The number of bootstraps n and the number of samples used for Monte Carlo integration is chosen to be subdominant to the uncertainty due to finite sample size; we use $n = 50$ and 10^5 samples for Monte Carlo integration, as increasing these by an order of magnitude did not significantly change GMM-MI results on test cases performed.

3.2.2 Contribution towards understanding the impact of initialisation

A known issue of Gaussian mixture models is that depending on the initialisation, their optimisation can become stuck in local optima, resulting in a poor fit. This issue can be mitigated through multiple initialisations ($n_{\text{init}} > 1$), and/or using several initialisation schemes (see e.g. Baudry and Celeux, 2015; Shireman et al., 2016). I encountered this issue as one of the first users of GMM-MI, when I used it to measure the mutual information between two latent variables of a prototype model trained to predict the halo mass function (the final model for predicting the halo mass function is presented in the next chapter in Sec. 4.1).

In some cases, I found the fitted Gaussian mixture model and the resulting mutual information estimate can vary significantly depending on the initialisation scheme. Fig. 3.6 shows an extreme case where two different initialisations lead to mutual information estimates differing by an order of magnitude, from $\mathcal{O}(10^{-3})$ nat to $\mathcal{O}(10^{-2})$ nat, and the difference is not accounted for by the error estimate from bootstrap. Fig. 3.6a shows the fit obtained using



(a) Random-sklearn initialisation. Validation log-likelihood: -1.443; MI $= (21 \pm 0.9) \times 10^{-4}$ nat

(b) k-means initialisation. Validation log-likelihood: -1.409; MI $= (22 \pm 0.3) \times 10^{-3}$ nat

Figure 3.6: Sample distributions of two latent variables fitted with the best Gaussian mixture model that GMM-MI found using two different initialisation schemes. *Left:* random-sklearn initialisation, described in Sec. 3.2.1. *Right:* initialisation using a k-means clustering algorithm (Lloyd, 1982). Black dots mark the means of each component and ellipses indicate the 95% contour. In both cases the best Gaussian mixture model has four components. The right initialisation leads to a better fit both visually and in terms of validation log-likelihood.

the default initialisation scheme "random-sklearn" (described in Sec. 3.2.1). There are four near-identical components that are visually indistinguishable, and the validation log-likelihood barely improved from using one component to four components (it only increased by 1.8×10^{-6} to a final value of -1.443). On the other hand, using a different initialisation scheme which utilises a k-means clustering algorithm (Lloyd, 1982) provided a better fit: this can be confirmed visually in Fig. 3.6b, and the validation log-likelihood of -1.409 is also higher.

Both cases in Fig. 3.6 are fitted to a large number of latent samples (400000 samples) using $n_{\text{init}} = 5$ and 3-fold cross-validation for each initialisation; each fit terminates if the log-likelihood does not improve by more than 10^{-5}, and the number of components c is determined using a component-threshold of 10^{-5}. The error in mutual information is estimated by bootstrapping the samples 100 times, and each Monte Carlo integration uses 10^5 samples. For the fit in Fig. 3.6a, I tried changing the number of random initialisations n_{init} from 5 to 10, and decreasing the fit-threshold from 10^{-5} to 10^{-7}; these did not significantly affect the results, but further lowering the fit-threshold to 10^{-10} improved the fit and brought the mutual information estimate into agreement with Fig. 3.6b.

With further experimentation, we find a poor fit of highly overlapping components tends to happen when the Gaussian mixture model is fitted to a large number of $\sim \mathcal{O}(10^5)$ samples. For a poor final model, during cross-validation the log-likelihood of each fit barely improves and the fit terminates after one or two iterations, a clear sign that the optimisation became stuck. GMM-MI now gives the user a warning when all fits in a cross-validation converged after their second iteration. My finding also highlights the importance of visually inspecting the fits using plots like Fig. 3.6 to check that GMM-MI uses a sensible model to produce the mutual information estimate.

When GMM-MI finds a poor fit, decreasing the fit-threshold helps to overcome this issue.[6] Alternatively, as Fig. 3.6 demonstrates, choosing an alternative initialisation scheme can also mitigate this issue. Because k-means clustering can lead to fits becoming stuck in local optima, we still use random-sklearn initialisation as the default scheme; we use the warning message together with visual inspection of the fits to decide if any mutual information estimates needs revisiting due to a poorly-fitted model.

3.2.3 Application of GMM-MI to 3D Shapes results

Piras et al. (2023b) show GMM-MI returns unbiased estimates for toy problems where the ground truth mutual information can be analytically calculated, and in those scenarios it performs equally well or better than established mutual information estimators. The paper also shows the uncertainty returned by GMM-MI decreases with the number of data samples as $\propto 1/\sqrt{N}$, agreeing with expectations. Apart from occasional initialisation issues which can be resolved as described in Sec. 3.2.2, Piras et al. (2023b) find GMM-MI is robust to hyperparameters. GMM-MI therefore serves as a valuable tool we can use to interpret latent representations learnt by neural networks.

Using GMM-MI, we revisit interpreting the best model trained on 3D Shapes (see Sec. 3.1.5). Fig. 3.7 shows the mutual information between each latent and each factor calculated using GMM-MI. Where mutual information > 0.01 nat, I also show the error estimated using GMM-MI (errors not shown are $\mathcal{O}(10^{-4} - 10^{-5})$). Comparing with Fig. 3.4, the results are qualitatively the same: the model disentangles four out of the six factors, and shape and scale are entangled. Looking more closely, the mutual information estimates along the diago-

[6] This is demonstrated in GMM-MI's GitHub repository in the Walkthrough Jupyter notebook.

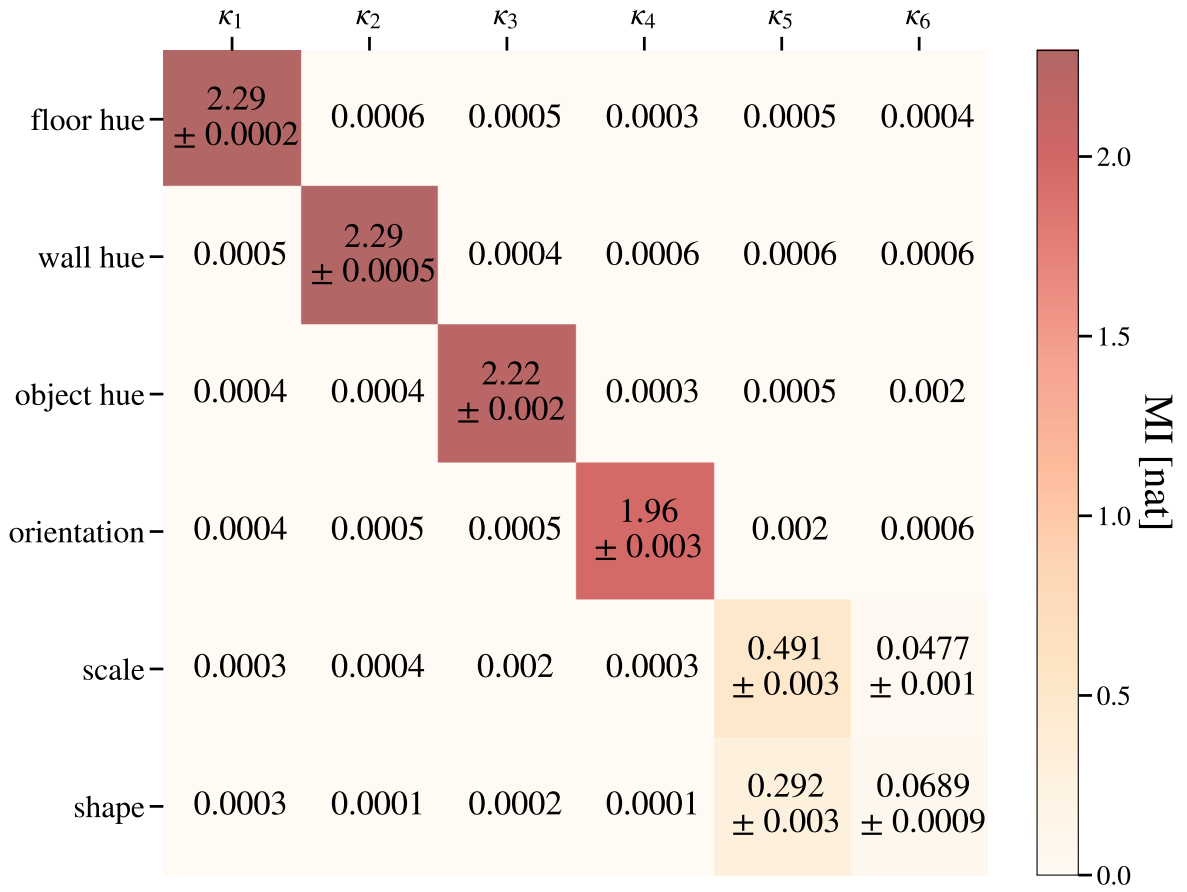


Figure 3.7: Mutual information between each factor varied to generate the 3D Shapes dataset and latents of the trained β-VAE as in Fig. 3.4 (see Sec. 3.1.5), but calculated using GMM-MI instead of using kernel density estimation. The mutual information estimates are the same as in figure 5 in Piras et al. (2023b), but I show here the error estimates returned by GMM-MI when the mutual information is above 0.01 nat (errors not shown are $\mathcal{O}(10^{-4} - 10^{-5})$ nat), and I re-label the latents as per Fig. 3.4..

nals for the four disentangled factors (floor hue, wall hue, object hue, and orientation) agree to two significant figures with the estimates I previously obtained using kernel density estimation; the mutual information between latent κ_5 and shape and scale respectively also agree well with previous estimates. Where latents encode low information on factors, e.g. κ_6 on scale, or most off-diagonal elements in Fig. 3.7, kernel density estimation generally leads to higher mutual information estimates than GMM-MI. This may be because we used a fixed bandwidth for all estimates (see Sec. 3.1.3) as opposed to optimising the bandwidth for each fit, and the chosen bandwidth may have overfitted where mutual information is low.

Using GMM-MI, we also check that the mutual information between two latent variables directly reflects how disentangled they are. We calculate the mutual information between pairs

of latent variables, and find this is below 10^{-4} nat for all pairs except between κ_5 and κ_6, which both encode information on scale and shape: $\mathrm{MI}(\kappa_5, \kappa_6) = (0.04 \pm 0.01)$ nat. Therefore, we can assess if the latent representation is disentangled without knowing any factors by directly measuring mutual information between all pairs of latents; we will use this method for assessing the latent representation of a model trained to predict the halo mass function in the next chapter.

3.3 Conclusions

In this chapter, we trained an encoder-decoder architecture using the β-VAE loss to reconstruct images from a synthetic 3D Shapes dataset produced by varying a known set of factors. We find the model can produce interpretable latent representations, where the latents encode high information on the factors and each factor is encoded in a different latent. While we did not find any general guidelines for choosing β and the number of latents L, we find empirically that we can estimate L as the smallest latent space dimension below which the model's accuracy significantly degrades. This makes it possible to first determine L while holding $\beta = 0$, and then optimise β and fine-tune both parameters further. We also find that without knowing the factors, we can still assess if latents are disentangled by calculating mutual information between pairs of latents. Then, as Sec. 3.1.5 shows, we can extract knowledge on the relevant factors by interpreting each latent through calculating its mutual information with candidate quantities that may be relevant to forming an explanation.

Mutual information plays a key role in this framework of extracting knowledge from a low-dimensional latent representation learnt by a deep learning model. To obtain robust mutual information estimates with errors, my collaborators developed GMM-MI (Piras et al., 2023b). As one of the earliest users of GMM-MI, I have contributed to its early test and application, leading to its improved reliability which is crucial for the use of mutual information for knowledge extraction. I have found that especially when GMM-MI is used to fit a large dataset, the Gaussian mixture model can become stuck in local optima, resulting in unreliable mutual information estimates. We find we can identify these cases both by monitoring the number of iterations in each fit of the Gaussian mixture model, and by visually inspecting the final fit. For those cases, we find the problem can be solved by using different initialisation schemes, or by requiring each fit to only terminate if log-likelihood ceases to improve by a smaller value. We

applied GMM-MI to our model trained on 3D Shapes and found agreement with the results in Sec. 3.1.5, but the errors on mutual information due to finite sample size are now quantified.

GMM-MI provides us with a robust mutual information estimator we require to complete our framework. In the next chapter, using the insights and tools developed in this chapter, we apply our framework to gain insights on what is required to model the cosmology dependence of the halo mass function.

Chapter 4

Deep learning insights into the non-universal halo mass function at z=0

As discussed in Ch. 1, the halo mass function is strongly sensitive to cosmological parameters. This makes galaxy cluster number-counts one of the main cosmological probes in current and near-future surveys (e.g. DES, The Dark Energy Survey Collaboration, 2005; eROSITA, Hofmann et al., 2017; *Euclid*, Sartoris et al., 2016; and LSST, Ivezić et al., 2019). Placing tight cosmological constraints using survey data requires modelling the halo mass function to 1% accuracy or better (Sartoris et al., 2016; McClintock et al., 2019b; Euclid Collaboration et al., 2023a) over a high-dimensional cosmological parameter space to probe the physics of dark energy and neutrinos (The Dark Energy Survey Collaboration, 2005; LSST Science Collaboration et al., 2009; Laureijs et al., 2011; Amendola et al., 2018).

We also motivated in Sec. 1.4 to 1.6 the need to account for non-universality when modelling the halo mass function. Recall from Sec. 1.4.3 that the halo mass function is often referred to as 'universal' if all of its cosmology and redshift dependence is accounted for by the mass variance $\sigma(M,z)$, which is calculated from the linear matter power spectrum $P(k,z)$ as per Eq. (1.59). The multiplicity function $f(\sigma)$ in Eq. (1.62) is therefore independent of cosmology or redshift for a universal halo mass function. Ignoring non-universality can introduce errors of a few percent or higher depending on the cosmological parameter space.

A promising way to model the (non-universal) halo mass function to the required level of accuracy is through emulators, introduced in Sec. 1.6.3.3. However, emulators do not offer physical insights into what determines the halo mass function and, in particular, what is driving its additional cosmology and redshift dependence beyond universality. This lack of under-

standing both limits the generalisability of emulators and their efficient training. Emulators are trained on a finite number of simulations placed in a high-dimensional cosmological parameter space, so understanding the factors determining the halo mass function would enable designing emulator training sets and/or architectures that cover these spaces with greater accuracy.

Studies investigating the source of non-universality point to the growth history and the shape of the power spectrum as relevant factors, as summarised in Sec. 1.6.1 and 1.6.2. We also saw that semi-analytical fitting functions based on the extended Press-Schechter (EPS) formalism but incorporating additional redshift or cosmology dependence have been proposed, but their accuracy over a large cosmological parameter space is not yet well-tested.

In this chapter, we use the deep learning method developed in the previous chapter to shed new light on this problem. We use a β-VAE-based model called the *interpretable variational encoder* (IVE, introduced in Sec. 2.3.2; Iten et al., 2020; Lucie-Smith et al., 2022b, 2024a) to learn a disentangled low-dimensional latent representation that encodes all the information needed to predict the halo mass function over a cosmological parameter space, and interpret the latents using mutual information (MI) calculated with GMM-MI (Piras et al., 2023b), introduced in Sec. 3.2. The IVE takes as inputs the linear matter power spectrum $P(k, z = 0)$ and optionally the linear growth function $D(z)$ (introduced in Sec. 1.2), and reproduces halo mass functions at $z = 0$ generated by the AEMULUS emulator (McClintock et al., 2019b) to within 0.25% residuals for $M = 10^{13.2-15}\,h^{-1}\mathrm{M}_{\odot}$. We focus on understanding how to model the cosmology dependence of the halo mass function at a single redshift $z = 0$ first, and leave modelling both the redshift and cosmology dependence for future work, outlined in Ch. 5. The work in this chapter is published in Guo et al. (2024).

Our IVE approach enables us to determine whether the growth history (given through $D(z)$) provides additional independent information to $P(k, z = 0)$ for predicting the halo mass function. It also enables us to establish the dimensionality of the latent space required, and provides us with a disentangled latent representation which we can interpret through MI to extract physical knowledge. This approach incorporates minimal prior assumptions, thereby providing a new perspective on what drives the non-universality of the halo mass function that is difficult to obtain using existing approaches.

This chapter is organised as follows. Sec. 4.1 describes the IVE model. In particular, the training data is described in Sec. 4.1.1, the model architecture in Sec. 4.1.2, and the loss

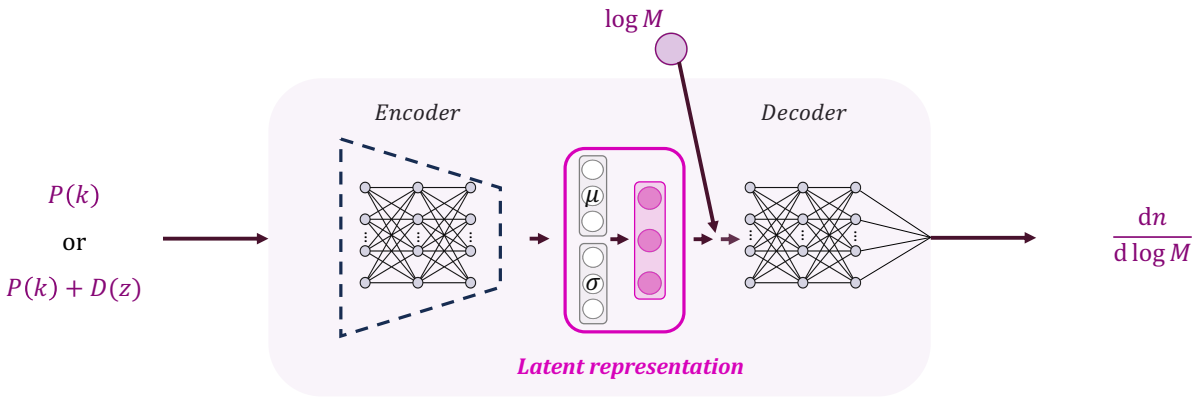


Figure 4.1: The interpretable variational encoder (IVE) consists of an encoder compressing the input linear matter power spectrum $P(k)$ at $z = 0$ (or $P(k)$ concatenated with the linear growth function $D(z)$) into a low-dimensional latent representation, followed by a decoder that maps the latent representation and a given halo mass $\log M$ into the differential halo number density $\frac{\mathrm{d}n}{\mathrm{d}\log M}$. The latent representation encodes the information required to predict the halo mass function at $z = 0$ as independent components, allowing us to interpret each component to gain physical understanding.

function and training procedure in Sec. 4.1.3 and Sec. 4.1.4, respectively. We present results on the dimensionality of the latent space and inputs required to model the halo mass function in Sec. 4.1.5 and 4.1.6 respectively. We interpret the latent representation in Sec. 4.2, in particular focussing on understanding the source of non-universality in Sec. 4.2.4. We discuss our results and conclude in Sec. 4.3.

4.1 The interpretable variational encoder

We investigate the origin of the universal and non-universal information in the halo mass function using the interpretable variational encoder (IVE) model (Lucie-Smith et al., 2022b, 2024a); a schematic is shown in Fig. 4.1. The IVE learns to predict the halo mass function at $z = 0$ given the input for a range of cosmologies. In doing so, it generates a low-dimensional latent representation, which captures all the information in the input that is required to predict the halo mass function. Interpreting the latent space allows us to gain insights on how independent factors controlling the halo mass function relate to cosmological parameters, and how these relate to non-universal information (this will become clearer in Sec. 4.2).

Since predicting the non-universal halo mass function requires information beyond $\sigma(M)$, we provide as inputs to the IVE the linear matter power spectrum $P(k, z = 0)$ (the $z = 0$ argument

will be dropped from now on for brevity). Because Courtin et al. (2010); Ondaro-Mallea et al. (2021); Euclid Collaboration et al. (2023a) find that non-universality is related to the growth history, we optionally provide as input the linear growth function $D(z)$ too. This allows us to investigate whether $D(z)$ contains additional information about the halo mass function over the information already present in $P(k)$. The training and test data are described in Sec. 4.1.1.

The IVE, as described earlier and in Sec. 2.3.2, consists of an encoder-decoder architecture that learns a compressed latent representation, which we interpret to gain physical insights. Details on the architecture are described in Sec. 4.1.2. For the latent representation to be interpretable, we require it to be *disentangled*, i.e. each latent variable encodes an independent factor of variation in the halo mass function. Following Iten et al. (2020); Lucie-Smith et al. (2022b), and the work done in the previous chapter, we achieve disentanglement by training the IVE using the β-VAE loss in Eq. (2.7). Details on the training procedure are given in Sec. 4.1.4.

4.1.1 Dataset

We produce the halo mass functions to train our IVE models using the state-of-the-art halo mass function emulator AEMULUS (McClintock et al., 2019b), introduced in Sec. 1.6.3.3. AEMULUS can achieve 1% precision over its seven-dimensional cosmological parameter space varying $\Omega_b h^2$, $\Omega_c h^2$, w_0, n_s, $\ln 10^{10} A_s$, H_0, and $N_{\rm eff}$. Using AEMULUS is particularly convenient for a first exploration of the cosmology dependence of halo mass functions, because it can generate unlimited training samples. In contrast, the largest available simulation suites only have data for a few thousand cosmological parameter samples (Villaescusa-Navarro et al., 2020), which may limit the accuracy our IVE model can achieve. The cosmological parameter space we use to produce our data is described in Sec. 4.1.1.1, and Sec. 4.1.1.2 describes the construction of the ground truth halo mass function outputs.

To test whether growth history information is required in addition to $P(k)$, we train two IVEs: one is given only $P(k)$ as the input, and the other is additionally given $D(z)$ as input. We compare the prediction accuracy of the two IVEs; if growth history provides additional relevant information, the IVE also given $D(z)$ should predict the ground truth halo mass function to higher accuracy. The IVE inputs are described in Sec. 4.1.1.3.

We additionally train a model to predict the Tinker et al. (2008) halo mass function introduced in Sec. 1.6.2.1, which is universal with cosmology. In Sec. 4.2, comparing the informa-

tion in the latent space of the Tinker IVE model to that of the AEMULUS IVE model helps with identifying the non-universal information captured by the AEMULUS IVE latent space.

4.1.1.1 Cosmological parameters

To ensure we train on reliable halo mass functions, we choose cosmological parameter samples for our dataset to lie within the domain of validity of AEMULUS. Following the approach that DeRose et al. (2019) used to construct the AEMULUS training set, we project a 7D Latin hypercube into the cosmological parameter space such that the resulting samples cover approximately the $\pm 3\sigma$ region allowed by CMB+BAO+SNIa (Anderson et al., 2014), and follow degeneracies among cosmological parameters. Details on the cosmological parameter sampling process are in Appendix B.1. Using this process, we generate a dataset with 10^5 cosmological parameters samples, where 48000 samples are used for the training set, 12000 samples for the validation set, and 40000 samples are used for the test set. Fig. 4.2 shows in blue the cosmological parameter samples we generate, and in black the AEMULUS training cosmologies.[1] Note that our samples always fall within the region covered by AEMULUS training cosmologies, so the emulator never has to extrapolate when generating our ground truth halo mass functions.

4.1.1.2 Outputs to the IVE: query ground truth pairs

For each cosmological parameters sample in our dataset, we initially evaluate the halo mass function in units of $(h\text{Mpc}^{-1})^3$ on a grid of 500 mass points linearly spaced between $\log(M/h^{-1}\text{M}_\odot) = [13.2, 15.0]$[2] using AEMULUS (recall that AEMULUS uses the M_{200b} mass definition). The mass range is chosen where AEMULUS is most reliable (see Sec. 1.6.3.3). We use the same mass range for all cosmologies since this mass range is covered by the binned halo counts from all AEMULUS training simulations at $z = 0$.[3] Fig. 4.3 shows the set of halo mass functions we generate using AEMULUS for training; the left column shows the ground truth halo mass functions, which can vary by $\gtrsim 100\%$ depending on cosmology. The right column shows $f(\sigma)$ as a function of $\sigma(M)$. In particular, the lower right panel shows the distribution of fractional variations in $f(\sigma)$ for the training set. The effect of non-universality is

[1] Accessed from https://github.com/tmcclintock/Aemulus_data.

[2] Recall we use log for logarithms to base 10, and ln for natural logarithms.

[3] The binned halo data were accessed from https://github.com/tmcclintock/Aemulus_data/tree/master/aemulus_data/mass_functions.

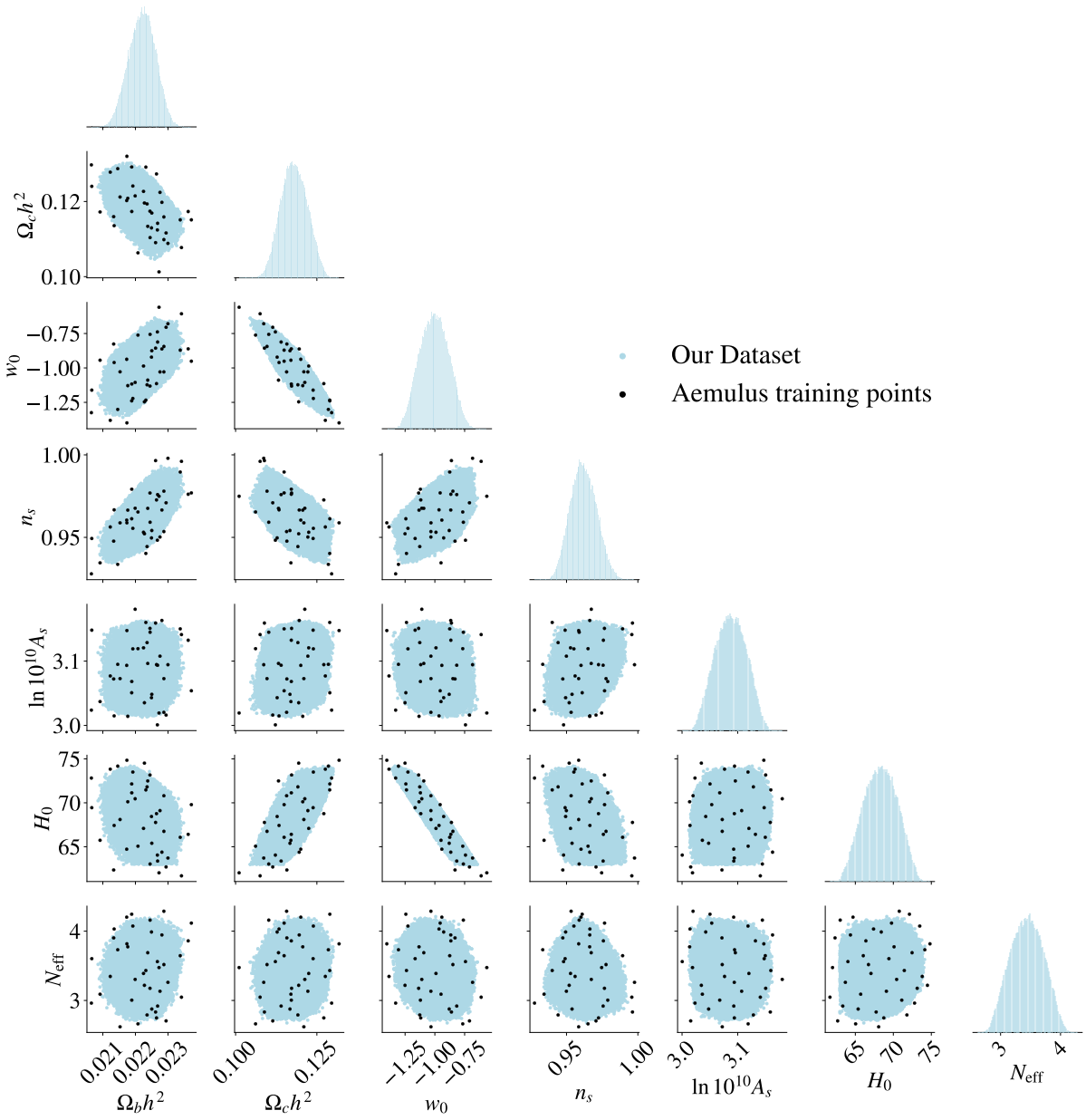


Figure 4.2: Cosmological parameter space covered by the dataset we use to train and test our IVE models (blue). This completely lies within the convex hull of the black points used to train the AEMULUS emulator (DeRose et al., 2019; McClintock et al., 2019b) to ensure we do not extrapolate the emulator beyond its domain of validity.

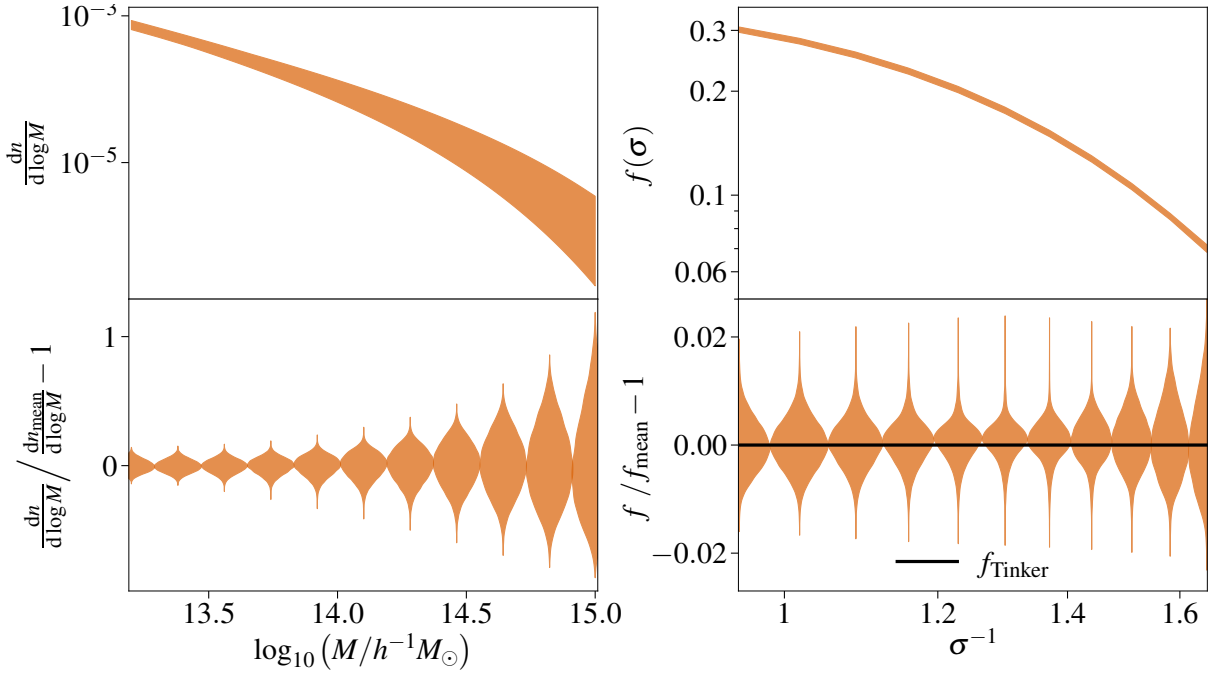


Figure 4.3: *Left:* Ground truth halo mass functions we train our IVE models on, produced using the AEMULUS halo mass function emulator (McClintock et al., 2019b). The shaded region in the upper panel shows the minimum to maximum range of the halo mass functions. The lower panel shows a violin plot of the distribution of fractional differences in the mass functions with respect to that of the mean cosmology (see Sec. 4.1.1.2). *Right:* The same halo mass functions as on the left, but plotting the $f(\sigma)$ component from Eq. (1.62) as a function of $\sigma(M)$. The Tinker mass function is shown in black in the lower right panel for reference; in contrast, AEMULUS captures non-universality, i.e. variations in $f(\sigma)$ at fixed $\sigma(M)$, which is up to $\sim 2\%$ for the cosmological parameters used in this work. Note the violins are only for visualisation; the actual training set samples the mass and σ range much more finely. The lower and upper ends of the mass range correspond to different $\sigma(M)$ depending on cosmology; we only show the range of σ^{-1} covered by all training set mass functions.

at a maximum of $\pm \sim 2\%$ level within the parameter space we consider.[4] Learning this small but non-trivial effect places stringent demands on the IVE prediction accuracy; we will return to this in Sec. 4.1.4.

To normalise our halo mass functions for machine learning, we calculate a mean cosmology, for which each of the seven cosmological parameters is the mean of its training set distribution. Specifically, the parameters are $\Omega_b h^2 = 0.022$, $\Omega_c h^2 = 0.118$, $w_0 = -1$, $n_s = 0.962$,

[4]During our investigation, we found that AEMULUS misconfigured CLASS (Lesgourgues, 2011; Blas et al., 2011) such that the power spectrum used to evaluate $\sigma(M)$ in Eq. (1.72) is always calculated using $w_0 = -1$, even for $w_0 \neq -1$ cosmologies. This however does not affect the overall accuracy of AEMULUS , which can be seen in Fig. 1.9. Further details are provided in Appendix B.2.

$\ln 10^{10} A_s = 3.087$, $H_0 = 68.3\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $N_{\mathrm{eff}} = 3.44$. We divide out the set of halo mass functions by that of the mean cosmology, then take the logarithm to reduce the dynamic range.

To produce query-ground-truth pairs, for the training set halo mass functions, we randomly sample 50 masses per halo mass function from the initial grid of 500 halo masses. We tested that further increasing the number of queries per halo mass function did not significantly improve the model's prediction accuracy. For the validation set, we randomly sample 10 masses per halo mass function (model training is not sensitive to this specific choice). For the test set, to evaluate the IVE predictions across the mass range, we subsample the initial grid of 500 masses by taking every 50-th mass. We additionally include the maximum mass of $M = 10^{15}\,h^{-1}\mathrm{M}_\odot$ to test the IVE performance at the upper mass bound. The distribution of $\log\left(\frac{\mathrm{d}n}{\mathrm{d}\log M}\right)_{\mathrm{normalised}}$ for all M in the training set is rescaled to $[-1,1]$.

As described earlier, to better understand what non-universal information is needed to predict the AEMULUS halo mass functions, we also train a comparison model on the universal Tinker et al. (2008) mass functions for the same set of cosmologies. We generate the ground truth Tinker mass functions using the CCL package[5] (Chisari et al., 2019), following the same steps as for producing the AEMULUS halo mass functions.

4.1.1.3 Inputs to the IVE

For each cosmological parameters sample, we generate the IVE inputs: the linear matter power spectrum $P(k)$ at $z = 0$ and the linear growth function $D(z)$, using CAMB (Lewis et al., 2000, introduced in Sec. 1.2.3). For modelling dark energy with $w_0 \neq -1$, we use the Parameterized Post-Friedmann (PPF) approximation as implemented in CAMB (Hu and Sawicki, 2007).

The linear matter power spectrum $P(k)$ in units of $(h^{-1}\mathrm{Mpc})^3$ at $z = 0$ is evaluated for $k \in [10^{-4}, 10]\,h\mathrm{Mpc}^{-1}$ at 200 points logarithmically spaced in k. The broad range of k-values allows the IVE access to information at all scales, so we can discover where additional information may be needed beyond $\sigma(M)$. The growth function is derived from CAMB's output using the redshift-dependent transfer function values as $D(z) = T(z, k = 1\,h\mathrm{Mpc}^{-1})/T(z = z_{\mathrm{norm}}, k = 1\,h\mathrm{Mpc}^{-1})$, where $z \in [0,5]$, and z_{norm} is the redshift where the growth is normalised as in Eq. (1.43). We choose $z_{\mathrm{norm}} = 50$ to give a broad distribution of $D(z)$ at low redshifts, when growths differ the most due to dark energy. The growth function is evaluated at 100 values of $z \in [0,5]$ linearly spaced in scale factor, which samples the lower redshifts more finely.

[5] https://ccl.readthedocs.io/en/latest/

To normalise our inputs for model training, we first calculate $P_{\rm mean}(k)$ and $D_{\rm mean}(z)$ of the mean cosmology. We then divide out the set of power spectra we obtain by $P_{\rm mean}(k)$, and similarly divide out the set of growth functions by $D_{\rm mean}(z)$. We find this division is especially important in allowing the IVE model to pick up enough information from the power spectrum. We also take the logarithm of the inputs to reduce their dynamic range. We rescale the distribution of $\log P_{\rm normalised}(k)$ for all k in the training set to the range $[-1,1]$, and rescale the distribution of $\log D_{\rm normalised}(z)$ for all z similarly.

4.1.2 Model architecture

As introduced earlier and in Sec. 2.3.1 and 2.3.2, the IVE architecture consists of an encoder, a latent space, and a decoder. The encoder compresses the input $P(k)$ (and optionally $D(z)$ too) into a low-dimensional latent representation, which is an L-dimensional Gaussian distribution. The decoder takes a random sample from the latent distribution and a *query* halo mass $\log M$, and maps this to the corresponding differential halo number density $\mathrm{d}n/\mathrm{d}\log M(M)$.

More specifically, the IVE takes as input a 1D array of either $P(k)$ at $z=0$, or a concatenation of $P(k)$ with $D(z)$. This input $\mathbf{x}$ passes through the encoder, which is a neural network with six fully-connected layers. Each fully-connected layer has 512 neurons, and each neuron follows $y = h(\mathbf{w}\cdot\mathbf{x}+\mathbf{b})$, where $\mathbf{w}$ and $\mathbf{b}$ are the weights and biases that are optimised when training the network, and h is the activation function. We use leaky ReLU activation (Maas et al., 2013) for all hidden layers except the last layer, for which we use a hyperbolic tangent (tanh) activation to ensure smoothness of the encoder output. Weights are initialised with He uniform initialisation scheme (He et al., 2015) for layers with leaky ReLU activation, and Glorot uniform initialisation for the layer with tanh activation (Glorot and Bengio, 2010, see Sec. 2.1).

In short, the encoder consists of weights and biases that learn a non-linear mapping from inputs to a multivariate distribution in the latent space $q(\boldsymbol{\kappa}|\mathbf{x})$, where $\boldsymbol{\kappa}$ denotes the latent representation. We assume it is possible to achieve the latent distribution using a factorised Gaussian, $q(\boldsymbol{\kappa}|\mathbf{x}) = \prod_i^L \mathcal{N}(\mu_i(\mathbf{x}), \sigma_i(\mathbf{x}))$, where L is the dimensionality of the latent space and μ_i, σ_i are the means and standard deviations of the Gaussian distribution for each component of the latent representation. Using this assumption, the encoder maps the inputs to μ_i and σ_i.

The decoder consists of another neural network with six fully-connected layers, with the same number of neurons per layer, the same activation functions, and the same initialisation schemes as the encoder. It maps a latent vector sampled from the latent distribution κ, together with the query $\log M$, to a single estimate of $(\log)\frac{\mathrm{d}n}{\mathrm{d}\log M}$ at the given halo mass.

For the comparison model trained to predict the Tinker halo mass function (Tinker et al., 2008), the architecture is the same except the encoder and decoder each have seven hidden layers. The number of layers in the encoder and decoder, as well as the number of neurons per layer were separately optimised for each IVE model. They are chosen such that further widening or deepening the network does not lead to significant improvements in the model's prediction accuracy.

4.1.3 Loss function

Training the IVE amounts to optimising parameters of the encoder and decoder such that the model achieves high prediction accuracy and the latent representation is disentangled. This is done via minimising a β-VAE loss function like in Eq. (2.7) (Higgins et al., 2017; Burgess et al., 2018), which reads for our IVE

$$\mathscr{L} = \frac{1}{N}\sum_{N}\left(\mathscr{L}_{\mathrm{pred}} + \beta \cdot \mathcal{D}_{KL}(q(\kappa|\mathbf{x})\|p(\kappa))\right) . \tag{4.1}$$

Here, N is the number of training samples per batch, and the first term $\mathscr{L}_{\mathrm{pred}}$ measures the accuracy of the model's predictions, for which we use the mean squared error:

$$\mathscr{L}_{\mathrm{pred}} = \left(\log\frac{\mathrm{d}n_{\mathrm{pred}}}{\mathrm{d}\log M} - \log\frac{\mathrm{d}n_{\mathrm{truth}}}{\mathrm{d}\log M}\right)^2 . \tag{4.2}$$

The second term $\mathcal{D}_{KL}(q(\kappa|\mathbf{x})\|p(\kappa))$ regularises the latent space to encourage disentanglement, as described in Sec. 2.3.1. The balance between the two terms is controlled by the hyperparameter β, which must be carefully tuned while training to achieve both high prediction accuracy and disentanglement.

To verify that the latents are disentangled, since we do not know the relevant factors in advance (unlike for the 3D shapes dataset in Sec. 3.1), we take the approach discussed at the end of Sec. 3.2 and measure the MI between pairs of latent variables. Specifically, we check

that the information shared between latent variables is negligible compared to the information each latent encodes about the ground truth halo mass function.

4.1.4 Optimisation

Since our goal is to discover what factors govern the non-universal halo mass function at $z = 0$, our IVE models (trained on AEMULUS) must be accurate enough to learn non-universality in our training data. Fig. 4.3 shows this is at a maximum of $\sim 2\%$, with a 95% confidence interval of $\pm \sim 0.9\%$ at low σ^{-1} (low mass) and $\pm 1.3\%$ at high σ^{-1} (high mass). This requires the IVE to predict the halo mass function to sub-percent-level accuracy while achieving disentanglement. This is a much more stringent requirement than in previous works by Lucie-Smith et al. (2022b, 2024a); to achieve this, we adopt carefully tuned annealing strategies.

We first aim to achieve disentanglement while improving prediction accuracy. We find models trained with a constant β in the loss function, Eq. (4.1), either do not have disentangled latents when the value of β is small, or for high values of β the latents are disentangled at the cost of losing prediction accuracy. We therefore use β-annealing, where we start with a high value of β that disentangles the latents, and then slowly decrease the value of β with each training epoch t to improve prediction accuracy (Shao et al., 2020). The slow decrease in β allows the latents to remain disentangled while prediction accuracy improves. We minimise the loss function in Eq. (4.1) using the AMSGrad optimiser (Reddi et al., 2018) with a training data batch size of 200 and a learning rate of 5×10^{-4}. We vary β according to a generalised logistic function of the form

$$\beta(t) = \beta_i + \frac{\beta_f - \beta_i}{1 + \exp(-k(t - t_c))}, \tag{4.3}$$

where β_i and β_f are the initial and final values that β asymptotes to, k controls the rate of decrease, and t_c is the epoch where the rate of decrease in β is maximum. We run experiments with different β_i, β_f, k, and t_c, and choose as our final models the ones with a prediction accuracy that is comparable to when the model is trained with $\beta = 0$ (i.e. when only the prediction accuracy is optimised), while having the lowest MI between latent variables (calculated as in Sec. 4.2.1). In particular, we ensure that the MI between pairs of latents remains below the MI between latents and the ground truth halo mass function.

Table 4.1: Hyperparameters used for β-annealing of the final models presented (trained on AEMULUS).

	$P(k)+D(z)$		$P(k)$ only model		
	$t < 5400$	$t > 5400$	$t < 5000$	$5000 < t < 6000$	$t > 6000$
β_i	3.6×10^{-5}	1.4×10^{-6}	3.6×10^{-5}	1.4×10^{-6}	8.485×10^{-7}
β_f	10^{-8}	10^{-7}	10^{-8}	10^{-7}	10^{-7}
k	0.001	0.0015	0.001	0.0015	0.002
t_c	2000	6400	2000	6000	9000

The parameter values for our models, trained on AEMULUS halo mass functions, are in Table 4.1. In each case, using the starting set of parameters, we are able to achieve the required accuracy at the cost of insufficient disentanglement. In particular, we find the MI between latent pairs was higher than the MI between the third latent and the ground truths (we find a total of three latents is required, with the third latent learning a small but significant amount of information on the halo mass function; see Sec. 4.1.5). We found performing a second stage of annealing improved disentanglement. We restore the model to when its accuracy matched that of a two-latent IVE and the MI between all latents was negligible at $< \mathcal{O}(10^{-4})$ nats; this occurred at $t = 5400$ for the $P(k)+D(z)$ model, and at $t = 5000$ for the $P(k)$-only model. We then resume training with a new set of parameters, where β_i was chosen to match $\beta(t)$ at the epoch when training is resumed, and the other parameters were chosen to ensure that $\beta(t)$ decreased at a slower rate than in the first stage, which we find is crucial for achieving disentanglement. While the $P(k)+D(z)$ model reached the high accuracy and the required level of disentanglement when β plateaued to β_f in the second stage, the $P(k)$-only model required a third stage of β-annealing. We again resume training from when the model accuracy was comparable to a two-latent IVE, but when the MI between latents was $\sim \mathcal{O}(10^{-4})$ nats; this occurred at epoch 6000. The third stage allowed the $P(k)$-only model to also reach the required levels of accuracy and disentanglement. In total, β-annealing lasted 11650 epochs for the $P(k)+D(z)$ model and 12250 epochs for the $P(k)$-only model. The IVE trained on Tinker halo mass functions, on the other hand, reached disentanglement with just one round of β-annealing, lasting 12800 epochs, using $\beta_i = 4 \times 10^{-5}$, $\beta_f = 10^{-7}$, $k = 0.001$, $t = 1500$.

Once the β-annealing procedure finished, we further refined the prediction accuracy of the model to ensure it is comparable to the same model trained with $\beta = 0$. We increased the batch size from 200 to 500 and halved the learning rate, and continued to double the batch size and halve the learning rate every time the loss on the validation set ceased to improve over 40

training epochs, until a maximum batch size of 8000 (limited by GPU memory). Decreasing the learning rate reduces the possibility that the optimiser will overstep the minimum (Alsing et al., 2020), while increasing the batch size increases the accuracy of gradient direction estimates. Our models reach convergence at the epoch where the loss evaluated on the validation set is minimum using the final batch size.

In total, training the IVE model (including all β-annealing and prediction accuracy refinement) lasted 12053 epochs for the $P(k)+D(z)$, 12561 epochs for the $P(k)$-only model, and 13042 epochs for the Tinker-IVE model. Note that the second stage lasts only a small number of epochs compared to the first stage; if we only optimise prediction accuracy without the need of disentangling (β-annealing), the total number of epochs required for each model to reach the prediction accuracy in Fig. 4.4 would reduce to $\sim 600-700$ epochs.

4.1.5 Determining latent space dimensionality

We first determine the number of latents required for each IVE. Following the method established in Sec. 3.1.4, we train IVEs with different numbers of latent variables L first setting $\beta = 0$, and determine the L below which prediction accuracy significantly worsens. The top panel of Fig. 4.4 shows the IVE predictive accuracy for $P(k)$-only models with different L; the plot for the $P(k)+D(z)$ model is very similar. Note that although the model predicts $\log \frac{\mathrm{d}n}{\mathrm{d}\log M}$, residuals are quoted as a fraction of the original halo mass function to facilitate comparisons with literature.

We find that an IVE with three latent variables predicts the halo mass function to similar accuracy as an IVE with seven latent variables; the latter is the maximum number of latent variables we expect to need, since this is the number of cosmological parameters varied to generate the training set. Further decreasing the number of latents to two, however, significantly increases the residuals; this persists when we use instead a deeper network of seven hidden layers each in the encoder and decoder, and/or a larger number of 1024 neurons per layer. We therefore conclude that three latent variables is required to predict AEMULUS halo mass functions at $z = 0$, and train these models using the annealing approach described in Sec. 4.1.4 to disentangle the latents. The residuals of the resulting three-latent-variable models are $\leq$ 0.25%, subdominant to the percent-level error in AEMULUS, so our IVE essentially reproduces

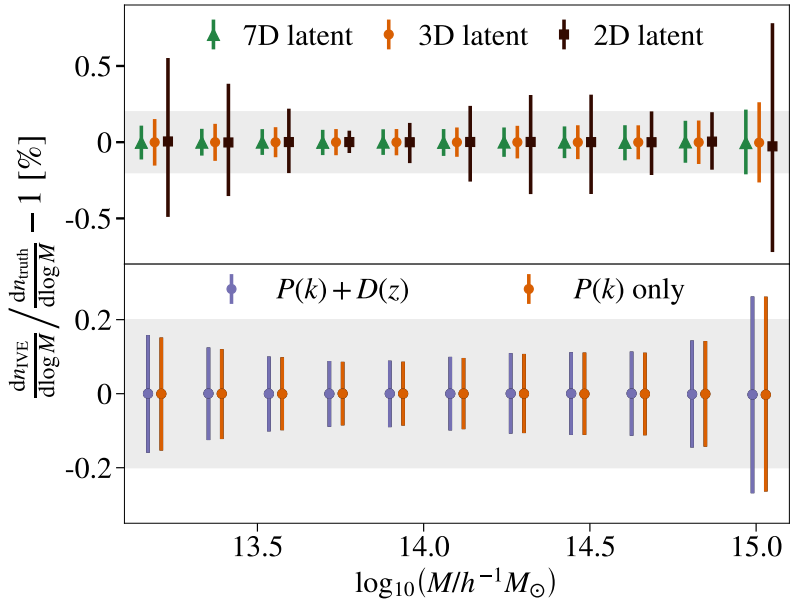


Figure 4.4: Mean and 95% confidence interval of the residuals of the predicted halo mass function for IVE models trained with different numbers of latent variables (*upper panel*), and different inputs (*lower panel*). Grey shaded bands indicate $\pm 0.2\%$. *Upper panel:* three latent variables contain enough information to predict the halo mass function to similar accuracy as seven latent variables (the maximum number expected). Decreasing to two latents significantly worsens the prediction accuracy. *Lower panel:* Adding the linear growth function $D(z)$ does not improve the prediction accuracy compared to training on the linear matter power spectrum $P(k)$ alone.

AEMULUS outputs. As non-universality varies our training set halo mass functions by up to $\sim 2\%$, this accuracy ensures our models have learnt information on non-universality.

4.1.6 Determining the inputs required

We then test whether including growth history improves prediction accuracy compared to using $P(k)$ alone. We separately tune the width and depth of the neural networks for each model given different inputs to ensure the prediction accuracy is converged with architecture. The lower panel of Fig. 4.4 shows both models trained with and without the linear growth function $D(z)$ reach comparable prediction accuracy on the test set halo mass functions. This suggests $D(z)$ does not contain additional information needed to predict the non-universal halo mass function at $z = 0$ compared to $P(k)$. Because of this, for the comparison model trained on Tinker halo mass functions, we only provide $P(k)$ as input. Although not shown, we conduct a similar search and find that the Tinker IVE model also requires three latent variables, and it reaches comparable prediction accuracy to the $P(k)$-only (AEMULUS) model with three latents.

The finding that $D(z)$ does not contain additional information on the halo mass function seems at first to contradict Courtin et al. (2010); Ondaro-Mallea et al. (2021); Euclid Collaboration et al. (2023a), which find different halo mass functions from simulations ran with the same power spectrum but different growth histories (see Sec. 1.6.1.2). However, in order to decouple the effects of growth history from the shape of the power spectrum, these studies initialised their simulations using power spectra calculated from a different set of cosmological parameters to those governing the growth history. For example, Ondaro-Mallea et al., 2021 initialised their simulations with $P(k)$ calculated using a fixed $\Omega_m = 0.307$, while the growth history evolves according to $\Omega_m = [0.148, 1]$. Therefore, by construction, the power spectrum they use does not contain information needed to recover growth history. On the other hand, our results show that within the AEMULUS parameter space and where $P(k)$ and $D(z)$ are calculated from consistent cosmological parameters, growth history information can be inferred from the linear matter power spectrum at $z = 0$, at least when $P(k)$ is evaluated over a wide k-range as we have provided. Ondaro-Mallea et al., 2021 showed on a separate set of wCDM simulations (where $P(k)$ and growth follow consistent cosmological parameters) that including growth history in addition to $\sigma(M)$ (given in Eq. 1.59) improves the modelling of non-universality. However, $\sigma(M)$ evaluated over a finite mass range does not contain all the information in the linear matter power spectrum. Therefore, there is no tension between existing literature and our finding that once $P(k, z = 0)$ is known, $D(z)$ adds no further information to predicting the halo mass function.

4.2 Interpreting the latent space

In the previous section, we have established that predicting AEMULUS halo mass functions at $z = 0$ requires only $P(k)$ as input, and three latent variables are sufficient to capture all the cosmology-dependent information that is required. We therefore use this model as the baseline model, and examine the information encoded within the latent space, which is disentangled as can be seen in Fig. 4.5. The three latent variables are named the 'universal', 'mapping', and 'non-universal' latent respectively, which will become clear in Sec. 4.2.2 and 4.2.3.

We first use latent traversal to visualise the effect that each latent variable has on the predicted halo mass function. Fig. 4.6 shows in the upper panels the decoded halo mass functions

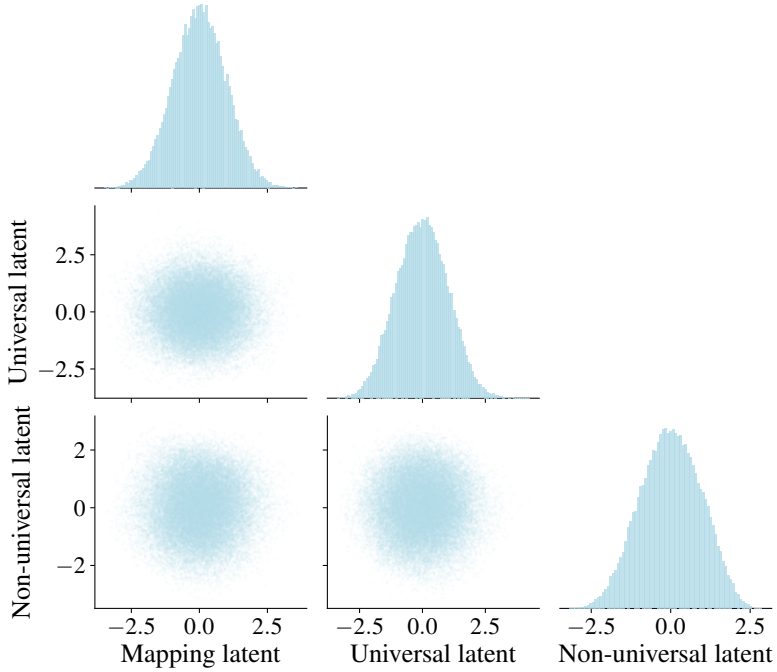


Figure 4.5: Joint and marginal distributions of the disentangled model given only $P(k)$ as input (predicting AEMULUS). The latent distributions closely resemble the standard Gaussian prior, and the MI between latents is $\lesssim 0.002$ nats.

when we systematically vary one latent at a time, while keeping the other latents fixed; their fractional differences with respect to the halo mass function of the reference mean cosmology (defined in Sec. 4.1.1.2) are shown in the lower panels. Each panel from left to right shows the impact that individually varying each latent has on the predicted halo mass function. In each case, the chosen latent is varied between ± 3 std of its mean value, where std is the standard deviation in the latent's marginal distribution.

The universal latent primarily controls the high-mass tail of the halo mass function, the mapping latent controls the normalisation at the low-mass end (note that our low-mass end still corresponds to group-sized halos) and pivots around $M \simeq 10^{14.6}\, h^{-1}\mathrm{M}_{\odot}$, while the non-universal latent controls the curvature of the halo mass function around intermediate mass scales, pivoting at two locations. Interestingly, the latents exhibit a clear hierarchy: varying the universal latent varies the halo mass function by up to 100%, varying the mapping latent varies the halo mass function at $\sim 10\%$, while varying the non-universal latent causes $\sim 1\%$ change. Note however that because the latent traversals show the effect of changing one latent while holding the other two fixed, the actual quantitative effect of each latent is somewhat larger than that illustrated in the latent traversal. In particular, while the non-universal latent appears to have a small effect

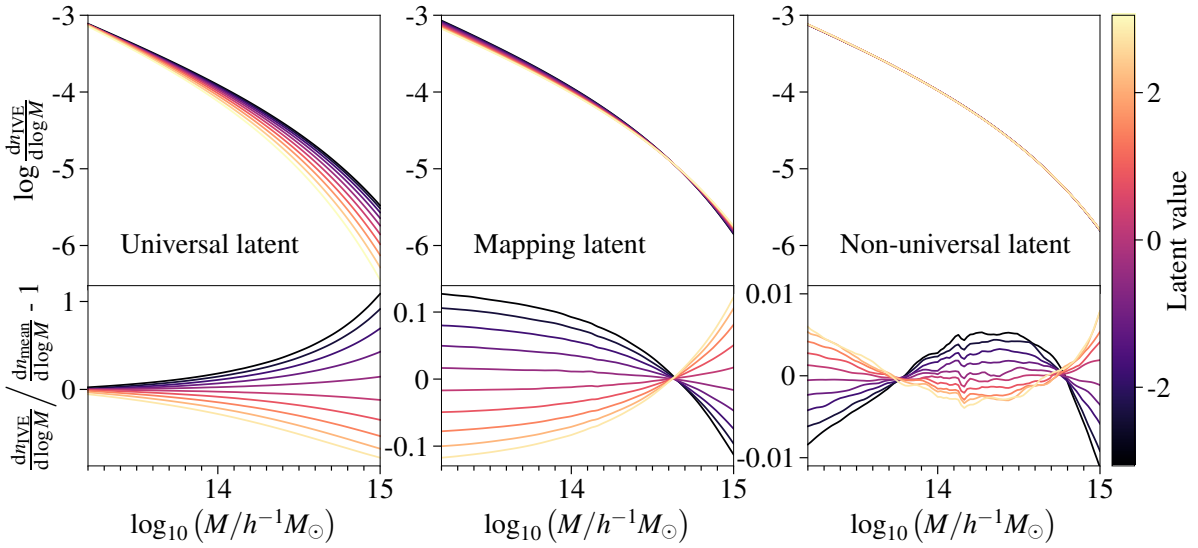


Figure 4.6: *Upper panels:* Variations in the predicted halo mass function when systematically varying the value of one latent variable, while keeping the others fixed. Each panel from left to right varies the universal, mapping, and non-universal latent, respectively. *Lower panels:* The same halo mass functions as in the upper panels, but plotted as fractional differences with respect to that of the reference mean cosmology (defined in Sec. 4.1.1.2) instead.

on the halo mass function that seems comparable to the level of intrinsic errors in AEMULUS, we will see in Sec. 4.2.2 and Sec. 4.2.4 that it encodes physically meaningful information.

4.2.1 Mutual information calculation

As discussed earlier in this chapter and in Ch. 3, we use MI (introduced in Sec. 2.4) both to determine whether latent variables are disentangled, and to interpret what the latent variables encode. Additionally, we make use of the *conditional* MI, also introduced in Sec. 2.4, as a further tool to assist with interpretation. The conditional MI quantifies the amount of information shared between two variables once the conditioned variables are already known.

Both MI and conditional MI are evaluated using GMM-MI (Piras et al., 2023b), introduced in Sec. 3.2. GMM-MI calculates MI through Eq. (2.8), and conditional MI through Eq. (3.5). To obtain the latent distributions which are fitted using GMM-MI, we pass samples from the test set through the encoder, and draw one realisation of the latent representation $\boldsymbol{\kappa}$ from the latent distribution for each test set sample, $q(\boldsymbol{\kappa}|\mathbf{x})$. When estimating the MI between two latent variables κ_i and κ_j, to ensure an accurate and precise evaluation on whether the latents are disentangled, we fit latent samples obtained from the entire test set. The disentangled models we analyse have $\mathrm{MI}(\kappa_i, \kappa_j) \lesssim 0.002$ nats. When interpreting the latent variables via calculating

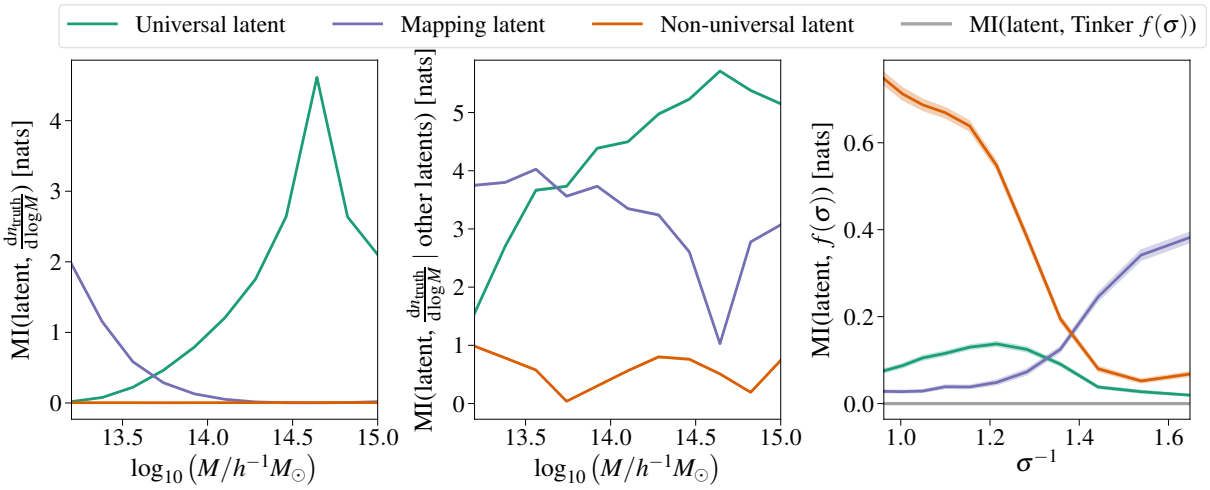


Figure 4.7: *Left panel:* MI between each latent variable and the ground-truth halo mass function in nats. Solid lines and shaded bands indicate the mean and standard deviation as estimated by GMM-MI (Piras et al., 2023b). *Middle panel:* MI between each latent variable and the ground-truth halo mass function, conditioned on the other two latent variables. *Right panel:* MI between each latent and $f(\sigma)$ as function of σ^{-1}, showing the amount of non-universal information, i.e. the information about variation in $f(\sigma)$ a a function of $\sigma(M)$, captured in each latent variable. The grey line shows as reference the MI between latent variables and the universal Tinker multiplicity function, which is consistent with MI $= 0$ nats as expected.

their MI with other quantities, e.g. the ground truth halo number density $\frac{dn}{d\log M}$ as a function of halo mass, we use a randomly sampled 10% subset of the test set to speed up calculation. The increased uncertainty in the MI estimate due to a smaller sample size is accounted for in the GMM-MI error estimate, and the estimate is unbiased (Piras et al., 2023b).

4.2.2 Information on the halo mass function

Fig. 4.7 shows the information each of the three latents encode on the halo mass function. The left panel shows the MI between latents and the ground truth $\frac{dn}{d\log M}$, the middle panel shows the MI between latents and the ground truth $\frac{dn}{d\log M}$ *conditioned on the other two latents*, and the right panel shows the MI between latents and $f(\sigma)$ as a function of σ, for the range of σ that is covered by all cosmologies in the test set.

The left panel shows there are two dominant latents, the universal latent and the mapping latent (the naming will become apparent as we analyse the latents), which learn most of the information on the halo mass function. This is expected from the top panel of Fig. 4.4, which shows that two latent variables allow the IVE model to predict the halo mass function to an accuracy of $\lesssim 0.5\%$. The third latent variable, however, is required to further improve the prediction accuracy to $\leq 0.25\%$.

The universal latent predominantly encodes information on the high-mass end of the halo mass function, and peaks at approximately $M = 10^{14.6}\,h^{-1}\mathrm{M}_\odot$, where MI reaches 4.6 nats. We only observe this clear peak when the latents are disentangled, so we conclude that the mass scale $10^{14.6}\,h^{-1}\mathrm{M}_\odot$ allows best disentanglement of the three factors of variation in the halo mass function. The mapping latent, on the other hand, has the highest MI at the low-mass end where $M = 10^{13.2}\,h^{-1}\mathrm{M}_\odot$, and the MI decreases for higher masses. The fact that the universal and mapping latents capture information on the high- and low-mass ends of the halo mass function respectively is consistent with the latent traversal results from Fig. 4.6. We also note that the peak in the universal latent's MI corresponds to the pivot point when varying only the mapping latent in Fig. 4.6, so predicting the halo mass function at that mass requires almost only the universal latent.

The middle panel of Fig. 4.7 shows the conditional MI between each latent and the halo mass function given the two other latents. This provides a complementary view of the information encoded in the latents; comparing it with Fig. 4.6, we see the latent traversal can be quantified by the conditional MI. The conditional MI therefore reveals the effect of a single latent when the other two latents are known. It confirms that the non-universal latent does carry additional information about the halo mass function; this was not visible from the left panel because information on the halo mass function is dominated by the other two latents. The two troughs in the non-universal latent's line corresponds to the two pivots seen in the latent traversal. Furthermore, the conditional MI reveals that the mapping latent also carries information on the high-mass end of the halo mass function, in agreement with Fig. 4.6. This was again not seen in the left panel because the high-mass end is dominated by the universal latent. The mapping latent's line has a trough at the previously identified scale $M = 10^{14.6}\,h^{-1}\mathrm{M}_\odot$, where both the MI and the conditional MI for the universal latent peaks. This again reflects that learning the halo mass function at $M = 10^{14.6}\,h^{-1}\mathrm{M}_\odot$ allows the IVE to disentangle the latent representation.

The right most panel of Fig. 4.7 focusses on the information that the latents encode on non-universality, showing the MI between latents and $f(\sigma)$. As introduced earlier in this chapter and in Sec. 1.4.3, non-universality manifests as a cosmology-dependent variation in $f(\sigma)$ beyond what is captured by $\sigma(M)$. A non-zero MI in the right panel of Fig. 4.7 therefore indicates the latent encodes information related to non-universality. For reference, we show the mutual

information between latents and the Tinker multiplicity function $f_T(\sigma)$ defined in Eq. (1.67) and used to train our Tinker IVE model (as described in Sec. 4.1.1). Because the Tinker halo mass function is universal with cosmology, the MI between latents and $f_T(\sigma)$ is consistent with 0.[6]

Interestingly, the hierarchy observed in the left panel of Fig. 4.7 is reversed in the right-most panel. The non-universal latent carries the most information on non-universality despite its overall subdominant effect on the halo mass function, giving its name. This is especially true at the low-mass end corresponding to low σ^{-1}. Additional information on non-universality at the high-mass end (high σ^{-1}) is carried by the mapping latent. On the other hand, the universal latent encodes little information on non-universality (less than 0.2 nats), which is very low compared to the maximum of 4.6 nats of information it encodes on the halo mass function. For this reason, we refer to it as the universal latent. The mapping latent also mostly encodes universal information: it encodes up to 2.0 nats of information on the halo mass function, but only about 0.4 nats on non-universality.

The following Sec. 4.2.3 and 4.2.4 examine the information related to the universal and non-universal parts of the halo mass function respectively in more detail. Before this, as we have established that an IVE provided with only $P(k)$ extracts all the relevant information from the growth function $D(z)$ needed to predict the halo mass function at $z = 0$, we quantify this information in the latent space. The left panel of Fig. 4.8 shows the mutual information between each latent variable and $D(z)$ (normalised at $z = 0$) conditioned on the other two latents. We find that the disentangled latents exhibit a trade-off in the information before and after the redshift of matter-dark-energy equality $z_{\rm m=DE}$, which is particularly prominent for the universal and the mapping latent. All three latents have similar amounts of conditional MI at $z_{\rm m=DE}$. Furthermore, we find that while the non-universal latent encodes information on growth history at all redshifts, its conditional MI peaks at the redshift where the conditional MI of the universal latent is lowest; we label this redshift $z_{\rm late} \simeq 0.11$. We stress that this was not imposed by us during training, but instead discovered by the IVE when provided with only $P(k)$ as input.

For comparison, the right panel of Fig. 4.8 shows the conditional mutual information between latents of the Tinker IVE and $D(z)$. These latents, labelled A, B and C, encode only

[6]While this is true by construction, we also confirmed numerically using GMM-MI that MI(latents, $f_T(\sigma)$)= 0 nats.

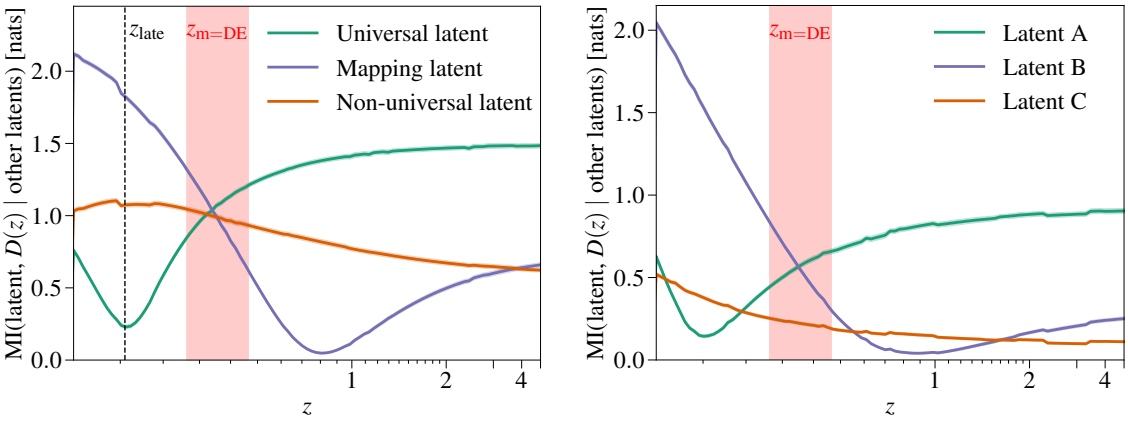


Figure 4.8: *Left:* Conditional MI between each latent variable and the growth function $D(z)$ normalised at $z = 0$ given other latent variables for the baseline model, i.e. the $P(k)$-only IVE trained on AEMULUS. The black dashed line indicates $z_{\rm late} = 0.11$, and the red shaded region shows the range of the redshifts of matter-dark energy equality $z_{\rm m=DE}$ in the training set. *Right:* The same as the left, except the latents are of the IVE trained to predict the Tinker halo mass function. Latents A, B and C of the Tinker IVE therefore only learn universal information by construction.

universal information by construction. For the Tinker latents, we still observe a trade-off in the information before and after $z_{\rm m=DE}$ particularly for latent A and B. However, the conditional MI for latent C does not peak where the conditional MI for latent A is lowest; instead it peaks at $z = 0$. While the conditional mutual information in this panel cannot be quantitatively compared with the left panel because the two panels condition on latents from different IVE models, this qualitative difference suggests a relation between $D(z_{\rm late})$ and non-universality. We will return to this in Sec. 4.2.4.

4.2.3 Universal information

We now discuss the information related to the universal part of the halo mass function, i.e. information that can be understood within the extended Press-Schechter formalism (EPS; see Sec. 1.4.1). As discussed in the last subsection, this information is primarily captured in the universal and mapping latents.

We begin by examining the information content in the universal latent. The abundance of high-mass galaxy clusters is known to be a sensitive probe of approximately $S_8 = \sqrt{\Omega_m/0.3}\sigma_8$ (e.g. White et al., 1993; Eke et al., 1996; Rozo et al., 2010; Norton et al., 2024), which relates to the amplitude of matter fluctuations within a certain redshift range (Jain and Seljak, 1997). Therefore, we expect the universal latent learning primarily the abundance of high mass halos

to have a similar cosmological parameter dependence. The scatter plot in the top panel of Fig. 4.9 confirms this: the universal latent is tightly correlated with the parameter combination $\Omega_m^{0.46}\sigma_8$. This has almost the same information as S_8: the exponent is nearly the same, and the $\sqrt{0.3}$ normalisation has no impact on the information content.

To better understand the power on $\Omega_m^{0.46}$, recall that the MI between the universal latent and the HMF peaks at $M = 10^{14.6}\,h^{-1}\mathrm{M}_\odot$. These halos form half their mass by approximately $z_{\rm form} \sim 0.46$, indicated by the narrow red band in the lower panel of Fig. 4.9. At this redshift, the background matter density of the universe for our mean cosmology ($\Omega_m = 0.3\,, w_0 = -1$) is approximately $\Omega_m(z_{\rm form}) \approx \Omega_m^{0.46}$, as illustrated in the lower panel of Fig. 4.9. This suggests that the parameter combination for the universal latent can be predicted from the formation history of $M = 10^{14.6}\,h^{-1}\mathrm{M}_\odot$ halos. We note that when estimating $z_{\rm form}$, we use the mass accretion history formulae in Correa et al., 2015 corrected for the M_{200b} mass definition. Because these formulae were fitted on ΛCDM simulations, they strictly apply only to our mean cosmology with $w_0 = -1$, so it would be interesting to further test our interpretation when simulation-verified models of half-mass formation times for wCDM cosmologies become available.

We next examine the mapping latent. Under EPS, we expect that once $f(\sigma)$ is known, the next important piece of information needed to predict the halo mass function is the mapping from $f(\sigma)$ to $\frac{\mathrm{d}n}{\mathrm{d}\log M}$. Recalling that the halo mass function in Eq. (1.62) can be written as

$$\frac{\mathrm{d}n}{\mathrm{d}\log M} \propto \Omega_m \frac{\mathrm{d}\log\sigma}{\mathrm{d}\log M} f(\sigma)\,, \tag{4.4}$$

we expect the latent to encode information on the mapping factor $\Omega_m \frac{\mathrm{d}\log\sigma}{\mathrm{d}\log M}$. The top panel of Fig. 4.10 shows the latent in fact learns a parameter combination very similar to the mapping factor: $\Omega_m^{0.35} \frac{\mathrm{d}\log\sigma}{\mathrm{d}\log M}$ evaluated at $M = 10^{13.2}\,h^{-1}\mathrm{M}_\odot$ (where the latent's MI with the halo mass function peaks in the left panel of Fig. 4.7), with an additional small dependence on $N_{\rm eff}$. The role of this latent can therefore be seen as primarily mapping between $f(\sigma)$ and $\frac{\mathrm{d}n}{\mathrm{d}\log M}$, hence its name.

Because the abundance of lower mass halos depends more strongly on $\Omega_m \frac{\mathrm{d}\log\sigma}{\mathrm{d}\log M}$ than $f(\sigma)$, this result is also consistent with the mapping latent learning mostly information on the abundance of lower mass halos, as seen in the left panel of Fig. 4.7.

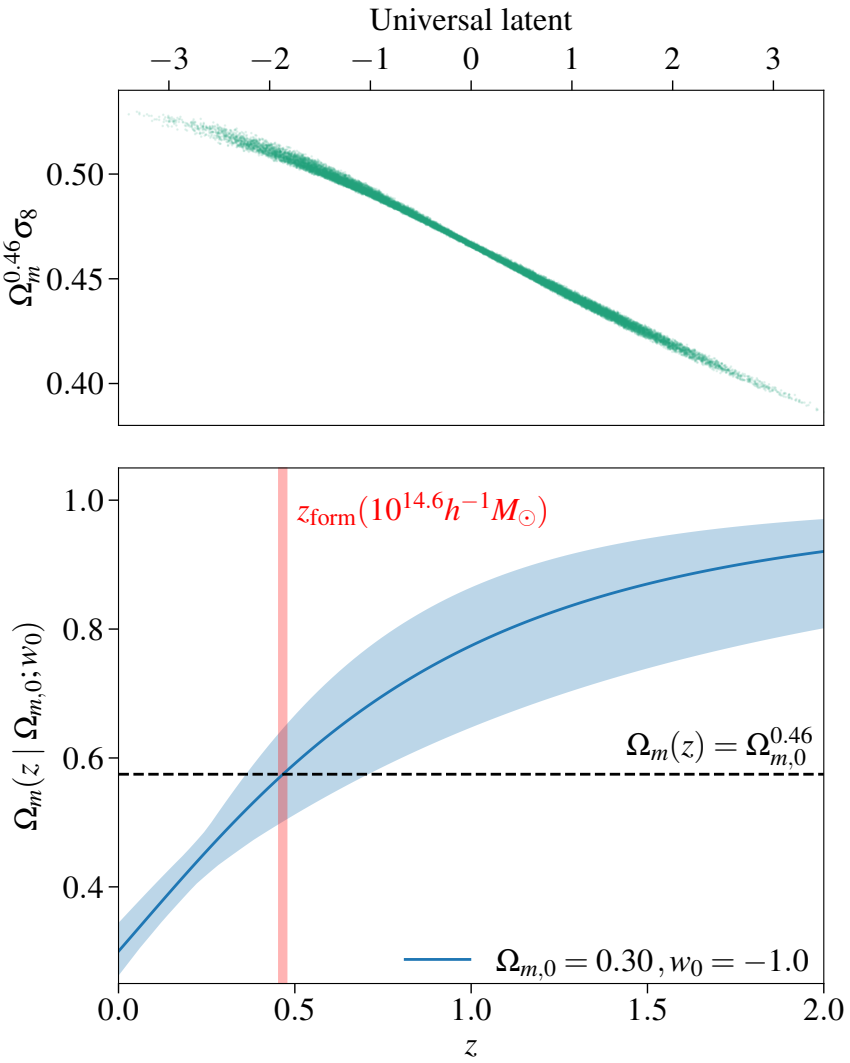


Figure 4.9: Cosmological parameter dependence of the universal latent. *Upper panel:* the universal latent is well-approximated by $\Omega_m^{0.46}\sigma_8$, which is close to the standard $S_8 = \sqrt{\Omega_m/0.3}\sigma_8$ parameter that the abundance of high mass halos is known to be sensitive to. *Lower panel:* The index of $\Omega_m^{0.46}$ can be predicted from $\Omega_m(z)$ of our mean cosmology (with $\Omega_m = 0.3$) evaluated at the half-mass formation redshift $z_{\rm form}$ of $M = 10^{14.6}\,h^{-1}\mathrm{M}_\odot$ halos. The blue band shows the range of $\Omega_m(z)$ in the training set, and the red band indicates the half-mass formation redshifts $z_{\rm form}$ estimated from Correa et al., 2015, corrected for the M_{200b} mass definition. The black dashed line shows $\Omega_m^{0.46}$.

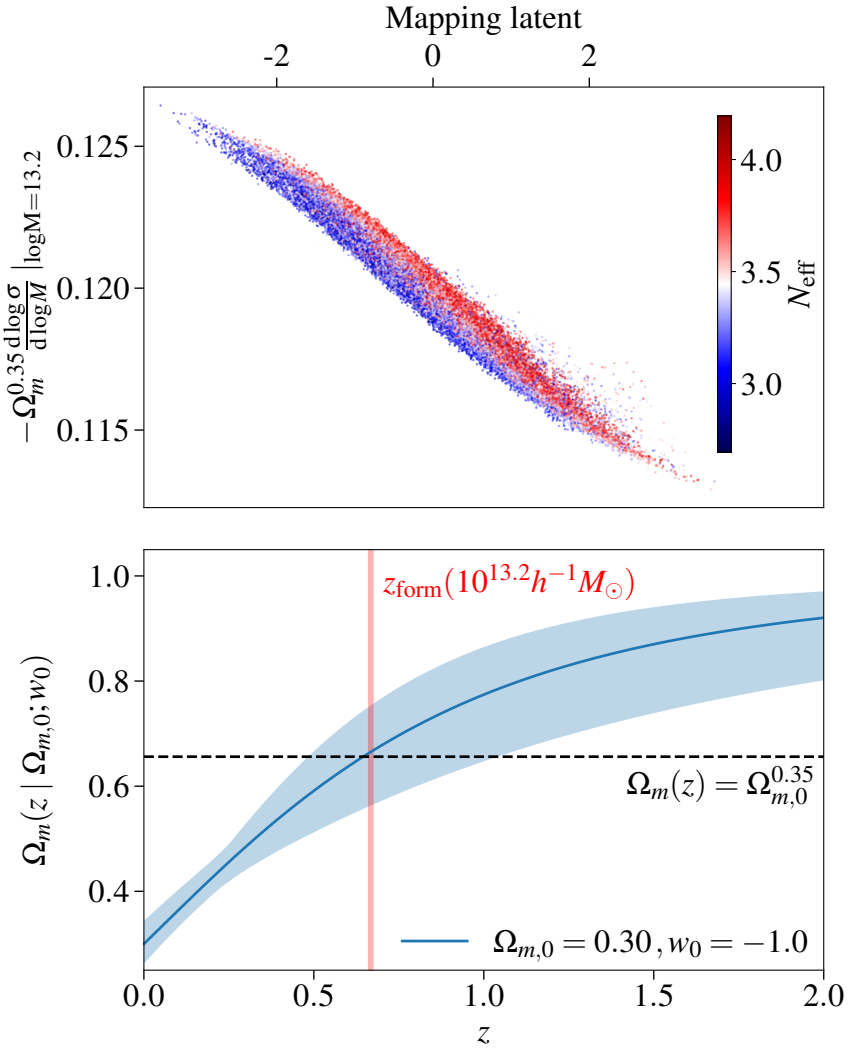


Figure 4.10: Cosmological parameter dependence of the mapping latent. *Upper panel:* the mapping latent is strongly correlated with $\Omega_m^{0.35}\frac{d\log\sigma}{d\log M}$ evaluated at $M = 10^{13.2}\,h^{-1}\mathrm{M}_\odot$, and has an additional small dependence on $N_{\rm eff}$. *Lower panel:* The index of $\Omega_m^{0.35}$ can be predicted from $\Omega_m(z)$ of our mean cosmology (with $\Omega_{m,0}=0.3$) evaluated at the half-mass formation redshift $z_{\rm form}$ of $M = 10^{13.2}\,h^{-1}\mathrm{M}_\odot$ halos. The blue band shows the range of $\Omega_m(z)$ in the training set, and the red band indicates the half-mass formation redshifts $z_{\rm form}$ estimated from Correa et al. (2015), corrected for the M_{200b} mass definition.

Similar to the universal latent, we find that for the mapping latent, the power on $\Omega_m^{0.35}$ can be predicted from the formation history of low mass halos. This is shown in the lower panel of Fig. 4.10: at the redshift $z_{\rm form}$ where low-mass halos with $M = 10^{13.2}\,h^{-1}{\rm M}_\odot$ form half their mass, $\Omega_m(z_{\rm form}) \approx \Omega_m^{0.35}$ for the mean cosmology in the training set. Again, it would be interesting to further test this finding when simulation-calibrated half-mass halo formation redshifts for wCDM cosmologies become available.

In principle, we could improve the approximation $\Omega_m^{0.35}\frac{{\rm d}\log\sigma}{{\rm d}\log M}$ by incorporating the dependence on $N_{\rm eff}$ through manual search or symbolic regression (introduced in Sec. 2.2). However, we choose not to pursue this to focus on interpreting the latent using physically meaningful, rather than empirically-motivated, parameter combinations.

In Appendix B.3, we present additional confirmation that the parameter proxies described in this subsection capture the main information content in the universal and mapping latents.

4.2.4 Non-universal information

As discussed in Sec. 4.2.2, in addition to universal information discussed in the last section, the latents also capture information on non-universality, i.e. information beyond $\sigma(M)$. This was shown in the right panel of Fig. 4.7 through the MI between each latent variable and $f(\sigma)$ as a function of σ. Now, we focus on interpreting the non-universal information learnt by each latent, by conditioning on the other two latents. Fig. 4.11 shows in each row the mutual information between each latent and $f(\sigma)$, but conditioned on the other two latents to isolate the effect of each latent.

The upper panel of Fig. 4.11 shows the MI between the universal latent and $f(\sigma)$ given the mapping latent and the non-universal latent (the solid line). We find that its non-universal information is well-described by its dependence on σ_8; the dashed line showing the MI between $f(\sigma)$ and σ_8 given the other two latents agrees well with that of the universal latent. Furthermore, the grey dotted line shows the MI between $f(\sigma)$ and the universal latent given σ_8, in addition to the other two latents. Recalling that a decrease in MI when conditioning on a variable means the variable explains the information between the latent and $f(\sigma)$, the close-to-zero grey dotted line confirms there is little information on non-universality once σ_8 is known. Furthermore, because we expect σ_8 to be relevant even for a universal halo mass function, we conclude that there is no significant non-universal information in the universal latent.

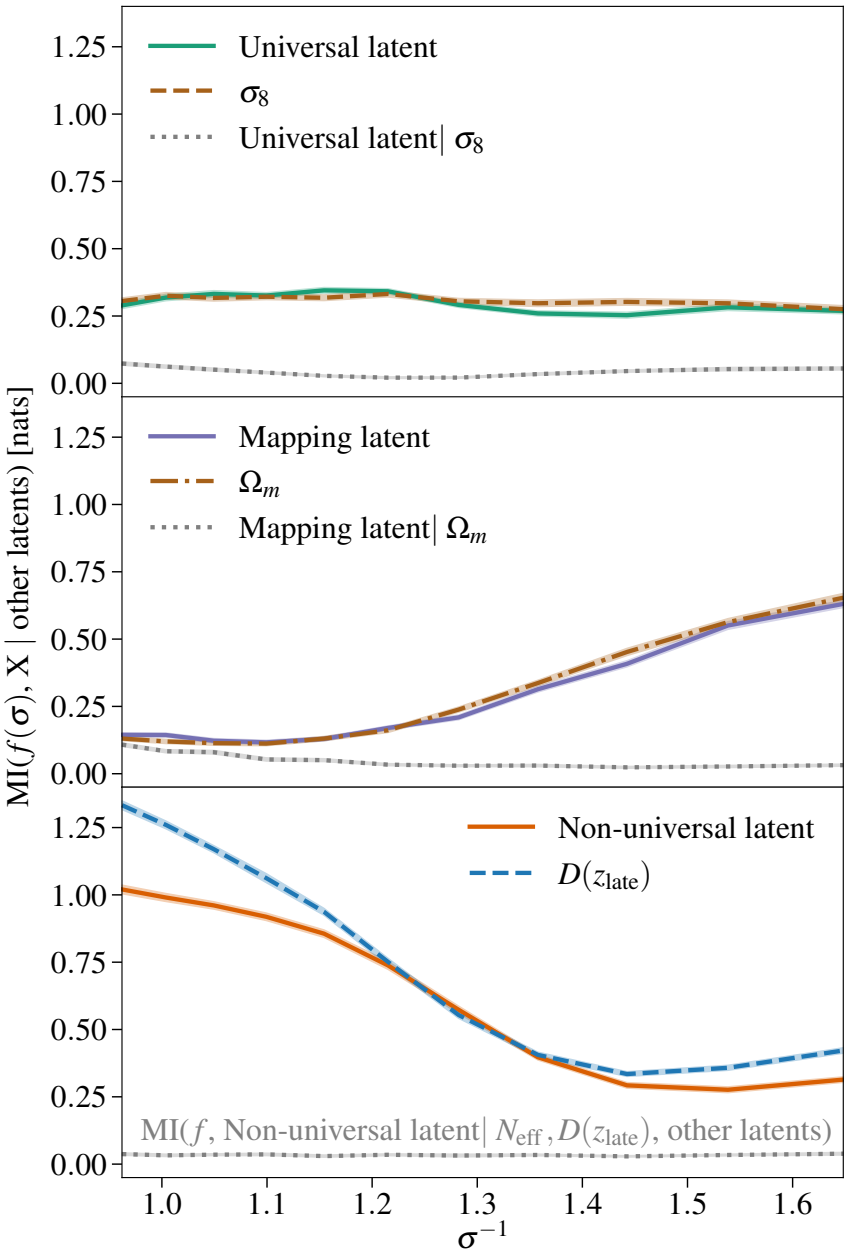


Figure 4.11: *Upper panel:* Conditional MI between X (as specified in the legend) and $f(\sigma)$ as function of σ^{-1}, given the mapping latent and the non-universal latent. The non-universality picked up by the universal latent can be explained by its dependence on σ_8. This can be seen from the grey dotted line showing the MI between the universal latent and $f(\sigma)$ given σ_8 and the other two latents. *Middle panel:* Conditional MI given the universal latent and the non-universal latent. The main source of non-universality in the mapping latent is Ω_m; further conditioning on Ω_m gives the grey dotted line, which is close to zero. *Lower panel:* Conditional MI given the universal latent and the mapping latent. The non-universal information encoded is due to a combination of $D(z_{\rm late})$ and $N_{\rm eff}$. This can be seen from the grey dotted line showing the MI between the non-universal latent and $f(\sigma)$ given $D(z_{\rm late})$, $N_{\rm eff}$ and the other two latents.

The middle panel of Fig. 4.11 shows the MI between the mapping latent and $f(\sigma)$ given the universal and non-universal latents. The dash-dotted line shows the information that Ω_m carries about $f(\sigma)$ given the same two latents, which is near-identical to the mapping latent line. Hence, we conclude the information on non-universality for high σ^{-1} (high-mass halos) that is encoded by the mapping latent is mostly due to the information it carries on Ω_m. This is again confirmed by the grey dotted line showing the conditional MI between the mapping latent and $f(\sigma)$ once Ω_m is known in addition to the other two latents.

The lower panel of Fig. 4.11 shows the mutual information between the non-universal latent and $f(\sigma)$ given the universal and mapping latents. As noted in Sec. 4.2.2, this latent dominates the information on non-universality, especially at low σ^{-1} (low mass). In Fig. 4.8, we saw that the non-universal latent has information on the growth function, which peaks at $z_{\rm late} \simeq 0.11$ (unlike the Tinker IVE's latents). Accordingly, we find that the information contained in the non-universal latent on $f(\sigma)$ can be largely explained by the growth factor at this redshift: the dashed line in Fig. 4.11 is similar to the non-universal latent line. However, because it does not closely trace the conditional MI curve for the non-universal latent, we investigate the information in the non-universal latent further.

Fig. 4.12 shows the mutual information between the non-universal latent and the most relevant physical quantities given the other two latents. Also shown in this figure is the mutual information between latent C of the Tinker model and the same quantities, given latents A and B of the Tinker model. Comparing the two allows us to assess the non-universal information encoded (note in this case a quantitative comparison is reasonable because latents A and B of the Tinker IVE are very similar to the universal and mapping latents of the Aemulus IVE). We find that $D(z_{\rm late})$ is a better description of the non-universal information compared to $\alpha_{\rm eff}$, in Eq. (1.69), proposed by Ondaro-Mallea et al. (2021). We also find the non-universal latent has high information on the redshift of matter-dark-energy equality, and non-trivial information on w_0, both of which are related to growth. In comparison, the Tinker latent C only encodes a small amount of the information on these growth-related quantities.

Additionally, the non-universal latent also contains significant information on the effective number of neutrino species $N_{\rm eff}$. We find that a combination of $D(z_{\rm late})$ and $N_{\rm eff}$ explains almost all the conditional information in the non-universal latent on $f(\sigma)$: the grey dotted line in the

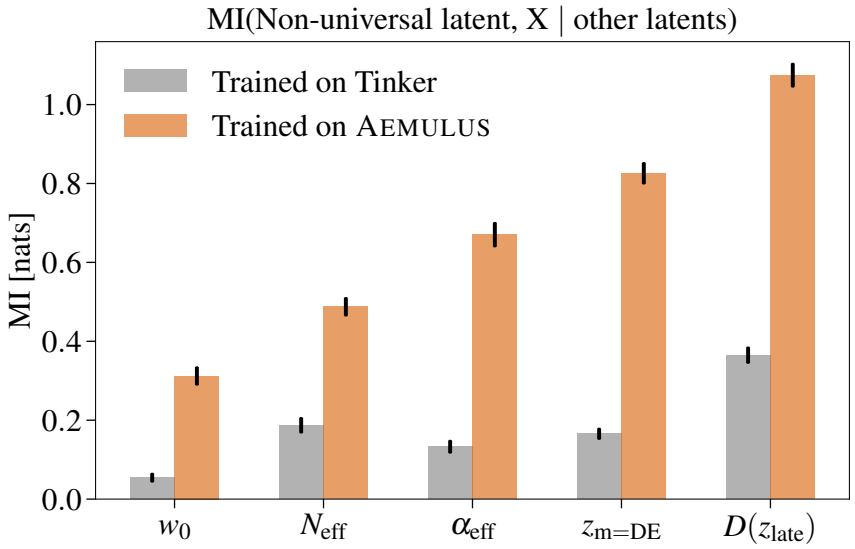


Figure 4.12: Conditional MI between the non-universal latent and X given the other two latent variables, where X is each of the cosmological quantities on the x-axis. We compare the MI of latents of two IVE models, separately trained on Tinker (grey) and AEMULUS HMFs (orange); the former does not contain any non-universal information by construction. The non-universal latent has high information on parameters associated with the recent growth history, as well as $N_{\rm eff}$. The parameter $\alpha_{\rm eff}$ is the effective growth rate parametrisation used by Ondaro-Mallea et al. (2021).

lower panel of Fig. 4.11 shows that when additionally conditioned on both $D(z_{\rm late})$ and $N_{\rm eff}$, the non-universal information left in the latent is close to zero.

4.3 Conclusion and discussion

In this chapter, we have used deep learning and mutual information methods developed in the previous chapter to generate new insights into accurately modelling the M_{200b} halo mass function at $z = 0$. We have trained a deep learning model (the IVE) to learn a low-dimensional latent representation, which captures all the cosmological information required from the linear matter power spectrum (and optionally the linear growth function) to predict the halo mass function produced by AEMULUS (McClintock et al., 2019b). We have then interpreted the latent representation using mutual information (MI) to extract knowledge on the physical quantities required to model the halo mass function. We summarise our main findings below.

1. The linear growth function $D(z)$ does not carry information needed to predict the halo mass function in addition to $P(k, z = 0)$ evaluated over $10^{-4} \leq k/(h\text{Mpc}^{-1}) \leq 10$. While we find $D(z)$ plays a role in the halo mass function, our architecture is able to extract

all the relevant growth history information from $P(k, z = 0)$ alone. This would not be possible had we only used $\sigma(M, z = 0)$, as it does not contain all the information in the power spectrum.

2. Only three independent latent variables are required to predict the halo mass function at $z = 0$ to $\leq 0.25\%$ residuals for $M_{200b} = 10^{13.2-15}\,h^{-1}\mathrm{M}_\odot$ in the wCDM+$N_{\rm eff}$ parameter space we consider. At this accuracy, the IVE essentially reproduces AEMULUS predictions perfectly, as AEMULUS itself produces halo mass functions to approximately 1% accuracy.

3. One latent variable primarily encodes universal information. It has high information on the abundance of high-mass halos, and has a cosmological parameter dependence of $\Omega_m^{0.46}\sigma_8$, which is very similar to $S_8 = \sqrt{\Omega_m/0.3}\sigma_8$. Its cosmological parameter dependence can be predicted from the formation history and mass fluctuation variance of high-mass halos.

4. The second latent variable (the 'mapping latent') encodes mostly $\Omega_m^{0.35}\frac{\mathrm{d}\log\sigma}{\mathrm{d}\log M}$. This closely relates to the 'mapping' factor in Eq. (4.4), which maps a function of mass variance $f(\sigma)$ to a function of halo mass. The power on $\Omega_m^{0.35}$ relates to the formation time of low-mass halos for our mean cosmology. The latent also contains some information on the effective number of neutrino species $N_{\rm eff}$. Furthermore, it learns information on non-universality for high σ^{-1} (high-mass halos), which is largely related to Ω_m.

5. The third latent variable learns primarily non-universal information, especially for lower σ^{-1} (lower-mass halos). This information can be parametrised by the linear growth factor in the recent past, $D(z_{\rm late} = 0.11)$, together with a contribution from $N_{\rm eff}$.

Our finding that non-universality is related to growth history agrees with existing literature (Courtin et al., 2010; Ondaro-Mallea et al., 2021; Euclid Collaboration et al., 2023a). As described in Sec. 1.6.1.2, to explore non-universality, existing literature has used simulations designed to isolate the effect of growth history and that of the effective slope of the power spectrum. In doing so, they consider toy cosmologies with values of e.g. Ω_m and n_s that are well outside current data constraints. In contrast, our methodology allows us to explore non-

universality within a cosmological parameter space relevant for forthcoming surveys, without the need of constructing extreme test cases.

Like Ondaro-Mallea et al. (2021), we also find that recent growth history is required to model halo mass functions with varying w_0. However, Ondaro-Mallea et al. (2021) suggest accounting for non-universality using the growth rate at redshift $z_{\rm ev}$ (see Eq. 1.69), which for our cosmologies evaluates to $z_{\rm ev} \approx 0.43$, earlier than the redshift of matter-dark-energy equality. Ondaro-Mallea et al., 2021 came to their finding through trying different definitions of $z_{\rm ev}$ such that their resulting fitting function agreed well with halo mass functions from simulations ran with varying w_0 (see Sec. 1.6.2). Unlike their approach, our deep-learning based method avoids the need to assume any functional form for parametrising the recent growth (in fact we do not even assume that recent growth is relevant when training our IVE model), thus avoiding any limitations or biases that may be incorporated. Using mutual information, we find that the growth factor $D(z)$ at a considerably lower redshift $z_{\rm late} = 0.11$, which contains equivalent information to the growth rate at $z = 0.05$ (see Appendix B.4), is more informative. This is particularly true for lower σ^{-1} (lower masses).

Unlike Ondaro-Mallea et al., 2021, Euclid Collaboration et al., 2023a accounted for the effect of growth on non-universality using $\Omega_m(z)$ evaluated at the redshift of interest (which is at $z = 0$ for our work). They use $\Omega_m(z)$ as the equivalent of the growth rate at redshift z following Eq. (1.42). We also find that non-universality at high σ^{-1} (high mass) learnt by the mapping latent is related to Ω_m. This agrees with the expectation that the most recent growth should affect the most massive halos that formed very recently, more so than lower mass halos which formed earlier.

We note that in our study, we have defined non-universality as information beyond $\sigma(M)$ as is common in literature. The extended Press-Schechter formalism more precisely predicts the halo mass function to be universal when expressed in terms of peak height $\nu = \delta_c/\sigma$, where the spherical collapse threshold has a very weak dependence on cosmology (see Sec. 1.3.2). As described in Sec. 1.6.1.2, Courtin et al., 2010; Bhattacharya et al., 2011 find accounting for the cosmology dependence of the collapse threshold δ_c reduces non-universality at high peaks (although it does not explain all non-universality there). For wCDM cosmology, the collapse barrier depends only on Ω_m and w_0 (Bhattacharya et al., 2011). Since we find the non-universality captured by the mapping latent affects the high peaks and is related to Ω_m,

it would be interesting to determine whether this can be (partly) understood in terms of the spherical collapse framework as relating to δ_c.

In addition to growth history, Ondaro-Mallea et al., 2021; Euclid Collaboration et al., 2023a also find that the shape of the power spectrum contributes to non-universality in the halo mass function. Both papers parametrise power spectrum shape using a parameter based on $\frac{\mathrm{d}\log\sigma}{\mathrm{d}\log R}$, which gives the local spectral tilt at the Lagrangian radius of the halo, and can serve as a proxy for the local density profile of the collapsing region (Ondaro-Mallea et al., 2021, see Sec. 1.6.2.3). Ondaro-Mallea et al., 2021 find however that non-universality is much more strongly related to the growth history than to the shape of the power spectrum. Within the currently viable cosmological parameter space we consider, we do not find the local slope of the power spectrum to play any significant role in non-universality at $z = 0$.

Apart from $D(z_{\mathrm{late}})$, we find N_{eff} also contributes to non-universality. The effective number of neutrino species primarily modifies the shape of the power spectrum as it is not involved in late time growth. Future work could explore which of the effects that N_{eff} has on the power spectrum (discussed at the end of in Sec. 1.2.2) contributes to non-universality.

We also find the cosmological parameter combinations that approximate the first two latents can be related to the halo formation history (conclusion points 3 and 4). While to our knowledge, there are no existing formulae of halo formation redshifts that are calibrated against simulations ran with wCDM cosmologies, it would be interesting to further test our finding once such formula becomes available.

In this chapter, we focussed on training IVE models to predict the halo mass function at a single redshift $z = 0$. Since the halo mass function exhibits strong non-universality with redshift evolution (Reed et al., 2007; Tinker et al., 2008; Bhattacharya et al., 2011; Diemer, 2020) and forthcoming surveys will observe cluster out to $z = 2$ (Euclid Collaboration et al., 2019; Cerardi et al., 2024), future work will need to accurately model both the cosmology and the redshift dependence of the halo mass function (e.g. McClintock et al., 2019b; Bocquet et al., 2020; Artis et al., 2021). The IVE framework we have presented in this paper can be straightforwardly extended to address this problem, which will be discussed in Sec. 5.1 in the next chapter.

Because we use AEMULUS to generate training data, our exploration is limited by the mass range, halo definition, cosmological parameter space, and accuracy of the emulator. It

would be interesting to apply our framework to training data generated from other emulators such as DARK EMULATOR (Nishimichi et al., 2019) and Mira-Titan (Bocquet et al., 2020), introduced in Sec. 1.6.3.3. The DARK EMULATOR is trained on wCDM cosmologies with a different cosmological parameter range, but it uses the same halo definition and halo finder as AEMULUS (Nishimichi et al., 2019). It would therefore serve as a valuable cross-check of how our findings on AEMULUS generalise, and to gauge whether there are spurious effects due to differences in emulators. The Mira-Titan emulator, on the other hand, is trained on a w_0-w_a cosmological parameter space varying also the neutrino mass, and uses a different halo definition M_{200c} defined with respect to the critical density, as well as a different halo finder (Bocquet et al., 2020). As non-universality is also known to depend on halo definition (see Sec. 1.6.1.1), a comparison with Mira-Titan results may also shed light on what causes the differences in the non-universality depending on halo definitions. In particular, our results on the relation between late-time growth and non-universality offers a possible explanation for why mass definitions probing inner regions of halos tend to exhibit more non-universality than definitions that probe out to halo outskirts (Diemer, 2020). Since inner regions of halos are less sensitive to late-time growth than outer regions (Lucie-Smith et al., 2024a), mass definitions that probe out to halo outskirts would incorporate information on halo accretion history, leading to improved universality. Under this hypothesis, we would expect modelling the M_{200c} halo mass function to require more information on growth, since at $z = 0$, M_{200c} tends to probe more inner regions of the halo compared to M_{200b} (see e.g. Ondaro-Mallea et al., 2021). In addition to training on different emulators, future work should also investigate training IVEs on halo mass functions directly from simulations.

In addition to enabling physical understanding, the latent representation that our IVE model has learnt can also be applied to assist with the design of emulator training sets. As described in Sec. 1.6.3.2, the training sets of existing emulators are generally designed to uniformly cover the cosmological parameter space in some way (e.g. Heitmann et al., 2016; DeRose et al., 2019). This does not take into account which directions in the parameter space are more or less informative for predicting the halo mass function. By contrast, the three-dimensional latent space found by our IVE model corresponds to independent factors governing changes in the halo mass function. Therefore, by sampling the 3D latent space more uniformly, the accuracy of existing emulators could be improved using a small number of additional train-

ing simulations. Sec. 5.2 in the next chapter discusses a proposal for achieving this. Combined with the proposals to extend the IVE approach to model the redshift dependence of the halo mass function, and to consider alternative cosmological spaces and different halo definitions, the deep learning approach we have developed can offer important insights that will contribute towards robust and accurate modelling of the halo mass function, which is crucial to analysing forthcoming survey data.

Chapter 5

Future work

In Ch. 4, we trained an interpretable variational encoder (IVE) and interpreted the disentangled latent representation it learnt to discover that in addition to the mass variance $\sigma(M, z = 0)$, the halo mass function at $z = 0$ also depends on the recent growth history after matter-dark-energy equality and $N_{\rm eff}$. Forthcoming surveys will not just be measuring low-redshift galaxy clusters; e.g. *Euclid* expects to observe clusters out to $z = 2$ (Euclid Collaboration et al., 2019). This requires accurate and robust halo mass function models across a range of redshifts for cosmological models beyond ΛCDM, which in turn requires an understanding of the factors that govern the redshift dependence of the halo mass function in addition to its cosmology dependence. As we saw in Sec. 1.6, such understanding is still far from complete.

Like for the problem of understanding which factors govern the cosmology dependence of the halo mass function at single redshift ($z = 0$), we can use the IVE to gain novel insights on the factors governing the halo mass function across a range of redshifts over a given cosmological parameter space. In Sec. 5.1, I outline how this extension would work in practice. The latent space of an IVE which also learns the redshift-dependence can be compared with that in Ch. 4 to determine if the factors change when redshift evolution is taken into account, and to identify the additional information that is required. As mentioned at the end of Ch. 4, the findings would not only improve our understanding of the physics of halo formation and evolution; knowing what factors dictate the halo mass function can also inform the design of halo mass function emulators and improve halo mass function modelling to meet future survey demands. Sec. 5.2 outlines one way in which the IVE's latent space may be used to optimise the training sets of existing emulators to further improve their accuracy.

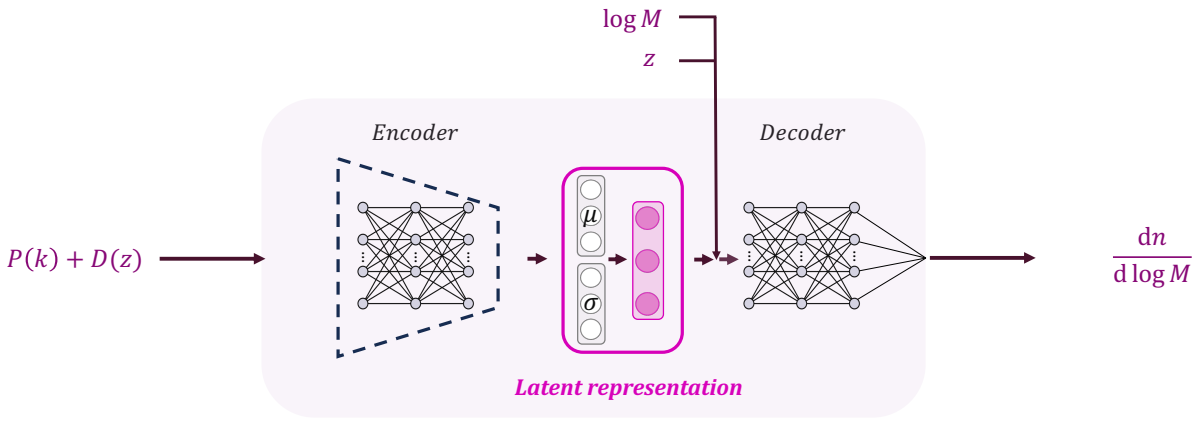


Figure 5.1: The interpretable variational encoder (IVE) that additionally models the redshift dependence of the halo mass function is similar to that in Fig. 4.1, except we train the IVE to predict the halo number density at a given redshift z in addition to a given halo mass $\log M$, and provide as input both $P(k)$ and $D(z)$. The latent representation will encode the information required to predict the halo mass function across a range of redshifts over a cosmological parameter space.

5.1 The non-universal halo mass function across a range of redshifts

5.1.1 The IVE model

We aim to gain physical insights on what is needed to model the cosmology and redshift dependence of the halo mass function through interpreting the IVE latent space. For this, we plan to use the model illustrated in Fig. 5.1. It is similar to the model used to learn the halo mass function at a single redshift $z = 0$ in Ch. 4, except we provide the redshift z as an additional input to the decoder along with halo mass $\log M$. This asks the latent space to encode all the information required to predict the halo mass function at multiple redshifts across the range of cosmological parameters we consider. Interpreting the latent space then allows us to gain insights into what determines the halo mass function across a range of redshifts and cosmological parameters.

Another difference in the model is we plan to always input both $P(k)$ and $D(z)$ to the IVE. Although we found in Ch. 4 that $D(z)$ information needed to model the halo mass function at $z = 0$ can be inferred from $P(k)$, it is not clear if *all* information in $D(z)$ can be inferred from $P(k)$. Therefore, we provide $D(z)$ because the linear growth factor is expected to be needed even in the extended Press-Schechter (EPS) formalism: the EPS halo mass function depends

on the mass variance $\sigma(M,z) = \sigma(M,z_{\rm norm})D(z)$, where $D(z)$ is defined as in Eq. 1.43 and $z_{\rm norm}$ is the redshift at which the growth factor is normalised to unity. We may later investigate whether the required $D(z)$ information can be completely extracted from the power spectrum itself.

Otherwise, the model is as described in Sec. 4.1. The number of layers in the encoder and decoder, the number of neurons per layer, and the number of latents will need to be re-tuned to ensure the model's prediction accuracy is not limited by the architecture.

5.1.2 Training data

As in Ch. 4, we plan to train the IVE in Fig. 5.1 to predict the halo mass functions produced by the AEMULUS emulator (McClintock et al., 2019b). We saw in Sec. 1.6.3.3 that AEMULUS is trained on binned halo counts from simulation snapshots ranging from $z = 0$ to $z = 3$, and it emulates the fitting parameters to $f(\sigma)$ given in Eq. 1.72. McClintock et al. (2019b) found the fitting parameters evolve approximately linearly with the scale factor, so they simultaneously fitted data from all snapshots by imposing linear redshift evolution in the fitting parameters. As such, AEMULUS models the redshift evolution of the halo mass function between $z = 0-3$. Using AEMULUS allows us to compare the latent spaces between IVEs trained to predict the halo mass function at $z = 0$ from Ch. 4 with that trained to predict the halo mass function across a range of redshifts. We can compare the number of latents required, as well as the information encoded in each latent, to gain insights into whether and what extra information is needed to model the redshift dependence of the halo mass function.

We plan to use the same cosmological parameters as in the $z = 0$ dataset described in Sec. 4.1.1. There will be 48000 samples of cosmological parameters in the training set, 12000 in the validation set, and 40000 in the test set, chosen from within the domain of validity of AEMULUS.

The encoder inputs will be the same as described in Sec. 4.1.1.3: for each sample of cosmological parameters, we calculate the linear matter power spectrum $P(k,z = 0)$, and the linear growth function $D(z)$ normalised at $z_{\rm norm} = 50$. We again divide each function by that of the mean cosmology to facilitate the extraction of all available information from the inputs (see Sec. 4.1.1 for details).

For each sample of cosmological parameters, the ground truths are now the halo mass functions at a number of query redshifts z_q. At $z_q = 0$, the halo mass function will be evaluated at query masses drawn from $\log(M/h^{-1}\mathrm{M}_\odot) = [13.2, 15.0]$ as in Ch. 4. At higher redshifts, however, the upper bound of the mass range will need to decrease, as the most massive halos have not yet had enough time to form. We choose the upper mass range by requiring the Poisson noise in the binned halo counts data[1] used to train AEMULUS to be $\lesssim 10\%$. This ensures we only train the IVE using AEMULUS where it is reliable. The lower mass range is limited by the mass resolution and therefore remains fixed with redshift. This means for the highest redshift $z = 3$, we would only be using halo mass functions evaluated over $\log(M/h^{-1}\mathrm{M}_\odot) = [13.2, 13.4]$. Nevertheless, we still plan to include $z = 3$ in the training set to allow the IVE to learn information across the largest possible redshift range. Also, while for the single redshift case in Ch. 4 we divided the halo mass functions by that of the mean cosmology, here we plan to start without doing so. This is because if we divide the halo mass functions at z_q by that of the mean cosmology at z_q, the IVE would not need to predict the overall decrease in the amplitude of the halo mass function with redshift. This means interpreting the latent space may not reveal all the information needed to model the halo mass function's redshift evolution. We may later experiment with normalising all halo mass functions by the halo mass function of the mean cosmology evaluated at an intermediate redshift within the query redshift range (although this requires extrapolating AEMULUS beyond the mass range it is trained on).

We will need to carefully choose the redshifts at which to produce the ground truth halo mass functions to ensure the IVE learns the redshift evolution correctly. Otherwise, the IVE may reach high accuracy at the training redshifts but generalise poorly to other redshifts, as shown in Fig. 5.2. For this figure, I train an IVE as shown in Fig. 5.1 on halo mass functions at $z = [0.00, 0.59, 3.00]$. The IVE has six layers each in the encoder and decoder as for the $z = 0$ model described in Sec. 4.1.2, and I use seven latent variables which is the maximum number expected (see Sec. 4.1.5). The IVE is trained using only the MSE loss (i.e. with $\beta = 0$) to determine if the architecture and the dataset allows the IVE to reach a high prediction accuracy, and the batch size and learning rates change every time the validation loss ceases to improve over 40 epochs, with initial values as described in Sec. 4.1.4. For this model, the training redshifts are chosen to be linearly spaced between $z = 0 - 3$ in scale factor; this

[1] Accessed from https://github.com/tmcclintock/Aemulus_data/tree/master/aemulus_data/mass_functions.

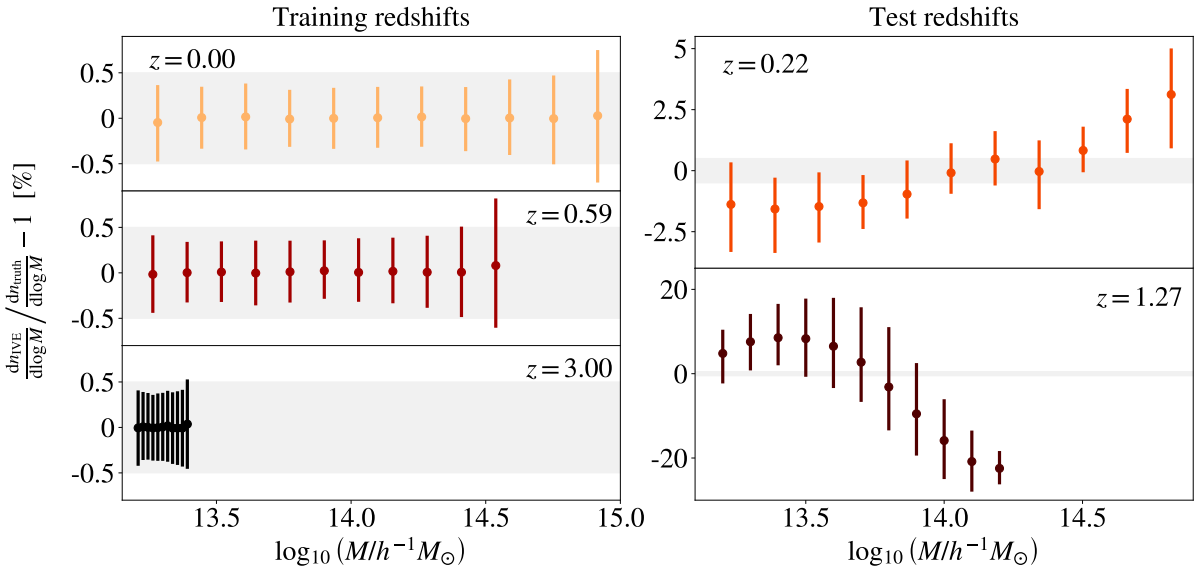


Figure 5.2: Mean and 95% confidence interval of the residuals, $\frac{dn_{pred}}{d\log M} / \frac{dn_{truth}}{d\log M} - 1$, for an IVE model like in Fig. 5.1 trained to predict the halo mass function at multiple redshifts. The residuals are evaluated on halo mass functions of the test set cosmologies (not seen by the IVE during training) at *the left column*: the training redshifts as annotated in each panel, and *the right column*: two additional redshifts lying in between the training redshifts. The grey shaded band in each panel shows the $\pm 0.5\%$ region.

samples the lower redshifts more finely where the evolution is most affected by different dark energy parameters. The two test redshifts at $z = 0.22, 1.27$ are chosen as the midpoint between the training redshifts in scale factors. The validation set is generated at $z = 0.39, 1.93$ lying between the training and test redshifts.

The left column of Fig. 5.2 shows the mean and 95% confidence interval of the residuals evaluated on the halo mass functions calculated for the test set cosmologies (the mass functions were not seen by the IVE during training) at the training redshifts. The IVE mostly reaches $\lesssim 0.5\%$ residuals at each of the three training redshifts. This shows the IVE has learnt to correctly predict the halo mass function at multiple redshifts. However, the right column of Fig. 5.2 reveals the IVE has not correctly learnt the redshift evolution of the halo mass function. The residuals evaluated on the halo mass functions calculated for the test set cosmologies at two additional test redshifts are significantly worse; it can be at $\lesssim 5\%$ at $z = 0.22$ and $\sim 20\%$ at $z = 1.27$. For comparison, AEMULUS residuals at $z = 2$ is still $\lesssim 5\%$ at $z = 2$.[2] Therefore, while simply providing z_q as an additional query to the IVE allows the IVE to predict the halo

[2]Compared to binned halo counts data from https://github.com/tmcclintock/Aemulus_data/tree/master/aemulus_data/mass_functions.

mass function at multiple redshifts, further investigation is needed to determine how to choose training redshifts so the IVE learns to correctly model the redshift evolution of the halo mass function. As remarked, the architecture and/or training schedule would also require further tuning, but we plan to start with determining the training data first.

We may start by increasing the number of training redshifts, as the brief test in Fig. 5.2 only used a small number of three training redshifts. We may also need to test alternative ways of sampling training redshifts from $z = 0 - 3$. The training redshifts for Fig. 5.2 were chosen to be evenly spaced in scale factors, but the residuals at the test redshifts suggest this results in worse prediction accuracy at higher redshifts. We may therefore test sampling evenly in redshifts instead.

Producing halo mass functions at more redshifts per training set cosmology has a disadvantage that the training set will become very large, which will increase the training time per model and can become computationally expensive when we later anneal β to disentangle the latents. An alternative possibility is to randomly sample n redshifts from $z = 0 - 3$ at which to calculate the ground truth halo mass function for each training set cosmology. This ensures the redshift range is well-covered, and could reduce the training set size if $n \leq 3$. Further experimentation will be required to determine n. We will also need to re-tune the number of query masses per halo mass function to ensure the model accuracy is converged with respect to the number of query masses.

5.1.3 Further steps

Once we obtain a model that generalises well with redshift and predicts the halo mass function to high accuracy, we will then follow the same procedure as in Ch. 4 and tune the latent space dimensionality. Comparing the latent space dimensionality to the single-redshift case can indicate whether there are additional independent factors that contribute to the redshift evolution of the halo mass function.

Then, we will train the IVE with the chosen number of latents using β-annealing to disentangle the latents, and interpret the latent representation using similar plots and calculations as in Ch 4. Comparing the information in the latent spaces of IVEs trained to predict the halo mass function at a single redshift versus across a range of redshifts can offer us insights into what information is required to model the redshift evolution of the halo mass function. Furthermore,

the latent space of the IVE can be used to improve halo mass function emulators; a proposal for this is outlined in the next section.

5.2 Practical Application: improving emulator accuracy

As we saw in Sec. 1.6, the state of the art on modelling the halo mass function beyond ΛCDM is achieved with emulators. The AEMULUS emulator (McClintock et al., 2019b) can achieve $\lesssim 1\%$ residuals within the parameter space it is trained on at $z = 0$; by $z = 2$ this loosens to $\lesssim 5\%$.[3] The MIRA-TITAN emulator (Bocquet et al., 2020), trained over a wider cosmological parameter space that also varies w_a and neutrino mass M_ν (although it does not vary $N_{\rm eff}$), can also reach percent-level precision for $M_{200c} = 10^{13} - 10^{14}\,h^{-1}\mathrm{M}_\odot$ at $z \lesssim 0.4$, though its precision on the high-mass end ($M \simeq 10^{15}\,h^{-1}\mathrm{M}_\odot$) that is most sensitive to cosmology can be at $\sim 10\%$. At $z = 2$, its precision at $M = 10^{14}\,h^{-1}\mathrm{M}_\odot$ becomes $\sim 10\%$. Hence, while these emulators' precisions meet the 1% level required for analysing forthcoming survey data at low redshifts, at higher redshifts, their precision could be further improved in order to remain negligible compared to the total error budget (McClintock et al., 2019b). This is important as the abundance of clusters at higher redshifts ($z > 1$) allows for improved constraints on dark energy equation of state parameters (Cerardi et al., 2024).

In Sec. 1.6.3.2, we saw that training sets of existing emulators are designed so that the cosmological parameter samples are space-filling, i.e. they are approximately uniformly distributed in the cosmological parameter space. Further improving the emulator accuracy is currently achieved by adding samples so the overall training set remains space-filling, causing the training set size to approximately double (Bocquet et al., 2020; Heitmann et al., 2016). Further improving the accuracy of emulators therefore quickly becomes computationally expensive.

However, the current emulator training set designs are not specifically optimised for emulating the halo mass function (Nishimichi et al., 2019). As mentioned in Sec. 1.6.3.2, the emulator training set can be optimised by adding a small number (a fraction of the existing training set size) of training samples to the current space-filling design. Bayesian optimisation is one way for choosing additional training samples, and has been applied to improve emulators

[3] Compared to binned halo counts data from https://github.com/tmcclintock/Aemulus_data/tree/master/aemulus_data/mass_functions.

of the Lyman-alpha forest flux power spectrum to achieve tighter posterior constraints (Rogers et al., 2019; Rogers and Peiris, 2021). While Bayesian optimisation can also be used to improve the accuracy of halo mass function emulators, it requires specifying a likelihood given a dataset. Applying the emulator to different datasets may require re-optimising each time, which would increase the overall number of simulations that need to be run. Instead, we propose to optimise emulator training sets using the latent space found by an interpretable variational encoder (IVE). As we saw in Ch. 4, the IVE finds a latent space that is lower-dimensional than the cosmological parameter space, and each dimension of the latent space encodes independent information that is needed to accurately predict the halo mass function. Therefore, the accuracy of halo mass function emulators could be improved if their training sets cover the latent space well. The latent space of the IVE in Ch. 4 encodes all the cosmological information required to accurately predict the halo mass function at $z = 0$ for M_{200b} halos; combined with the extension to model the redshift dependence proposed in Sec. 5.1, the latent space would encode all the information required to model the halo mass function over $z = 0 - 3$ and a wCDM+$N_{\rm eff}$ parameter space. Well-sampling this latent space could improve the accuracy of emulators that use the same mass definition (e.g. DARK EMULATOR presented in Nishimichi et al., 2019; in addition to AEMULUS presented in McClintock et al., 2019b). As discussed at the end of the last chapter, the IVE can also learn the halo mass function for other halo mass definitions; its latent space can then be used to improve e.g. the Mira-Titan emulator, which uses a different halo mass definition (Bocquet et al., 2020).

It is not straightforward to choose cosmological parameter samples that are uniformly distributed in the latent space, because the mapping between the cosmological parameter space and the latent space is non-linear and non-bijective. While the encoder maps cosmological parameters to latents, there is no way to uniquely map latents back to a set of cosmological parameters. When developing the framework used in the last chapter, we also developed a sampling algorithm for adding simulations to existing emulator training sets such that the training set samples the latent space more uniformly. I describe the algorithm in Sec. 5.2.1, and test this on a toy problem described in Sec. 5.2.2. The results are presented in Sec. 5.2.3 and discussed in Sec. 5.2.4.

5.2.1 Algorithm

We devise a sampling algorithm that improves the emulator accuracy by adding a small number – $\mathcal{O}(20)$ – of simulations to existing emulator training sets; this is only half the size of AEMULUS's training set (McClintock et al., 2019b), and an even smaller fraction of the training set sizes of DARK EMULATOR (Nishimichi et al., 2019) and Mira-Titan (Bocquet et al., 2020). The algorithm should balance between sampling in the cosmological parameter space to reduce emulator uncertainty, which is proportional to the distance from training points (see Sec. 1.6.3.1), and sampling so that the training set distribution in the latent space tends towards a uniform distribution.

To achieve a uniform distribution in the latent space, we would like to add samples where the sample density in latent space κ would be low if samples in the cosmological parameter space θ are distributed uniformly. We estimate sample density in the latent space by calculating Voronoi tessellations. We add new samples by drawing from a distribution of cosmological parameters weighted by the volume they occupy in the latent space, as well as by their associated emulator uncertainty. The algorithm has the following steps:

1. Start with a large number of Latin hypercube (LH) samples in the cosmological parameter space, and map the samples to the latent space using the trained IVE's encoder. For each sample it suffices to consider only the mean of the corresponding latent distribution.
2. Estimate the volume each sample occupies in the latent space by calculating the Voronoi tessellation in the latent space using the mapped samples as seeds (the centres for each Voronoi cell). The tessellation needs to be clipped at the boundary (see Sec. 5.2.2) to avoid boundary cells having large or infinite volume.
3. Weight each cosmological parameter sample θ_i with the volume it occupies in the latent space, i.e. by the volume of its associated Voronoi cell V_i. We would like to increase the probability of choosing samples in sparsely sampled regions of the latent space; such samples have larger Voronoi cells. The probability of choosing the sample is therefore $p(\theta_i) \propto V_i$. The emulator uncertainty given each sample can be accounted for by adding it as another weight to each cosmological parameters sample.
4. Add new points to the emulator training set by drawing from the weighted samples.

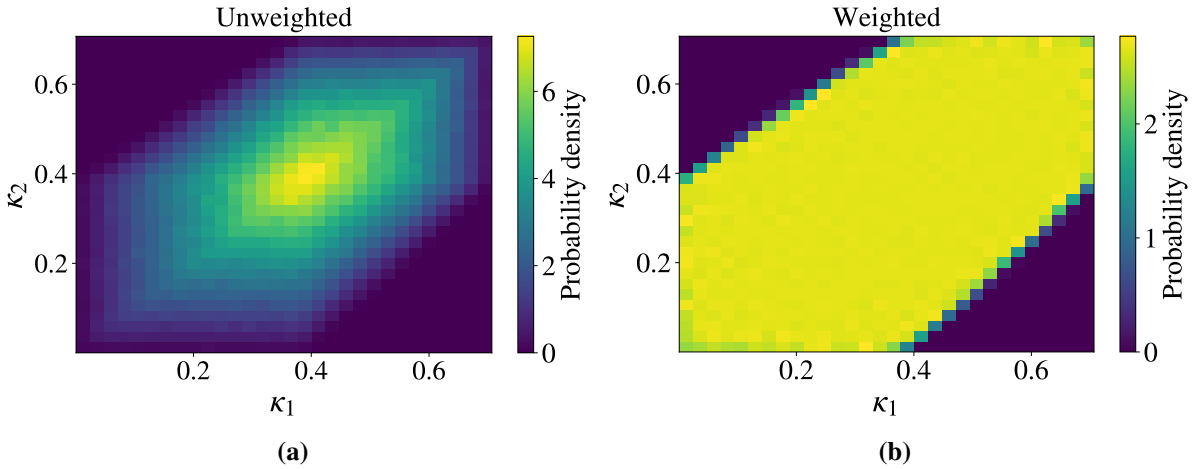


Figure 5.3: Distribution of κ if *left*: θ were unweighted, and *right*: if θ samples are weighted by their Voronoi cell volumes.

We test this algorithm on a simple toy problem with a two-dimensional latent space described in the next section (5.2.2). We ignore weighting samples by the emulator uncertainty in step 3, and check if the proposed algorithm leads to a uniform distribution in the latent space. We also test if the algorithm can be scaled to higher dimensions, as the boundary clipping in step 2 can quickly become expensive (more details in Sec. 5.2.3.2).

5.2.2 Testing the algorithm on a toy problem

To test the algorithm proposed in the last section, we use a simple toy problem with an arbitrarily chosen non-linear and non-bijective mapping from a 'parameter space' θ to a 'latent space' κ. To keep things simple, we choose a three-dimensional θ space, and a two-dimensional κ space (which also facilitates its visualisation). We draw 10^6 samples of θ from a unit Latin hypercube with sides $\theta_i \in [0,1]$, and map them to κ via:

$$\kappa_1 = \sin\left(\frac{\pi}{8}(\theta_1+\theta_2)\right), \qquad \kappa_2 = \sin\left(\frac{\pi}{8}(\theta_2+\theta_3)\right). \tag{5.1}$$

The distribution of samples in κ space is shown in Fig. 5.3a.

To calculate the Voronoi tessellation in κ space, we use `scipy`'s implementation of the quickhull algorithm (Barber et al., 1996) and the `shapely` package for boundary clipping. We define a boundary as the convex hull of the point cloud consisting all samples, and clip the

tessellation at the boundary by calculating the intersection between the convex hull and the surfaces (or lines in 2D) defining each Voronoi cell.

Reweighting the samples in κ space by their Voronoi cell volumes gives the 2D histogram Fig. 5.3b, which visually appears approximately uniform. To quantitatively assess if the samples chosen using our algorithm are uniformly distributed in the latent space, we quantify the dispersion of the samples using the generalised variance, i.e. the determinant of the covariance matrix, $|\Sigma|$. Generalised variance is large if the samples are more uniformly distributed, i.e. if the variance along each latent space dimension is large, *and* if there is no strong correlation between the samples (otherwise the additional samples would only sample a degenerate subspace of the latent space).

5.2.3 Results

5.2.3.1 Distribution of emulator samples chosen

As described earlier, to improve the accuracy of existing emulators, we would like to add $\sim \mathcal{O}(20)$ simulations to, say, AEMULUS's training set. The cosmological parameters of each simulation should be unique to avoid running simulations with identical cosmological parameters. To test whether additional samples chosen using our proposed algorithm are more uniformly distributed in the latent space, we compare the dispersion between 20 unique samples chosen using our proposed algorithm with that between 20 unique samples drawn randomly from the Latin hypercube samples of the cosmological parameter space obtained from step 1 of our algorithm.

We repeatedly draw samples and calculate the generalised variance 10^5 times (the 20 samples drawn are replaced after each trial). The resulting distributions are plotted in Fig. 5.4. The distribution in purple shows the generalised sample variances for samples chosen without weighting, i.e. it shows the distribution if we Latin-hypercube sample the cosmological parameter space. The distribution in orange shows instead generalised variances from samples weighted by their Voronoi cell volumes, which is larger as expected. Furthermore, for comparison, we Latin-hypercube sample the latent space (within the range covered by the samples, which is between $[0, 0.71]$ along each dimension as shown in Fig. 5.3) using 10^6 samples. We then keep only the samples that lie in the convex hull of the point cloud, i.e. the points that do not lie in the deep blue regions in Fig. 5.3b. From the samples kept, we repeatedly perform

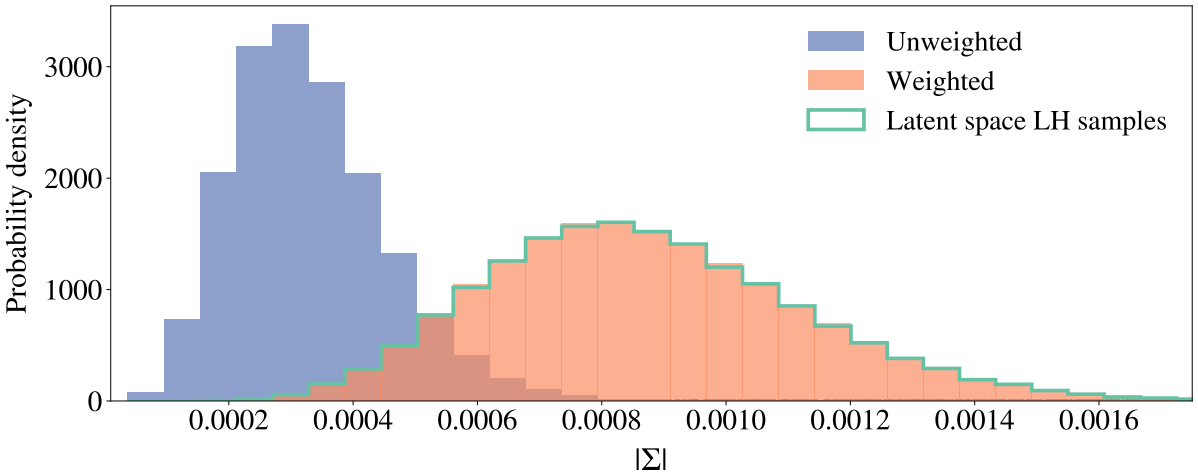


Figure 5.4: Distribution of generalised variances $|\Sigma|$, each calculated from 20 samples chosen from unweighted samples vs. samples weighted by their Voronoi cell volume in the latent space (10^5 trials per distribution). Green shows the distribution for when the 20 samples are drawn from a Latin hypercube (LH) in the latent space.

trials where we draw 20 unique samples and calculate their generalised variance. The distribution of generalised variances for 10^5 trials is shown in green in Fig. 5.4. It is very similar to the distribution obtained by drawing from samples weighted by their Voronoi volumes. Hence, sampling using our proposed algorithm is comparable to sampling uniformly in the latent space. Recall, however, that we cannot directly add simulations by sampling uniformly in the latent space, because we cannot straightforwardly map latent samples to the cosmological parameter space.

5.2.3.2 Scalability to higher dimensions

For the 2D toy problem, we calculated a bounded Voronoi tessellation by clipping at the convex hull of the point cloud consisting all samples. Without boundary clipping, Voronoi cells near the convex hull can be unbounded or have very large volumes, such that weighting by Voronoi cell volumes results in samples always being drawn from close to the boundary, rather than uniformly in the latent space.

As described in Sec. 5.2.2, boundary clipping requires calculating the intersection between the convex hull and surfaces defining each Voronoi cell. For the 3D latent space found in Ch. 4, computing a clipped Voronoi tessellation already becomes challenging (see e.g. Liu et al., 2022, and references therein). For a redshift-dependent IVE whose latent space also contains the information required to model the redshift evolution of the halo mass function, the latent space

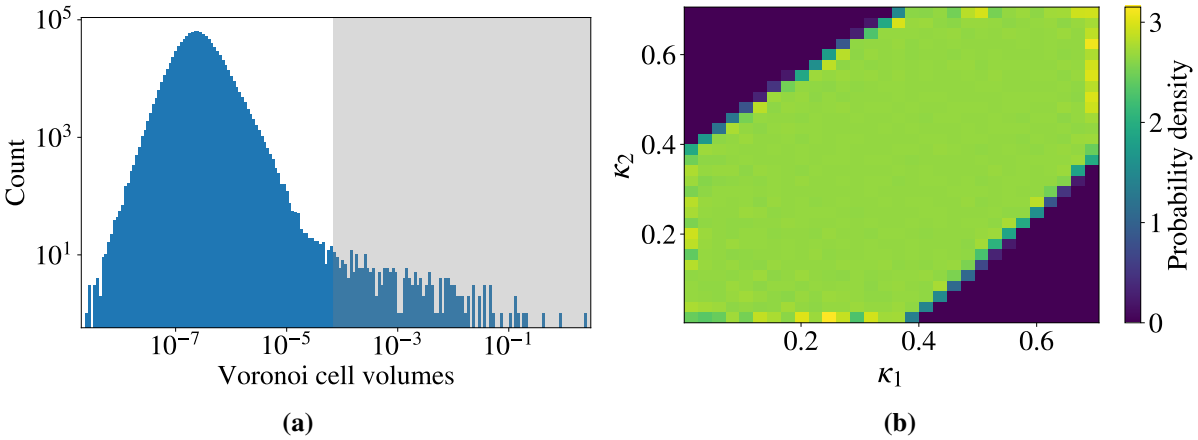


Figure 5.5: *Left*: The distribution of Voronoi cell volumes for an unbounded Voronoi tessellation calculated on samples from step 1 of the algorithm. We discard samples with a Voronoi cell volume $> 6 \times 10^{-5}$, i.e. samples in the grey region. *Right*: Distribution in the latent space if the remaining samples are weighted by their Voronoi cell volume.

dimensionality may be higher to capture the additional information. This would make our algorithm impractical if we continue to require a clipped Voronoi tessellation. We propose as an alternative calculating unbounded Voronoi tessellations, then filtering out samples with large Voronoi cell volumes.

We test this on the same toy problem. We repeat the experiment in the previous section (Sec. 5.2.3.1), except that instead of boundary clipping in step 2 of the algorithm, the samples are filtered so that only samples with a Voronoi volume $< 6 \times 10^{-5}$ are kept, illustrated in Fig. 5.5a. The threshold is chosen by eye as the value above which the distribution of Voronoi volumes is no longer smoothly varying. A weighted histogram of the samples after filtering is shown in Fig. 5.5b, which looks approximately uniform. Fig. 5.6 shows the distributions of generalised variances as in Fig. 5.4, but the weighted samples only consist of samples with a Voronoi volume $< 6 \times 10^{-5}$ from an unbounded Voronoi tessellation. Similar to in Sec. 5.2.3.1, the generalised variance is larger after weighting by Voronoi cell volumes filtered to remove very large cells, and is comparable to sampling from Latin hypercube samples in the latent space. Hence, boundary clipping can be replaced with filtering out samples with large Voronoi volumes.

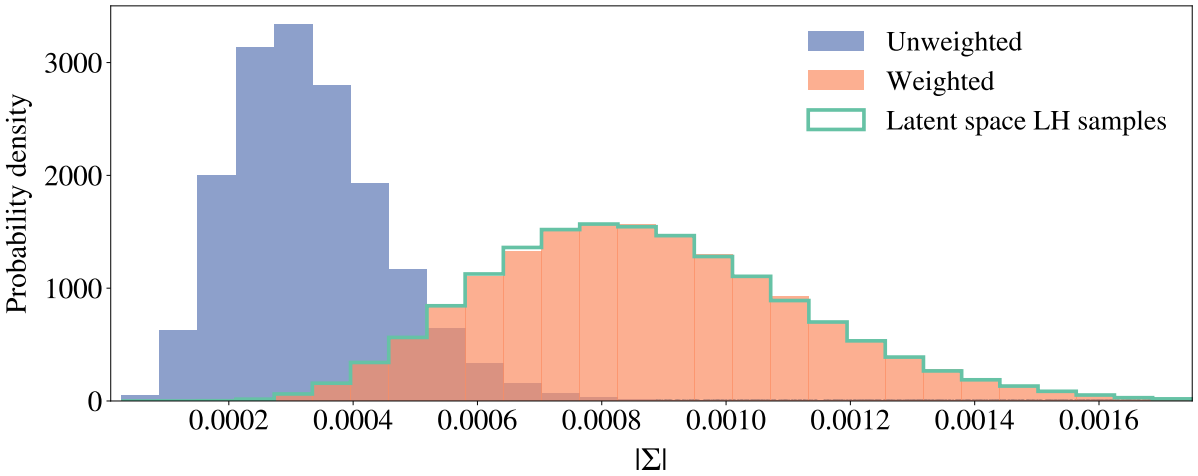


Figure 5.6: Distributions of generalised variances $|\Sigma|$ as in Fig. 5.4, except instead of boundary clipping, Voronoi cell volumes are calculated from unbounded Voronoi tessellations, but the weighted samples only consist of samples with Voronoi volume $< 6 \times 10^{-5}$.

5.2.4 Discussion

Since the IVE latent space encodes the information required to accurately predict the halo mass function, the accuracy of existing halo mass function emulators can be improved if their training sets sample the latent space more uniformly. This is not straightforward because the mapping between the cosmological parameter space and the latent space is non-bijective.

In this section, we have proposed an algorithm that adds samples to the training sets of existing emulators to achieve a more uniform distribution in the latent space. The algorithm chooses additional samples from Latin hypercube samples in the cosmological parameter space weighted by their Voronoi cell volumes in the latent space. We have demonstrated using a toy problem that drawing samples using our algorithm is comparable to drawing samples from Latin hypercube samples in the latent space. Furthermore, the algorithm scales to three- and higher-dimensions: we do not require a bounded Voronoi tessellation to calculate Voronoi cell volumes, and we can avoid always drawing samples from the boundary by filtering out samples with a large Voronoi cell volume.

We note that since our algorithm requires randomly drawing a small number of samples, the resulting samples from a single draw may not always be more uniformly distributed in the latent space; this can be seen from the overlap in the weighted and unweighted distributions in Fig. 5.4 and 5.6. However, since the computational cost of performing each draw is low, one

can simply perform a large number of draws, and choose the draw with the highest generalised variance for running additional simulations to the emulator training set.

These results on the toy problem show that our proposed algorithm is a viable means of choosing additional training set samples to sample the latent space more uniformly. It would therefore be interesting to next test whether adding samples using our algorithm does improve the accuracy of existing halo mass function emulators.

5.3 Summary and further steps

This chapter discusses how the IVE can be extended to model the halo mass function across a range of redshifts. Interpreting the latent space of the redshift-dependent IVE will allow us to discover the information that is needed to model both the cosmology and the redshift dependence of the halo mass function. Like for the single redshift ($z = 0$) case in Ch. 4, because the IVE learns the latent representation with minimal prior assumptions imposed, we expect the IVE to offer novel insights on the sources of non-universality in the halo mass function, which are difficult to obtain using existing approaches based on controlled numerical simulations and semi-analytical halo mass function modelling (see Sec. 1.6.1.2 and 1.6.2).

Understanding what determines the cosmology and redshift dependence of the halo mass function is important for improving its modelling to meet demands of future survey analyses. The second half of this chapter has presented an algorithm that uses the IVE latent space to select cosmological parameters of additional training simulations for existing halo mass function emulators. This has the potential to improve emulator accuracy by adding a much smaller number of training simulations compared to the number that would be required under current strategies – by adding (less than) half the size of current emulators' training sets instead of doubling the training set size (Heitmann et al., 2016). We showed on a toy problem that the algorithm can sample the latent space more uniformly; future work will test whether the algorithm leads to improved emulator accuracy compared to current design strategies. This would require running N-body simulations and retraining existing emulators (e.g. the AEMULUS emulator McClintock et al., 2019b) on binned halo counts data. One could potentially also explore alternative methods of mapping from a point in the latent space to an array of (non-unique) cosmological parameters. This would enable designing training sets of emulators that lie on a grid

or a Latin hypercube in the latent space, which would sample the latent space even better than our algorithm does.

More broadly, the IVE latent space can be used for a variety of downstream tasks to improve the halo mass function modelling across a large cosmological parameter space and redshift range. In addition to improving the design of halo mass function emulators, the IVE latent space could potentially be used for understanding the robustness of existing emulators. By training IVEs on halo mass functions produced by different emulators that use the same halo definition (e.g. AEMULUS McClintock et al., 2019b and DARK EMULATOR Nishimichi et al., 2019) and restricting both to a common cosmological parameter space, redshift and mass range, we can compare the information contents in the latent spaces. Any differences in the information captured by the latent spaces would then point to emulator-specific effects (including effects related to numerical simulations discussed in Sec. 1.5.1) not associated with the underlying physics of halo formation and evolution, as the latter should be the same across the two emulators. The IVE latents could also be used to find alternative, more interpretable halo mass function models compared to emulators. As the latents capture independent factors governing the halo mass function, one could explore modelling the halo mass function as a combination of the latents and a set of basis functions. One could additionally use e.g. symbolic regression (introduced in Sec. 2.2) to find accurate mathematical expressions for the latents in terms of cosmological parameters. Combining these could lead to a halo mass function model whose behaviour over the parameter space of interest is easier to understand than emulators. This more-robust halo mass function model could then be used in survey analyses instead of emulators. The representation learning approach we take towards modelling the halo mass function therefore provides us with a latent space that enables improving both our physical understanding and the practical modelling of the halo mass function.

Chapter 6

Conclusions

6.1 Summary

This thesis has developed a novel framework for using deep learning to gain physical insights into problems in cosmology. In particular, we have focussed on the problem of understanding what information is needed to accurately model the dark matter halo mass function, which plays an important role in cosmological inference using forthcoming survey data. The framework combines deep representation learning with mutual information to achieve model interpretability and enable knowledge extraction. We have trained a deep learning model with an encoder-decoder architecture to learn a complex, non-linear mapping (between the linear matter power spectrum $P(k)$ and the halo mass function) by first finding a low-dimensional latent representation that captures all the relevant information needed to predict the output. The latent representation is disentangled, i.e. each component of the latent representation captures independent information on the output, facilitating interpretation.

The framework developed in this thesis contributes both to improving halo mass function modelling and to the methods of interpreting latent representations. For the former, existing approaches to modelling the halo mass function and its non-universality either need to assume the relevant quantities and their parametrisation (see Sec. 1.6.2 and the discussion in Sec. 4.3), or they take a fully data-driven approach using emulators that lacks interpretability (see Sec. 1.6.3). The novel approach in this thesis enables *discovering* the relevant quantities needed to model the halo mass function to emulator precision through interpreting the latent representation. Such insight potentially leads to a novel way of optimising emulator designs, as discussed in Sec. 5.2. For the latter, this thesis contributes to the effort towards using more prin-

cipled metrics of disentanglement based on mutual information, and using mutual information to interpret the latent space without prior knowledge of the underlying factors (work in Ch. 3 and 4). Notably, Ch. 3 has contributed towards improving the robustness and reliability of the mutual information estimator GMM-MI (Piras et al., 2023b), and Ch. 4 presents a novel application of conditional mutual information to assist with interpreting the latent representation. We go beyond just using conditional mutual information to identify redundant information like in Bhattacharjee et al. (2020): we also use it to determine the information each latent encodes given knowledge of other latents, and to identify additional dependencies in the latents.

In Ch. 4, we have applied this framework to understand which factors govern the non-universal halo mass function at $z = 0$. We find that the present-day halo mass function is controlled by three factors for the mass range $M_{200b} = 10^{13.2-15.0}\,h^{-1}\mathrm{M}_\odot$ in the wCDM+$N_{\rm eff}$ cosmological parameter space investigated. One latent primarily encodes universal information, and is well-approximated by $\Omega_m^{0.46}\sigma_8$ which is very similar to S_8. Another latent learns the parameter combination $\Omega_m^{0.35}\frac{\mathrm{d}\log\sigma}{\mathrm{d}\log M}$, which is very similar to the factor in Eq. (4.4) that maps from the multiplicity function $f(\sigma)$ to the halo mass function. It also has additional dependencies on $N_{\rm eff}$. Furthermore, this latent learns non-universal information for high σ^{-1} corresponding to high-mass halos, which is largely related to Ω_m. The third latent primarily encodes non-universal information, especially for low σ^{-1}. Given information in the first two latents, the non-universal information in the third latent is strongly related to growth after matter-dark-energy equality, which can be parametrised by $D(z = 0.11)$. There is also a contribution to its non-universal information from $N_{\rm eff}$. Our finding that non-universality is related to growth history agrees with existing literature (Courtin et al., 2010; Ondaro-Mallea et al., 2021; Euclid Collaboration et al., 2023a). However, as discussed in Sec. 4.3, our approach allows us to explore non-universality without resorting to extreme toy cosmologies (unlike existing literature), and without limiting us to specific functional forms or parametrisations. Our approach enables us to identify a more informative parametrisation of the growth history information compared to the proxy suggested by Ondaro-Mallea et al. (2021), and reveals that within the currently viable cosmological parameter space we consider, the local slope of the power spectrum does not play a significant role in non-universality at $z = 0$.

6.2 Outlook

6.2.1 Non-universality and halo mass function modelling

The findings from Ch. 4 point to several avenues for further investigations. Sec. 4.3 discussed further investigating the role of $N_{\rm eff}$ in non-universality, and whether the parameter dependences of the first two latents are indeed related to $\Omega_m(z)$ at the halo formation redshift. Ch. 5 proposed how the IVE can be extended to additionally learn the redshift dependence of the halo mass function. The latent space of the redshift-dependent IVE can offer unique insights into the source of non-universality with redshift and cosmology, and it can be used to improve halo mass function modelling across a wide cosmological parameter space and redshift range to meet the demands of future surveys. Sec. 5.2 proposes one way in which the accuracy of existing emulators could be improved by sampling the IVE latent space more uniformly using a small number of additional simulations to the training set. Furthermore, as discussed at the end of Ch. 5, training IVEs on halo mass functions predicted by different emulators and comparing their latent spaces could lead to valuable insights on the robustness of the emulators.

6.2.2 Knowledge extraction with representation learning

Deep learning can be a powerful tool for extracting physical insights on complex problems that are difficult to obtain using traditional approaches like semi-analytical modelling. However, they are difficult to interpret, and the model's flexibility means it could be learning spurious features (overfitting the dataset) rather than features truly relevant to the underlying mapping. To this end, the representation learning approach we have taken is especially promising: it has the flexibility of deep learning models, but the low-dimensional latent space serves as an information bottleneck (Burgess et al., 2018), so the latent space only retains the relevant, important information needed for the mapping rather than spurious features specific to the dataset. Interpreting the low-dimensional, disentangled latent space therefore allows extracting information on the relevant features to the problem.

Wider applications of our framework would, however, benefit from further developments in methods of achieving disentanglement. During training, I have found it difficult to disentangle the latent representation while keeping high prediction accuracy using β-VAE loss. This is because β-VAE loss achieves disentanglement by increasing β, which impacts the model's

prediction accuracy (see Sec. 2.3.1). For learning the non-universal halo mass function which requires sub-percent-level accuracy, achieving disentanglement required large amounts of tuning for β-annealing (see Sec. 4.1.4) and long model training times, which is not needed for applications with less strict accuracy requirements (e.g. for learning halo profiles where log-residuals errors can be at $\gtrsim 10\%$, Lucie-Smith et al., 2022b). It would therefore be extremely useful to devise other methods of achieving a continuous, disentangled latent space with minimal impacts on prediction accuracy. For example, one could incorporate terms in the loss function that specifically penalise high total correlation (one way of generalising mutual information to multiple variables; Kim and Mnih, 2018; Chen et al., 2018). One could also explore combining IVEs with normalising flows (Dinh et al., 2014; Rezende and Mohamed, 2015), which can learn an invertible mapping between an arbitrary distribution and the diagonal Gaussian distribution. Further developments are still necessary to devise an efficient method (with a small number of hyperparameters) for disentangling the latent space. Interpretation would be further facilitated if the latents follow an information hierarchy like we found in Ch. 4; achieving this may require exploring efficient ways of gradually increasing the number of latents that can encode information (Burgess et al., 2018; Ho et al., 2023).

Additionally, the current IVE requires a large training set in order to reach high precision; this prevented us from directly applying the IVE to halo mass functions from simulations to investigate non-universality. The largest available simulation suites to date would only provide a few thousand halo mass functions, which is orders of magnitude smaller than our training set. As mentioned in Sec. 4.3, future work will need to explore ways of training the IVE with less samples to dispense with the need of using an intermediate emulator. This could be done through limiting the number of trainable parameters in the IVE, which in turn requires simplifying the architecture and/or incorporating physical priors. Such approach could, however, impact the ability of IVEs to extract relevant information and form a disentangled compact representation; it could also impact the advantage of IVEs being agnostic to prior assumptions. Another possibility that seems more promising is to pre-train the IVE, e.g. on halo mass functions from emulators, and then fine tune the network to learn halo mass functions directly from numerical simulations.

6.2.3 Knowledge extraction using mutual information

In this thesis, we have made extensive use of mutual information to interpret the model. Mutual information provides an information-theoretic metric that quantifies the amount of information in the latents, and it summarises how latents and variables relate to each other over the dataset. By contrast, some other methods of interpretation, e.g. saliency maps discussed in Sec. 2.2.1, require examining many examples to identify a relation that generalises across the dataset. Furthermore, mutual information allows us to examine the relation between latent variables and a wide variety of quantities beyond just the model inputs and outputs: to interpret the latent space in Ch. 4, we measure mutual information with cosmological parameters, as well as with functions like $f(\sigma)$. We have also identified additional dependencies in latents, and the parameters associated with non-universality, using conditional mutual information. Therefore, mutual information is a very powerful tool towards achieving interpretability.

It is, however, not without flaws. Like other interpretation methods, the mutual information results themselves require some interpretation. This becomes less straightforward if several candidate quantities are correlated. For example, the cosmological parameter space used in Ch. 4 incorporates parameter degeneracies, so one may measure a high mutual information between a latent and a cosmological parameter not because it actually learns this parameter, but because the latent learns another parameter that is degenerate. While this problem can be alleviated somewhat by calculating conditional mutual information (see Sec. 2.4), it points more generally to the difficulty of using mutual information in higher dimensions. Extending mutual information to measure the shared information between more than two variables is complicated; several generalisations of mutual information exist, but their interpretation is difficult (see Finn and Lizier, 2020, and references therein). Another feature of mutual information is that it tests for any relations between variables, not just causal. We therefore require additional tests to establish causality. Ways to do so include checking if the results agree with (or can be explained by) domain knowledge, and performing designed experiments (e.g. running simulations with different $N_{\rm eff}$ values to confirm its role in non-universality). In this regard, the framework developed in this thesis is a helpful tool in the scientific process: it helps with proposing new hypotheses that can then be further tested, leading to the advancement of our scientific knowledge.

As data from surveys such as *Euclid* and LSST become available in the near future, stress testing ΛCDM and probing physics beyond it will require highly accurate theoretical models, and also models of systematics e.g. associated with survey strategy and instrumental noise, which become important due to the data's high precision and low statistical noise. With the small scales probed by forthcoming surveys, baryonic modelling will also become essential. Building accurate models of observables over a large parameter space and running inference will therefore become computationally demanding. The ability of machine learning to accelerate models by emulating, as well as their general ability to deal with large amounts of data, means they will likely form an integral part of survey analyses in the future. Interpretability of the machine learning models used in the pipelines will be important for scientists to fully trust the analysis results. More generally, understanding how a machine learning model produces its outputs enables insights that can lead to building more accurate and robust models both in science (as discussed in this thesis) and beyond. Furthermore, interpretability is crucial in areas where credibility and accountability is key, such as in medicine, finance, and policy making. The methods of representation learning and interpretation developed and used in this thesis have the potential to lead to an understanding of black-box models in a wide range of settings. As discussed at the end of Ch. 5, the learnt representation can potentially also be used to build interpretable models that could replace the black-box models. With these, this thesis also contributes towards the crucial quest of gaining trust in machine learning as the society enters the age of artificial intelligence.

Appendix A

Appendix to Chapter 3

A.1 Deriving expression for mutual information between factor and latent

We derive here Eq. (3.1) used to evaluate the mutual information between latents and factors for the models trained on 3D Shapes (Burgess and Kim, 2018).

The mutual information between two continuous variables x and y is defined as

$$\mathrm{MI}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x\mid y)p(y)\ln\frac{p(x\mid y)}{p(x)}\mathrm{d}x\mathrm{d}y\,, \quad \text{(A.1)}$$

where $p(x,y)$ is the joint probability density function, $p(x)$ and $p(y)$ are the marginals, and $p(x\mid y)$ is the conditional probability density of x given y.

Each latent variable is continuous, whereas each 3D Shapes factor takes on discrete values with equal probability. For a factor f, we can write a generalised probability density function given a population as

$$p(f) = \sum_{i=1}^{M} P(f_i)\delta(f-f_i) = \frac{1}{M}\sum_{i=1}^{M}\delta(f-f_i)\,. \quad \text{(A.2)}$$

Here, δ is the Dirac delta function, $f_i = f_1,\ldots,f_M$ are values that the factor can take, and $P(f_i) = \frac{1}{M}$ because the dataset has equal numbers of each factor value.

The mutual information between one factor and one latent variable κ is then

$$\mathrm{MI}(f,\kappa) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(\kappa \mid f)p(f)\ln\frac{p(\kappa \mid f)}{p(\kappa)}\mathrm{d}f\mathrm{d}\kappa \tag{A.3}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{M}\sum_{i=1}^{M}\delta\left(f-f_i\right)p(\kappa \mid f)\ln\frac{p(\kappa \mid f)}{p(\kappa)}\mathrm{d}f\mathrm{d}\kappa \tag{A.4}$$

$$= \frac{1}{M}\sum_{i=1}^{M}\int_{-\infty}^{\infty} p(\kappa \mid f_i)\ln\frac{p(\kappa \mid f_i)}{p(\kappa)}\mathrm{d}\kappa. \tag{A.5}$$

We can further express $p(\kappa) = \int_{-\infty}^{\infty} p(\kappa \mid f)p(f)df = \frac{1}{M}\sum_{j=1}^{M} p(\kappa \mid f_j)$, so the mutual information between one factor f and one latent variable κ is

$$\mathrm{MI}(f,\kappa) = \frac{1}{M}\sum_{i=1}^{M}\int_{-\infty}^{\infty} p(\kappa \mid f_i)\left[\ln p(\kappa \mid f_i) - \ln\frac{1}{M}\sum_{j=1}^{M} p(\kappa \mid f_j)\right]\mathrm{d}\kappa. \tag{A.6}$$

Appendix B

Appendix to Chapter 4

B.1 Cosmological parameters of the dataset

The cosmological parameter samples we use in our dataset (shown in Fig. 4.2) are generated following the steps below:

1. Draw 10^7 samples from a 7D unit Latin hypercube with sides $[0,1]$.
2. Concatenate two MCMC chains for the constraints which determine the AEMULUS parameter space: Planck13+BAO+SNe (189623 samples) and *e*WMAP9+BAO+SNe (72299 samples) for *w*CDM from Anderson et al. (2014), where *e*WMAP combines WMAP9 data with temperature power spectra from SPT and ACT.
3. Obtain the eigenvectors and eigenvalues σ_i^2 of the concatenated chain by centring the data and performing principle component analysis; the eigenvalues give the variance in the direction of each eigenvector.
4. Project the Latin hypercube into the cosmological parameter space. Each dimension of the Latin hypercube is an eigenvector, and the range for each dimension is scaled from $[0,1]$ to $[-3,3]\times\sigma_i$. Note we choose $\pm3\sigma$ rather than $\pm4\sigma$ because we sample our Latin hypercube much more densely, so if we project to $\pm4\sigma$ a large number of samples will fall outside the region covered by AEMULUS . The MCMC chains were ran varying six cosmological parameters; we include $N_{\rm eff}$ by letting it correspond to the seventh dimension of the Latin hypercube.[1] The range of $N_{\rm eff}$ is determined from the minimum and maximum $N_{\rm eff}$ values in AEMULUS training cosmologies.

[1]The same is done by AEMULUS (DeRose et al., 2019).

5. Discard all samples lying outside the convex hull of AEMULUS training cosmologies. This eliminates the possibility for our network to have trouble learning the halo mass function because of incorrect ground truths produced by extrapolating the emulator outside its domain of validity.

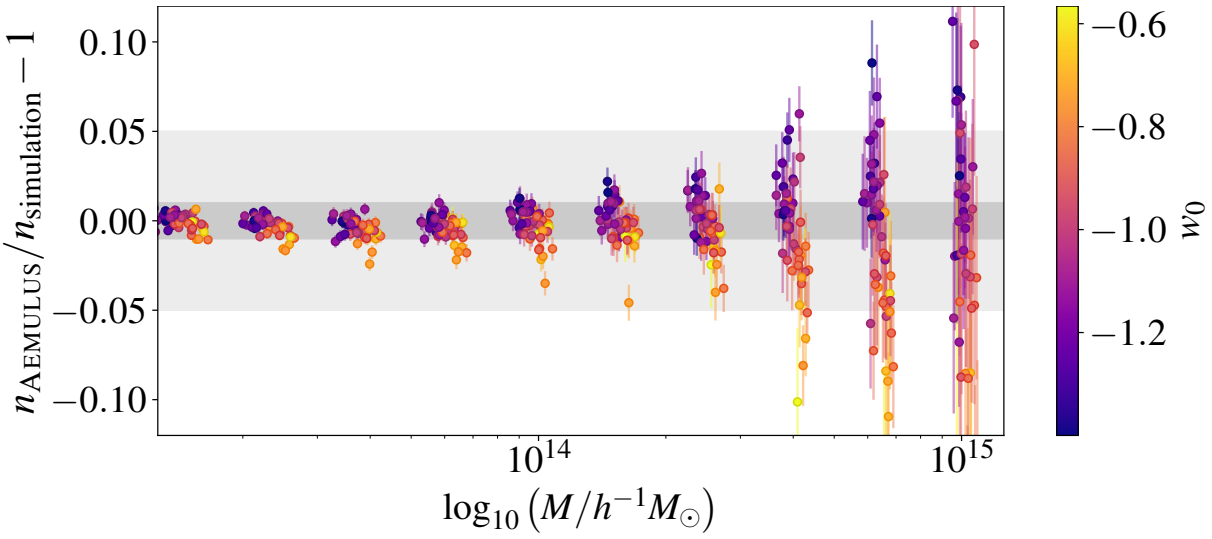
The procedure gives us a list of 628849 cosmological parameter samples for our dataset, from which we randomly sample 10^5 cosmological parameters samples to produce our dataset.

B.2 Bug in AEMULUS

As described in Sec. 1.6.3.3, AEMULUS (McClintock et al., 2019b) predicts the halo mass function by emulating fitting parameters to the multiplicity function $f(\sigma)$, and converts this into the halo mass function by calculating $\sigma(M)$ using CLASS (Lesgourgues, 2011). During our investigation, we have found that AEMULUS misconfigured CLASS such that the power spectrum used to evaluate $\sigma(M)$ is always calculated using $w_0 = -1$, even for $w_0 \neq -1$ cosmologies. Because this misconfiguration was already present when training AEMULUS , this does not affect the overall accuracy of AEMULUS, but it does result in a correlation between the emulator's residuals and w_0. We find for $w_0 > -1$, the emulator underpredicts the halo mass function especially at the high mass tail; the opposite is true for when $w_0 < -1$. This can be seen in Fig. B.1, which plots the halo mass functions and the residuals coloured by w_0 for the AEMULUS training and test set cosmologies.[2]

We expect this to have only a small effect on our results relating to non-universality, as this issue mostly affects high mass halos, whereas we find that non-universality affecting the high mass halos is mostly related to Ω_m rather than w_0. The recent growth history, on which w_0 has a stronger effect, affects mostly the non-universality for lower mass halos where the residuals of AEMULUS is small.

[2]Binned halo counts data and cosmological parameters for the training and test sets are obtained from https://github.com/tmcclintock/Aemulus_data/tree/master/aemulus_data/mass_functions.



(a)

Figure B.1: The residuals of AEMULUS as in Fig. 1.9, but coloured by w_0 for each cosmology. The clear correlation with w_0 suggests that not correctly accounting for w_0 inside AEMULUS when calculating the linear matter power spectrum, and hence σ, affects the emulator's learning of how the halo mass function depends on w_0 at the high-mass end, though the overall magnitude of the errors is unaffected. Error bars indicate Poisson error only.

B.3 Comparing information in the universal and mapping latents with that of their proxies

In Sec. 4.2.3, we have presented the cosmological parameter dependencies of the universal and mapping latents. Here we present additional confirmation showing that the parameter dependencies capture the main information contents in the latents about various quantities of interest: $P(k)$, $D(z)$, and the halo mass function, through the use of mutual information (MI).

In the upper row of Fig. B.2, we compare MI(universal latent, X) in solid line with that of its main parameter dependence, namely with MI($\Omega_m^{0.46}\sigma_8$, X) in dashed line, where X is given by $P(k)$, $D(z)$, and the halo mass function in the three panels from left to right, respectively. We find that the MI curves of the latent and the chosen cosmological parameter combination overlap almost perfectly (except with the halo mass function at $M = 10^{14.6}\,h^{-1}\mathrm{M}_{\odot}$). This agrees with the tight correlation seen between $\Omega_m^{0.46}\sigma_8$ and the universal latent in the top panel of Fig. 4.9, and confirms that this parameter combination is an excellent proxy of the universal latent.

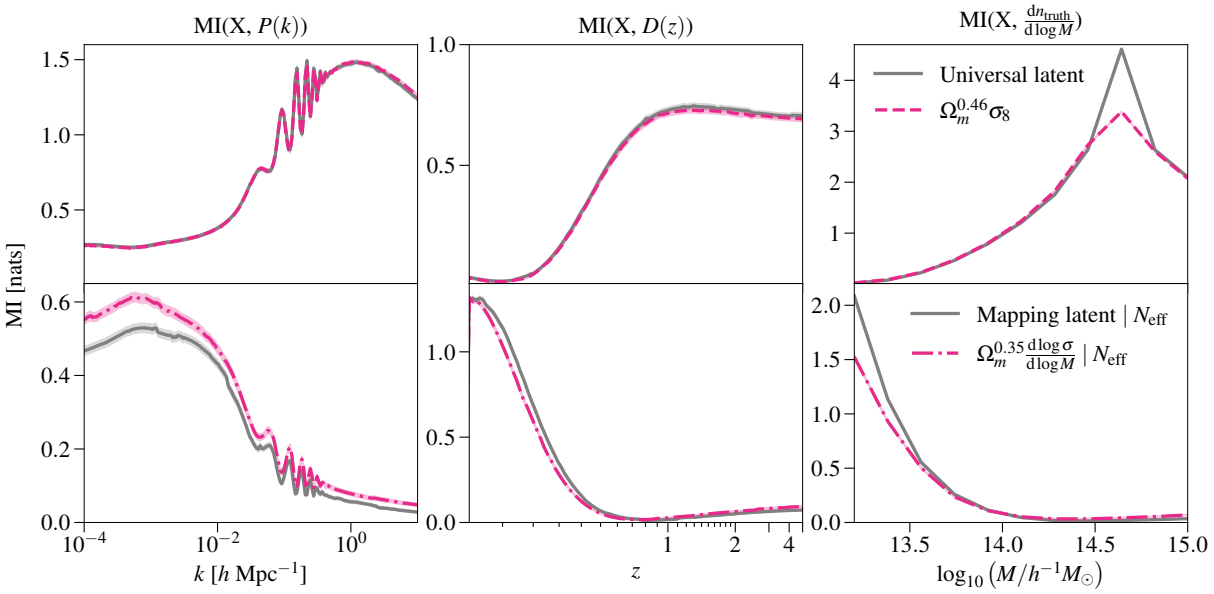


Figure B.2: *Upper row:* MI between the universal latent and three different functions: the linear matter power spectrum $P(k)$ (*left*), the growth function $D(z)$ normalized to unity at $z = 0$ (*middle*), and the ground truth halo mass function (*right*). In pink dashed line, we show the same MI quantities but using $\Omega_m^{0.46}\sigma_8$ instead of the mapping latent. *Lower row:* MI between the mapping latent and the same three functions, conditioned on $N_{\rm eff}$. In pink dash-dotted line, we show the same conditional MI quantities but using $\Omega_m^{0.35}\frac{d\log\sigma}{d\log M}$ evaluated at $M = 10^{13.2}\,h^{-1}\mathrm{M}_\odot$ instead of the mapping latent.

The lower row of Fig. B.2 compares the information content of the mapping latent (solid line) and $\Omega_m^{0.35}\frac{d\log\sigma}{d\log M}$ evaluated at $M = 10^{13.2}\,h^{-1}\mathrm{M}_\odot$ (dash-dotted line) about $P(k)$, $D(z)$ and the halo mass function. Because the correlation between $\Omega_m^{0.35}\frac{d\log\sigma}{d\log M}$ and the mapping latent is not as tight as between the universal latent and its proxy $\Omega_{\rm m}^{0.46}\sigma_8$, here we account for the latent's additional dependence on $N_{\rm eff}$ by conditioning on it. We compare MI(mapping latent, X | $N_{\rm eff}$) and MI(parameter combination, X | $N_{\rm eff}$). We find that although the cosmological parameter combination does not perfectly trace the MI curve for the mapping latent, the two curves overlap closely. This shows that $\Omega_m^{0.35}\frac{d\log\sigma}{d\log M}$ evaluated at $M = 10^{13.2}\,h^{-1}\mathrm{M}_\odot$ has similar amounts of information about $P(k)$, $D(z)$ and the halo mass function as the mapping latent when conditioned on $N_{\rm eff}$, and confirms that it is a good proxy of the cosmological information in the mapping latent.

We note that the mapping latent, which has the highest MI with the HMF at the low-mass end, encodes high information on $P(k)$ at small k (large scales). In contrast, the universal latent, which encodes the most information on the HMF at the high-mass end, has high information with $P(k)$ at larger k (smaller scales). This may seem counter-intuitive at first, but it can be

explained by the main parameter dependence of each latent. The universal latent has high MI with σ_8; from Eq. (1.59), we expect both very large scales and small scales to contribute minimally to the integrand; this is indeed the case in the top left panel of Fig. B.2. On the other hand, the mapping latent has high MI with Ω_m. The power spectrum is affected on all scales by Ω_m, but on small scales it is additionally affected by e.g. the baryon density and $N_{\rm eff}$. Hence, $P(k)$ is most informative of Ω_m on large scales, leading to a high MI between the mapping latent and $P(k)$ on those scales. This also leads to the high MI between the mapping latent and $D(z)$ at low redshifts, when the growth function normalized at $z=0$ is most strongly dependent on Ω_m.

B.4 Relation between recent growth history and growth rate

In Sec. 4.3, we mentioned that the growth factor $D(z_{\rm late})$ where $z_{\rm late}=0.11$ has equivalent information to the growth rate at $z=0.05$. This can be seen from approximating the linear growth factor through integrating the linear growth rate. The latter is approximated by Eq. (1.42) (Linder, 2005).

As we consider $D(z)$ normalised at $z=0$ when interpreting latents, the growth factor can be found with (see e.g. Huterer, 2023)

$$D(z) \simeq \exp\left(\int_0^{\ln a} \mathrm{d}\ln a' \left(\Omega_m(z')\right)^{\gamma}\right), \tag{B.1}$$

where $z' = \frac{1}{a'} - 1$ and $\gamma = 0.55 + 0.05(1+w_0)$ (or $0.55+0.02(1+w_0)$ for $w_0 < -1$, Linder, 2005).

When evaluating Eq. (B.1) at low redshifts, $\ln a \simeq 0$. The range of the integral is small and the integrand does not vary significantly, so it may be factored out. This gives the approximation

$$D(z_{\rm late}) \simeq \exp\left(\Omega_m^{\gamma}(\bar{z})\ln a_{\rm late}\right) = a_{\rm late}^{\Omega_m^{\gamma}(\bar{z})}, \tag{B.2}$$

where $\bar{z}$ is a representative redshift within the integral range. We choose $\bar{z}$ by defining a representative scale factor,

$$\bar{a} \equiv \frac{\int_0^{\ln a_{\rm late}} \mathrm{d}\ln a' a'}{\int_0^{\ln a_{\rm late}} \mathrm{d}\ln a'} = \frac{a_{\rm late}-1}{\ln a_{\rm late}}. \tag{B.3}$$

Inserting $z_{\rm late} = 0.11$ gives $\bar{a} \simeq 0.95$, corresponding to $\bar{z} \simeq 0.05$. Since $a_{\rm late}$ is a constant, the cosmological parameter dependence of the growth factor at $z_{\rm late}$ depends only on $\Omega_m^{\gamma}(\bar{z} = 0.05)$, which approximates the growth rate at $z = 0.05$.

For completeness, we note that

$$\Omega_m(\bar{z}) = \left(1 + \frac{1 - \Omega_{m,0}}{\Omega_{m,0}} \bar{a}^{-3w}\right)^{-1} . \tag{B.4}$$

References

Abazajian, K. N. et al. (2016). "CMB-S4 Science Book, First Edition". *arXiv e-prints*. DOI: 10.48550/arXiv.1610.02743. arXiv: 1610.02743 [astro-ph.CO].

Abbott, B. P. et al. (2016). "Observation of Gravitational Waves from a Binary Black Hole Merger". *Phys. Rev. Lett.* **116** (6), p. 061102. DOI: 10.1103/PhysRevLett.116.061102.

Abbott, T. M. C. et al. (2020). "Dark Energy Survey Year 1 Results: Cosmological constraints from cluster abundances and weak lensing". *Phys. Rev. D* **102**.2, p. 023509. DOI: 10.1103/PhysRevD.102.023509. arXiv: 2002.11124 [astro-ph.CO].

Abbott, T. M. C. et al. (2022). "Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing". *Phys. Rev. D* **105** (2), p. 023520. DOI: 10.1103/PhysRevD.105.023520.

Abdalla, E. et al. (2022). "Cosmology intertwined: A review of the particle physics, astrophysics, and cosmology associated with the cosmological tensions and anomalies". *JHEAP* **34**, pp. 49–211. DOI: 10.1016/j.jheap.2022.04.002.

Abdullah, M. H., Klypin, A., and Wilson, G. (2020). "Cosmological Constraints on Ω_m and σ_8 from Cluster Abundances Using the GalWCat19 Optical-spectroscopic SDSS Catalog". *ApJ* **901**.2, p. 90. DOI: 10.3847/1538-4357/aba619. arXiv: 2002.11907 [astro-ph.CO].

Ade, P. et al. (2019). "The Simons Observatory: science goals and forecasts". *J. Cosmology Astropart. Phys.* **2019**.2, p. 056. DOI: 10.1088/1475-7516/2019/02/056. arXiv: 1808.07445 [astro-ph.CO].

Adhikari, S., Dalal, N., and Chamberlain, R. T. (2014). "Splashback in accreting dark matter halos". *J. Cosmology Astropart. Phys.* **2014**.11, p. 019. DOI: 10.1088/1475-7516/2014/11/019.

Ahmad, Q. R. et al. (2001). "Measurement of the Rate of ν_e + d –> p + p + e^- Interactions Produced by ^{8}B Solar Neutrinos at the Sudbury Neutrino Observatory". *Phys. Rev. Lett.* **87**.7, p. 071301. DOI: 10.1103/PhysRevLett.87.071301. arXiv: nucl-ex/0106015 [nucl-ex].

Aihara, H. et al. (2018). "The Hyper Suprime-Cam SSP Survey: Overview and survey design". *PASJ* **70**, S4. DOI: 10.1093/pasj/psx066. arXiv: 1704.05858 [astro-ph.IM].

Akita, K. and Yamaguchi, M. (2020). "A precision calculation of relic neutrino decoupling". *J. Cosmology Astropart. Phys.* **2020**.8, p. 012. DOI: 10.1088/1475-7516/2020/08/012. arXiv: 2005.07047 [hep-ph].

Alam, S. et al. (2017). "The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample". *MNRAS* **470**.3, pp. 2617–2652. DOI: 10.1093/mnras/stx721. arXiv: 1607.03155 [astro-ph.CO].

Alam, S. et al. (2021). "Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Cosmological implications from two decades of spectroscopic surveys at the Apache Point Observatory". *Phys. Rev. D* **103**.8, p. 083533. DOI: 10.1103/PhysRevD.103.083533. arXiv: 2007.08991 [astro-ph.CO].

Allen, S. W., Evrard, A. E., and Mantz, A. B. (2011). "Cosmological Parameters from Observations of Galaxy Clusters". *ARA&A* **49**.1, pp. 409–470. DOI: 10.1146/annurev-astro-081710-102514. arXiv: 1103.4829 [astro-ph.CO].

Alpher, R. A., Bethe, H., and Gamow, G. (1948). "The Origin of Chemical Elements". *Phys. Rev.* **73** (7), pp. 803–804. DOI: 10.1103/PhysRev.73.803.

Alpher, R. A. and Herman, R. C. (1948). "On the Relative Abundance of the Elements". *Phys. Rev.* **74** (12), pp. 1737–1742. DOI: 10.1103/PhysRev.74.1737.

— (1949). "Remarks on the Evolution of the Expanding Universe". *Phys. Rev.* **75** (7), pp. 1089–1095. DOI: 10.1103/PhysRev.75.1089.

Alsing, J., Charnock, T., Feeney, S., and Wandelt, B. (2019). "Fast likelihood-free cosmology with neural density estimators and active learning". *MNRAS* **488**.3, pp. 4440–4458. DOI: 10.1093/mnras/stz1960. arXiv: 1903.00007 [astro-ph.CO].

Alsing, J. et al. (2020). "SPECULATOR: Emulating Stellar Population Synthesis for Fast and Accurate Galaxy Spectra and Photometry". *ApJS* **249**.1, p. 5. DOI: 10.3847/1538-4365/ab917f.

Amaro-Seoane, P. et al. (2017). "Laser Interferometer Space Antenna". *arXiv e-prints*. DOI: 10.48550/arXiv.1702.00786. arXiv: 1702.00786 [astro-ph.IM].

Amendola, L. et al. (2018). "Cosmology and fundamental physics with the Euclid satellite". *Living Reviews in Relativity* **21**.1, p. 2. DOI: 10.1007/s41114-017-0010-3. arXiv: 1606.00180 [astro-ph.CO].

Amon, A. et al. (2023). "Consistent lensing and clustering in a low-S_8 Universe with BOSS, DES Year 3, HSC Year 1, and KiDS-1000". *MNRAS* **518**.1, pp. 477–503. DOI: 10.1093/mnras/stac2938. arXiv: 2202.07440 [astro-ph.CO].

Anderson, L. et al. (2014). "The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Releases 10 and 11 Galaxy samples". *MNRAS* **441**.1, pp. 24–62. DOI: 10.1093/mnras/stu523.

Angulo, R. E. et al. (2012). "Scaling relations for galaxy clusters in the Millennium-XXL simulation". *MNRAS* **426**.3, pp. 2046–2062. DOI: 10.1111/j.1365-2966.2012.21830.x.

Angulo, R. E. et al. (2021). "The BACCO simulation project: exploiting the full power of large-scale structure for cosmology". *MNRAS* **507**.4, p. 5869. DOI: 10.1093/mnras/stab2018.

Angulo, R. E. and Hahn, O. (2022). "Large-scale dark matter simulations". *Living Reviews in Computational Astrophysics* **8**.1. DOI: 10.1007/s41115-021-00013-z.

Arcadi, G. et al. (2018). "The waning of the WIMP? A review of models, searches, and constraints". *Eur. Phys. J. C* **78**, p. 203. DOI: 10.1140/epjc/s10052-018-5662-y.

Aricò, G. et al. (2021). "The BACCO simulation project: a baryonification emulator with neural networks". *MNRAS* **506**.3, p. 4070. DOI: 10.1093/mnras/stab1911.

Armitage, T. J., Kay, S. T., and Barnes, D. J. (2019). "An application of machine learning techniques to galaxy cluster mass estimation using the MACSIS simulations". *MNRAS* **484**.2, pp. 1526–1537. DOI: 10.1093/mnras/stz039. arXiv: 1810.08430 [astro-ph.CO].

Artis, E., Melin, J.-B., Bartlett, J. G., and Murray, C. (2021). "Impact of the calibration of the halo mass function on galaxy cluster number count cosmology". *A&A* **649**, A47. DOI: 10.1051/0004-6361/202140293. arXiv: 2101.02501 [astro-ph.CO].

Asgari, M., Mead, A. J., and Heymans, C. (2023). "The halo model for cosmology: a pedagogical review". *OJAp* **6**. DOI: 10.21105/astro.2303.08752.

Babcock, H. W. (1939). "The rotation of the Andromeda Nebula". *Lick Observatory Bulletin* **498**, pp. 41–51. DOI: 10.5479/ADS/bib/1939LicOB.19.41B.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (1996). "The quickhull algorithm for convex hulls". *ACM Trans. Math. Softw.* **22**.4, pp. 469–483. DOI: 10.1145/235815.235821.

Bardeen, J. M., Steinhardt, P. J., and Turner, M. S. (1983). "Spontaneous creation of almost scale-free density perturbations in an inflationary universe". *Phys. Rev. D* **28**.4, pp. 679–693. DOI: 10.1103/PhysRevD.28.679.

Barnes, J. and Hut, P. (1986). "A hierarchical O(N log N) force-calculation algorithm". *Nature* **324**.6096, pp. 446–449. DOI: 10.1038/324446a0.

Barnes, J. and Efstathiou, G. (1987). "Angular Momentum from Tidal Torques". *ApJ* **319**, p. 575. DOI: 10.1086/165480.

Barrett, D. G. T. and Dherin, B. (2020). "Implicit Gradient Regularization". *arXiv e-prints*. DOI: 10.48550/arXiv.2009.11162. arXiv: 2009.11162 [cs.LG].

Battaglia, P. W. et al. (2018). "Relational inductive biases, deep learning, and graph networks". *arXiv e-prints*. DOI: 10.48550/arXiv.1806.01261. arXiv: 1806.01261 [cs.LG].

Baudry, J.-P. and Celeux, G. (2015). "EM for mixtures: Initialization requires special care". *Statistics and computing* **25**, pp. 713–726.

Baumann, D. (2009). "TASI Lectures on Inflation". *arXiv e-prints*. DOI: 10.48550/arXiv.0907.5424. arXiv: 0907.5424 [hep-th].

Behroozi, P. S., Wechsler, R. H., and Wu, H.-Y. (2013). "The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores". *ApJ* **762**.2, p. 109. DOI: 10.1088/0004-637X/762/2/109. arXiv: 1110.4372 [astro-ph.CO].

Belghazi, M. I. et al. (2018). "Mutual Information Neural Estimation". *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 531–540.

Beringer, J. et al. (2012). "Review of Particle Physics". *Phys. Rev. D* **86** (1), p. 010001. DOI: 10.1103/PhysRevD.86.010001.

Bertin, E. (1994). "Classification of Astronomical Images with a Neural Network". *Ap&SS* **217**.1-2, pp. 49–51. DOI: 10.1007/BF00990023.

Bertone, G. and Hooper, D. (2018). "History of dark matter". *Rev. Mod. Phys.* **90**.4, p. 045002. DOI: 10.1103/RevModPhys.90.045002.

Bertotti, B., Iess, L., and Tortora, P. (2003). "A test of general relativity using radio links with the Cassini spacecraft". *Nature* **425**.6956, pp. 374–376. DOI: 10.1038/nature01997.

Bertschinger, E. (1995). *Cosmological Dynamics*. arXiv: astro-ph/9503125 [astro-ph].

Bhambra, P., Joachimi, B., and Lahav, O. (2022). "Explaining deep learning of galaxy morphology with saliency mapping". *MNRAS* **511**.4, pp. 5032–5041. DOI: 10.1093/mnras/stac368. arXiv: 2110.08288 [astro-ph.IM].

Bhattacharjee, S., Pandey, B., and Sarkar, S. (2020). "Can a conditioning on stellar mass explain the mutual information between morphology and environment?" *J. Cosmology Astropart. Phys.* **2020**.09, p. 039. DOI: 10.1088/1475-7516/2020/09/039.

Bhattacharya, S. et al. (2011). "Mass Function Predictions Beyond ΛCDM". *ApJ* **732**.2, p. 122. DOI: 10.1088/0004-637X/732/2/122. arXiv: 1005.2239 [astro-ph.CO].

BICEP/Keck Collaboration et al. (2022). "The Latest Constraints on Inflationary B-modes from the BICEP/Keck Telescopes". *arXiv e-prints*. DOI: 10.48550/arXiv.2203.16556. arXiv: 2203.16556 [astro-ph.CO].

Biswas, A. and Mani, K. R. S. (2008). "Relativistic perihelion precession of orbits of Venus and the Earth". *Cent. Eur. J. Phys.* **6**.3, pp. 754–758. DOI: 10.2478/s11534-008-0081-6. arXiv: 0802.0176 [physics.gen-ph].

Blas, D., Lesgourgues, J., and Tram, T. (2011). "The Cosmic Linear Anisotropy Solving System (CLASS). Part II: Approximation schemes". *J. Cosmology Astropart. Phys.* **2011**.7, p. 034. DOI: 10.1088/1475-7516/2011/07/034. arXiv: 1104.2933 [astro-ph.CO].

Bleem, L. E. et al. (2015). "Galaxy Clusters Discovered via the Sunyaev-Zel'dovich Effect in the 2500-Square-Degree SPT-SZ Survey". *ApJS* **216**.2, p. 27. DOI: 10.1088/0067-0049/216/2/27. arXiv: 1409.0850 [astro-ph.CO].

Bocquet, S. et al. (2019). "Cluster Cosmology Constraints from the 2500 deg^2 SPT-SZ Survey: Inclusion of Weak Gravitational Lensing Data from Magellan and the Hubble Space Telescope". *ApJ* **878**.1, p. 55. DOI: 10.3847/1538-4357/ab1f10. arXiv: 1812.01679 [astro-ph.CO].

Bocquet, S. et al. (2024). "SPT Clusters with DES and HST Weak Lensing. II. Cosmological Constraints from the Abundance of Massive Halos". *arXiv e-prints*. DOI: 10.48550/arXiv.2401.02075. arXiv: 2401.02075 [astro-ph.CO].

Bocquet, S., Saro, A., Dolag, K., and Mohr, J. J. (2016). "Halo mass function: baryon impact, fitting formulae, and implications for cluster cosmology". *MNRAS* **456**.3, pp. 2361–2373. DOI: 10.1093/mnras/stv2657. arXiv: 1502.07357 [astro-ph.CO].

Bocquet, S. et al. (2020). "The Mira-Titan Universe. III. Emulation of the Halo Mass Function". *ApJ* **901**.1, p. 5. DOI: 10.3847/1538-4357/abac5c. arXiv: 2003.12116 [astro-ph.CO].

Bode, P., Ostriker, J. P., and Turok, N. (2001). "Halo Formation in Warm Dark Matter Models". *ApJ* **556**.1, pp. 93–107. DOI: 10.1086/321541. arXiv: astro-ph/0010389 [astro-ph].

Bond, J. R., Cole, S., Efstathiou, G., and Kaiser, N. (1991). "Excursion Set Mass Functions for Hierarchical Gaussian Fluctuations". *ApJ* **379**, p. 440. DOI: 10.1086/170520.

Bond, J. R. and Myers, S. T. (1996). "The Peak-Patch Picture of Cosmic Catalogs. II. Validation". *ApJS* **103**, p. 41. DOI: 10.1086/192268.

Bose, S., Deason, A. J., Belokurov, V., and Frenk, C. S. (2020). "The little things matter: relating the abundance of ultrafaint satellites to the hosts' assembly history". *MNRAS* **495**.1, pp. 743–757. DOI: 10.1093/mnras/staa1199. arXiv: 1909.04039 [astro-ph.GA].

Bowden, H., Behroozi, P., and Hearin, A. (2023). "Halo Properties from Observable Measures of Environment: I. Halo and Subhalo Masses". *OJAp* **6**, p. 37. DOI: 10.21105/astro.2307.07549. arXiv: 2307.07549 [astro-ph.GA].

Brout, D. and Scolnic, D. (2021). "It's Dust: Solving the Mysteries of the Intrinsic Scatter and Host-galaxy Dependence of Standardized Type Ia Supernova Brightnesses". *ApJ* **909**.1, p. 26. DOI: 10.3847/1538-4357/abd69b. arXiv: 2004.10206 [astro-ph.CO].

Bryan, G. L. and Norman, M. L. (1998). "Statistical Properties of X-Ray Clusters: Analytic and Numerical Comparisons". *ApJ* **495**.1, pp. 80–99. DOI: 10.1086/305262. arXiv: astro-ph/9710107 [astro-ph].

Bullock, J. S. and Boylan-Kolchin, M. (2017). "Small-Scale Challenges to the ΛCDM Paradigm". *ARA&A* **55**.1, pp. 343–387. DOI: 10.1146/annurev-astro-091916-055313. arXiv: 1707.04256 [astro-ph.CO].

Burgess, C. and Kim, H. (2018). *3D Shapes Dataset*. https://github.com/deepmind/3dshapes-dataset/.

Burgess, C. P. et al. (2018). "Understanding disentangling in β-VAE". *arXiv e-prints*. DOI: 10.48550/arXiv.1804.03599. arXiv: 1804.03599 [stat.ML].

Burles, S., Nollett, K. M., and Turner, M. S. (2000). "Big-Bang Nucleosynthesis Predictions for Precision Cosmology". *ApJ* **552**.1, pp. L1–L5. DOI: 10.1086/320251. arXiv: 0010171 [astro-ph].

Burles, S. and Tytler, D. (1998a). "The Deuterium Abundance toward Q1937-1009". *ApJ* **499**.2, pp. 699–712. DOI: 10.1086/305667. arXiv: astro-ph/9712108 [astro-ph].

— (1998b). "The Deuterium Abundance toward QSO 1009+2956". *ApJ* **507**.2, pp. 732–744. DOI: 10.1086/306341. arXiv: astro-ph/9712109 [astro-ph].

Calabrese, E., Slosar, A., Melchiorri, A., Smoot, G. F., and Zahn, O. (2008). "Cosmic microwave weak lensing data as a test for the dark universe". *Phys. Rev. D* **77**.12, p. 123531. DOI: 10.1103/PhysRevD.77.123531. arXiv: 0803.2309 [astro-ph].

Calderon, V. F. and Berlind, A. A. (2019). "Prediction of galaxy halo masses in SDSS DR7 via a machine learning approach". *MNRAS* **490**.2, pp. 2367–2379. DOI: 10.1093/mnras/stz2775. arXiv: 1902.02680 [astro-ph.GA].

Capozzi, F. et al. (2021). "Unfinished fabric of the three neutrino paradigm". *Phys. Rev. D* **104** (8), p. 083031. DOI: 10.1103/PhysRevD.104.083031.

Carbonneau, M.-A., Zaidi, J., Boilard, J., and Gagnon, G. (2022). "Measuring Disentanglement: A Review of Metrics". arXiv: 2012.09276 [cs.LG].

Carlstrom, J. E. et al. (2011). "The 10 Meter South Pole Telescope". *PASP* **123**.903, p. 568. DOI: 10.1086/659879. arXiv: 0907.4445 [astro-ph.IM].

Castorina, E., Paranjape, A., Hahn, O., and Sheth, R. K. (2016). "Excursion set peaks: the role of shear". *arXiv e-prints*. DOI: 10.48550/arXiv.1611.03619. arXiv: 1611.03619 [astro-ph.CO].

Cerardi, N., Pierre, M., Valageas, P., Garrel, C., and Pacaud, F. (2024). "The cosmological analysis of X-ray cluster surveys. V. The potential of cluster counts in the $1 < z < 2$ range". *A&A* **682**, A138. DOI: 10.1051/0004-6361/202347699. arXiv: 2312.04253 [astro-ph.CO].

Chen, R. T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018). "Isolating Sources of Disentanglement in Variational Autoencoders". *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc.

Chevallier, M. and Polarski, D. (2001). "Accelerating Universes with Scaling Dark Matter". *IJMPD* **10**.2, pp. 213–223. DOI: 10.1142/S0218271801000822. arXiv: gr-qc/0009008 [gr-qc].

Chisari, N. E. et al. (2019). "Core Cosmology Library: Precision Cosmological Predictions for LSST". *ApJS* **242**.1, p. 2. DOI: 10.3847/1538-4365/ab1658. arXiv: 1812.05995 [astro-ph.CO].

Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., and Griguta, V. (2020). "Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra". *A&A* **639**, A84. DOI: 10.1051/0004-6361/201936770. arXiv: 1909.10963 [astro-ph.GA].

Clemence, G. M. (1947). "The Relativity Effect in Planetary Motions". *Rev. Mod. Phys.* **19** (4), pp. 361–364. DOI: 10.1103/RevModPhys.19.361.

Colless, M. et al. (2001). "The 2dF Galaxy Redshift Survey: spectra and redshifts". *MNRAS* **328**.4, pp. 1039–1063. DOI: 10.1046/j.1365-8711.2001.04902.x. arXiv: astro-ph/0106498 [astro-ph].

Correa, C. A., Wyithe, J. S. B., Schaye, J., and Duffy, A. R. (2015). "The accretion history of dark matter haloes – II. The connections with the mass power spectrum and the density profile". *MNRAS* **450**.2, pp. 1521–1537. DOI: 10.1093/mnras/stv697.

Costanzi, M. et al. (2021). "Cosmological constraints from DES Y1 cluster abundances and SPT multiwavelength data". *Phys. Rev. D* **103**.4, p. 043522. DOI: 10.1103/PhysRevD.103.043522. arXiv: 2010.13800 [astro-ph.CO].

Courtin, J. et al. (2010). "Imprints of dark energy on cosmic structure formation - II. Non-universality of the halo mass function". *MNRAS*, no–no. DOI: 10.1111/j.1365-2966.2010.17573.x.

Crain, R. A. and van de Voort, F. (2023). "Hydrodynamical Simulations of the Galaxy Population: Enduring Successes and Outstanding Challenges". *ARA&A* **61**, pp. 473–515. DOI: 10.1146/annurev-astro-041923-043618. arXiv: 2309.17075 [astro-ph.GA].

Cranmer, M. et al. (2020). "Discovering Symbolic Models from Deep Learning with Inductive Biases". arXiv: 2006.11287 [cs.LG].

Crocce, M., Pueblas, S., and Scoccimarro, R. (2006). "Transients from initial conditions in cosmological simulations". *MNRAS* **373**.1, pp. 369–381. DOI: 10.1111/j.1365-2966.2006.11040.x.

Crocce, M., Fosalba, P., Castander, F. J., and Gaztañaga, E. (2010). "Simulating the Universe with MICE: the abundance of massive clusters". *MNRAS* **403**.3, pp. 1353–1367. DOI: 10.1111/j.1365-2966.2009.16194.x. arXiv: 0907.0019 [astro-ph.CO].

Cybenko, G. V. (1989). "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals and Systems* **2**, pp. 303–314.

Dai, B. and Seljak, U. (2021). "Learning effective physical laws for generating cosmological hydrodynamics with Lagrangian deep learning". *Proceedings of the National Academy of Sciences* **118**.16, e2020324118. DOI: 10.1073/pnas.2020324118.

Dai, Z., Moews, B., Vilalta, R., and Davé, R. (2023). "Physics-informed neural networks in the recreation of hydrodynamic simulations from dark matter". *MNRAS* **527**.2, pp. 3381–3394. DOI: 10.1093/mnras/stad3394. eprint: https://academic.oup.com/mnras/article-pdf/527/2/3381/53668171/stad3394.pdf.

Darbellay, G. and Vajda, I. (1999). "Estimation of the information by an adaptive partitioning of the observation space". *IEEE Transactions on Information Theory* **45**.4, pp. 1315–1321. DOI: 10.1109/18.761290.

Dark Energy Survey and Kilo-Degree Survey Collaboration et al. (2023). "DES Y3 + KiDS-1000: Consistent cosmology combining cosmic shear surveys". *OJAp* **6**, p. 36. DOI: 10.21105/astro.2305.17173. arXiv: 2305.17173 [astro-ph.CO].

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". *ApJ* **292**, pp. 371–394. DOI: 10.1086/163168.

Davis, M., Huchra, J., Latham, D. W., and Tonry, J. (1982). "A survey of galaxy redshifts. II. The large scale space distribution." *ApJ* **253**, pp. 423–445. DOI: 10.1086/159646.

de Jong, J. T. A. et al. (2013). "The Kilo-Degree Survey". *The Messenger* **154**, pp. 44–46.

de Salas, P. F. et al. (2021). "2020 global reassessment of the neutrino oscillation picture". *JHEP* **2021**.2, p. 71. DOI: 10.1007/JHEP02(2021)071. arXiv: 2006.11237 [hep-ph].

de Salas, P. F. and Pastor, S. (2016). "Relic neutrino decoupling with flavour oscillations revisited". *J. Cosmology Astropart. Phys.* **2016**.7, p. 051. DOI: 10.1088/1475-7516/2016/07/051. arXiv: 1606.06986 [hep-ph].

Del Popolo, A. and Le Delliou, M. (2017). "Small Scale Problems of the ΛCDM Model: A Short Review". *Galaxies* **5**.1, p. 17. DOI: 10.3390/galaxies5010017.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *J. R. Stat. Soc. Ser. B Methodol.* **39**.1, pp. 1–38.

DeRose, J. et al. (2019). "The Aemulus Project. I. Numerical Simulations for Precision Cosmology". *ApJ* **875**.1, p. 69. DOI: 10.3847/1538-4357/ab1085.

DESI Collaboration et al. (2016). "The DESI Experiment Part I: Science,Targeting, and Survey Design". DOI: 10.48550/arXiv.1611.00036. arXiv: 1611.00036 [astro-ph.IM].

DESI Collaboration et al. (2024). "DESI 2024 VI: Cosmological Constraints from the Measurements of Baryon Acoustic Oscillations". *arXiv e-prints*. DOI: 10.48550/arXiv.2404.03002. arXiv: 2404.03002 [astro-ph.CO].

Despali, G. et al. (2016). "The universality of the virial halo mass function and models for non-universality of other halo definitions". *MNRAS* **456**.3, pp. 2486–2504. DOI: 10.1093/mnras/stv2842. arXiv: 1507.05627 [astro-ph.CO].

Dicke, R. H., Peebles, P. J. E., Roll, P. G., and Wilkinson, D. T. (1965). "Cosmic Black-Body Radiation." *ApJ* **142**, pp. 414–419. DOI: 10.1086/148306.

Diemer, B. (2020). "Universal at Last? The Splashback Mass Function of Dark Matter Halos". *ApJ* **903**.2, p. 87. DOI: 10.3847/1538-4357/abbf52.

Diemer, B. and Kravtsov, A. V. (2014). "Dependence of the Outer Density Profiles of Halos on Their Mass Accretion Rate". *ApJ* **789**.1, p. 1. DOI: 10.1088/0004-637X/789/1/1. arXiv: 1401.1216 [astro-ph.CO].

Dinh, L., Krueger, D., and Bengio, Y. (2014). "NICE: Non-linear Independent Components Estimation". *arXiv e-prints*. DOI: 10.48550/arXiv.1410.8516. arXiv: 1410.8516 [cs.LG].

Dodelson, S. and Schmidt, F. (2020). *Modern Cosmology*. DOI: 10.1016/C2017-0-01943-2.

Dodelson, S. et al. (2016). *Cosmic Visions Dark Energy: Science*. arXiv: 1604 . 07626 [astro-ph.CO].

Doeser, L. et al. (2024). "Bayesian inference of initial conditions from non-linear cosmic structures using field-level emulators". *MNRAS* **535**.2, pp. 1258–1277. DOI: 10.1093/mnras/stae2429. arXiv: 2312.09271 [astro-ph.CO].

Domínguez-Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., and Fischer, J. L. (2018). "Improving galaxy morphologies for SDSS with Deep Learning". *MNRAS* **476**.3, pp. 3661–3676. DOI: 10.1093/mnras/sty338. arXiv: 1711.05744 [astro-ph.GA].

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo". *Phys. Lett. B* **195**.2, pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.

Dubey, S. R., Singh, S. K., and Chaudhuri, B. B. (2021). "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark". *arXiv e-prints*. DOI: 10.48550/arXiv.2109.14545. arXiv: 2109.14545 [cs.LG].

Dumoulin, V. and Visin, F. (2016). "A guide to convolution arithmetic for deep learning". *arXiv e-prints*. DOI: 10.48550/arXiv.1603.07285. arXiv: 1603.07285 [stat.ML].

Dyson, F. W., Eddington, A. S., and Davidson, C. (1920). "A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919". *Philosophical Transactions of the Royal Society of London Series A* **220**, pp. 291–333. DOI: 10.1098/rsta.1920.0009.

Eddington, A. S. (1923). *The mathematical theory of relativity*.

Efstathiou, G., Davis, M., White, S. D. M., and Frenk, C. S. (1985). "Numerical techniques for large cosmological N-body simulations". *ApJS* **57**, pp. 241–260. DOI: 10.1086/191003.

Efstathiou, G., Frenk, C. S., White, S. D., and Davis, M. (1988). "Gravitational clustering from scale-free initial conditions". *MNRAS* **235**, pp. 715–748.

Eisenstein, D. J. and Hu, W. (1998). "Baryonic Features in the Matter Transfer Function". *ApJ* **496**.2, pp. 605–614. DOI: 10.1086/305424.

— (1999). "Power Spectra for Cold Dark Matter and Its Variants". *ApJ* **511**.1, pp. 5–15. DOI: 10.1086/306640. arXiv: astro-ph/9710252 [astro-ph].

Eke, V. R., Cole, S., and Frenk, C. S. (1996). "Cluster evolution as a diagnostic for Omega". *MNRAS* **282**, pp. 263–280. DOI: 10.1093/mnras/282.1.263. arXiv: astro-ph/9601088 [astro-ph].

Esteban, I., Gonzalez-Garcia, M. C., Maltoni, M., Schwetz, T., and Zhou, A. (2020). "The fate of hints: updated global analysis of three-flavor neutrino oscillations". *JHEP* **2020**.9, p. 178. DOI: 10.1007/JHEP09(2020)178. arXiv: 2007.14792 [hep-ph].

Euclid Collaboration et al. (2019). "Euclid preparation. III. Galaxy cluster detection in the wide photometric survey, performance and algorithm selection". *A&A* **627**, A23. DOI: 10.1051/0004-6361/201935088. arXiv: 1906.04707 [astro-ph.CO].

Euclid Collaboration et al. (2023a). "Euclid preparation. XXIV. Calibration of the halo mass function in $\Lambda(\nu)$CDM cosmologies". *A&A* **671**, A100. DOI: 10.1051/0004-6361/202244674. arXiv: 2208.02174 [astro-ph.CO].

Euclid Collaboration et al. (2023b). "Euclid preparation TBD. The effect of baryons on the Halo Mass Function". *arXiv e-prints*, arXiv:2311.01465. DOI: 10.48550/arXiv.2311.01465. arXiv: 2311.01465 [astro-ph.CO].

Euclid Collaboration et al. (2024). "Euclid. V. The Flagship galaxy mock catalogue: a comprehensive simulation for the Euclid mission". *arXiv e-prints*. DOI: 10.48550/arXiv.2405.13495. arXiv: 2405.13495 [astro-ph.CO].

Ferreira, E. G. M. (2021). "Ultra-light dark matter". *A&A Rev.* **29**.1, p. 7. DOI: 10.1007/s00159-021-00135-6. arXiv: 2005.03254 [astro-ph.CO].

Finn, C. and Lizier, J. T. (2020). "Generalised Measures of Multivariate Information Content". *Entropy* **22**.2. DOI: 10.3390/e22020216.

Fixsen, D. J. (2009). "The Temperature of the Cosmic Microwave Background". *ApJ* **707**.2, pp. 916–920. DOI: 10.1088/0004-637X/707/2/916. arXiv: 0911.1955 [astro-ph.CO].

Flores, R. A. and Primack, J. R. (1994). "Observational and Theoretical Constraints on Singular Dark Matter Halos". *ApJ* **427**, p. L1. DOI: 10.1086/187350. arXiv: astro-ph/9402004 [astro-ph].

Fluke, C. J. and Jacobs, C. (2019). "Surveying the reach and maturity of machine learning and artificial intelligence in astronomy". *WIREs Data Mining and Knowledge Discovery* **10**.2. DOI: 10.1002/widm.1349.

Fontenla-Romero, O., Erdogmus, D., Principe, J. C., Alonso-Betanzos, A., and Castillo, E. (2003). "Linear Least-Squares Based Methods for Neural Networks Learning". *Artificial Neural Networks and Neural Information Processing — ICANN/ICONIP 2003*. Ed. by O. Kaynak, E. Alpaydin, E. Oja, and L. Xu. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 84–91.

Francis, P. J., Hewett, P. C., Foltz, C. B., and Chaffee, F. H. (1992). "An Objective Classification Scheme for QSO Spectra". *ApJ* **398**, p. 476. DOI: 10.1086/171870.

Fraser, A. M. and Swinney, H. L. (1986). "Independent coordinates for strange attractors from mutual information". *Phys. Rev. A* **33** (2), pp. 1134–1140. DOI: 10.1103/PhysRevA.33.1134.

Freedman, W. L. et al. (2024). "Status Report on the Chicago-Carnegie Hubble Program (CCHP): Three Independent Astrophysical Determinations of the Hubble Constant Using the James Webb Space Telescope". *arXiv e-prints*. DOI: 10.48550/arXiv.2408.06153. arXiv: 2408.06153 [astro-ph.CO].

Frenzel, S. and Pompe, B. (2007). "Partial Mutual Information for Coupling Analysis of Multivariate Time Series". *Phys. Rev. Lett.* **99** (20), p. 204101. DOI: 10.1103/PhysRevLett.99.204101.

Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine". *Annals of statistics*, pp. 1189–1232.

Friedmann, A. (1922). "Über die Krümmung des Raumes". *Zeitschrift fur Physik* **10**, pp. 377–386. DOI: 10.1007/BF01332580.

— (1924). "Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes". *Zeitschrift fur Physik* **21**.1, pp. 326–332. DOI: 10.1007/BF01328280.

Fukuda, Y. et al. (1998). "Evidence for Oscillation of Atmospheric Neutrinos". *Phys. Rev. Lett.* **81**.8, pp. 1562–1567. DOI: 10.1103/PhysRevLett.81.1562. arXiv: hep-ex/9807003 [hep-ex].

Fukushima, K. (1969). "Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements". *IEEE Transactions on Systems Science and Cybernetics* **5**.4, pp. 322–333. DOI: 10.1109/TSSC.1969.300225.

Gaia Collaboration et al. (2016). "The Gaia mission". *A&A* **595**, A1. DOI: 10.1051/0004-6361/201629272. arXiv: 1609.04153 [astro-ph.IM].

Gamow, G. (1946). "Expanding Universe and the Origin of Elements". *Phys. Rev.* **70** (7-8), pp. 572–573. DOI: 10.1103/PhysRev.70.572.2.

Gariazzo, S., de Salas, P. F., and Pastor, S. (2019). "Thermalisation of sterile neutrinos in the early universe in the 3+1 scheme with full mixing matrix". *J. Cosmology Astropart. Phys.* **2019**.7, p. 014. DOI: 10.1088/1475-7516/2019/07/014. arXiv: 1905.11290 [astro-ph.CO].

Gatti, M. et al. (2021). "Dark energy survey year 3 results: weak lensing shape catalogue". *MNRAS* **504**.3, pp. 4312–4336. DOI: 10.1093/mnras/stab918. arXiv: 2011.03408 [astro-ph.CO].

Ghirardini, V. et al. (2024). "The SRG/eROSITA all-sky survey: Cosmology constraints from cluster abundances in the western Galactic hemisphere". *A&A* **689**, A298. DOI: 10.1051/0004-6361/202348852. arXiv: 2402.08458 [astro-ph.CO].

Giblin, B. et al. (2021). "KiDS-1000 catalogue: Weak gravitational lensing shear measurements". *A&A* **645**, A105. DOI: 10.1051/0004-6361/202038850. arXiv: 2007.01845 [astro-ph.CO].

Giles, D. and Walkowicz, L. (2019). "Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection". *MNRAS* **484**.1, pp. 834–849. DOI: 10.1093/mnras/sty3461. arXiv: 1812.07156 [astro-ph.IM].

Glorot, X. and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks". *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Y. W. Teh and M. Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 249–256.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Gott J. Richard, I. et al. (2005). "A Map of the Universe". *ApJ* **624**.2, pp. 463–484. DOI: 10.1086/428890. arXiv: astro-ph/0310571 [astro-ph].

Greenacre, M. et al. (2022). "Principal component analysis". *Nature Reviews Methods Primers* **2**.1, p. 100. DOI: 10.1038/s43586-022-00184-w.

Gunn, J. E. and Gott J. Richard, I. (1972). "On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution". *ApJ* **176**, p. 1. DOI: 10.1086/151605.

Guo, N., Lucie-Smith, L., Peiris, H. V., Pontzen, A., and Piras, D. (2024). "Deep learning insights into non-universality in the halo mass function". *MNRAS* **532**.4, pp. 4141–4156. DOI: 10.1093/mnras/stae1696. arXiv: 2405.15850 [astro-ph.CO].

Guth, A. H. and Pi, S. .-Y. (1982). "Fluctuations in the New Inflationary Universe". *Phys. Rev. Lett.* **49**.15, pp. 1110–1113. DOI: 10.1103/PhysRevLett.49.1110.

Guth, A. H. (1981). "Inflationary universe: A possible solution to the horizon and flatness problems". *Phys. Rev. D* **23**.2, pp. 347–356. DOI: 10.1103/PhysRevD.23.347.

Handley, W. J., Hobson, M. P., and Lasenby, A. N. (2015a). "polychord: nested sampling for cosmology." *MNRAS* **450**, pp. L61–L65. DOI: 10.1093/mnrasl/slv047. arXiv: 1502.01856 [astro-ph.CO].

— (2015b). "POLYCHORD: next-generation nested sampling". *MNRAS* **453**.4, pp. 4384–4398. DOI: 10.1093/mnras/stv1911. arXiv: 1506.00171 [astro-ph.IM].

Hawking, S. W. (1982). "The development of irregularities in a single bubble inflationary universe". *Phys. Lett. B* **115**.4, pp. 295–297. DOI: 10.1016/0370-2693(82)90373-2.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". *2015 IEEE Int'l Conf. on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123. arXiv: 1502.01852 [cs.CV].

Heitmann, K. et al. (2016). "The Mira-Titan Universe: Precision Predictions for Dark Energy Surveys". *ApJ* **820**.2, p. 108. DOI: 10.3847/0004-637X/820/2/108. arXiv: 1508.02654 [astro-ph.CO].

Higgins, I. et al. (2017). "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". *ICLR*.

Hildebrandt, H. et al. (2021). "KiDS-1000 catalogue: Redshift distributions and their calibration". *A&A* **647**, A124. DOI: 10.1051/0004-6361/202039018. arXiv: 2007.15635 [astro-ph.CO].

Hirashima, K. et al. (2023). "Surrogate Modeling for Computationally Expensive Simulations of Supernovae in High-Resolution Galaxy Simulations". *arXiv e-prints*. DOI: 10.48550/arXiv.2311.08460. arXiv: 2311.08460 [astro-ph.GA].

Ho, M., Farahi, A., Rau, M. M., and Trac, H. (2021). "Approximate Bayesian Uncertainties on Deep Learning Dynamical Mass Estimates of Galaxy Clusters". *ApJ* **908**.2, p. 204. DOI: 10.3847/1538-4357/abd101. arXiv: 2006.13231 [astro-ph.CO].

Ho, M., Zhao, X., and Wandelt, B. (2023). "Information-Ordered Bottlenecks for Adaptive Semantic Compression". *arXiv e-prints*. DOI: 10.48550/arXiv.2305.11213. arXiv: 2305.11213 [cs.LG].

Ho, M. et al. (2019). "A Robust and Efficient Deep Learning Method for Dynamical Mass Measurements of Galaxy Clusters". *ApJ* **887**.1, p. 25. DOI: 10.3847/1538-4357/ab4f82. arXiv: 1902.05950 [astro-ph.CO].

Ho, T. K. (1995). "Random Decision Forests". *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*. ICDAR '95. M: IEEE Computer Society, pp. 278–282.

Hofmann, F. et al. (2017). "eROSITA cluster cosmology forecasts: Cluster temperature substructure bias". *A&A* **606**, A118. DOI: 10.1051/0004-6361/201730742. arXiv: 1708.05205 [astro-ph.CO].

Holmes, C. M. and Nemenman, I. (2019). "Estimation of mutual information for real-valued data with error bars and controlled bias". *Phys. Rev. E* **100** (2), p. 022404. DOI: 10.1103/PhysRevE.100.022404.

Homma, D. et al. (2024). "Final results of the search for new Milky Way satellites in the Hyper Suprime-Cam Subaru Strategic Program survey: Discovery of two more candidates". *PASJ* **76**.4, pp. 733–752. DOI: 10.1093/pasj/psae044. arXiv: 2311.05439 [astro-ph.GA].

Hu, W. (1995). "Wandering in the Background: A CMB Explorer". *arXiv e-prints*. DOI: 10.48550/arXiv.astro-ph/9508126. arXiv: astro-ph/9508126 [astro-ph].

— (2001). *Ringing in the New Cosmology: Intermediate Guide to the Acoustic Peaks and Polarization*. URL: https://pontzen.co.uk/PHAS0067/.

— (2017). *Astro 321 Set 7: Spherical Collapse & Halo Model [Lecture slides]*. URL: https://background.uchicago.edu/~whu/Courses/Ast321_17/ast321_7.pdf.

Hu, W. and Dodelson, S. (2002). "Cosmic Microwave Background Anisotropies". *ARA&A* **40**, pp. 171–216. DOI: 10.1146/annurev.astro.40.060401.093926. arXiv: astro-ph/0110414 [astro-ph].

Hu, W. and Sawicki, I. (2007). "Parametrized post-Friedmann framework for modified gravity". *Phys. Rev. D* **76**.10. DOI: 10.1103/physrevd.76.104043.

Hubble, E. (1929). "A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae". *Proceedings of the National Academy of Science* **15**.3, pp. 168–173. DOI: 10.1073/pnas.15.3.168.

Huertas-Company, M. and Lanusse, F. (2023). "The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys". *PASA* **40**, e001. DOI: 10.1017/pasa.2022.55. arXiv: 2210.01813 [astro-ph.IM].

Huertas-Company, M. et al. (2018). "Deep Learning Identifies High-z Galaxies in a Central Blue Nugget Phase in a Characteristic Mass Range". *ApJ* **858**.2, p. 114. DOI: 10.3847/1538-4357/aabfed. arXiv: 1804.07307 [astro-ph.GA].

Hui, L., Ostriker, J. P., Tremaine, S., and Witten, E. (2017). "Ultralight scalars as cosmological dark matter". *Phys. Rev. D* **95**.4, p. 043541. DOI: 10.1103/PhysRevD.95.043541. arXiv: 1610.08297 [astro-ph.CO].

Huterer, D. (2023). "Growth of cosmic structure". *A&A Rev.* **31**.1, p. 2. DOI: 10.1007/s00159-023-00147-4. arXiv: 2212.05003 [astro-ph.CO].

Huterer, D. and Shafer, D. L. (2017). "Dark energy two decades after: observables, probes, consistency tests". *Reports on Progress in Physics* **81**.1, p. 016901. DOI: 10.1088/1361-6633/aa997e.

Iten, R., Metger, T., Wilming, H., Rio, L. del, and Renner, R. (2020). "Discovering Physical Concepts with Neural Networks". *Phys. Rev. Lett.* **124**.1. DOI: 10.1103/physrevlett.124.010508.

Ivezić, Ž. et al. (2019). "LSST: From Science Drivers to Reference Design and Anticipated Data Products". *ApJ* **873**.2, p. 111. DOI: 10.3847/1538-4357/ab042c. arXiv: 0805.2366 [astro-ph].

Jaffe, A. H. (2012). *Cosmology 2012: Lecture Notes*. URL: https://www.sr.bham.ac.uk/~smcgee/ObsCosmo/Jaffe_cosmology.pdf.

Jain, B. and Seljak, U. (1997). "Cosmological Model Predictions for Weak Lensing: Linear and Nonlinear Regimes". *ApJ* **484**.2, pp. 560–573. DOI: 10.1086/304372. arXiv: astro-ph/9611077 [astro-ph].

Jamieson, D. et al. (2023). "Field-level Neural Network Emulator for Cosmological N-body Simulations". *ApJ* **952**.2, p. 145. DOI: 10.3847/1538-4357/acdb6c. arXiv: 2206.04594 [astro-ph.CO].

Jeffrey, N. et al. (2024). "Dark Energy Survey Year 3 results: likelihood-free, simulation-based *w*CDM inference with neural compression of weak-lensing map statistics". *arXiv e-prints*. DOI: 10.48550/arXiv.2403.02314. arXiv: 2403.02314 [astro-ph.CO].

Jenkins, A. et al. (2001). "The mass function of dark matter haloes". *MNRAS* **321**.2, pp. 372–384. DOI: 10.1046/j.1365-8711.2001.04029.x.

Joachimi, B. et al. (2015). "Galaxy Alignments: An Overview". *Space Sci. Rev.* **193**.1-4, pp. 1–65. DOI: 10.1007/s11214-015-0177-4. arXiv: 1504.05456 [astro-ph.GA].

Kaiser, N. (1986). "Evolution and clustering of rich clusters." *MNRAS* **222**, pp. 323–345. DOI: 10.1093/mnras/222.2.323.

Kim, H. and Mnih, A. (2018). "Disentangling by Factorising". arXiv: 1802.05983 [stat.ML].

Kingma, D. P. and Ba, J. (2017). "Adam: A method for stochastic optimization". arXiv: 1412.6980 [cs.LG].

Kingma, D. P. and Welling, M. (2013). "Auto-Encoding Variational Bayes". *arXiv e-prints*, arXiv:1312.6114. DOI: 10.48550/arXiv.1312.6114. arXiv: 1312.6114 [stat.ML].

— (2019). "An Introduction to Variational Autoencoders". *Foundations and Trends® in Machine Learning* **12**.4, pp. 307–392. DOI: 10.1561/2200000056.

Kinney, J. B. and Atwal, G. S. (2014). "Equitability, mutual information, and the maximal information coefficient". *Proceedings of the National Academy of Sciences* **111**.9, pp. 3354–3359. DOI: 10.1073/pnas.1309933111. eprint: https://www.pnas.org/content/111/9/3354.full.pdf.

Kleinebreil, F. et al. (2024). "The SRG/eROSITA All-Sky Survey: Weak-Lensing of eRASS1 Galaxy Clusters in KiDS-1000 and Consistency Checks with DES Y3 & HSC-Y3". *arXiv e-prints*. DOI: 10.48550/arXiv.2402.08456. arXiv: 2402.08456 [astro-ph.CO].

Klypin, A., Kravtsov, A. V., Valenzuela, O., and Prada, F. (1999). "Where Are the Missing Galactic Satellites?" *ApJ* **522**.1, pp. 82–92. DOI: 10.1086/307643. arXiv: 9901240 [astro-ph].

Kodi Ramanah, D., Wojtak, R., Ansari, Z., Gall, C., and Hjorth, J. (2020). "Dynamical mass inference of galaxy clusters with neural flows". *MNRAS* **499**.2, pp. 1985–1997. DOI: 10.1093/mnras/staa2886. arXiv: 2003.05951 [astro-ph.CO].

Kodi Ramanah, D., Wojtak, R., and Arendse, N. (2021). "Simulation-based inference of dynamical galaxy cluster masses with 3D convolutional neural networks". *MNRAS* **501**.3, pp. 4080–4091. DOI: 10.1093/mnras/staa3922. arXiv: 2009.03340 [astro-ph.CO].

Koza, J. R. (1994). "Genetic programming as a means for programming computers by natural selection". *Statistics and Computing* **4**.2, pp. 87–112. DOI: 10.1007/BF00175355.

Kramer, M. (1992). "Autoassociative neural networks". *Computers and Chemical Engineering* **16**.4. Neutral network applications in chemical engineering, pp. 313–328. DOI: https://doi.org/10.1016/0098-1354(92)80051-A.

Kramer, M. A. (1991). "Nonlinear principal component analysis using autoassociative neural networks". *AIChE Journal* **37**.2, pp. 233–243. DOI: 10.1002/aic.690370209.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). "Estimating mutual information". *Phys. Rev. E* **69** (6), p. 066138. DOI: 10.1103/PhysRevE.69.066138.

Kravtsov, A. V. and Borgani, S. (2012). "Formation of Galaxy Clusters". *ARA&A* **50**.1, pp. 353–409. DOI: 10.1146/annurev-astro-081811-125502. eprint: https://doi.org/10.1146/annurev-astro-081811-125502.

Kravtsov, A. V., Vikhlinin, A., and Nagai, D. (2006). "A New Robust Low-Scatter X-Ray Mass Indicator for Clusters of Galaxies". *ApJ* **650**.1, pp. 128–136. DOI: 10.1086/506319. arXiv: astro-ph/0603205 [astro-ph].

Kullback, S. and Leibler, R. A. (1951). "On Information and Sufficiency". *Ann. Math. Statist.* **22**.1, pp. 79–86.

Kwak, N. and Choi, C.-H. (2002). "Input feature selection by mutual information based on Parzen window". *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**.12, pp. 1667–1671. DOI: 10.1109/TPAMI.2002.1114861.

Lacey, C. and Cole, S. (1994). "Merger Rates in Hierarchical Models of Galaxy Formation - Part Two - Comparison with N-Body Simulations". *MNRAS* **271**, p. 676. DOI: 10.1093/mnras/271.3.676. arXiv: astro-ph/9402069 [astro-ph].

Lahav, O. et al. (1995). "Galaxies, Human Eyes, and Artificial Neural Networks". *Science* **267**.5199, pp. 859–862. DOI: 10.1126/science.267.5199.859. arXiv: astro-ph/9412027 [astro-ph].

Lau, E. T., Kravtsov, A. V., and Nagai, D. (2009). "Residual Gas Motions in the Intracluster Medium and Bias in Hydrostatic Measurements of Mass Profiles of Clusters". *ApJ* **705**.2, pp. 1129–1138. DOI: 10.1088/0004-637X/705/2/1129. arXiv: 0903.4895 [astro-ph.CO].

Laureijs, R. et al. (2011). "Euclid Definition Study Report". *arXiv e-prints*, arXiv:1110.3193. DOI: 10.48550/arXiv.1110.3193. arXiv: 1110.3193 [astro-ph.CO].

Lee, A. J. et al. (2024). "The Chicago-Carnegie Hubble Program: The JWST J-region Asymptotic Giant Branch (JAGB) Extragalactic Distance Scale". *arXiv e-prints*, arXiv:2408.03474. DOI: 10.48550/arXiv.2408.03474. arXiv: 2408.03474 [astro-ph.GA].

Lee, K.-G. et al. (2013). "The BOSS Lyα Forest Sample from SDSS Data Release 9". *AJ* **145**.3, p. 69. DOI: 10.1088/0004-6256/145/3/69. arXiv: 1211.5146 [astro-ph.CO].

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. (2023a). "Sampling-Based Accuracy Testing of Posterior Estimators for General Inference". *40th International Conference on Machine Learning* **202**, pp. 19256–19273. DOI: 10.48550/arXiv.2302.03026. arXiv: 2302.03026 [stat.ML].

Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., and Battaglia, P. (2023b). "Rediscovering orbital mechanics with machine learning". *Machine Learning: Science and Technology* **4**.4, p. 045002. DOI: 10.1088/2632-2153/acfa63. arXiv: 2202.02306 [astro-ph.EP].

Lesgourgues, J. (2011). *The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview*. arXiv: 1104.2932 [astro-ph.IM].

Lesgourgues, J., Mangano, G., Miele, G., and Pastor, S. (2013). *Neutrino Cosmology*.

Lewis, A., Challinor, A., and Lasenby, A. (2000). "Efficient computation of CMB anisotropies in closed FRW models". *ApJ* **538**, pp. 473–476. DOI: 10.1086/309179. arXiv: astro-ph/9911177 [astro-ph].

Li, J. et al. (2016). "Feature Selection: A Data Perspective". *arXiv e-prints*, arXiv:1601.07996. DOI: 10.48550/arXiv.1601.07996. arXiv: 1601.07996 [cs.LG].

Li, S.-S. et al. (2023a). "KiDS-1000: Cosmology with improved cosmic shear measurements". *A&A* **679**, A133. DOI: 10.1051/0004-6361/202347236.

Li, X. et al. (2023b). "Hyper Suprime-Cam Year 3 results: Cosmology from cosmic shear two-point correlation functions". *Phys. Rev. D* **108**.12, p. 123518. DOI: 10.1103/PhysRevD.108.123518. arXiv: 2304.00702 [astro-ph.CO].

Liang, Y., Melchior, P., Lu, S., Goulding, A., and Ward, C. (2023). "Autoencoding Galaxy Spectra. II. Redshift Invariance and Outlier Detection". *AJ* **166**.2, p. 75. DOI: 10.3847/1538-3881/ace100. arXiv: 2302.02496 [astro-ph.IM].

Lin, K., von wietersheim-Kramsta, M., Joachimi, B., and Feeney, S. (2023). "A simulation-based inference pipeline for cosmic shear with the Kilo-Degree Survey". *MNRAS* **524**.4, pp. 6167–6180. DOI: 10.1093/mnras/stad2262. arXiv: 2212.04521 [astro-ph.CO].

Lin, Y.-T. et al. (2012). "Baryon Content of Massive Galaxy Clusters at z = 0-0.6". *ApJ* **745**.1, p. L3. DOI: 10.1088/2041-8205/745/1/L3. arXiv: 1112.1705 [astro-ph.CO].

Linder, E. V. and Jenkins, A. (2003). "Cosmic structure growth and dark energy: Cosmic structure growth and dark energy". *MNRAS* **346**.2, pp. 573–583. DOI: 10.1046/j.1365-2966.2003.07112.x.

Linder, E. V. (2003). "Exploring the Expansion History of the Universe". *Phys. Rev. Lett.* **90**.9. DOI: 10.1103/physrevlett.90.091301.

— (2005). "Cosmic growth history and expansion history". *Phys. Rev. D* **72**.4. DOI: 10.1103/physrevd.72.043529.

— (2007). "The Mirage of w=-1". *arXiv e-prints*. DOI: 10.48550/arXiv.0708.0024. arXiv: 0708.0024 [astro-ph].

Liu, X., Ma, L., Guo, J., and Yan, D.-M. (2022). "Parallel Computation of 3D Clipped Voronoi Diagrams". *IEEE Transactions on Visualization and Computer Graphics* **28**.2, pp. 1363–1372. DOI: 10.1109/TVCG.2020.3012288.

Lloyd, S. (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory* **28**.2, pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

Locatello, F. et al. (2019). *Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations*. arXiv: 1811.12359 [cs.LG].

LSST Science Collaboration et al. (2009). "LSST Science Book, Version 2.0". *arXiv e-prints*, arXiv:0912.0201. DOI: 10.48550/arXiv.0912.0201. arXiv: 0912.0201 [astro-ph.IM].

Lucie-Smith, L., Adhikari, S., and Wechsler, R. H. (2022a). "Insights into the origin of halo mass profiles from machine learning". *MNRAS* **515**.2, pp. 2164–2177. DOI: 10.1093/mnras/stac1833. arXiv: 2205.04474 [astro-ph.CO].

Lucie-Smith, L., Peiris, H. V., and Pontzen, A. (2019). "An interpretable machine-learning framework for dark matter halo formation". *MNRAS* **490**.1, pp. 331–342. DOI: 10.1093/mnras/stz2599.

Lucie-Smith, L., Peiris, H. V., Pontzen, A., and Lochner, M. (2018). "Machine learning cosmological structure formation". *MNRAS* **479**.3, pp. 3405–3414. DOI: 10.1093/mnras/sty1719.

Lucie-Smith, L., Peiris, H. V., and Pontzen, A. (2024a). "Explaining Dark Matter Halo Density Profiles with Neural Networks". *Phys. Rev. Lett.* **132**.3, p. 031001. DOI: 10.1103/PhysRevLett.132.031001. arXiv: 2305.03077 [astro-ph.CO].

Lucie-Smith, L., Peiris, H. V., Pontzen, A., Nord, B., and Thiyagalingam, J. (2024b). "Deep learning insights into cosmological structure formation". *Phys. Rev. D* **109**.6. DOI: 10.1103/physrevd.109.063524.

Lucie-Smith, L. et al. (2022b). "Discovering the building blocks of dark matter halo density profiles with neural networks". *Phys. Rev. D* **105**.10. DOI: 10.1103/physrevd.105.103533.

Lukić, Z., Heitmann, K., Habib, S., Bashinsky, S., and Ricker, P. M. (2007). "The Halo Mass Function: High-Redshift Evolution and Universality". *ApJ* **671**.2, pp. 1160–1181. DOI: 10.1086/523083.

Lundberg, S. and Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions". *arXiv e-prints*. DOI: 10.48550/arXiv.1705.07874. arXiv: 1705.07874 [cs.AI].

Ma, C.-P. and Bertschinger, E. (1995). "Cosmological Perturbation Theory in the Synchronous and Conformal Newtonian Gauges". *ApJ* **455**, p. 7. DOI: 10.1086/176550.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). "Rectifier Nonlinearities Improve Neural Network Acoustic Models". *Proc. ICML*. Vol. 30.

Machado Poletti Valle, L. F. et al. (2021). "SHAPing the gas: understanding gas shapes in dark matter haloes with interpretable machine learning". *MNRAS* **507**.1, pp. 1468–1484. DOI: 10.1093/mnras/stab2252. arXiv: 2011.12987 [astro-ph.CO].

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Madhavacheril, M. S. et al. (2024). "The Atacama Cosmology Telescope: DR6 Gravitational Lensing Map and Cosmological Parameters". *ApJ* **962**.2, p. 113. DOI: 10.3847/1538-4357/acff5f. arXiv: 2304.05203 [astro-ph.CO].

Mandelbaum, R. (2018). "Weak Lensing for Precision Cosmology". *ARA&A* **56**, pp. 393–433. DOI: 10.1146/annurev-astro-081817-051928. arXiv: 1710.03235 [astro-ph.CO].

Mangano, G., Miele, G., Pastor, S., and Peloso, M. (2002). "A precision calculation of the effective number of cosmological neutrinos". *Phys. Lett. B* **534**.1–4, pp. 8–16. DOI: 10.1016/s0370-2693(02)01622-2.

Mangano, G. et al. (2005). "Relic neutrino decoupling including flavour oscillations". *Nucl. Phys. B* **729**.1–2, pp. 221–234. DOI: 10.1016/j.nuclphysb.2005.09.041.

Mantz, A., Allen, S. W., Ebeling, H., Rapetti, D., and Drlica-Wagner, A. (2010). "The observed growth of massive galaxy clusters - II. X-ray scaling relations". *MNRAS* **406**.3, pp. 1773–1795. DOI: 10.1111/j.1365-2966.2010.16993.x. arXiv: 0909.3099 [astro-ph.CO].

Marques, G. A. et al. (2024). "Cosmology from weak lensing peaks and minima with Subaru Hyper Suprime-Cam Survey first-year data". *MNRAS* **528**.3, pp. 4513–4527. DOI: 10.1093/mnras/stae098. arXiv: 2308.10866 [astro-ph.CO].

Marriage, T. A. et al. (2011). "The Atacama Cosmology Telescope: Sunyaev-Zel'dovich-Selected Galaxy Clusters at 148 GHz in the 2008 Survey". *ApJ* **737**.2, p. 61. DOI: 10.1088/0004-637X/737/2/61. arXiv: 1010.1065 [astro-ph.CO].

Martin, J. (2012). "Everything you always wanted to know about the cosmological constant problem (but were afraid to ask)". *Comptes Rendus. Physique* **13**.6–7, pp. 566–665. DOI: 10.1016/j.crhy.2012.04.008.

Martín Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org.

McClintock, T. et al. (2019a). "Dark Energy Survey Year 1 results: weak lensing mass calibration of redMaPPer galaxy clusters". *MNRAS* **482**.1, pp. 1352–1378. DOI: 10.1093/mnras/sty2711. arXiv: 1805.00039 [astro-ph.CO].

McClintock, T. et al. (2019b). "The Aemulus Project. II. Emulating the Halo Mass Function". *ApJ* **872**.1, p. 53. DOI: 10.3847/1538-4357/aaf568.

McClintock, T. et al. (2019c). "The Aemulus Project IV: Emulating Halo Bias". *arXiv e-prints*. DOI: 10.48550/arXiv.1907.13167. arXiv: 1907.13167 [astro-ph.CO].

Melchior, P., Liang, Y., Hahn, C., and Goulding, A. (2023). "Autoencoding Galaxy Spectra. I. Architecture". *AJ* **166**.2, p. 74. DOI: 10.3847/1538-3881/ace0ff. arXiv: 2211.07890 [astro-ph.IM].

Merloni, A. et al. (2012). "eROSITA Science Book: Mapping the Structure of the Energetic Universe". *arXiv e-prints*. DOI: 10.48550/arXiv.1209.3114. arXiv: 1209.3114 [astro-ph.HE].

Mészáros, P. (1974). "The behaviour of point masses in an expanding cosmological substratum." *A&A* **37**.2, pp. 225–228.

Michaux, M., Hahn, O., Rampf, C., and Angulo, R. E. (2020). "Accurate initial conditions for cosmological N-body simulations: minimizing truncation and discreteness errors". *MNRAS* **500**.1, pp. 663–683. DOI: 10.1093/mnras/staa3149.

Mo, H., van den Bosch, F. C., and White, S. (2010). *Galaxy Formation and Evolution*.

Moon, Y.-I., Rajagopalan, B., and Lall, U. (1995). "Estimation of mutual information using kernel density estimators". *Phys. Rev. E* **52** (3), pp. 2318–2321. DOI: 10.1103/PhysRevE.52.2318.

Moore, B. et al. (1999). "Dark Matter Substructure within Galactic Halos". *ApJ* **524**.1, pp. L19–L22. DOI: 10.1086/312287. arXiv: astro-ph/9907411 [astro-ph].

Moran, K. R. et al. (2022). "The Mira–Titan Universe – IV. High-precision power spectrum emulation". *MNRAS* **520**.3, pp. 3443–3458. DOI: 10.1093/mnras/stac3452.

More, S., Diemer, B., and Kravtsov, A. V. (2015). "The Splashback Radius as a Physical Halo Boundary and the Growth of Halo Mass". *ApJ* **810**.1, p. 36. DOI: 10.1088/0004-637X/810/1/36. arXiv: 1504.05591 [astro-ph.CO].

Mukhanov, V. F. (1985). "Gravitational instability of the universe filled with a scalar field". *JETPL* **41**, p. 493.

Mukhanov, V. F. and Chibisov, G. V. (1981). "Quantum fluctuations and a nonsingular universe". *JETPL* **33**, p. 532.

— (1982). "Vacuum energy and large-scale structure of the Universe". *JETP* **56**.2, p. 258.

Nair, V. and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines". *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.

Navarro, J. F., Frenk, C. S., and White, S. D. M. (1996). "The Structure of Cold Dark Matter Halos". *ApJ* **462**, p. 563. DOI: 10.1086/177173. arXiv: astro-ph/9508025 [astro-ph].

Navas, S. et al. (2024). "Review of Particle Physics". *Phys. Rev. D* **110** (3), p. 030001. DOI: 10.1103/PhysRevD.110.030001.

Neal, R. M. (1996). "Introduction". *Bayesian Learning for Neural Networks*. New York, NY: Springer New York, pp. 1–28. DOI: 10.1007/978-1-4612-0745-0_1.

Nielsen, M. A. (2018). *Neural Networks and Deep Learning*. misc.

Nishimichi, T. et al. (2019). "Dark Quest. I. Fast and Accurate Emulation of Halo Clustering Statistics and Its Application to Galaxy aClustering". *ApJ* **884**.1, p. 29. DOI: 10.3847/1538-4357/ab3719.

Norton, C. E., Adams, F. C., and Evrard, A. E. (2024). "Cluster cosmology redux: a compact representation for the halo mass function". *MNRAS* **531**.1, pp. 1685–1703. DOI: 10.1093/mnras/stae1222.

Ntampaka, M. et al. (2015). "A Machine Learning Approach for Dynamical Mass Measurements of Galaxy Clusters". *ApJ* **803**.2, p. 50. DOI: 10.1088/0004-637X/803/2/50. arXiv: 1410.0686 [astro-ph.CO].

Ntampaka, M. and Vikhlinin, A. (2022). "The Importance of Being Interpretable: Toward an Understandable Machine Learning Encoder for Galaxy Cluster Cosmology". *ApJ* **926**.1, p. 45. DOI: 10.3847/1538-4357/ac423e. arXiv: 2112.05768 [astro-ph.IM].

Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S. (2018). "Activation Functions: Comparison of trends in Practice and Research for Deep Learning". *arXiv e-prints*. DOI: 10.48550/arXiv.1811.03378. arXiv: 1811.03378 [cs.LG].

O'Meara, J. M. et al. (2001). "The Deuterium to Hydrogen Abundance Ratio toward a Fourth QSO: HS 0105+1619". *ApJ* **552**.2, pp. 718–730. DOI: 10.1086/320579. arXiv: astro-ph/0011179 [astro-ph].

Odaibo, S. (2019). "Tutorial: Deriving the Standard Variational Autoencoder (VAE) Loss Function". *arXiv e-prints*. DOI: 10.48550/arXiv.1907.08956. arXiv: 1907.08956 [cs.LG].

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., and Zumach, W. A. (1992). "Automated Star/Galaxy Discrimination With Neural Networks". *AJ* **103**, p. 318. DOI: 10.1086/116063.

Oh, S.-H. et al. (2015). "High-resolution Mass Models of Dwarf Galaxies from LITTLE THINGS". *AJ* **149**.6, p. 180. DOI: 10.1088/0004-6256/149/6/180. arXiv: 1502.01281 [astro-ph.GA].

Ondaro-Mallea, L., Angulo, R. E., Zennaro, M., Contreras, S., and Aricò, G. (2021). "Non-universality of the mass function: dependence on the growth rate and power spectrum shape". *MNRAS* **509**.4, pp. 6077–6090. DOI: 10.1093/mnras/stab3337.

OpenAI et al. (2023). "GPT-4 Technical Report". *arXiv e-prints*, arXiv:2303.08774. DOI: 10.48550/arXiv.2303.08774. arXiv: 2303.08774 [cs.CL].

Padmanabhan, T. (2003). "Cosmological constant: The Weight of the vacuum". *Phys. Rept.* **380**, pp. 235–320. DOI: 10.1016/S0370-1573(03)00120-0. arXiv: hep-th/0212290.

Page, L. et al. (2003). "First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Interpretation of the TT and TE Angular Power Spectrum Peaks". *ApJS* **148**.1, pp. 233–241. DOI: 10.1086/377224. arXiv: astro-ph/0302220 [astro-ph].

Paranjape, A. and Sheth, R. K. (2012). "Peaks theory and the excursion set approach". *MNRAS* **426**.4, pp. 2789–2796. DOI: 10.1111/j.1365-2966.2012.21911.x.

Pasquato, M. et al. (2023). "Interpretable machine learning for finding intermediate-mass black holes". *arXiv e-prints*. DOI: 10.48550/arXiv.2310.18560. arXiv: 2310.18560 [astro-ph.GA].

Peacock, J. A. (2013). "Slipher, Galaxies, and Cosmological Velocity Fields". *Origins of the Expanding Universe: 1912-1932*. Ed. by M. J. Way and D. Hunter. Vol. 471. Astronomical Society of the Pacific Conference Series, p. 3. DOI: 10.48550/arXiv.1301.7286. arXiv: 1301.7286 [physics.hist-ph].

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". *JMLR* **12**.85, pp. 2825–2830.

Peebles, P. J. E. (1982). "Primeval adiabatic perturbations - Effect of massive neutrinos". *ApJ* **258**, p. 415. DOI: 10.1086/160094.

Peebles, P. J. E. and Ratra, B. (2003). "The cosmological constant and dark energy". *Rev. Mod. Phys.* **75** (2), pp. 559–606. DOI: 10.1103/RevModPhys.75.559.

Penzias, A. A. and Wilson, R. W. (1965). "A Measurement of Excess Antenna Temperature at 4080 Mc/s." *ApJ* **142**, pp. 419–421. DOI: 10.1086/148307.

Percival, W. J. et al. (2007). "Measuring the Baryon Acoustic Oscillation scale using the Sloan Digital Sky Survey and 2dF Galaxy Redshift Survey". *MNRAS* **381**.3, pp. 1053–1066. DOI: 10.1111/j.1365-2966.2007.12268.x. arXiv: 0705.3323 [astro-ph].

Perlmutter, S. et al. (1999). "Measurements of Ω and Λ from 42 High-Redshift Supernovae". *ApJ* **517**.2, pp. 565–586. DOI: 10.1086/307221. arXiv: astro-ph/9812133 [astro-ph].

Pichler, G., Colombo, P., Boudiaf, M., Koliander, G., and Piantanida, P. (2022). "KNIFE: Kernelized-Neural Differential Entropy Estimation". *arXiv e-prints*. arXiv: 2202.06618 [cs.LG].

Piras, D., Joachimi, B., and Villaescusa-Navarro, F. (2023a). "Fast and realistic large-scale structure from machine-learning-augmented random field simulations". *MNRAS* **520**.1, pp. 668–683. DOI: 10.1093/mnras/stad052. arXiv: 2205.07898 [astro-ph.CO].

Piras, D. and Lombriser, L. (2024). "Representation learning approach to probe for dynamical dark energy in matter power spectra". *Phys. Rev. D* **110**.2, p. 023514. DOI: 10.1103/PhysRevD.110.023514. arXiv: 2310.10717 [astro-ph.CO].

Piras, D. and Spurio Mancini, A. (2023). "CosmoPower-JAX: high-dimensional Bayesian inference with differentiable cosmological emulators". *OJAp* **6**, p. 20. DOI: 10.21105/astro.2305.06347. arXiv: 2305.06347 [astro-ph.CO].

Piras, D. et al. (2023b). "A robust estimator of mutual information for deep learning interpretability". *Machine Learning: Science and Technology* **4**.2, p. 025006. DOI: 10.1088/2632-2153/acc444.

Planck Collaboration et al. (2016a). "Planck 2015 results. XV. Gravitational lensing". *A&A* **594**, A15. DOI: 10.1051/0004-6361/201525941. arXiv: 1502.01591 [astro-ph.CO].

Planck Collaboration et al. (2016b). "Planck 2015 results. XXIV. Cosmology from Sunyaev-Zeldovich cluster counts". *A&A* **594**, A24. DOI: 10.1051/0004-6361/201525833. arXiv: 1502.01597 [astro-ph.CO].

Planck Collaboration et al. (2020a). "Planck 2018 results. I. Overview and the cosmological legacy of Planck". *A&A* **641**, A1. DOI: 10.1051/0004-6361/201833880. arXiv: 1807.06205 [astro-ph.CO].

Planck Collaboration et al. (2020b). "Planck 2018 results. VIII. Gravitational lensing". *A&A* **641**, A8. DOI: 10.1051/0004-6361/201833886. arXiv: 1807.06210 [astro-ph.CO].

Planck Collaboration et al. (2020c). "Planck 2018 results. VI. Cosmological parameters". *A&A* **641**, A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO].

Planck Collaboration et al. (2020d). "Planck 2018 results. IX. Constraints on primordial non-Gaussianity". *A&A* **641**, A9. DOI: 10.1051/0004-6361/201935891. arXiv: 1905.05697 [astro-ph.CO].

Planck Collaboration et al. (2020e). "Planck 2018 results. X. Constraints on inflation". *A&A* **641**, A10. DOI: 10.1051/0004-6361/201833887. arXiv: 1807.06211 [astro-ph.CO].

Pontzen, A. and Peiris, H. V. (2020). *PHAS0067: Advanced Physical Cosmology [Lecture notes]*. URL: https://pontzen.co.uk/PHAS0067/.

Press, W. H. and Schechter, P. (1974). "Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation". *ApJ* **187**, pp. 425–438. DOI: 10.1086/152650.

Punturo, M. et al. (2010). "The Einstein Telescope: a third-generation gravitational wave observatory". *Classical and Quantum Gravity* **27**.19, p. 194002. DOI: 10.1088/0264-9381/27/19/194002.

Rasia, E. et al. (2012). "Lensing and x-ray mass estimates of clusters (simulations)". *New J. Phys.* **14**.5, p. 055018. DOI: 10.1088/1367-2630/14/5/055018. arXiv: 1201.1569 [astro-ph.CO].

Rasmussen, C. E. (2003). "Gaussian processes in machine learning". *Summer school on machine learning*. Springer, pp. 63–71.

Reddi, S. J., Kale, S., and Kumar, S. (2018). "On the Convergence of Adam and Beyond". *Int'l Conf. on Learning Representations*, p. 239. DOI: 10.48550/arXiv.1904.09237. arXiv: 1904.09237 [cs.LG].

Reed, D. S., Bower, R., Frenk, C. S., Jenkins, A., and Theuns, T. (2007). "The halo mass function from the dark ages through the present day". *MNRAS* **374**.1, pp. 2–15. DOI: 10.1111/j.1365-2966.2006.11204.x.

Reed, D. et al. (2003). "Evolution of the mass function of dark matter haloes". *MNRAS* **346**.2, pp. 565–572. DOI: 10.1046/j.1365-2966.2003.07113.x.

Reeves, H., Audouze, J., Fowler, W. A., and Schramm, D. N. (1973). "On the Origin of Light Elements". *ApJ* **179**, p. 909. DOI: 10.1086/151928.

Reid, B. A. et al. (2010). "Cosmological constraints from the clustering of the Sloan Digital Sky Survey DR7 luminous red galaxies". *MNRAS* **404**.1, pp. 60–85. DOI: 10.1111/j.1365-2966.2010.16276.x. arXiv: 0907.1659 [astro-ph.CO].

Reitze, D. et al. (2019). *Cosmic Explorer: The U.S. Contribution to Gravitational-Wave Astronomy beyond LIGO*. arXiv: 1907.04833 [astro-ph.IM].

Renzi, F., Di Valentino, E., and Melchiorri, A. (2018). "Cornering the P l a n c k A_{lens} tension with future CMB data". *Phys. Rev. D* **97**.12, p. 123534. DOI: 10.1103/PhysRevD.97.123534. arXiv: 1712.08758 [astro-ph.CO].

Rezende, D. and Mohamed, S. (2015). "Variational Inference with Normalizing Flows". *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 1530–1538.

Riess, A. G. et al. (1998). "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant". *AJ* **116**.3, pp. 1009–1038. DOI: 10.1086/300499. arXiv: astro-ph/9805201 [astro-ph].

Riess, A. G. et al. (2021). "Cosmic Distances Calibrated to 1% Precision with Gaia EDR3 Parallaxes and Hubble Space Telescope Photometry of 75 Milky Way Cepheids Confirm Tension with ΛCDM". *ApJ* **908**.1, p. L6. DOI: 10.3847/2041-8213/abdbaf. arXiv: 2012.08534 [astro-ph.CO].

Riess, A. G. et al. (2022). "A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km s^{-1} Mpc^{-1} Uncertainty from the Hubble Space Telescope and the SH0ES Team". *ApJ* **934**.1, p. L7. DOI: 10.3847/2041-8213/ac5c5b. arXiv: 2112.04510 [astro-ph.CO].

Riess, A. G. et al. (2024a). "JWST Observations Reject Unrecognized Crowding of Cepheid Photometry as an Explanation for the Hubble Tension at 8σ Confidence". *ApJ* **962**.1, p. L17. DOI: 10.3847/2041-8213/ad1ddd. arXiv: 2401.04773 [astro-ph.CO].

Riess, A. G. et al. (2024b). "JWST Validates HST Distance Measurements: Selection of Supernova Subsample Explains Differences in JWST Estimates of Local H0". *arXiv e-prints*. DOI: 10.48550/arXiv.2408.11770. arXiv: 2408.11770 [astro-ph.CO].

Rimoldini, L. et al. (2023). "Gaia Data Release 3. All-sky classification of 12.4 million variable sources into 25 classes". *A&A* **674**, A14. DOI: 10.1051/0004-6361/202245591. arXiv: 2211.17238 [astro-ph.GA].

Rogers, K. K. and Peiris, H. V. (2021). "General framework for cosmological dark matter bounds using N -body simulations". *Phys. Rev. D* **103**.4. DOI: 10.1103/physrevd.103.043526.

Rogers, K. K. et al. (2019). "Bayesian emulator optimisation for cosmology: application to the Lyman-alpha forest". *J. Cosmology Astropart. Phys.* **2019**.02, pp. 031–031. DOI: 10.1088/1475-7516/2019/02/031.

Rosenblatt, M. (1956). "Remarks on Some Nonparametric Estimates of a Density Function". *The Annals of Mathematical Statistics* **27**.3, pp. 832–837. DOI: 10.1214/aoms/1177728190.

Rozo, E. et al. (2010). "Cosmological Constraints from the Sloan Digital Sky Survey maxBCG Cluster Catalog". *ApJ* **708**.1, pp. 645–660. DOI: 10.1088/0004-637X/708/1/645. arXiv: 0902.3702 [astro-ph.CO].

Rubin, V. C. and Ford W. Kent, J. (1970). "Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions". *ApJ* **159**, p. 379. DOI: 10.1086/150317.

Ruder, S. (2016). "An overview of gradient descent optimization algorithms". *arXiv e-prints*, arXiv:1609.04747. DOI: 10.48550/arXiv.1609.04747. arXiv: 1609.04747 [cs.LG].

Sakharov, A. D. (1968). "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation". *Soviet Physics Doklady* **12**, p. 1040.

Salvati, L. et al. (2022). "Combining Planck and SPT Cluster Catalogs: Cosmological Analysis and Impact on the Planck Scaling Relation Calibration". *ApJ* **934**.2, p. 129. DOI: 10.3847/1538-4357/ac7ab4. arXiv: 2112.03606 [astro-ph.CO].

Sartoris, B. et al. (2016). "Next generation cosmology: constraints from the Euclid galaxy cluster survey". *MNRAS* **459**.2, pp. 1764–1780. DOI: 10.1093/mnras/stw630. eprint: https://academic.oup.com/mnras/article-pdf/459/2/1764/8038954/stw630.pdf.

Saxe, A. M. et al. (2019). "On the information bottleneck theory of deep learning". **2019**.12, p. 124020. DOI: 10.1088/1742-5468/ab3985.

Schaye, J. et al. (2023). "The FLAMINGO project: cosmological hydrodynamical simulations for large-scale structure and galaxy cluster surveys". *MNRAS* **526**.4, pp. 4978–5020. DOI: 10.1093/mnras/stad2419. arXiv: 2306.04024 [astro-ph.CO].

Schmidt, M. and Lipson, H. (2009). "Distilling Free-Form Natural Laws from Experimental Data". *Science* **324**.5923, pp. 81–85. DOI: 10.1126/science.1165893. eprint: https://www.science.org/doi/pdf/10.1126/science.1165893.

Schramm, D. N. and Steigman, G. (1981). "Relic Neutrinos and the Density of the Universe". *ApJ* **243**, p. 1. DOI: 10.1086/158559.

Sedaghat, N., Romaniello, M., Carrick, J. E., and Pineau, F.-X. (2021). "Machines learn to infer stellar parameters just by looking at a large number of spectra". *MNRAS* **501**.4, pp. 6026–6041. DOI: 10.1093/mnras/staa3540. arXiv: 2009.12872 [astro-ph.IM].

Semelin, B. et al. (2023). "Accurate modelling of the Lyman-α coupling for the 21-cm signal, observability with NenuFAR, and SKA". *A&A* **672**, A162. DOI: 10.1051/0004-6361/202244722. arXiv: 2301.13178 [astro-ph.CO].

Sengottuvelan, P. and Arulmurugan, R. (2014). "Object Classification Using Substance Based Neural Network". *Mathematical Problems in Engineering* **2014**. Ed. by P. Karthigaikumar, p. 716782. DOI: 10.1155/2014/716782.

Sevilla-Noarbe, I. et al. (2021). "Dark Energy Survey Year 3 Results: Photometric Data Set for Cosmology". *ApJS* **254**.2, p. 24. DOI: 10.3847/1538-4365/abeb66. arXiv: 2011.03407 [astro-ph.CO].

Shah, P., Lemos, P., and Lahav, O. (2021). "A buyer's guide to the Hubble constant". *A&A Rev.* **29**.1, p. 9. DOI: 10.1007/s00159-021-00137-4. arXiv: 2109.01161 [astro-ph.CO].

Shao, H. et al. (2022). "Finding Universal Relations in Subhalo Properties with Artificial Intelligence". *ApJ* **927**.1, p. 85. DOI: 10.3847/1538-4357/ac4d30. arXiv: 2109.04484 [astro-ph.CO].

Shao, H. et al. (2020). "DynamicVAE: Decoupling Reconstruction Error and Disentangled Representation Learning". DOI: 10.48550/arXiv.2009.06795. arXiv: 2009.06795 [cs.LG].

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". *MNRAS* **323**.1, pp. 1–12. DOI: 10.1046/j.1365-8711.2001.04006.x. arXiv: astro-ph/9907024 [astro-ph].

Sheth, R. K. and Tormen, G. (1999). "Large-scale bias and the peak background split". *MNRAS* **308**.1, pp. 119–126. DOI: 10.1046/j.1365-8711.1999.02692.x. eprint: https://academic.oup.com/mnras/article-pdf/308/1/119/18409158/308-1-119.pdf.

Shireman, E. M., Steinley, D., and Brusco, M. J. (2016). "Local Optima in Mixture Modeling". *Multivariate Behavioral Research* **51**.4. PMID: 27494191, pp. 466–481. DOI: 10.1080/00273171.2016.1160359. eprint: https://doi.org/10.1080/00273171.2016.1160359.

Shirish Keskar, N., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima". *arXiv e-prints*. DOI: 10.48550/arXiv.1609.04836. arXiv: 1609.04836 [cs.LG].

Silk, J. (1968). "Cosmic Black-Body Radiation and Galaxy Formation". *ApJ* **151**, p. 459. DOI: 10.1086/149449.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. DOI: 10.48550/arXiv.1312.6034. arXiv: 1312.6034 [cs.CV].

Slipher, V. M. (1915). "Spectrographic Observations of Nebulae". *Popular Astronomy* **23**, pp. 21–24.

— (1917). "Nebulae". *Proceedings of the American Philosophical Society* **56**, pp. 403–409.

Spindler, A., Geach, J. E., and Smith, M. J. (2021). "AstroVaDEr: astronomical variational deep embedder for unsupervised morphological classification of galaxies and synthetic image generation". *MNRAS* **502**.1, pp. 985–1007. DOI: 10.1093/mnras/staa3670. arXiv: 2009.08470 [astro-ph.IM].

Springel, V. (2005). "The cosmological simulation code GADGET-2". *MNRAS* **364**.4, pp. 1105–1134. DOI: 10.1111/j.1365-2966.2005.09655.x. arXiv: astro-ph/0505010 [astro-ph].

Springel, V., Frenk, C. S., and White, S. D. M. (2006). "The large-scale structure of the Universe". *Nature* **440**.7088, pp. 1137–1144. DOI: 10.1038/nature04805. arXiv: astro-ph/0604561 [astro-ph].

Springel, V., Pakmor, R., Zier, O., and Reinecke, M. (2021). "Simulating cosmic structure formation with the GADGET-4 code". *MNRAS* **506**.2, pp. 2871–2949. DOI: 10.1093/mnras/stab1855. arXiv: 2010.03567 [astro-ph.IM].

Springel, V., Yoshida, N., and White, S. D. M. (2001). "GADGET: a code for collisionless and gasdynamical cosmological simulations". *New A* **6**.2, pp. 79–117. DOI: 10.1016/S1384-1076(01)00042-2. arXiv: astro-ph/0003162 [astro-ph].

Springel, V. et al. (2005). "Simulations of the formation, evolution and clustering of galaxies and quasars". *Nature* **435**.7042, pp. 629–636. DOI: 10.1038/nature03597. arXiv: astro-ph/0504097 [astro-ph].

Spurio Mancini, A., Piras, D., Alsing, J., Joachimi, B., and Hobson, M. P. (2022). "COSMOPOWER: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys". *MNRAS* **511**.2, pp. 1771–1788. DOI: 10.1093/mnras/stac064. arXiv: 2106.03846 [astro-ph.CO].

Starobinsky, A. A. (1982). "Dynamics of phase transition in the new inflationary universe scenario and generation of perturbations". *Phys. Lett. B* **117**.3-4, pp. 175–178. DOI: 10.1016/0370-2693(82)90541-X.

Steigman, G. and Turner, M. S. (1985). "Cosmological constraints on the properties of weakly interacting massive particles". *Nucl. Phys. B* **253**, pp. 375–386. DOI: 10.1016/0550-3213(85)90537-1.

Storrie-Lombardi, M. C., Lahav, O., Sodre L., J., and Storrie-Lombardi, L. J. (1992). "Morphological Classification of Galaxies by Artificial Neural Networks". *MNRAS* **259**, 8P. DOI: 10.1093/mnras/259.1.8P.

Sunayama, T. et al. (2023). "Optical Cluster Cosmology with SDSS redMaPPer clusters and HSC-Y3 lensing measurements". *arXiv e-prints*. DOI: 10.48550/arXiv.2309.13025. arXiv: 2309.13025 [astro-ph.CO].

Sundararajan, M., Taly, A., and Yan, Q. (2017). "Axiomatic attribution for deep networks". *International conference on machine learning*. PMLR, pp. 3319–3328.

Sunyaev, R. A. and Zeldovich, Y. B. (1970). "Small-Scale Fluctuations of Relic Radiation". *Ap&SS* **7**.1, pp. 3–19. DOI: 10.1007/BF00653471.

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. (2008). "Approximating Mutual Information by Maximum Likelihood Density Ratio Estimation". *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008*. Ed. by Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y. V. d. Pee. Vol. 4. Proceedings of Machine Learning Research. Antwerp, Belgium: PMLR, pp. 5–20.

Szandała, T. (2021). "Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks". *Bio-inspired Neurocomputing*. Ed. by A. K. Bhoi, P. K. Mallick, C.-M. Liu, and V. E. Balas. Singapore: Springer Singapore, pp. 203–224. DOI: 10.1007/978-981-15-5495-7_11.

Takada, M. et al. (2014). "Extragalactic science, cosmology, and Galactic archaeology with the Subaru Prime Focus Spectrograph". *PASJ* **66**.1, R1. DOI: 10.1093/pasj/pst019. arXiv: 1206.0737 [astro-ph.CO].

Tenachi, W., Ibata, R., and Diakogiannis, F. I. (2023). "Deep Symbolic Regression for Physics Guided by Units Constraints: Toward the Automated Discovery of Physical Laws". *ApJ* **959**.2, p. 99. DOI: 10.3847/1538-4357/ad014c. arXiv: 2303.03192 [astro-ph.IM].

The Dark Energy Survey Collaboration (2005). "The Dark Energy Survey". *arXiv e-prints*. DOI: 10.48550/arXiv.astro-ph/0510346. arXiv: astro-ph/0510346 [astro-ph].

The Event Horizon Telescope Collaboration et al. (2019). "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole". *ApJ* **875**.1, p. L1. DOI: 10.3847/2041-8213/ab0ec7. arXiv: 1906.11238 [astro-ph.GA].

Tinker, J. et al. (2008). "Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality". *ApJ* **688**.2, pp. 709–728. DOI: 10.1086/591439.

Tinker, J. L. et al. (2010). "The Large-scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests". *ApJ* **724**.2, pp. 878–886. DOI: 10.1088/0004-637X/724/2/878. arXiv: 1001.3162 [astro-ph.CO].

Tröster, T., Ferguson, C., Harnois-Déraps, J., and McCarthy, I. G. (2019). "Painting with baryons: augmenting N-body simulations with gas using deep generative models". *MNRAS: Letters* **487**.1, pp. L24–L29. DOI: 10.1093/mnrasl/slz075.

Troxel, M. A. et al. (2018). "Dark Energy Survey Year 1 results: Cosmological constraints from cosmic shear". *Phys. Rev. D* **98**.4, p. 043528. DOI: 10.1103/PhysRevD.98.043528. arXiv: 1708.01538 [astro-ph.CO].

Tsujikawa, S. (2013). "Quintessence: a review". *Classical and Quantum Gravity* **30**.21, p. 214003. DOI: 10.1088/0264-9381/30/21/214003. arXiv: 1304.1961 [gr-qc].

Tulin, S. and Yu, H.-B. (2018). "Dark matter self-interactions and small scale structure". *Phys. Rep.* **730**, pp. 1–57. DOI: 10.1016/j.physrep.2017.11.004. arXiv: 1705.02358 [hep-ph].

van den Bosch, F. (2022a). *Astr 610 Theory of Galaxy Formation Lecture 8: Non-Linear Collapse & Virialization [Lecture slides]*. URL: http://www.astro.yale.edu/vdbosch/astro610_lecture8.pdf.

— (2022b). *Astr 610 Theory of Galaxy Formation Lecture 20: Numerical Simulations [Lecture slides]*. URL: http://www.astro.yale.edu/vdbosch/astro610_lecture20.pdf.

Verde, L., Schöneberg, N., and Gil-Marín, H. (2023). "A tale of many H_0". *arXiv e-prints*. DOI: 10.48550/arXiv.2311.13305. arXiv: 2311.13305 [astro-ph.CO].

Vergara, J. R. and Estévez, P. A. (2013). "A review of feature selection methods based on mutual information". *Neural Computing and Applications* **24**.1, pp. 175–186. DOI: 10.1007/s00521-013-1368-0.

Viel, M., Becker, G. D., Bolton, J. S., and Haehnelt, M. G. (2013). "Warm dark matter as a solution to the small scale crisis: New constraints from high redshift Lyman-α forest data". *Phys. Rev. D* **88**.4, p. 043502. DOI: 10.1103/PhysRevD.88.043502. arXiv: 1306.2314 [astro-ph.CO].

Vikhlinin, A. et al. (2009). "Chandra Cluster Cosmology Project III: Cosmological Parameter Constraints". *ApJ* **692**.2, pp. 1060–1074. DOI: 10.1088/0004-637X/692/2/1060. arXiv: 0812.2720 [astro-ph].

Villaescusa-Navarro, F. et al. (2020). "The Quijote Simulations". *ApJS* **250**.1, p. 2. DOI: 10.3847/1538-4365/ab9d82.

Villanueva-Domingo, P. et al. (2022). "Inferring Halo Masses with Graph Neural Networks". *ApJ* **935**.1, p. 30. DOI: 10.3847/1538-4357/ac7aa3. arXiv: 2111.08683 [astro-ph.CO].

Vogelsberger, M., Marinacci, F., Torrey, P., and Puchwein, E. (2020). "Cosmological simulations of galaxy formation". *Nature Reviews Physics* **2**.1, pp. 42–66. DOI: 10.1038/s42254-019-0127-2. arXiv: 1909.07976 [astro-ph.GA].

von Wietersheim-Kramsta, M. et al. (2024). "KiDS-SBI: Simulation-Based Inference Analysis of KiDS-1000 Cosmic Shear". *arXiv e-prints*. DOI: 10.48550/arXiv.2404.15402. arXiv: 2404.15402 [astro-ph.CO].

Walmsley, M. et al. (2020). "Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning". *MNRAS* **491**.2, pp. 1554–1574. DOI: 10.1093/mnras/stz2816. arXiv: 1905.07424 [astro-ph.GA].

Warren, M. S., Abazajian, K., Holz, D. E., and Teodoro, L. (2006). "Precision Determination of the Mass Function of Dark Matter Halos". *ApJ* **646**.2, pp. 881–885. DOI: 10.1086/504962.

Warren, M. S., Quinn, P. J., Salmon, J. K., and Zurek, W. H. (1992). "Dark Halos Formed via Dissipationless Collapse. I. Shapes and Alignment of Angular Momentum". *ApJ* **399**, p. 405. DOI: 10.1086/171937.

Watson, W. A. et al. (2013). "The halo mass function through the cosmic ages". *MNRAS* **433**.2, pp. 1230–1245. DOI: 10.1093/mnras/stt791.

Wechsler, R. H. and Tinker, J. L. (2018). "The Connection Between Galaxies and Their Dark Matter Halos". *ARA&A* **56**, pp. 435–487. DOI: 10.1146/annurev-astro-081817-051756. arXiv: 1804.03097 [astro-ph.GA].

Weinberg, D. H., Bullock, J. S., Governato, F., De Naray, R. K., and Peter, A. H. (2015). "Cold dark matter: Controversies on small scales". *PNAS* **112**.40, pp. 12249–12255. DOI: 10.1073/pnas.1308716112. arXiv: 1306.0913.

Weinberg, S. (2008). *Cosmology*.

White, M. (2002). "The Mass Function". *ApJS* **143**.2, pp. 241–255. DOI: 10.1086/342752.

White, S. D. M., Efstathiou, G., and Frenk, C. S. (1993). "The amplitude of mass fluctuations in the universe". *MNRAS* **262**.4, pp. 1023–1028. DOI: 10.1093/mnras/262.4.1023.

Whitmore, B. C. (1984). "An objective classification system for spiral galaxies. I. The two dominant dimensions." *ApJ* **278**, pp. 61–80. DOI: 10.1086/161768.

Wu, H.-Y., Zentner, A. R., and Wechsler, R. H. (2010). "The Impact of Theoretical Uncertainties in the Halo Mass Function and Halo Bias on Precision Cosmology". *ApJ* **713**.2, pp. 856–864. DOI: 10.1088/0004-637X/713/2/856. arXiv: 0910.3668 [astro-ph.CO].

Wu, J. F., Kragh Jespersen, C., and Wechsler, R. H. (2024). "How the Galaxy-Halo Connection Depends on Large-Scale Environment". *arXiv e-prints*. DOI: 10.48550/arXiv.2402.07995. arXiv: 2402.07995 [astro-ph.GA].

Wyner, A. D. (1978). "A definition of conditional mutual information for arbitrary ensembles". *Information and Control* **38**.1, pp. 51–59.

Yip, K. H. et al. (2021). "Peeking inside the Black Box: Interpreting Deep-learning Models for Exoplanet Atmospheric Retrievals". *AJ* **162**.5, p. 195. DOI: 10.3847/1538-3881/ac1744.

Zel'Dovich, Y. B. (1970). "Gravitational instability: an approximate theory for large density perturbations." *A&A* **500**, pp. 13–18.

Zhai, Z. et al. (2019). "The Aemulus Project. III. Emulation of the Galaxy Correlation Function". *ApJ* **874**.1, p. 95. DOI: 10.3847/1538-4357/ab0d7b. arXiv: 1804.05867 [astro-ph.CO].

Zhang, Q. and Zhu, S.-C. (2018). *Visual Interpretability for Deep Learning: a Survey*. arXiv: 1802.00614 [cs.CV].

Zubeldia, Í. and Challinor, A. (2019). "Cosmological constraints from Planck galaxy clusters with CMB lensing mass bias calibration". *MNRAS* **489**.1, pp. 401–419. DOI: 10.1093/mnras/stz2153. arXiv: 1904.07887 [astro-ph.CO].

Zwicky, F. (1933). "Die Rotverschiebung von extragalaktischen Nebeln". *Helvetica Physica Acta* **6**, pp. 110–127.