Vers une meilleure interprétation de la connexion galaxies-halos pour les galaxies à raies d'émission du relevé spectroscopique DESI

Antoine Rocher

université
PARIS-SACLAY

Towards a comprehensive interpretation of the galaxy-halo connection for emission-line galaxies in the DESI spectroscopic survey

Vers une meilleure interprétation de la connexion galaxies-halos pour les galaxies à raies d'émission du relevé spectroscopique DESI

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°576, Particules, hadrons, énergie et noyau : instrumentation, imagerie, cosmos et simulation (PHENIICS)
Spécialité de doctorat : Science des astroparticules & Cosmologie
Graduate School : Physique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans le **Département de Physique des Particules** (Université Paris-Saclay, CEA), sous la direction de **Vanina RUHLMANN-KLEIDER**, Directrice de Recherche, et sous la co-direction de **Etienne BURTIN**, Directeur de Recherche

Thèse soutenue à Paris-Saclay, le 29 Septembre 2023, par

Antoine ROCHER

Composition du jury

Membres du jury avec voix délibérative

Delphine HARDIN Professeure des Université, Sorbonne Université, Laboratoire de physique nucléaire et de hautes énergies	Présidente
Shaun COLE Professor, University of Durham, Institute for Computational Cosmology	Rapporteur & Examinateur
Sylvain DE LA TORRE Astronome adjoint, HDR, Université Aix-Marseille, Laboratoire d'astrophysique de Marseille	Rapporteur & Examinateur
Violeta GONZALEZ-PEREZ Atracciòn de Talento senior Fellow, Universitat Autònoma Madrid, Department of Theoretical Physics	Examinatrice
Yann RASERA Chargé de recherche, Université Paris Sciences & Lettres, Laboratoire Univers et Théories	Examinateur

NNT : 2023UPASP115

université PARIS-SACLAY

ÉCOLE DOCTORALE
Particules, hadrons, énergie et noyau : instrumentation, imagerie, cosmos et simulation (PHENIICS)

Titre: Vers une meilleure interprétation de la connexion galaxies-halos pour les galaxies à raies d'émission du relevé spectroscopique DESI

Mots clés: galaxies à raies d'émissions, cosmologie, connexion galaxie-halo, simulations cosmologique, distribution spatiale des galaxies, relevé spectroscopique de galaxies

Résumé: Le Dark Energy Spectroscopic Instrument (DESI) vise à sonder la structuration à grande échelle de l'Univers en mesurant le décalage vers le rouge de ~ 40 millions de galaxies dont 17M de galaxies à raies d'émission (ELGs). Aux petites échelles, les mesures de distribution spatiale des galaxies sont essentielles pour étudier la connexion galaxie-halo, i.e. la façon dont les galaxies peuplent les halos de matière noire. Cette thèse est consacrée à l'analyse de la connexion galaxie-halo des ELGs de DESI et vise, d'une part, à donner une image complète de la façon dont les ELGs sont connectées au champ de matière noire, et d'autre part, à générer des catalogues simulés de galaxies de haute fidélité pour tester les analyses cosmologiques et corriger les effets systématiques observationnels et théoriques. Pour contraindre la connexion galaxie-halo des ELGs, nous utilisons les données des deux premiers mois de DESI, soit ∼270k ELGs à $0.8 < z < 1.6$. La grande complétude de cet échantillon permet de mesurer la distribution spatiale des galaxies jusqu'à de très petites échelles, jamais sondées auparavant. La caractéristique la plus frappante est un fort signal aux très petites échelles. Nous analysons ces données dans le cadre de la distribution d'occupation des halos (HOD), une approche empirique reliant les galaxies et les halos de matière noire dans les simulations à N-corps. Pour ce faire, nous avons développé et testé une méthode pour ajuster les modèles HOD basée sur des processus gaussiens, puis l'avons appliquée aux données. Nous considérons différentes distributions pour les galaxies centrales et des hypothèses standard pour les satellites en termes d'assignation, positionnement et dispersion de vitesse. Pour tous les modèles considérés, nous trouvons une masse moyenne de halo pour les ELGs de l'ordre de $10^{11.9} M_\odot$ et une dispersion des vitesses des satellites environ 50% plus grande que celle des particules de matière noire. Nous étudions diverses extensions de nos modèles de base, tels que le biais d'assemblage, la conformité centrale-satellite, un profil de positionnement des satellites modifié et varions la cosmologie de référence. La conformité permet de retrouver une compréhension plus physique de la HOD. Les autres extensions n'apportent pas de changement significatif à nos résultats, excepté quand nous permettons aux ELGs satellites de se situer en dehors du rayon viriel des halos. C'est avec cette hypothèse que nous obtenons la meilleure modélisation des mesures de distribution spatiale, correspondant à $\sim 0.5\%$ d'ELGs résidant à la périphérie des halos de matière noire.

Title: Towards a comprehensive interpretation of the galaxy-halo connection for emission-line galaxies in the DESI spectroscopic survey

Keywords: emission line galaxies, cosmology, galaxy-halo connection, cosmological simulations, galaxy clustering, spectroscopic galaxy surveys

Abstract: The Dark Energy Spectroscopic Instrument DESI aims to probe the large scale structure of the Universe by measuring ~ 40 million of galaxy/quasar redshifts including 17M redshifts of emission line galaxies (ELGs). At small scales, clustering measurements are invaluable to study the so-called galaxy-halo connection, i.e. the way galaxies populate dark matter halos. This thesis is dedicated to the analysis of the galaxy-halo connection of DESI ELGs and aims, on one hand, to give a complete picture of how ELGs are connected to the dark matter field, and, on the other hand, to generate high-fidelity simulated galaxy catalogues to test cosmological analysis pipelines and mitigate observational and theoretical systematic effects. To constrain the ELG galaxy-halo connection we focus on the first 2 months of DESI, which collected ∼270k ELGs at $0.8 < z < 1.6$. The high completeness of this sample made it possible to measure galaxy clustering down to very small scales, never probed before. The most striking feature of the measurements is a strong signal at the smallest scales. We analyse these data using the halo occupation distribution (HOD) framework, an empirical approach to link galaxies and dark matter halos in N-body simulations. To this end, we develop and test a method based on Gaussian processes to fit HOD models, which we then apply to data. We consider different distributions for the central galaxies and standard assumptions for satellite assignment, positioning and velocities, which we then vary. For all models considered, we report a mean halo mass of the ELG sample around $10^{11.9} M_\odot$ and satellite velocity dispersions about 50% higher than that of dark matter particles. We study various extensions of our baseline HOD models such as assembly bias, central-satellite conformity, modified satellite positioning and vary the fiducial cosmology. Conformity allows us to recover a more physical understanding of the HOD. The other extensions bring no significant change to our results, except when we allow satellite ELGs to lie outside of the halo virial radius. It is with this assumption that we obtain the best modelling of the measured clustering, corresponding to $\sim 0.5\%$ of ELGs residing in the halo outskirts.

Table of contents

Remerciements

Je tiens tout d'abord à remercier mon jury pour avoir pris le temps d'examiner mon travail, et à mes 2 rapporteurs Sylvain et Shaun pour vos commentaires sur le manuscrit.

Cette thèse n'aurait pas été la même sans le soutien indéfectible de ma directrice et mon directeur de thèse. Vanina et Etienne, je tiens à vous remercier pour tout ce que vous m'avez apporté et appris. Merci de toujours m'avoir accordé du temps pour nos longues, parfois un peu trop longues discussions, que ce soit au bureau ou sur zoom (à se reconnecter 3 fois à cause de la coupure à 40min) surtout pendant la periode de confinement. Vanina, merci pour ta rigueur et ton pointillisme qui, je l'avoue m'ont des fois fait perdre un peu patience, mais m'ont permis de toujours aller plus loin dans mes réflexions et dans mon travail. J'ai été très heureux d'être le presque dernier étudiant que tu as encadré et je te souhaite plein de belle choses pour les grandes vacances qui t'attendent ! Etienne merci pour ton enthousiasme et tes petites promenades matinales dans les bureaux des jeunes pour prendre le café. Tes petits bricolages de dessins m'ont toujours impressionné. Merci pour tout ce que vous m'avez apporté, je ressors grandi de cette expérience.

Je tiens à remercier tous les membres du groupe cosmo du DPhP, pour leur accueil et les discussions très variées que nous avons pu avoir pendant nos repas partagés. Je voulais particulièrement remercier Arnaud, pour ton aide et pour avoir eu le courage de suivre nos longues discussions avec Etienne et Vanina, ainsi que Jean-Baptiste, mon parrain de thèse, pour ta bienveillance, ton intérêt pour mon travail et pour avoir pris le temps de relire des parties de ma thèse. Merci à tous les jeunes du groupe avec qui j'ai pu partager mes nombreuses pauses café. Arnaud je t'ai remercié avant car tes surlunettes de soleil m'ont obligé de te déclasser de la catégorie jeune. Merci Mathilde et Marie-Lynn pour nos pause café au soleil et Marie-Lynn pour avoir été ma fournisseuses de produit libanais. Merci a toi, Alexandre, avec qui j'ai partagé mon bureau pendant 2 ans, même si, au début je ne savais pas trop si tu étais du groupe cosmo ou du groupe CMS. Merci pour ta bonne humeur, tes arrivées matinales et tes présentations annuelles en reunion de groupe. Le mouvement justice pour Alex Balancelamoula ne sera pas oublié ! Bon courage pour ta 3eme année de thèse !

Edmond, merci pour ces moments partagés pendant ces 3 ans, nos pauses café à rallonge et ces longs trajets en bus jusqu'à Paris. T'entendre te plaindre des aléas du bus va manquer a tout le monde je pense. Bravo pour ton travail de thèse, Dr Chaussidon (ou Chaussison pour les connaisseurs). Bon courage et bonne continuation pour tes aventures outre Atlantique. Je

te souhaite le meilleur pour la suite et j'espère aussi que nos chemin se recroiseront. Merci à Corentin pour 2 années passées au DPhP, pour ta bonne humeur même si elle n'était pas toujours au beau fixe pendant la rédaction de ta thèse. J'ai été heureux de faire la connaissance de Jean Louis Bory, un grand auteur que nous avons découvert ensemble à Biarritz. Félicitations pour ton poste et bonne continuation dans les contrés auvergnates. Et merci Romain, pour tes apparitions surprise dans mon bureau qui comme on dit chez toi "réchauffe les coeurs".

Je tiens maintenant à remercier mes proches, pour certains présents aujourd'hui. Merci à mes parents qui m'ont toujours soutenu dans mon projet de faire de l'astrophysique depuis tout petit, même si, le temps vous a parfois paru long, notamenent pendant mes longues années de fac. Merci de m'avoir permis d'arriver jusqu'ici, c'est un beau cadeau que vous vous faites pour fêter votre 34eme anniversaire de mariage. Merci à ma soeur de m'avoir supporté pendant toutes ces années, ce n'était pas de tout repos !

Merci à mes Xaxi, Yohan, Axel, Pierre et Timothée, Adrien qui ont toujours repondu présent pour le meilleur comme pour le pire. Merci pour votre soutien et tous nos moment partagés, lors nos soirée, week-end ou pèlerinage. Merci aussi à celles qui les accompagnent, Margot, Claire, Lucie, Clémentine et Lorène pour tous nos moment partagés sans oublier Laury et Benjamin.

Je voulais particulièrement remercier Yohan mon kipain depuis toujours, ou plutôt mon katse. A nos 26 ans d'amitiés, nos soirées passer à regarder les étoiles, à geeker, à jouer au tarot ou monopoly jusqu'à pas d'heure. La vie nous aura faire vivre plein de bonnes aventures, et plein d'autres sont encore a venir. Je vous souhaite à toi et à Margot plein de bonheur pour l'heureux événement qui vous attend !

Je tiens aussi à remercier ma belle famille, Alain et Claire, pour m'avoir accueilli, nourri, chouchouté, soutenu pendant le confinement et pendant mes 3 ans de thèse. J'ai passé des moments très agréables en votre compagnie qui m'ont permis de décompresser. Merci à Thibaud et Margaux pour votre soutien sans faille, nos soirée jeunes et votre accueil pendant ma rédaction de thèse. Et merci à votre petite merveille, Juliette, qui apporte du bonheur et de la chaleur dans nos cœurs. Merci aux parisiens, Florian et Céleste, pour nos escapade parisienne et les sorties au parc avec votre petite brioche Octave, qui lui aussi nous apporte beaucoup de chaleur et de bonheur.

Et pour terminer, je tiens particulièrement à remercier celle qui partage ma vie depuis plus de 3 ans maintenant, Ambre. On s'est rencontré juste avant de commencer ma thèse, et tu ne savais pas vraiment dans quoi tu t'embarquais. Même si parfois tu aurais voulu jeter mon ordinateur par la fenêtre, tu m'as toujours soutenu, encouragé, nourri, soutenu dans mon travail (et ce même pendant ma rédaction de thèse). Une aventure se termine et d'autres commencent. J'ai hâte de voir ce que la vie nous réserve. Mais une chose est sûre c'est qu'elle sera toujours plus belle avec toi, merci pour tout ce que tu m'as apporté, je t'aime.

1

Modern Cosmology

"La grandeur de l'univers nous rappelle l'infinité de possibilités qui existent dans notre propre vie."

– ChatGPT, *2023*

Contents

Cosmology is an extraordinary science. This is the sentence that Vanina, my advisor, told me when I asked her which sentence she would like to have in my manuscript. I agree with her and I will try to explain why. Cosmology (from the Greek: *kosmos*, Universe and logos, *theory*) is a fundamental science that aims to answer simple questions whose answers are difficult.

The Universe is constantly evolving, like knowledge. It is not immutable, but evolves with time, observation and progress. For a century, with Einstein's theory of general relativity, the discovery of the Universe expansion by Edwin Hubble and Georges Lemaître our understanding of the Universe has changed. From Einstein's hypothesis of a static Universe, to the observation of an accelerating expansion, our knowledge has continued to grow until today and will continue with new observational facilities and scientific discoveries. Throughout this chapter, I wish to express a global and comprehensive view of the Universe as we know it today. It might not be the real story, but it is a story built on a century of scientific research and observations.

The Universe is homogeneous and isotropic on large scales. This is the first cosmological principle. It means that its general appearance does not depend on the position of the observer. It might be difficult to admit, as we see billions of stars in our galaxy and billions of galaxies outside. We can even see super-structures, such as galaxy clusters or giant cosmic voids that can reach few tens of Megaparsec (Mpc) and form, altogether a web of knots and filaments, which is called the *cosmic web*. The parsec (pc) or even mega-pc (Mpc) is the standard unit of distances in cosmology. The parsec is defined by the distance of astronomical objects (i.e. stars), which have an angular displacement on the sky of 1 arc second(") when the Earth moves half an orbit of the Sun (also known as parallax). A parsec is approximately 3.26 light-years, or about 31 trillion (10^{12}) kilometres. To give an idea, if 1 km was the size of an atom, 1 parsec would be the Earth-Moon distance ! But these giant objects are small compared to the size of the observable Universe ($\sim$ 14,300 Mpc for the comoving distance between the Earth and the edge of the observable Universe). At this scale, the Universe looks the same everywhere (homogeneous) in every direction (isotropic). The main observable evidence of this principle is the cosmic microwave background, a homogeneous and isotropic radiation representing the first light of the Universe.

Today, the standard cosmological model, ΛCDM, describes the content and the dynamics of the Universe. This model is based on only six parameters, the gravity is ruled by general relativity (GR) and different contributions make up the energy content of the Universe today, as shown in Figure 1.1:

- **Baryonic matter**: it represents the *ordinary matter*, the one we can see, i.e. planets, stars, galaxies... and only represents $\sim$ 5% of the energy content of the Universe. The other part $\sim$ 95% is the dark side of the Universe, the one we can only guess by its effect on baryonic matter.

- **Cold dark matter (CDM)**: it is the major component of the mass in the Universe – $\sim$ **85**% of the mass – and $\sim$ 25% of the energy content of the Universe. Detected only by its impact through gravitational effects, its nature is still unknown. It could be particles beyond the standard model of particle physics or astrophysical objects that have to be formed before the primordial nucleosynthesis (e.g. primordial black holes).

- **Dark energy**: it is the main component of the energy content of the Universe $\sim$ 70%. It is a form of energy that is responsible for the late acceleration of the Universe expansion.

- **Radiation**: it encompasses all relativistic species (i.e. photons, neutrinos) from the hot and dense early Universe. Today its contribution is negligible.

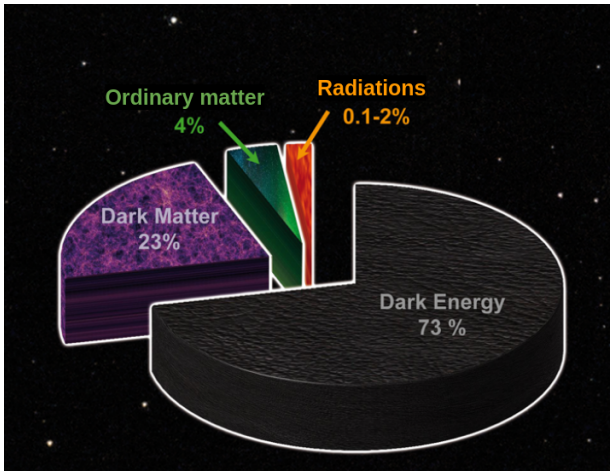

Figure 1.1: *Energy content of the Universe today. It is mainly dominated by a unknown form of energy called dark energy* $\sim 70\%$. *The other* $\sim 30\%$ *are matter components:* $\sim 25\%$ *of cold dark matter (CDM) and* $\sim 5\%$ *of baryonic (or ordinary) matter. A negligible part* ($< 10^{-4}$) *of the energy budget comes from radiation, i.e. photons and relativistic neutrinos, but at the early stage of the Universe it was the dominant part. This figure is adapted from this website: https://www.spacecentre.co.uk/news/space-now-blog/what-s-in-the-dark/*

The ΛCDM model can describe the Universe from the earliest moments, when baryons and dark matter were condensed into a very hot plasma, to the formation of galaxies and the large-scale structures we see today. It is based on 3 strong observational constraints, called the 3 cosmological pillars:

- **the Universe expansion**: a recession of galaxies at a speed proportional to their distance from us,
- **the primordial nucleosynthesis**: this explains the abundance of the chemical elements in the Universe,
- **the cosmic microwave background (CMB)**: it is the first light from the Universe, $\sim 380,000$ years after the Big Bang.

Let's get back in time and explore how through years of scientific research and discoveries, we ended up with this current understanding of our Universe.

Units and convention

Throughout the thesis, as often in cosmology, we use the natural such that $c = \hbar = k_B = 1$, where c is the speed of light, $\hbar$ the reduced Planck constant and k_B is the Boltzmann's constant, and we adopt the metric signature (+,-,-,-) for $g_{\mu\nu}$.

1.1 General relativity

The theoretical framework for gravitation and distance in cosmology is based on general relativity (GR).

1905 - Albert Einstein developed special relativity (Einstein, 1905). He introduced the notion of *space-time* and linked the mass of a particle to its rest-frame energy though $E = mc^2$. Special relativity introduced several assumptions. The first is the invariance of physics laws when changing from one Galilean frame to another. The second is the *universality of the speed of light*, i.e. the speed of light in vacuum is invariant and independent of the observer's motion. This notion is opposed to the old concept that space and time were fundamental, and that velocities were derived from them. The invariance of the speed of light in vacuum introduces a new way of thinking. Space and time become relative to the observer's frame and are no longer independent. They form a unified entity: *space-time*. This hypothesis was confirmed by Michelson & Morley in 1887[1]. According to special relativity, an object with a velocity v and a mass m has an energy given by $E = \gamma mc^2$, where γ is the Lorentz factor ($\gamma = 1$ when $v = 0$). γ goes to infinity when the velocity of the object is close to c. Which means that a massive object needs infinite energy to reach c. Therefore, special relativity imposes that the speed of light in the vacuum is an universal speed limit. Massive particles would need infinite energy to reach that limit, and only massless particles like photons travel at that speed. The notion of speed limit is incompatible with Newton's theory of gravitation. In Newtonian mechanics, two bodies are attracted by gravity according to their mass, and this force is instantaneously distributed.

1915 - Einstein extended his theory of special relativity into the theory of general relativity, taking into account the notion of speed limit for gravity. GR is based on the **Equivalence Principle**. It means that the inertial mass m_i, associated to the second law of Newton $\mathbf{F} = m_i\mathbf{a}$ is equivalent to the gravitational mass m_g associated to the gravitational interaction $\mathbf{F}_g = m_g\mathbf{g}$, i.e. $m_i = m_g$. This principle has been tested and has been confirmed down to the 10^{-15} precision level (MICROSCOPE Collaboration et al., 2022).

In the framework of GR, the space-time geometry is described as a 4-dimensional space (also called manifold $\mathcal{M}$), with a coordinate system $x^\mu = (x^0, x^1, x^2, x^3)$ where x^1, x^2, x^3 represent the 3 dimensional space coordinates, x^0 the 1-dimensional time coordinate and with a metric denoted g. The distance ds between two events in space-time separated by dx^μ is given by the *metric tensor* $g_{\mu\nu}$:

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu \tag{1.1}$$

The metric $g_{\mu\nu}$ is a 4×4 symmetric matrix. In special relativity the metric is described with the 4-dimensional Minkowski space-time metric:

[1] They tried to assess the existence of the *aether* by measuring the difference in the speed of light in perpendicular directions at two periods of 6 months apart. It was not conclusive and has the opposite effect, proving that the speed of light is invariant, constant with a velocity close to 300 000 km·s^{-1}.

$$g_{\mu\nu} = \eta_{\mu\nu} \equiv \begin{pmatrix} +1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{1.2}$$

In GR, gravity is no longer considered as an external force, but is directly included in the metric. Near massive objects, space-time will be curved and particles in free-motion (e.g. photons) will follow this curved path called *geodesics*, a generalisation of a straight line over non-plane surfaces. The geodesics of the free-moving particle (i.e. only submitted to gravitation) in a curved 4-dimensional space-time describes the minimal path between two space points and obeys the following equation known as the *geodesic equation*:

$$\frac{d^2x^\alpha}{dt^2} + \Gamma^\alpha_{\mu\nu}\frac{dx^\mu}{dt}\frac{dx^\nu}{dt} = 0 \tag{1.3}$$

with $\Gamma^\alpha_{\mu\nu}$ the *Christoffel symbol*:

$$\Gamma^\alpha_{\mu\nu} = \frac{1}{2}g^{\lambda\alpha}\left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu}\right) \tag{1.4}$$

We note the partial derivative $\partial_\mu = \partial/\partial_\mu$.

The evolution of the metric $g_{\mu\nu}$ can be derived following the principle of least action. The action used by Einstein (and first introduced by Hilbert) is the minimal action one can build from the metric and functions of the metric and is as follows:

$$S_{EH} = \frac{c^4}{16\pi G}\int \mathcal{R}\sqrt{-\det(g_{\mu\nu})}\, d^4x \tag{1.5}$$

where $\mathcal{R}$ is the Ricci scalar, $\mathcal{R} \equiv g^{\mu\nu}R_{\mu\nu}$ and $R_{\mu\nu}$ is the Ricci tensor, a function of Christoffel symbols. $\mathcal{R}$ describes a scalar curvature. Other actions verifying the equivalence principle can be used to describe a coherent theory of gravity. Among them, a simple case replaces the Ricci scalar $\mathcal{R}$ by a function of this scalar $f(\mathcal{R})$, leading to $f(\mathcal{R})$ theories.

Varying the Einstein-Hilbert action, $d\mathcal{S}_{EH} = 0$, one can derive the **Einstein equations** that govern the evolution of the metric and relate the geometry of the space-time (l.h.s) to its matter and energy content (r.h.s):

$$\boxed{G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}\mathcal{R} = 8\pi G T_{\mu\nu}} \tag{1.6}$$

$G_{\mu\nu}$ is called the *Einstein tensor*, G the gravitational constant and $T_{\mu\nu}$ is the *momentum-energy tensor*.

1917 - Einstein first applied GR under the assumption of a *static* universe that follows the cosmological principle, i.e. a homogeneous and isotropic universe (no preferred direction or orientation on the sky) on large scales $\gg 100$ Mpc.

$$\boxed{G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}\mathcal{R} = 8\pi G T_{\mu\nu} + \Lambda g_{\mu\nu}} \tag{1.7}$$

This constant acts as an opposite effect to gravitation and counterbalances the attractive effect of the gravity of matter. However, the hypothesis of a static Universe poses some problems: it

requires a positive curvature, the cosmological constant must take a very specific value for the Universe to remain static, and finally it is an unstable solution.

Note: *Einstein removed the cosmological constant Λ after the discovery of the Universe expansion, saying it was his biggest mistake. However, at the end of the 90s, the cosmological constant was reintroduced after the discovery of dark energy (see Section 1.6.3)*

In the meantime, the concept of an expanding universe emerged and was confirmed by observation. In the next section we will discuss, the concepts and evidence for the Universe expansion.

1.2 The story of the Universe expansion

1912 - A very important result from Henrietta Leavitt to build the evidence of the Universe expansion was the relation between the intrinsic brightness of variable stars, the *Cepheids*, and their pulsation period. *Cepheids* are giant bright stars with a periodic variation in luminosity. H. Leavitt studied the relationship between the period and brightness of 25 variable stars of the Small and Large Magellanic Clouds (Leavitt & Pickering, 1912). She found that the brightest Cepheids have the longest period of variation. Figure 1.2 shows the original diagram of the luminosity-period relation from Leavitt's paper in 1912. From these results, she could connect the apparent brightness of these stars to their intrinsic brightness. Knowing the intrinsic brightness allows the distance of these stars to be measured. Couple of decades later, Edwin Hubble will use Leavitt's results to measure the distance of galaxies to demonstrate the expansion of

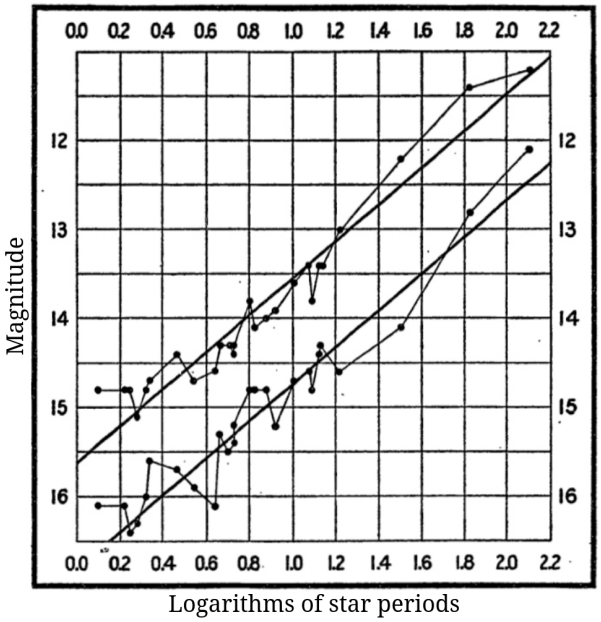

Figure 1.2: *Original plot from Leavitt's paper in 1912. It shows the period of 25 Cepheids as a function of their magnitude. Solid lines connect points corresponding to the Cepheid's minimum and maximum brightness, respectively.*

the Universe. *Cepheids* are nowadays used as *standard candles* (i.e. light sources with a known intrinsic luminosity) to derive distances accurately.

Note: *Henrietta Leavitt died of cancer in 1921. Unaware of her death, a Swedish mathematician, Gösta Mittag-Leffler, attempted to nominate her for the Nobel Prize in 1924. Unfortunately, she was not awarded the Nobel Prize posthumously. Edwin Hubble often said that she deserved the Nobel Prize. She would then have been the second woman to win the prize in physics after Marie Curie in 1920.*

1917 - Vesto Slipher (Slipher, 1917) measured the spectra of 25 nebulae – old name for galaxies, for at this time it was not excluded that nebulae were part of the Milky Way – and derived their recession velocities. He measured the spectral shift of emission lines (here H_α) due to velocities, and interpreted it as a *Doppler effect*. In practical terms, for a source moving away from the observer, the spectral lines will appear at a wavelength greater than that of the spectrum at rest. The spectral lines are *red-shifted*. Conversely, for a source moving closer to the observer, spectral lines will have a lower wavelength, i.e. will be *blue-shifted*. Slipher showed that out of 25 nebulae, only 4 are approaching us, the others are moving away. He measured redshifts with recession velocites up to 1100 $\mathrm{km{\cdot}s^{-1}}$, indicating that such objects could be outside our galaxy. This result was heavily discussed in the scientific community. It suggested that the nebulae were outside our galaxy and that the Universe was therefore much larger than scientists thought at the time. Doubts remained until the results obtained by Edwin Hubble in 1929. Slipher's result was one of the first hints for the expansion of the Universe.

1922 - Based on Einstein's GR, Alexander Friedmann published his theoretical work considering a homogeneous, isotropic Universe, with spherical or flat geometry (Friedman, 1922). Contrary to the results obtained by Einstein in 1917, he did not assume a static Universe but a dynamic one, taking into account any value of the cosmological constant Λ. Expanding or contracting universes are possible solutions. In this framework he derived a set of equations, so-called *Friedmann equations* that describe the dynamics of the Universe as a whole. His results show three possible geometries for the Universe: an open universe with a negative curvature, a closed universe with a positive curvature and a flat universe with a null curvature. He also mentioned that a non-static Universe can imply an original singularity. Friedmann's work is nowadays the theoretical basis of modern cosmology to describe the dynamics of an expanding, homogeneous and isotropic Universe. In Section 1.3 we review the mathematical framework introduced by Friedmann. Unfortunately he died in 1925, before the evidence of the Universe expansion.

1927 - Georges Lemaître measured the distance of nebulae (galaxies) and their velocities (Lemaître, 1927). He discovered that galaxies are moving away from us and that their velocity increases in proportion to their distance.

1929 - Edwin Hubble (Hubble, 1929) confirmed these results and introduced the **Hubble constant** H_0, coefficient of proportionality between the distance D and the velocity v of galaxies:

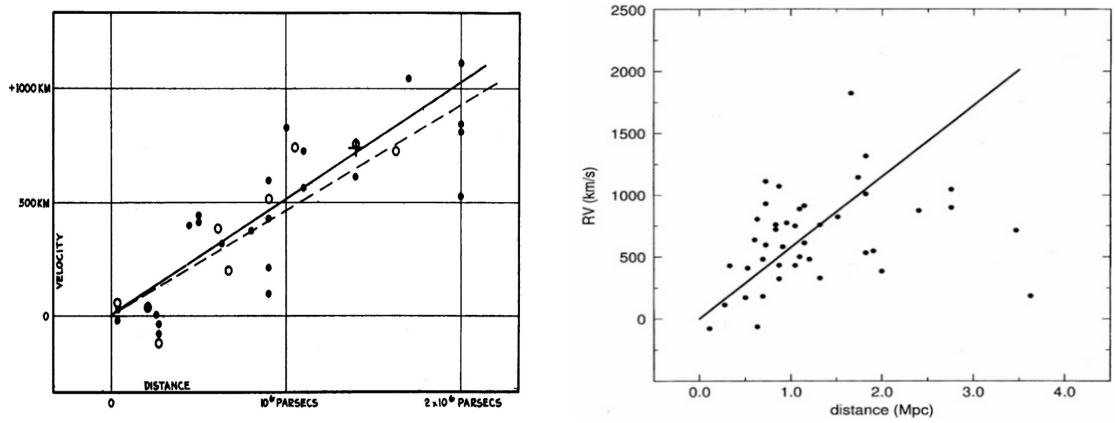

Figure 1.3: *Left: Original Hubble diagram (Hubble, 1929) showing the distance-velocity relation of nearby galaxies. Right: Same diagram made by Georges Lemaître (Lemaître, 1927).*

$$\boxed{v = H_0 D} \tag{1.8}$$

Figure 1.3 shows the original measurements from E. Hubble (*left panel*) and G. Lemaître (*right panel*). They found a value for H_0 of 530 and 570 km·s^{-1}·Mpc^{-1}, respectively. This was the first measurement/evidence for the Universe expansion, meaning that two galaxies separated by 1 Mpc today, are moving away from each other at velocity H_0 km·s^{-1}·Mpc^{-1}. The subscript 0 refers to the value of the Universe expansion at the present time. In cosmology the Hubble constant is usually expressed in the following way :

$$H_0 = h \cdot 100 \text{ km·s}^{-1}\text{·Mpc}^{-1} \tag{1.9}$$

where h is a dimensionless parameter. Using this parametrisation allows us to express the Universe expansion in unit of h without assuming a value for H_0. Today, H_0 value is measured $\sim$ 70km·s^{-1}·Mpc^{-1}, but there is strong disagreement – more than 5σ – between two independent methods to measure H_0. The first one is derived from a cosmological fit to early-Universe measurements from the *Comic Microwave Background* (CMB), and gives: $H_0 = 67.36 \pm 0.54$km·s^{-1}·Mpc^{-1} (Planck Collaboration et al., 2020). The second method is a direct distance-ladder measurement from late-Universe Cepheid and Supernova measurements, which gives: $H_0 = 73.04 \pm 1.04$km·s^{-1}·Mpc^{-1} (Riess et al., 2022).

The Hubble constant has the dimension of $[t^{-1}]$. The inversion of H_0 gives the so-called *Hubble time* t_H, which gives the characteristic time scale of an expanding Universe:

$$t_H = t_0 = H_0^{-1} = h^{-1} \cdot \frac{1}{100} \text{ s} \cdot \text{Mpc} \cdot \text{km}^{-1} = h^{-1} \cdot 9.78 \cdot 10^9 \text{ years.} \tag{1.10}$$

It is the time required for the Universe to expand to its present size, assuming that the Hubble parameter has remained unchanged since the Big Bang. Using $h = 0.7$ we find $t_H = 13.97$ billion years. In practice, the age of the Universe can be expressed in units of the Hubble time. In a universe without dark energy, the age of the observable universe today would be equal to 2/3 of the Hubble time. Within the current cosmological model of our Universe, the age of the Universe today is close to one Hubble time.

The distance travelled by light during one Hubble time t_H is called the *Hubble distance*:

$$D_H = t_H \cdot c = \frac{c}{H_0} \approx 3 \text{ Gpc}/h \tag{1.11}$$

In the next section, we mathematically describe the dynamics of a homogeneous, isotropic and expanding universe.

1.3 Friedmann-Lemaître–Robertson–Walker metric

The metric to mathematically describe a homogeneous, isotropic and expanding Universe is the **Friedmann-Lemaître–Robertson–Walker** (FLRW):

$$ds^2 = dt^2 - a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right] \tag{1.12}$$

This metric is defined by three spherical spatial coordinates $[r, \theta, \phi]$ and one temporal coordinate t, the cosmic time. The radial part of the metric can be affected by the curvature of space-time k. The Universe can be open, flat or closed (respectively $k < 0$, $k = 0$, $k > 0$). Figure 1.4 gives a representation of a 2D surface in the different cases.

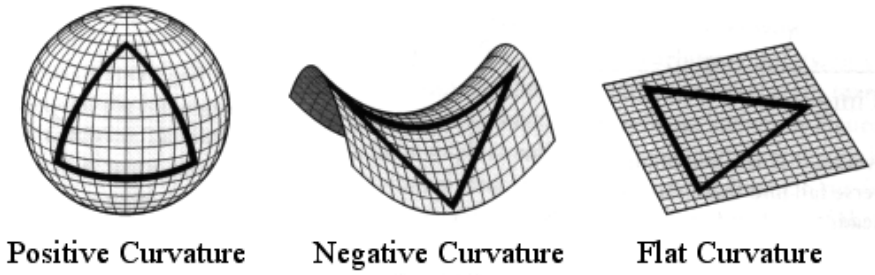

Figure 1.4: *Schematic representation of the curvature of the Universe. From left to right, in a closed Universe $k > 0$ it can be seen as a sphere, a saddle for an open Universe $k < 0$, and a plane for a flat Universe $k = 0$.*

In an expanding Universe, we consider object positions as fixed in an expanding space and we define a scaling factor $a(t)$ that describes the expansion of the space itself at a given time. In practice, objects keep the same coordinates (called *comoving* coordinates) at any time, so that their *comoving distance* will remain the same, whereas their *proper distance* (or physical distance), i.e. their distance that would be measured at a given time with a rigid ruler, will increase following the Universe expansion. Figure 1.5 shows a representation of the comoving distance. Points x_1 and x_2 at time $t_1 < t_2$ keep the same coordinates and their *comoving distance* $\mathbf{d_c}$ is the same at different scale factors, but their proper distance $\mathbf{d}$ increases as $\mathbf{d_c} = a \cdot \mathbf{d}$. The scale factor at present time t_0 is set to unity and denoted $a_0 \equiv a(t_0) \equiv 1$. We also introduce the conformal time:

$$\eta \equiv \int a^{-1} \mathrm{d}t \tag{1.13}$$

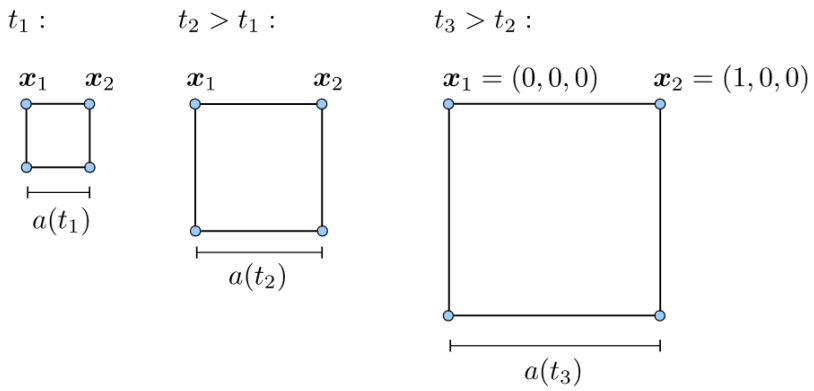

Figure 1.5: *Schematic representation of the comoving and proper distances taken from (Dodelson & Schmidt, 2020). The proper size of the square increases with time following the Universe expansion $a(t)$ but the comoving coordinates do not change.*

Assuming the FLRW metric, the energy-momentum tensor $T_{\mu\nu}$ of the background matter can be described by that of a perfect-fluid. We can decompose this tensor into different components (or species) (s), each characterized by a pressure term $P^{(s)}$ and a density term $\rho^{(s)}$:

$$T_{\mu\nu} = \sum_s T^{(s)}_{\mu\nu} = \sum_s \left(\rho^{(s)} + P^{(s)}\right) u_\mu u_\nu + P^{(s)} g_{\mu\nu} \tag{1.14}$$

with u_μ the four-velocity of the fluid in comoving coordinates. The homogeneity implies that P and ρ are only functions of the cosmic time t. Isotropy (and non-internal rotation of the fluid) implies that non-diagonal terms of the energy-momentum tensor are null.

From Einstein equation Equation (1.7), imposing local energy conservation, using Equation (1.14) and some math, one can derive the *Friedmann equations* (this derivation can be found in many textbooks, it will not be described in details here):

$$\boxed{\begin{aligned}\left(\frac{\dot{a}}{a}\right)^2 = H^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2} + \frac{\Lambda}{3} \\ \frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3P) + \frac{\Lambda}{3}\end{aligned}} \tag{1.15}$$

The dot $(\cdot)$ represents the cosmic time derivative. P and ρ are the total pressure $P = \sum_s P^{(s)}$ and total density $\rho = \sum_s \rho^{(s)}$, respectively. The first equation above relates the cosmological expansion *Hubble parameter*, $H = \dot{a}/a$, to the total energy density. The second equation describes the deceleration of the expansion of the Universe. From Equation (1.15), using the covariant conservation of the total energy-momentum tensor $\boldsymbol{\nabla}_\nu T^{\nu\mu} = 0$, we can derive the *continuity* or *conservation equation*:

$$\boxed{\dot{\rho} + 3H(\rho + P) = 0} \tag{1.16}$$

This equation describes how energy density of particles are diluted by the Hubble flow. To solve the system of equations in Equation (1.15), where we have 2 equations for 3 unknowns a, ρ, P, we need to introduce a third ingredient. We can write the equation of state of a given fluid as:

$$P = w\rho \tag{1.17}$$

For a single fluid with $w =$ constant, the evolution of ρ and P for a flat Universe ($k = 0$) can be derived by solving the Friedmann equation:

$$\rho \propto a^{-3(1+w)} \tag{1.18}$$

Solutions differ according to the type of fluid, corresponding to different values of the equation-of-state parameter, w. They are summarised in Table 1.1.

The set of equations Equation (1.15) and Equation (1.17) fully describe the dynamics of the cosmological expansion. From equations Equation (1.15), we introduce the *critical energy density* of the Universe at a given time:

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

The first equation in Equation (1.15), is equivalent to say that, at any time, the total energy density in the Universe verifies: $\sum_s \rho^{(s)}(a) = \rho_c(a)$. Today, the critical density of the Universe, $\rho_{c,0}$ is ~ 5 proton/m^3. In cosmology, we define a density parameter for every component normalised by the *critical density* as follows:

$$\Omega_s(a) \equiv \Omega^{(s)}(a) = \frac{\rho^{(s)}(a)}{\rho_c(a)} \tag{1.20}$$

At any time, we have: $\sum_s \Omega_s(a) = 1$. Parameter values today will be labelled as: $\Omega_{s,0}$. As $\rho_{c,0}$ depends on H_0^2 which is not perfectly known, cosmologists often report the density terms as a combination of the density $\Omega_{s,0}$ and reduced Hubble constant h, defined as $\omega_{s,0} \equiv \Omega_{s,0}h^2$.

From the solutions of the Friedmann equations in Table 1.1, we can get the relative contribution of each component of the Universe, $\rho^{(s)}(a) \propto a^n$ and rewrite the first Friedmann equation as:

$$\begin{aligned} H^2(a) &= H_0^2 \sum_s \Omega_{s,0} a^{-3(1+w_s)} \\ &\equiv H_0^2 E(a)^2 \end{aligned} \tag{1.21}$$

In the current Universe, the different components that contribute to the total energy density of the Universe are: dark energy (Ω_Λ), cold dark matter and baryonic matter ($\Omega_m = \Omega_{\rm cdm} + \Omega_b$), radiation (photons and relativistic neutrinos, $\Omega_{\rm rad} = \Omega_\gamma + \Omega_\nu$) and curvature ($\Omega_k$) (see Figure 1.1). We will describe in more details the energy content of the Universe in Section 1.6. In the next section we describe the cosmological distances.

Table 1.1: *Solutions of Friedmann equations considering different components in the Universe.*

Component	Equation of state	Energy density	Scale factor
Cosmological constant	$w = -1$	$\rho_\Lambda \propto$ constant	$a \propto \mathrm{e}^{Ht}$
Curvature	$w = -1/3$	$\rho_k \propto a^{-2}$	$a \propto t$
Non-relativistic matter	$w = 0$	$\rho_m \propto a^{-3}$	$a \propto t^{2/3}$
Radiations	$w = 1/3$	$\rho_r \propto a^{-4}$	$a \propto t^{1/2}$

1.4 Distances in cosmology

This thesis aims at use the spatial distribution of galaxies to constrain cosmological models. Therefore, we need to properly define distances. Based on the FLRW metric, i.e. an homogeneous and isotropic universe, we can determine distances in different ways.

➢ Redshift

To measure the distance of an object, one can measure its *redshift*. As the Universe expands, galaxies move away from us, and their spectra are *red-shifted*. This redshift z is related to the scale factor $a(t)$ by:

$$1 + z = \frac{\lambda_{obs}}{\lambda_{RF}} = \frac{a_0}{a} \tag{1.22}$$

where λ_{obs} and λ_{RF} are the observed and rest frame wavelenghts. Since galaxies move in a flow, driven by the expansion of the Universe, called the *Hubble flow*, by measuring the redshift, we can accurately determine the distance of an object in the Hubble flow. This is commonly used to get the galaxy distances. Using the above relation, we can rewrite $E(a)$ in Equation (1.21) as a function of the Universe components and the redshift (for ΛCDM cosmology):

$$E(z) = \sqrt{\Omega_{\rm rad}(1+z)^4 + \Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda} \tag{1.23}$$

However, other contributions affect the redshift measurement. In space-time, objects have also a peculiar velocity, v_{pec}, that has a small but non-negligible contribution to the redshift by the Doppler effect :

$$1 + z_{pec} = \sqrt{\frac{1 + v_{pec}/c}{1 - v_{pec}/c}} \approx 1 + \frac{v_{pec}}{c} \tag{1.24}$$

Peculiar velocities are of the order of a few hundred km·s^{-1}, corresponding to $z_{pec} \approx 0.001$. So, when the measured redshift $z \ll 1$, the contribution from peculiar velocities is dominant, otherwise what prevails is the contribution of the Hubble flow, called *cosmological redshift*. The effect from peculiar velocities may be small, but as described later in the manuscript, it is very useful for measuring and constraining cosmological models through redshift space distortions (RSD).

Another contribution to the redshift is the *gravitational redshift* also called *Einstein redshift*. It is caused by the difference in magnitude of the gravitational potential between the observer and the photon source. A photon going through a strong gravitational field looses energy when leaving this gravitational well, which translates in an additional redshift, z_g. This effect arises only in very strong gravitational fields, e.g. near black holes, neutron stars, white dwarf stars... and is therefore negligible in other cases. We ignore this effect throughout the manuscript.

At large distances, the redshift is dominated by the effect of the scale factor and can therefore also be used as a time indicator. The further away an object is, the higher its redshift and the more it is observed in a young Universe because of the finite speed of light. Whereas an object observed in the local Universe close to us, is observed almost as it is today.

Note: *As light travels at a finite speed, the distance light could have travelled since the beginning of time ($t = 0$) is also finite. We define the comoving cosmological horizon, $\chi_H \equiv \int_0^t \mathrm{dt}'/a(t')$. This is the maximum distance at which information is accessible (assuming no interactions between photons).*

Comoving distance

We have already mentioned the difference between the proper distance and the comoving distance, which takes into account the expansion of the Universe. We introduce the *radial comoving distance*, the distance travelled by light emitted by a distant object, following the geodesic (the fastest path) until the observer. The radial comoving distance D_C is computed by integrating along the geodesic from today at $z = 0$ to the object position at the time light was emitted z_e:

$$D_C = \int_{z=0}^{z_e} \frac{dz}{H(z)} \tag{1.25}$$

At the same redshift, the comoving distance between two objects separated in the sky by an angle $d\theta$ is defined to be $D_M d\theta$ with the *comoving transverse distance* D_M given by:

$$D_M = \begin{cases} D_\mathrm{H} \frac{1}{\sqrt{\Omega_k}} \sinh\left[\sqrt{\Omega_k} D_\mathrm{C}/D_\mathrm{H}\right], & \text{for } \Omega_k > 0, \\ D_\mathrm{C}, & \text{for } \Omega_k = 0, \\ D_\mathrm{H} \frac{1}{\sqrt{|\Omega_k|}} \sin\left[\sqrt{|\Omega_k|} D_\mathrm{C}/D_\mathrm{H}\right], & \text{for } \Omega_k < 0. \end{cases} \tag{1.26}$$

where $D_H = c/H_0$ is the Hubble distance defined in Equation (1.11) and Ω_k the energy density of the Universe curvature. This distance can be known only if we have access to the emitter spectrum which is not always the case.

Angular diameter distance

We can also use the *angular diameter distance*, D_A, defined as the ratio of an object proper transverse size to its angular size, which is related to D_M via:

$$D_A = \frac{D_M}{1+z} \tag{1.27}$$

It is used to obtain the proper separation between two sources from the angular separation measured from imaging. The interesting property of the angular distance is that it does not increase indefinitely when $z \to \infty$, but reaches its maximum at $z \sim 1$ and then decreases, which means that objects at $z > 1$ appear larger in angular size (see Figure 1.6).

Luminosity distance

Another distance used in cosmology, especially to measure distances of standard candles (Cepheids, supernovae...), is the *luminosity distance*. It is defined by the relationship between the bolometric (i.e. integrated over all frequencies) flux, ϕ and the bolometric luminosity, L:

$$D_L = \sqrt{\frac{L}{4\pi\phi}} \tag{1.28}$$

The luminosity distance is related to the transverse comoving distance (and angular diameter distance) by:

$$D_L = (1+z)D_M = (1+z)^2 D_A \tag{1.29}$$

Specifying $\Omega_{s,0}$ and H_0 we can compute the distance using Equation (1.21) and Equation (1.23). Figure 1.6 shows the evolution of the cosmological distances previously defined as a function of the redshift for a flat Universe with a cosmological constant.

Lookback time

Finally, we introduce the *lookback time* t_L, the difference between t_0, the age of the Universe today and the age of the Universe at the time photons were emitted from the source, z_e:

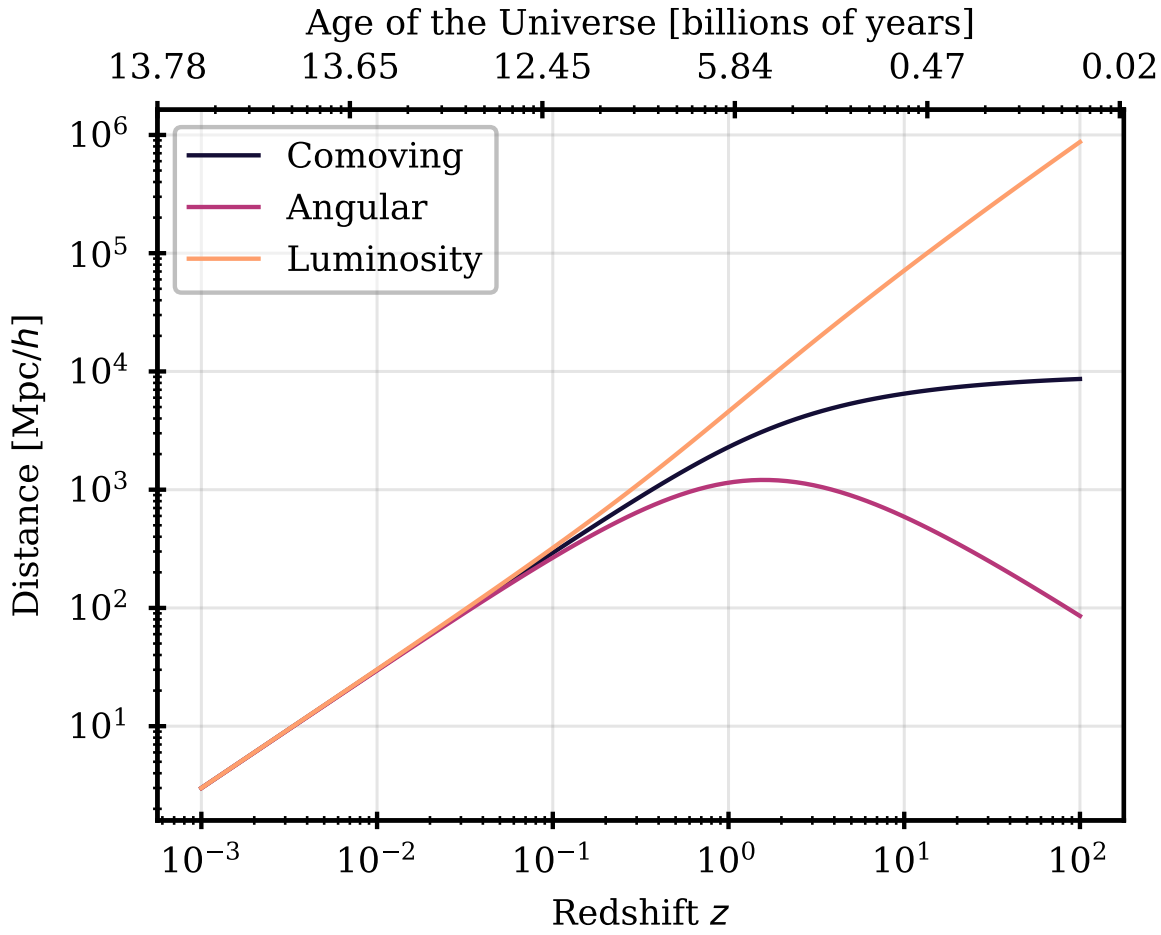

Figure 1.6: *Evolution of the cosmological distances as a function of redshift. The comoving transverse distance, angular and luminosity distances are represented by the solid dark blue, purple and orange lines, respectively.*

$$t_L = \frac{1}{H_0}\int_0^{z_e} \frac{dz}{(1+z)E(z)} \tag{1.30}$$

Comoving volume

Using the distances previously described, we can derive the *comoving volume* at a given redshift z for a given solid angle $d\Omega$ on the sky:

$$dV_C = D_H \frac{(1+z)^2 D_A^2}{E(z)} d\Omega dz \tag{1.31}$$

Integrating the above equation from $z = 0$ to a redshift z gives the all sky comoving volume. Considering the case of an open ($\Omega_k < 0$), flat ($\Omega_k = 0$) and closed ($\Omega_k > 0$) Universe, we have:

$$V_{\rm C} = \begin{cases} \left(\frac{4\pi D_{\rm H}^3}{2\Omega_k}\right)\left[\frac{D_{\rm M}}{D_{\rm H}}\sqrt{1+\Omega_k\frac{D_{\rm M}^2}{D_{\rm H}^2}} - \frac{1}{\sqrt{|\Omega_k|}}\,\text{arcsinh}\left(\sqrt{|\Omega_k|}\frac{D_{\rm M}}{D_{\rm H}}\right)\right] & \text{for } \Omega_k > 0 \\ \frac{4\pi}{3}D_{\rm M}^3 & \text{for } \Omega_k = 0 \\ \left(\frac{4\pi D_{\rm H}^3}{2|\Omega_k|}\right)\left[\frac{D_{\rm M}}{D_{\rm H}}\sqrt{1+\Omega_k\frac{D_{\rm M}^2}{D_{\rm H}}} - \frac{1}{\sqrt{|\Omega_k|}}\,\text{arcsin}\left(\sqrt{|\Omega_k|}\frac{D_{\rm M}}{D_{\rm H}}\right)\right] & \text{for } \Omega_k < 0 \end{cases} \tag{1.32}$$

We can also determine the *Hubble volume*, which is the volume of a sphere of radius D_H:

$$V_H = \frac{4}{3}\pi D_H^3 \approx 113\ [\text{Gpc}/h]^3 \tag{1.33}$$

1.5 The early Universe

1.5.1 The story of elements: Big Bang nucleosynthesis

1948 - Ralph Alpher, Hans Bethe and George Gamow published a paper entitled *The Origin of Chemical Elements* (Alpher et al., 1948a), commonly called the $\alpha\beta\gamma$ paper. This paper is the first that described the formation of elements in the Universe, and states that a process, the *Big Bang nucleosynthesis or primordial nucleosynthesis*, should create light elements, i.e. hydrogen, helium, lithium and beryllium in the correct proportions to explain their abundance in the early Universe.

The Big Bang Nucleosynthesis (BBN) is one of the three pillars of the cosmological model. It is the process that produces the light elements we can see in the Universe today. BBN happens in the early Universe starting when the temperature of the Universe cools down to $T < 1$ MeV and lasts for a few minutes. Before that time, the Universe is too hot and too dense to allow the formation of bound nuclei. Protons, neutrons and neutrinos are in thermal equilibrium through:

$$\begin{aligned} n + \nu_e &\leftrightarrow p + e^- \\ p + \bar{\nu}_e &\leftrightarrow e^+ + n \\ n &\leftrightarrow p + e^- + \bar{\nu}_e \end{aligned} \tag{1.34}$$

This equilibrium lasts until *freeze-out*, which happens when the reaction time becomes longer than the Hubble time $t_H = H^{-1}$, the characteristic time for temperature and density changes

due to the expansion. At $T \simeq 0.8$ MeV the equilibrium is lost, neutrinos stop interacting with the rest of the matter and start to freely propagate through the Universe. At this time, the proton over neutron ratio has a value around 1/6 (Particle Data Group et al., 2022).

After freeze-out, the neutrons are free to decay into protons and the neutron fraction decreases at a rate governed by the neutron lifetime, $\tau_n = 879.4 \pm 0.6$ s (Particle Data Group et al., 2022). Simultaneously, the first phase of deuterium formation starts but is counter-balanced by photo-dissociation due to the high number density of photons, $n + p \leftrightarrow D + \gamma$. This delays the production of deuterium until $T \sim 60$ keV. At that time, the neutron fraction has decreased to ~ 0.1. The primordial nucleosynthesis chain then starts: deuterium (D), tritium (T), helium-4 (^{4}He), lithium-7 (^{7}Li) are formed. Below we list the main processes that arise during this period:

$$\begin{aligned} n + p &\rightarrow D + \gamma \\ D + p &\rightarrow {}^3\mathrm{He} + \gamma \\ D + D &\rightarrow {}^3\mathrm{He} + n \\ D + D &\rightarrow T + p \\ T + D &\rightarrow {}^4\mathrm{He} + n \\ {}^3\mathrm{He} + n &\rightarrow T + p \\ {}^3\mathrm{He} + D &\rightarrow {}^4\mathrm{He} + p \\ {}^3\mathrm{He} + {}^4\mathrm{He} &\rightarrow {}^7\mathrm{Be} \\ {}^4\mathrm{He} + T &\rightarrow {}^7\mathrm{Li} \\ {}^7\mathrm{Li} + p &\rightarrow {}^4\mathrm{He} + {}^4\mathrm{He} \\ {}^7\mathrm{Be} + n &\rightarrow {}^7\mathrm{Li} + p \end{aligned} \tag{1.35}$$

The primordial nucleosynthesis lasts until $T \sim 30$ keV (so for a few minutes in total) when the Universe is no longer hot and dense enough to continue the reaction processes. The formation of light elements stops at ^{7}Li because of the absence of stable elements at A = 5,8 and because the temperature and density conditions are no longer satisfied to get heavier nuclei. The synthesis of heavier elements will start again once stars are formed and initiate the stellar nucleosynthesis, creating elements up to iron ^{56}Fe. Then, high energy events, such as supernovae or neutron star collapses will create elements heavier than ^{56}Fe. Their abundances are very low compared to those of the light elements formed during BBN. In astrophysics, elements heavier than lithium (and even helium) are commonly called metals.

The abundance of elements created during the primordial nucleosynthesis can be calculated using dedicated codes that require input from nuclear physics. Predicted abundances depend primarily on the ratio between the baryon and photon number densities, $\eta \equiv n_b/n_\gamma$. In the standard cosmological model, the present value of η was set a few seconds after Big Bang and has not changed till the present epoch. Figure 1.7 shows the evolution of these abundances as a function of temperature (or time). The result of the BBN is that the Universe is composed of hydrogen H at $\sim 75\%$, helium-4 ^{4}He at $\sim 25\%$ and very few other light nuclei up to lithium-7 ^{7}Li. Theoretical predictions of light element abundances in the early Universe agree well with the measurements (Cooke et al., 2018), except for ^{7}Li, for which observations find a lower abundance (Particle Data Group et al., 2022). It could be due to astrophysical effects related to stellar nucleosynthesis that can affect the measurements of primordial abundances, and is still an important question to solve.

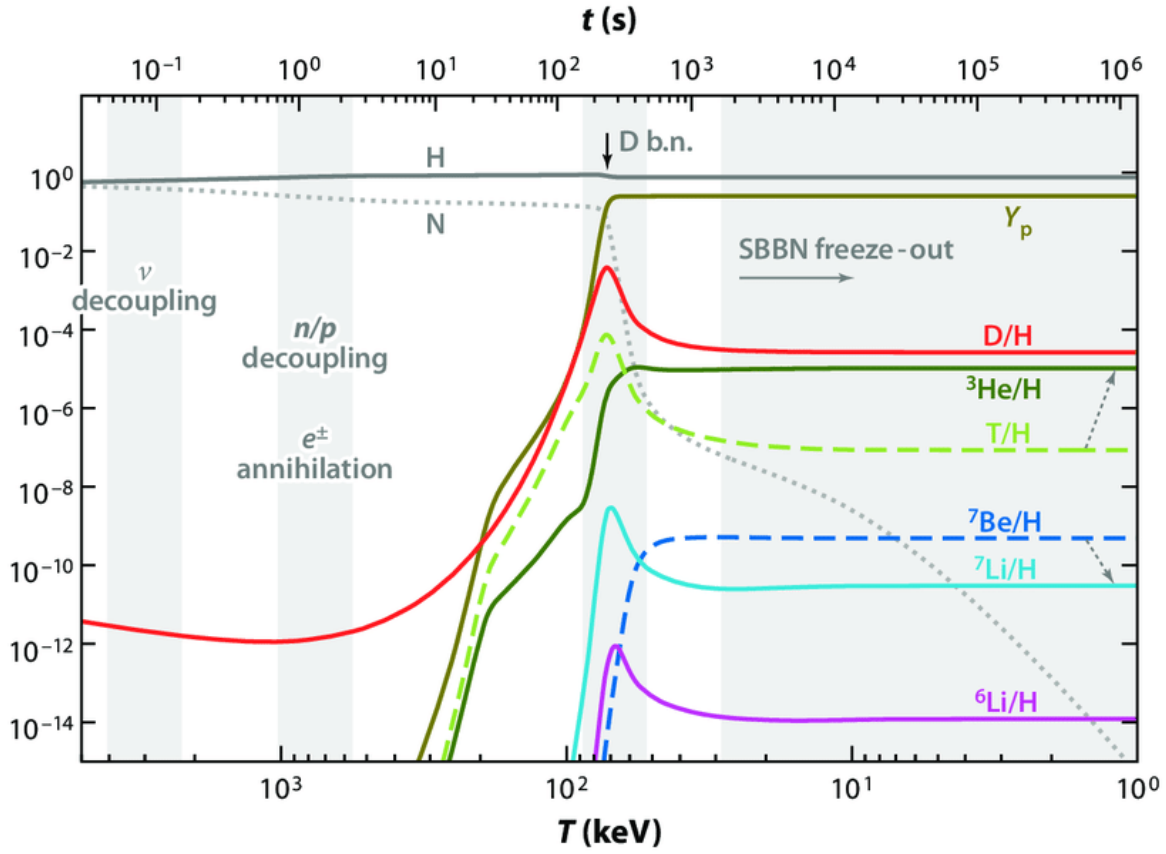

Figure 1.7: *Evolution of the mass fraction of light elements during BBN as a function of temperature (lower x-axis) and time (upper x-axis). From (Pospelov & Pradler, 2010).*

1.5.2 Birth of light: the cosmic microwave background

1948 - Continuing the reasoning behind the BBN, George Gamow published a paper entitled *The Evolution of the Universe* (Gamow, 1948). Ralph Alpher and Robert Herman published an erratum of this paper and predicted the temperature of the present Universe to be ~ 5 K (Alpher et al., 1948b). One year later, they published a paper entitled: *Remarks on The Evolution of an Expending Universe* where they introduce a model of expanding universe "in which there is a homogeneous and isotropic mixture of radiation and matter, assumed to be non-interconverting". They confirm their calculation of a present temperature of 5K for the Universe, due to a black body radiation coming from early times (Alpher & Herman, 1949).

1964 - A light coming from the very early Universe was discovered by chance by Arno Penzias and Robert Wilson who were initially looking for neutral hydrogen and were disturbed by a faint, noisy and isotropic signal while calibrating their microwave antenna in New Jersey. Meanwhile, Robert H. Dicke, Jim Peebles and David Wilkinson were preparing to search for microwave radiation from the Big Bang. Luckily, they were at Princeton University, just 60 km from the Penzias and Wilson radio telescope. The two teams interacted and published their results jointly, indicating that they had measured a residual background that could be a possible observation of the cosmic background radiation predicted by R.Alpher, R.Herman and G.Gamow (Dicke et al.,

1965, Penzias & Wilson, 1965). Penzias and Wilson were awarded the Nobel Prize in Physics for their joint detection in 1978, and J. Peebles recently (in 2019) received the Nobel Prize in Physics "for theoretical discoveries in physical cosmology".

Note: *In 1941, Andrew McKellar studied the absorption lines in the spectra of B-type stars produced by cyano radicals (CN) in the interstellar medium. He determined that they must be bathed in a* ~ 2.3 *K radiation, and associated this temperature with that of the interstellar medium but this was potentially the first observation of the CMB.*

1970s - After the discovery of the CMB, a question remained: if the Universe was perfectly isotropic in its early stage (at the time of CMB emission), how does it come from that the matter distribution in the Universe has large anisotropies, namely the large scale structures that we see today ? This questioning led theorists (J. Peebles, Y. Zel'dovich and R. Sunyaev) to predict that the CMB should have anisotropies to serve as seeds for the cosmic structure we observe today. The first anisotropy found in the CMB was the observation of a dipole, due to the displacement of the Earth w.r.t. the CMB rest frame (Conklin, 1969, Henry, 1971).

1989 - After the launch of the COBE satellite, the first anisotropies in the temperature power spectrum of the CMB were detected. This measure was a striking evidence and a confirmation of the validity of the Big Bang model. It also confirmed *the first principle of cosmology*: the Universe is homogeneous and isotropic on large scales. These temperature fluctuations have been measured very precisely by the Planck satellite (see Figure 1.9) and are of the order of $\delta T/T \sim 10^{-5}$.

Since its discovery, the CMB has been widely studied and measured with very high precision by space missions: COBE (Cosmic Background Explorer) from 1989 to 1993 (Smoot et al., 1992), WMAP (Wilkinson Microwave Anisotropies Probe) from 2001 to 2007 (Bennett et al., 2013), and the Planck satellite from 2008 to 2013 (Planck Collaboration et al., 2020). Figure 1.8 shows the measurement of the CMB spectrum by COBE, compared to a black body spectrum with $T = 2.728$ K. The data point are invisible on the figure because their error bars are smaller than the width of the line! Indeed, the CMB provides the best black body spectrum ever measured.

The cosmological interpretation of the CMB is as follows. During the period following the BBN, the Universe is dominated by radiation and the temperature of the Universe is still high enough to ionise the recently-formed atoms. Light elements (also called *baryons*) and photons are strongly coupled and form an ionised plasma called the *baryon-photon plasma*. The Universe is totally opaque, protons and electrons interact permanently by the photo-ionisation process, creating hydrogen which re-ionises instantaneously, generating photons:

$$p^+ + e^- \leftrightarrow H + \gamma \tag{1.36}$$

The Universe is in thermal equilibrium and its energy distribution follows a black body spectrum. During this period, the Universe continues to expand and cool, so that at some point, the Universe is no longer dense and hot enough to maintain the above equilibrium, which therefore breaks down: it is the start of the *recombination epoch*. When the Universe reaches a temperature $T \sim 0.26$ eV, the density of hydrogen (or protons) n_p is low enough for the mean free time of photons, τ_γ, to exceed the Hubble time $t_H = H^{-1}$. The photons then decouple from the plasma

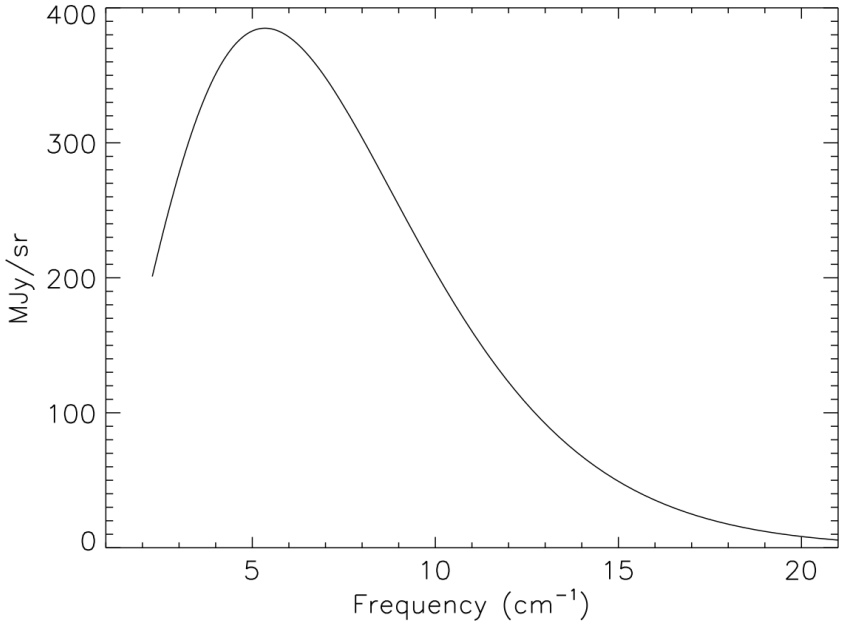

Figure 1.8: *Intensity of the cosmic microwave radiations measured by COBE as a function of frequency. The solid line shows a theoretical black body spectrum with $T = 2.728$ K (Fixsen et al., 1996). The data point are invisible, and their errors are smaller than the width of the line.*

and travel freely through the Universe, while the electrons and protons recombine to form atoms. Since then, these photons at equilibrium before recombination have been redshifted and form the cosmic microwave background (CMB). The thermal character (black body) of the CMB spectrum is conserved by the Universe expansion. This is what we observe today, i.e. a redshifted black body spectrum of photons produced at the *last-scattering surface*. This radiation offers a snapshot of the first "free" light of the early Universe, which is very valuable to study cosmology. The recombination occurs at redshift $z_{rec} \sim 1100$ when the Universe was $\sim 380,000$ years old.

Note: *Similarly to the CMB (Section 1.5.2), the Big Bang model predicts the existence of relic neutrinos, a radiation from these neutrinos that freely travel since the neutrino decoupling in the early Universe. Unfortunately, a direct detection of relic neutrinos is very difficult, and there is no observation evidence of this signal yet. However, the success of BBN predictions gives a real theoretical evidence for the existence of relic neutrinos.*

Today the CMB is a major source of information in cosmology. It is one of the most important probes that confirm the hot Big Bang theory. The anisotropies in the CMB map can be compressed into an angular power spectrum, which, through fits by cosmological models, provides the most significant constraints on cosmological parameters today. The CMB temperature power spectrum measured by (Planck Collaboration et al., 2020) is shown in Figure 1.10. Best fitting cosmological parameter values from the CMB measurement (TT, TE, EE, lowE, lensing, see note below), assuming a flat ΛCDM model with two massless and one massive neutrino ($m_\nu = 0.06$ eV), are summarized in Table 1.2.

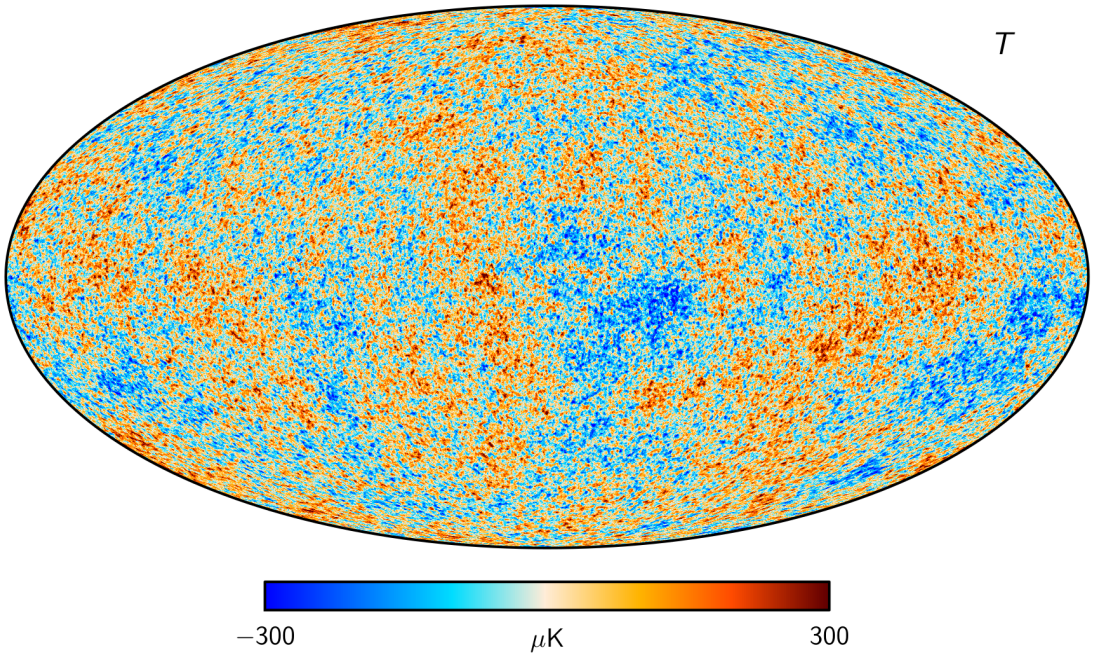

Figure 1.9: *CMB temperature fluctuation map from (Planck Collaboration et al., 2020) after foreground subtraction.*

reduced Hubble constant	h	0.6736 ± 0.0054
baryon density	$\omega_{b,0}$	0.02237 ± 0.00015
cold dark matter density	$\omega_{c,0}$	0.1200 ± 0.0012
dark energy density	$\Omega_{\Lambda,0}$	0.6847 ± 0.0073
optical depth at reionisation	τ_{rei}	0.0544 ± 0.0073
redshift of reionisation	z_{rei}	7.67 ± 0.73
index of the primordial power spectrum	n_s	0.9649 ± 0.0042
amplitude of the primordial power spectrum	$\ln(10^{10} A_s)$	3.044 ± 0.014
normalisation of the matter power spectrum	$\sigma_{8,0}$	0.8111 ± 0.0060
redshift of matter-radiation equality	z_{eq}	3402 ± 26
last scattering redshift	$z_\star$	1089.92 ± 0.25
drag redshift	z_{drag}	1059.94 ± 0.30
sound horizon at the drag epoch	r_{drag} [Mpc]	147.09 ± 0.26

Table 1.2: *Cosmological parameter 68% intervals as measured from Planck TT, TE, EE, lowE and lensing data, within the flat* ΛCDM *model (Planck Collaboration et al., 2020)*

Note: *In addition to the temperature (T) fluctuations, the polarisation of the CMB light can be measured in two projected modes: E and B. B modes are only caused by tensor modes, i.e. gravitational waves. The detection of primordial B modes would give major insights for inflation. E modes are produced by scalar and tensor modes and have been measured. Constraints from the CMB on cosmological parameters come from the combined fit to the three measured power spectra: TT, TE and EE. lowE adds low-ℓ data (large scales, $2 <$ low-$\ell < 29$) to the EE constraint. In addition, results are also derived adding the measured CMB lensing power spectrum: the large*

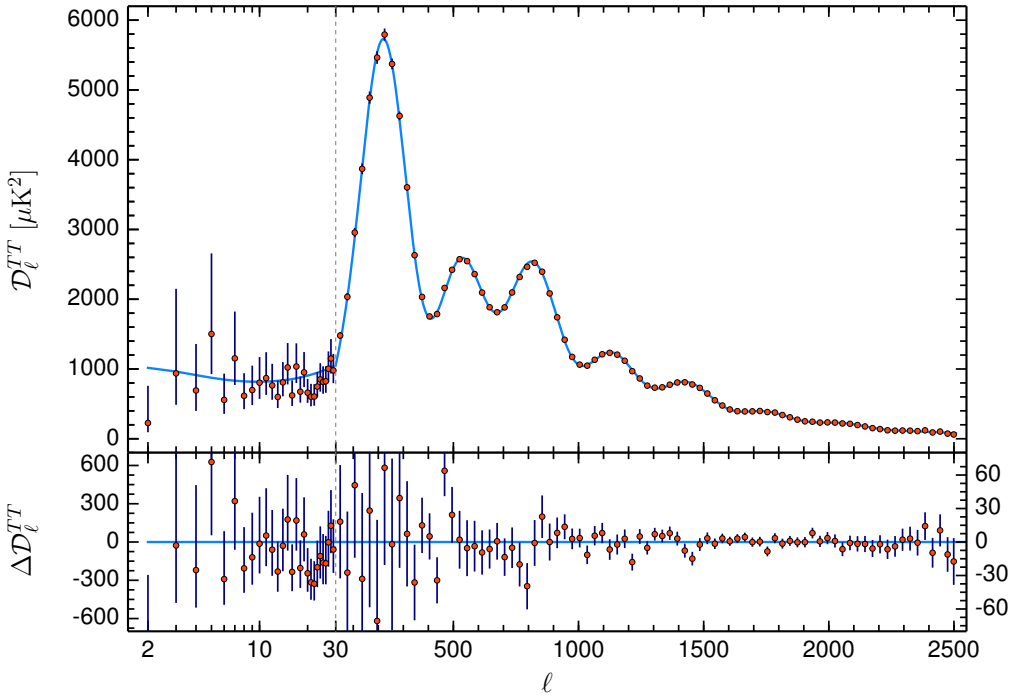

Figure 1.10: *CMB temperature anisotropy power spectrum from Planck. x-axis bins are logarithmic for* $l < 30$. *: Top: Red dots with error bars are data and the blue line shows the best-fit model from (Planck Collaboration et al., 2020). Bottom Best fit residuals.*

scale structures lens the CMB light, leaving a signal on the CMB (detected at 40σ *by the Planck collaboration).*

The Baryonic Acoustic Oscillation

Before recombination, baryons are submitted to radiative pressure forces due to their coupling with photons. As there are anisotropies in the CMB spectrum, regions with higher temperature are denser whereas under dense regions are cooler. Baryons are also submitted to gravity and fall into over-dense regions. Therefore, a competition between radiative pressure and gravity occurs and leads to the propagation of acoustic waves in the baryon-photon plasma. These acoustic waves propagate until the time of photon-baryon decoupling when acoustic waves (or oscillations) of baryons are frozen. This time is called the *drag epoch*, z_d. Acoustic waves travel at the sound speed in the baryon-photon plasma:

$$c_s(z) = \frac{c}{\sqrt{3}}\left[1 + \frac{3\rho_b(z)}{4\rho_\gamma(z)}\right] \tag{1.37}$$

At z_d, $\rho_b \ll \rho_\gamma$, so the sound speed in the baryon-photon plasma is $\sim c/\sqrt{3}$. At the time of the drag epoch t_d, the sound waves have travelled a distance $r_d(z_d) \approx c_s \cdot t_d$ called the *sound horizon*. Once the acoustic waves are frozen, they leave an imprint at this characteristic distance $r_d(z_d)$ in the baryon-photon plasma, corresponding to a small over-density in the spatial distribution

of matter. This feature is called the *Baryonic Acoustic Oscillation* or BAO. We can compute the size of the sound horizon today $r_d(z_d \sim 1060)$ using the cosmological parameter values from Planck:

$$r_d\left(z_d\right) = \int_{z_d}^{\infty} \frac{c_s(z)}{H(z)} dz = 99.08 \pm 0.33 \text{ Mpc}/h \tag{1.38}$$

The BAO scale $r_d(z_d)$ is fixed at the drag epoch z_d, and since then evolves only through the Universe expansion. It is today used as a *standard ruler* to constrain the cosmological parameters. This imprint can be measure at different epochs of the evolution of the Universe in the clustering (spatial distribution) of galaxies. The BAO scale was first measured in the spatial distribution of galaxies by e.g. the Sloan Digital Sky Survey (SDSS) collaboration (Eisenstein et al., 2005). Then, measurements of BAO scales were performed at different redshifts. The latest measurements are from the extended Baryon Oscillation Spectroscopic Survey (eBOSS) collaboration that performed a precise measurement of the BAO at different redshifts using different galaxy populations (called tracers) (Alam et al., 2021a). Figure 1.11 shows the measurement of the BAO peak in the clustering of galaxies and quasars at 6 different epochs. The peak corresponds to a statistical overdensity between pairs of galaxies separated by the characteristic scale $r_d(z_d)$.

One of the main goals of the Dark Energy Spectroscopic Instument (DESI) is to improve the precision of the BAO measurement, by covering a wider redshift range and increasing by a factor ~ 13 the number of observed galaxies and quasars.

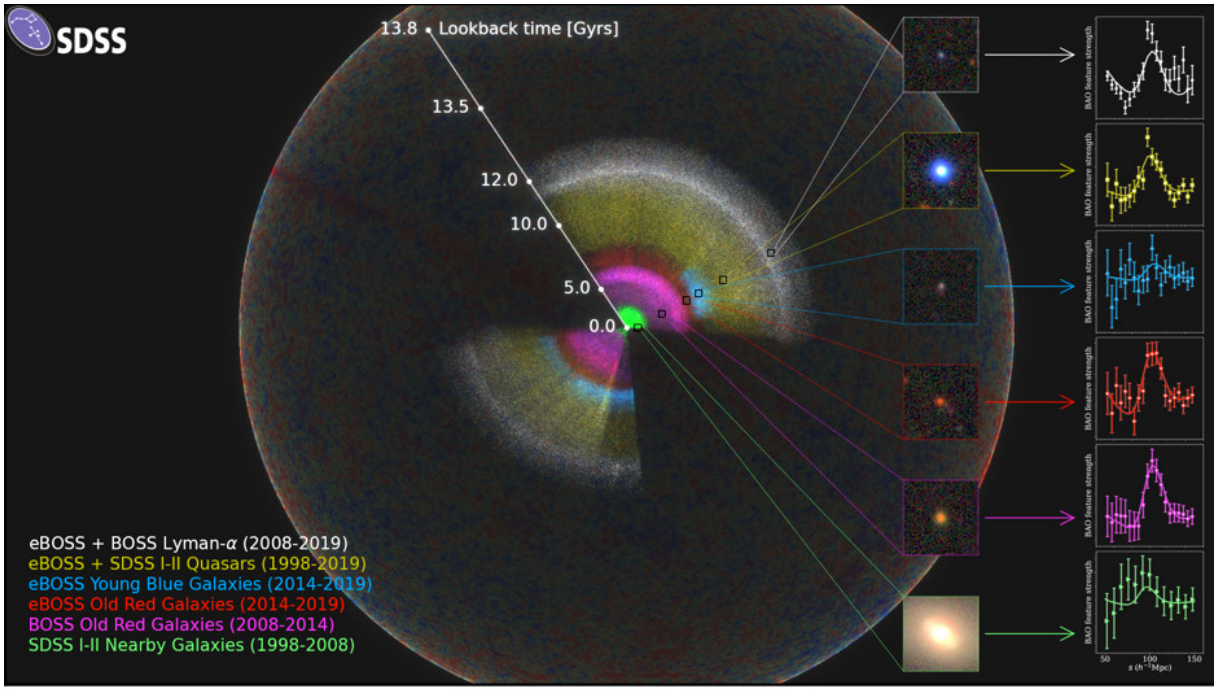

Figure 1.11: *Representation of the spatial galaxy distribution measured by the eBOSS survey (Alam et al., 2021a). For each type of galaxies, the BAO peak is represented on the right in the 2 two point correlation function.*

1.5.3 Inflation

The CMB measurement shows that the Universe looks similar in every direction. However this finding leads to several problems. The speed of light imposes a limit on the causality between two events. In a radiation dominated Universe, the physical scale for two events to be causally related from $t = 0$ to recombination corresponds to a solid angle $\sim 1\text{deg}^2$. Therefore, a first question arises: how can we explain that regions separated by scales which seem uncorrelated today, have almost the same temperature in the early Universe? This is known as the *horizon problem*. Another problem arises during the radiation era of the Universe and is known as the *flateness problem*. From Friedmann equations, in a universe dominated by radiation, the curvature increases exponentially with time (see Table 1.1). However, measurements from many probes tend towards a flat Universe, $\Omega_k = 0$ today.

1981 - Alan H. Guth proposed a mechanism that could reconcile the CMB measurements and the dynamics of the primordial Universe (Guth, 1981). By assuming a very brief but incredibly rapid expansion of the Universe that occurred in the first fraction of a second after Big Bang, it is possible to address both problems. This period is called *inflation*. During this period, the Universe, which was initially small ($\sim$ Planck length), exponentially expands and its volume becomes order of magnitudes larger than the observable Universe today in less than a second. This can explain the flatness, homogeneity and isotropy of the Universe we see today. It is also thought that the inflationary period seeded the Universe with tiny quantum fluctuations that evolved into the large-scale structures we observe today.

As of today, many models describing the inflation phase exist. The simplest one introduces a single scalar inflation field in the slow-roll regime. Slow-roll regime means that the inflation field ϕ evolves very slowly compared to its potential $V(\phi)$, leading to an inflationary phase during which the scale factor evolves exponentially $a \propto e^{Ht}$. This period is very difficult to probe experimentally. However, inflationary models predict the existence of primordial gravitational waves. These waves would affect the temperature anisotropy and polarization of the CMB but have not yet been detected. The discovery of primordial gravitational waves would confirm the existence of an inflationary period and allow the expansion rate of inflation $H_{\text{inflation}}$ to be determined. The future space telescope LiteBird (scheduled for launch $\sim$ 2030) and the ground-based experiment CMB-S4 aim to measure the signature produced by primordial gravitational waves on the CMB (Abazajian et al., 2016, Collaboration LiteB I R D et al., 2023).

1.6 Energy content of the Universe

In this section we describe the composition of our Universe as we know it today from Planck Collaboration et al. (2020) results. We make an inventory of the different species and their evolution over time. We refer to Figure 1.1 for a pie chart of the Universe components today. The evolution of the energy density of each species Ω_s is shown in Figure 1.12. The early Universe is dominated by radiation, then it transitions to a matter dominated era. We define z_{eq}, the time of radiation-matter equality ($\Omega_{\text{rad}} = \Omega_m$). After the matter dominated era, the Universe enters a phase of accelerated expansion where dark energy Ω_Λ dominates. In this section, we assume as fiducial cosmology the final Planck ΛCDM results described in Table 1.2.

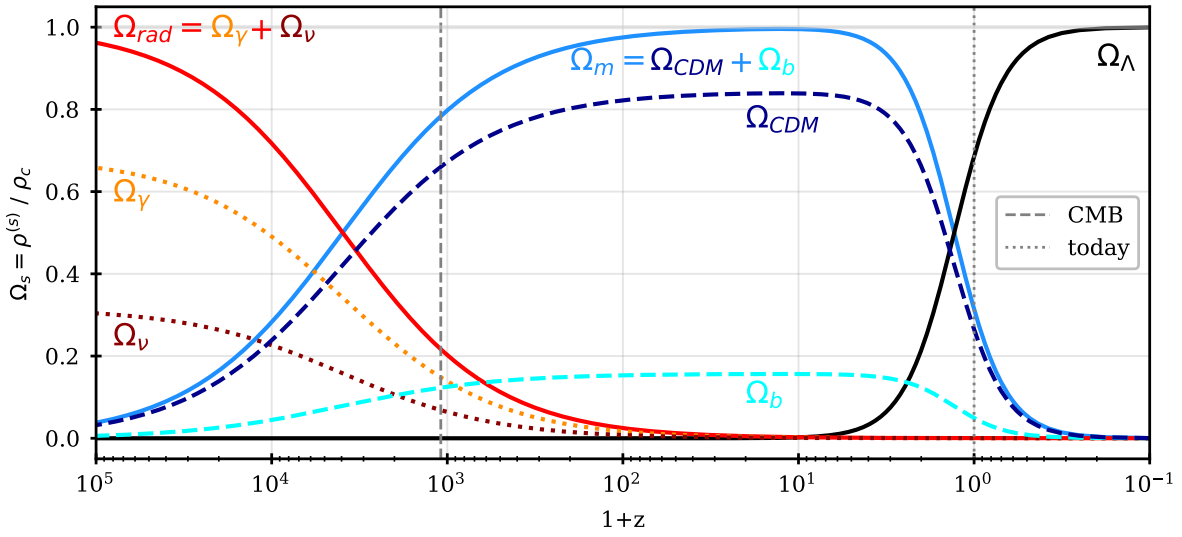

Figure 1.12: *The density parameters Ω_s as a function of redshift. The solid red line represents the contribution from radiations $\Omega_{\rm rad} = \Omega_\gamma + \Omega_\nu$ where γ is the contribution from photons (dashed yellow line) and ν from massless neutrinos (dashed brown line). The solid blue line represents the contribution from matter $\Omega_M = \Omega_{\rm cdm} + \Omega_b$ where cdm is the contribution from dark matter (dashed darkblue line) and b from baryons (dashed cyan line). The solid blue line represents the contribution from dark energy Ω_Λ. We consider a flat Universe (no contribution from curvature) with cosmological parameters from Planck Collaboration et al. (2020). The vertical grey dotted line indicates the present time, $z = 0$, and the vertical grey dashed line is z_{rec} the time of recombination, when photons from the CMB were emitted.*

1.6.1 Radiation

The term radiation includes all relativistic species present at a given redshift z. At present, radiation comes mostly from CMB photons, relativistic neutrinos and any other possible thermal relic. All other sources of photons, i.e. from stars or galaxies, are negligible.

Photons

Given the black body nature of the CMB spectrum, the energy density of photons follows a Bose-Einstein distribution without chemical potential:

$$\rho_\gamma = g_s \int \frac{p}{e^{p/T} - 1} \frac{\mathrm{d}^3 p}{(2\pi)^3} \tag{1.39}$$

where g_s is the number of degrees of freedom (here $g_s = 2$ for the two spin states of photons), p_γ is the photon momentum and $T = 2.7255 \pm 0.0006$ is the temperature of the CMB (Fixsen, 2009). Solving the integral in Equation (1.39) (see Dodelson & Schmidt (2020) for computation details) we get:

$$\rho_\gamma = \frac{\pi^2}{15} T^4 \tag{1.40}$$

Therefore, we can calculate the photon density parameter Ω_γ today:

$$\Omega_{\gamma,0} = \frac{\rho_\gamma}{\rho_c} \sim 5.45 \cdot 10^{-5} \tag{1.41}$$

which is negligible.

Neutrinos

Neutrinos, similarly to photons, decouple from electrons and the Hot Big Bang model predicts the existence of a cosmic relic background of neutrinos. Unlike the CMB, it has not been directly observed. However, strong theoretical arguments based on very well-understood physics suggest that this radiation exists and predict its contribution to the energy density of the Universe.

From what we know about neutrinos from particle physics, there are three *flavors* of neutrinos in the standard model of particle physics: ν_e, ν_μ and ν_τ, associated to electrons, muons and taus, respectively. Neutrinos are fermions, i.e. follow the Fermi-Dirac distribution at equilibrium. The precise mass of each neutrino is still unknown, but constraints exist on the value of the sum of the neutrino masses, $\sum_\nu m_\nu$. As neutrino decoupling arises before photon decoupling, neutrinos do not get reheated by the electron-positron annihilation reaction from the BBN. However, photons get reheated because of entropy conservation. Thus, the photon temperature T_γ increases proportionally to the temperature of neutrinos T_ν as (see Dodelson & Schmidt (2020) for derivation details):

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma \tag{1.42}$$

Considering relativistic neutrinos (i.e. massless neutrinos $m_\nu = 0$), their energy density is given by:

$$\rho_\nu = \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} \rho_\gamma \cdot N_\nu \tag{1.43}$$

with N_ν the number of neutrino types (3 in the standard model). Actually, neutrino decoupling is not instantaneous and a small fraction of neutrinos get reheated in the primordial plasma. To take this into account, we introduce an effective number of neutrinos N_{eff} that replaces the number of massless neutrinos. This number is predicted to be $N_{\text{eff}} = 3.044$ from (Particle Data Group et al., 2022) if we consider only 3 neutrino types as thermal relics. Planck Collaboration et al. (2020) measure the effective number of thermal relic to be $N_{\text{eff}} = 2.99 \pm 0.17$, consistent with $N_{\text{eff}} = 3.044$. In this manuscript we consider that thermal relics are only composed of 3 massless neutrinos (and the transition from relativistic to non-relativistic neutrinos will be ignored). Considering Equation (1.43) with $N_{\text{eff}} = 3.044$ the energy density of massless neutrinos is:

$$\Omega_\nu = 2.514 \cdot 10^{-5} \tag{1.44}$$

Therefore, the total contribution from radiation today is:

$$\Omega_{\text{rad}} = \Omega_\gamma + \Omega_\nu = 7.964 \cdot 10^{-5} \tag{1.45}$$

Transition from radiation to matter domination

The energy contribution from radiation dominates in the early Universe until the latter moves into the matter dominated era. At this transition, the energy density contributions from radiation and matter are equal, $\rho_{\text{rad}}(z_{\text{eq}}) = \rho_m(z_{\text{eq}})$, and the redshift is z_{eq}. Using Equation (1.43) and the time evolution of radiation ($\propto a^{-4}$) and matter ($\propto a^{-3}$) energy densities from the solutions of Friedmann equations (Table 1.1), we can determine z_{eq}:

$$1 + z_{\text{eq}} = \frac{\Omega_m}{1.68 \cdot \Omega_\gamma} \approx 3405 \tag{1.46}$$

which corresponds to the cosmic time $t_{\rm eq} \approx 50,897$ years.

1.6.2 Non-relativistic matter

After the radiation domination epoch ($\sim$ after CMB emission), the non-relativistic matter dominates the total energy density of the Universe until redshift ~ 2. The two known components of the non-relativistic matter are baryons, i.e the *ordinary matter* and *dark matter*, a form of matter only detectable by gravitational interaction.

Baryons

Baryonic matter constitutes the ordinary matter: gas, dust, planets, stars... The relative abundance of baryons in the Universe is defined by the Big Bang Nucleosynthesis process described in Section 1.5.1. To estimate the energy density of baryons we can measure the abundances of the different light nuclei produced by the BBN. Several ways can be used to infer the baryon density, the most precise being the measure of the imprint of baryonic acoustic oscillations in the CMB anisotropies:

$$\Omega_b = 0.04936 \pm 0.00033 \tag{1.47}$$

as measured by Planck Planck Collaboration et al. (2020). Other methods to determine the energy contribution of baryons agree with this measurement (Particle Data Group et al., 2022). In particular, this result is in remarkable agreement with the value deduced from a comparison between light element abundance measurements and BBN predictions.

Cold dark matter

1933 - Fritz Zwicky measured the radial velocity dispersion of galaxies in the Coma galaxy cluster and inferred the total mass of this cluster applying simple Newtonian dynamics relating the Keplerian velocity v to the mass M inside the circular orbit r by:

$$v(r) = \sqrt{\frac{GM(<r)}{r}} \tag{1.48}$$

He found that the total mass calculated from radial velocities was $\sim$ 400 times higher than what was visually observable. Given the high velocity dispersion he observed, most galaxies would have escaped the cluster. From these observations, he concluded that there could be some *unseen matter* that provided the mass required to hold the whole cluster together by gravitation. Zwicky was one of the first to mention the missing matter in the Universe, called nowadays *dark matter*. Previous measurements from Jan Oort measuring stellar motion in the Milky Way also suggested that the mass in the galactic plane must be greater than observed.

1962 - Vera Rubin measured the rotation curve from $\sim$ 1100 stars in the Milky Way, and was the first to show a flat rotation curve (Rubin et al., 1962). Few years later in 1970, with W. Kent Ford, they made a precise measurement of the Andromeda's rotation curve (Rubin & Ford, 1970) showing that the curve was flat at high radii, up to $\sim$ 24 kpc (see left panel of Figure 1.13). These results advocated for the presence of *unseen* matter around galaxies, otherwise stars would escape and the galaxies would not be stable. However, Vera Rubin never mentioned the term *dark matter* in her papers.

1980s - The existence of dark matter wasn't completely accepted until the 80's. At this time, numerous measurements of galaxy rotation curves confirmed the existence/need of dark matter. The right panel of Figure 1.13 taken from Begeman et al. (1991) shows the velocity measurements of galaxy NGC 3198. The contribution from the disk (visible matter) is clearly not sufficient to explain the observation. In order to explain the constant velocities at high radii, the presence of a non-visible massive halo around the galaxy, called a *dark matter halo* was introduced. This finding was confirmed on the theoretical side, with the emergence of numerical simulations. Using N-body simulations with ~ 300 mass points, Ostriker & Peebles (1973) found that adding a spherical halo component to galaxies was required to reproduce the measured velocities at high radii and get stable galaxies.

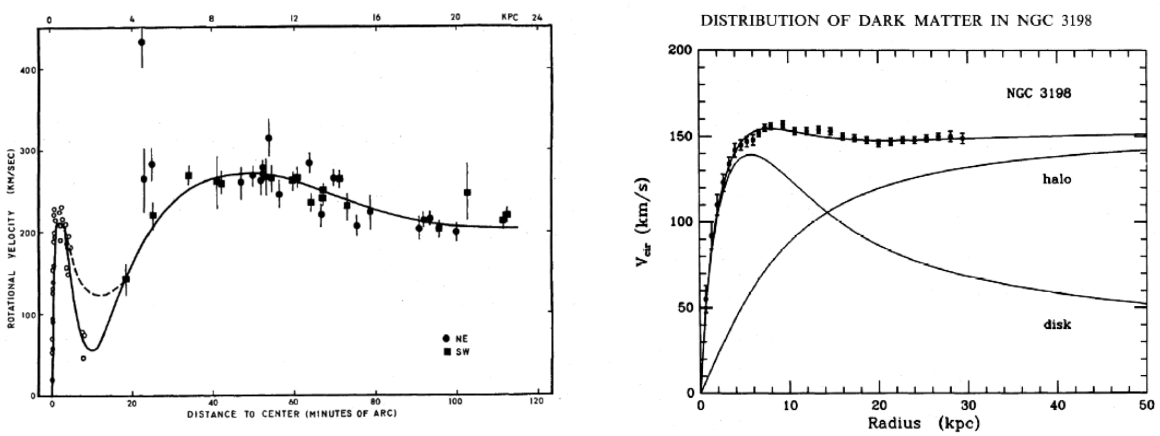

Figure 1.13: *Left: Rotation curve of Andromeda measured by Rubin & Ford (1970). Points with error bars correspond to the radial velocity measurement. Right: Rotation curve of NGC 3198 from Begeman et al. (1991). Points with error bars correspond to the radial velocity measurement. The lower curve, labelled disk, represents the expected radial velocity of the galaxy from its observed mass. The middle curve represents the expected radial velocity of the (non-visible) halo around the galaxy. The upper curve, that fits the data, represents the sum of the two other curves, suggesting that adding a (non-visible) massive halo around the galaxy could explain the velocity measurement.*

2006 - One of the recent observational evidences for the existence of dark matter is the discovery of the bullet cluster shown in Figure 1.14. This figure shows two merging clusters (orange circles). Lensing measurements indicate that most of the mass is located in the blue regions, whereas X-ray measurements in pink show the gas (baryonic matter). The gas interacted strongly during the merger, whereas the mass does not seem to have interacted. This suggests that dark matter interacts weakly with baryons and does not seem to interact through electromagnetic interaction.

Today, the existence of dark matter has been demonstrated on many cosmological scales and for very different probes. The formation of large scale structures, the lensing effect and the measurement from CMB represent a striking evidence for dark matter. Figure 1.15 illustrates how the amount of dark matter in the Universe modifies the shape of the CMB temperature power spectrum. In Chapter 3, I will present how structures were formed in the Universe and we will see that without dark matter the Universe would not appear as it is today.

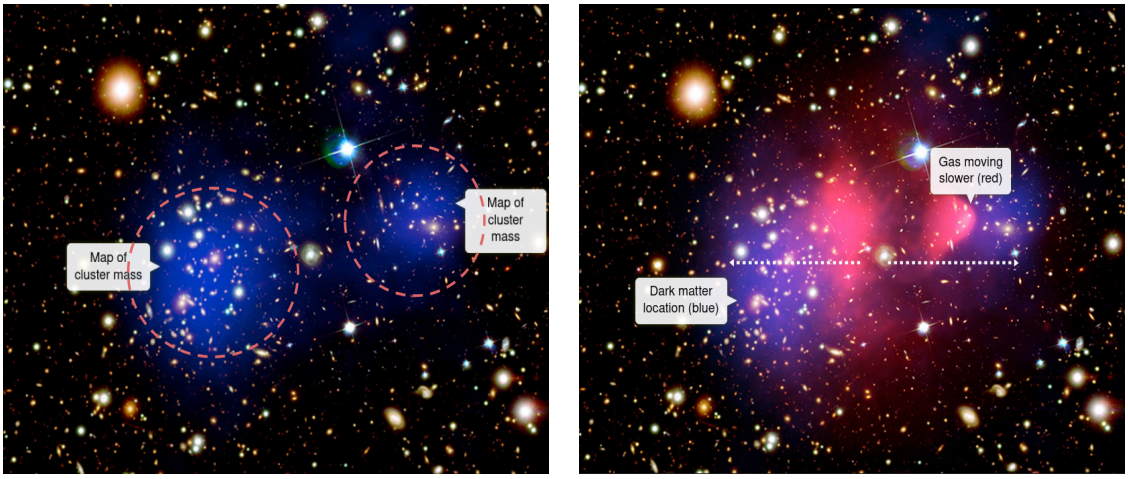

Figure 1.14: *Composite image of the Bullet cluster. Left: lensing measurement representing the mass (dark matter) of the bullet cluster in blue. Right: X-ray measurements representing the gas (baryonic matter) are added in pink. The gas is delayed compared to the mass, meaning that the gas strongly interacted during the merger, whereas the mass (dark matter) does not seem to have interacted. It suggests that dark matter does not interact or very weakly. This figure is originally from Markevitch et al. (2004) and taken from viewspace.org.*

Since the *"discovery"* of dark matter by its gravitational impact, many efforts have been made to detect it, either as astrophysical objects or in particle physics experiments. However, despite all the effort put from the particle physics side, dark matter has never been directly observed and its nature is still unknown. In the standard model of cosmology, dark matter is considered to be *cold*, i.e. is a non-relativistic fluid, *collisionless*, i.e. dark matter does not interact (or interactions are small enough to remain undetected), *stable*, i.e. dark matter does not decay or its lifetime is longer than the age of the Universe today and its field has *adiabatic inhomogeneities*, i.e. dark matter follows the same primordial density field as other components of the Universe. Due to its unknown nature, extensions to the standard model of particle physics can predict the existence of potential dark matter candidates. Along them, popular candidates are the *WIMP* (weakly interacting massive particle) motivated by supersymmetry, or the *axion*, a scalar particle introduced originally to solve the strong CP problem in quantum chromodynamics (QCD) (Peccei & Quinn, 1977).

Transition from matter domination to dark energy domination

After $z \sim 2$, a new and unknown form of energy rises and begin to dominate the energy budget of the Universe. As for the transition between radiation and matter, we can derive the redshift $z_{\rm de}$ at which the matter-dark energy transition occurs. At redshift $z < 2$ we can neglect the contribution from radiation to the energy content of the Universe. Using Equation (1.15) and Equation (1.21) in a flat Universe ($\Omega_k = 0$) we find $z_{\rm de}$ to be:

$$z_{\rm de} = \left(\frac{2\Omega_{\Lambda,0}}{\Omega_{M,0}}\right)^{1/3} - 1 \approx 0.63 \tag{1.49}$$

which corresponds to the cosmic time $t_{\rm eq} \approx 7.8$ billion years.

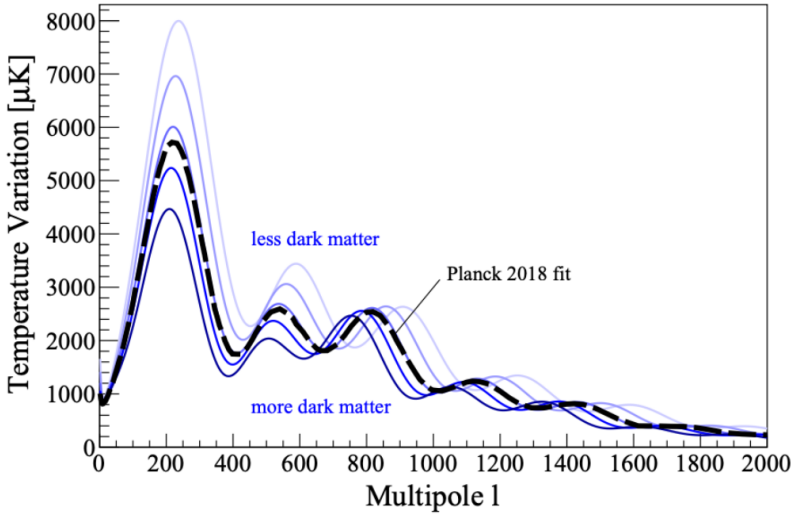

Figure 1.15: *Impact of changing the dark matter energy density on the shape of the CMB temperature anisotropy power spectrum. The dashed line shows the best-fit from Planck Collaboration et al. (2020). This figure is taken from Schumann (2019).*

1.6.3 Dark energy

Dark energy is the last missing piece of the ΛCDM model. Initially, the cosmological constant was introduced by Einstein to solve the equation of GR for a static Universe. Then, with the discovery of the Universe expansion, there was no more need for a cosmological constant. In a Universe of matter, GR predicts that the cosmic expansion will slow down over time due to the gravitational forces.

1998 - The Supernova Cosmology Project (SCP) (Perlmutter et al., 1999), and the High-redshift Supernova Search Team (HSST) (Riess et al., 1998) measured the luminosity distance of supernova Ia samples and reported independently a late-time acceleration of the Universe expansion. Therefore, a new type of energy is needed to drive this acceleration: *dark energy*. Previous works (in late 80s and 90s) on large scale structures also concluded that in a flat Universe (as required by inflation) with cold dark matter, a positive cosmological constant is needed, and should contribute "as much as 80% of the critical density " (Efstathiou et al., 1990), but not evidence had yet been found.

Type Ia supernovae (SNe Ia) are likely to come from the disruption of white dwarf stars. In a binary system of a white dwarf and a massive companion, the white dwarf accretes matter from its companion. When its mass approaches the *Chandrasekhar* mass $m_c = 1.44 M_\odot$, the electronic pressure cannot withstand the gravity forces. The density of the white dwarf core increases and its temperature reaches the ignition temperature for carbon fusion. Then, the white dwarf undergoes a suite of runaway nuclear reactions leading to its disruption. SN Ia spectra are characterised by the lack of hydrogen and the presence of a singly ionised silicon SiII absorption line at 615 nm (near peak). The explosion being driven by the Chandrasekhar mass, the amount of energy released by the supernova is relatively similar for different SNe Ia and so is the supernova intrinsic luminosity. However, there is some diversity between SNe Ia, which does not make them perfect *standard candles*. But, it has been established there is a relationship

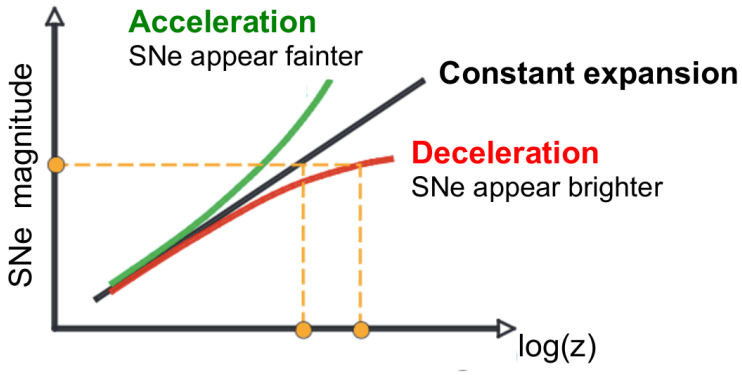

Figure 1.16: *Pedagogical plot of the apparent brightness of SNe Ia as a function of redshift. SNe Ia appear fainter in an accelerating Universe and brighter in a decelerating Universe.*

between their optical peak luminosity and decay time (bluer-slower relation) (Phillips, 1993) as well as with their colour (bluer-brighter relation) (Astier et al., 2006), making them excellent standardizable candles.

As the amount of energy released by a SN Ia is tremendous - its luminosity is $\sim 5 \cdot 10^9$ times higher than that of the Sun - SNe Ia are ideal for cosmology up to high redshifts.

Neglecting SN Ia diversity, we can assume the SN Ia intrinsic luminosity to be the same for all events. Then, measuring the apparent magnitude m of a SN Ia allows us to measure its luminosity distance:

$$m - M = 5 \log_{10}(D_L) + 25 \tag{1.50}$$

where M is the SN Ia absolute magnitude, i.e. its magnitude at a distance of 10 pc, which is measured to be $M \sim -19$ at brightness peak. The luminosity distance being related to the underlying cosmology, see Equation (1.29), SNe Ia can be used as cosmological probes. If the only component of the Universe that opposes its expansion is gravity, the more time passes, the less expansion there is. Consequently, in the case of deceleration, SNe Ia should appear brighter because their light travels a shorter distance than in the case of constant expansion. Conversely, SNe Ia appear fainter in a Universe that is expanding at an accelerating rate. Figure 1.16 illustrates this effect.

Observations from both SCP and HSST teams showed that distant SNe Ia appear fainter, meaning that the Universe expansion is accelerating. This acceleration is driven by an unknown form of energy called *dark energy*. Figure 1.17 from Perlmutter et al. (1999) shows the measurements from a sample of 42 distant SNe Ia at redshifts $0.18 < z < 0.83$ combined with a set of nearby SNe Ia. Frome these data, they measure the *dark energy* contribution to be $\Omega_\Lambda \sim 0.7$ (in the ΛCDM model).

Theoretically, dark energy can be associated to a cosmological constant Λ in the Friedmann equations shown in Section 1.3. It is considered as a fluid with a negative pressure (repulsive gravitational effect). Its equation of state is $P = w\rho$ with $w = -1$ (see Equation (1.17)). By allowing the dark energy equation of the state to vary over time, one can explore deviations to the cosmological constant case. The parametrisation from (Chevallier & Polarski, 2001)

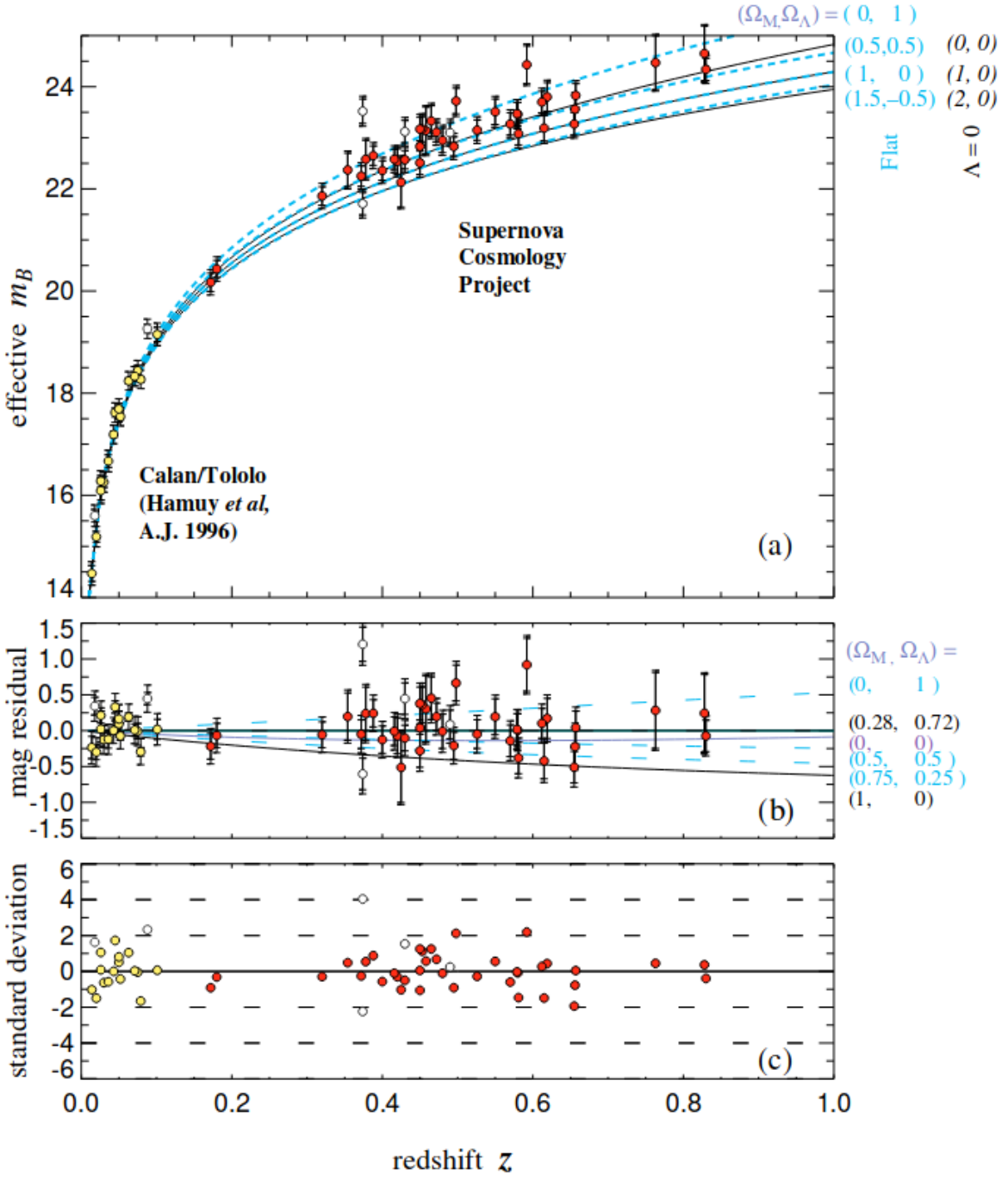

Figure 1.17: *Hubble diagram (apparent magnitude vs redshift) of 42 high redshift SNe Ia from the Supernova Cosmology Project (red dots with error bars), and 18 low redshift supernovae from the Cálan/Tololo Supernova Survey (yellow dots with error bars). Solid and dashed lines represent different ΛCDM predictions with different* $(\Omega_m, \Omega_\Lambda)$ *values (Perlmutter et al., 1999). The residuals favour a cosmological model with* $\Omega_m = 0.28$ *and* $\Omega_\Lambda = 0.72$.

and (Linder, 2003) (CPL) is generally used to test such deviations:

$$w(a) = w_0 + (1 - a)w_a \tag{1.51}$$

This parametrisation is valid in many scalar field models of dark energy (*quintessence models).* In the case of ΛCDM cosmology $w_0 = -1$ and $w_a = 0$. The amount of dark energy today as measured by Planck Collaboration et al. (2020) is :

$$\Omega_\Lambda = 0.6847 \pm 0.0073 \tag{1.52}$$

As of today, the nature of the dark energy is still unknown. Many models beyond ΛCDM try to explain this accelerated expansion by the introduction of a new cosmological fluid (e.g. quintessence, Tsujikawa (2013)) or by considering deviations from general relativity over cosmological distances (e.g. $f(R)$ theories (De Felice & Tsujikawa, 2010) or the galileon model (Nicolis et al., 2009)). More details on this can be found in Sami & Myrzakulov (2015) and Tsujikawa (2010). Another unresolved question about dark energy is to understand why it starts to dominate the energy content of the Universe at late time, as shown in Figure 1.12.

Dark energy has been one of the major fundamental questions in cosmology over the last two decades. Many past and future experiments aim to understand the nature of dark energy. With DESI, we aim to constrain the equation of state of dark energy at the percentage level.

1.7 Current status of ΛCDM

In this chapter I have given a general overview of the standard (or concordance) model of cosmology, ΛCDM. We described how, thanks to almost a century of theoretical predictions and observational confirmations, we moved from a vision of our Universe in which we wondered whether there were nebulae inside the Milky Way to a global vision of our Universe, from the Big Bang and the first seconds to today, almost 14 billion years later.

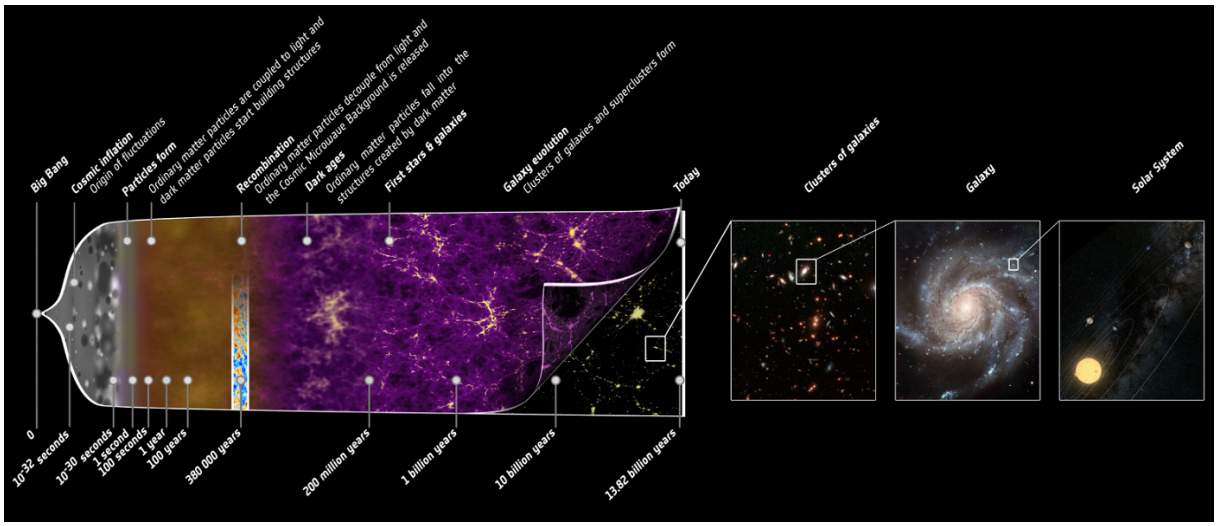

Figure 1.18: *A schematic view of the evolution of the Universe since the Big Bang. Credit: NASA/ESA*

This concordance model is based on the theory of general relativity and on three observational pillars. The first one is the observation of the expansion of the Universe in the 1930s by G.

Lemaître and E. Hubble (Section 1.2). The second is the understanding of the origin of light elements in the framework of the primordial nucleosynthesis, developed in the late 1940s by Alpher, Bethe and Gamow (Section 1.5.1). Finally, the third pillar is the fortuitous observation of the cosmic microwave background in 1964 by Penzias & Wilson (Section 1.5.2), which confirmed the prediction of a relic radiation by Alpher, Herman and Gamow in the late 1940s. The CMB is still today an invaluable source of information on the physical properties of the Universe. Two pieces are still missing from the standard model of cosmology: cold dark matter and dark energy. The need for cold dark matter was first demonstrated by the observation of galaxy rotation curves in the 1970s/1980s and confirmed by the observation of the CMB (Section 1.6.2). Dark energy is responsible for the late acceleration of the expansion of the Universe, and was revealed through the observation of distant SNe Ia (Section 1.6.3). These two elements today make up $\sim 95\%$ of the energy density of the Universe, while the remaining $\sim 5\%$ is baryonic matter, which composes the visible structures of the Universe, i.e. galaxies, stars, planets...

The description of the evolution of the Universe is given by the hot Big Bang model. The Universe was born from a hot and dense initial state and began to expand around 13.8 billion years ago. It first underwent a rapid and gigantic expansion phase, called *inflation*, where initial quantum fluctuations generated small density fluctuations thought to be the seeds of the structures we observe today. These tiny perturbations then evolved and grew under gravity in different regimes depending on the energy content of the Universe (Section 1.6) until the perturbations were dense enough to collapse and create the structure of the Universe. A schematic representation of the evolution of the Universe is shown in Figure 1.18. I will describe how these tiny fluctuations generated during the inflation phase evolved to create the large-scale structure of the Universe in Chapter 3. In order to describe the evolution and properties of the Universe, the ΛCDM model depends on just six free parameters (in the framework of CMB analyses):

- A_s: the amplitude of the primordial power spectrum,
- n_s: the spectral index of the primordial power spectrum,
- θ_*: the angular scale on the sky corresponding to the comoving sound horizon at recombination,
- $\Omega_b h^2$: the baryon density in the Universe today,
- $\Omega_{\rm cdm} h^2$: the dark matter density in the Universe today,
- τ: the optical depth at reionisation.

These parameters are constrained by cosmological fits to CMB data from the Planck satellite in Table 1.2. Today, the values of the cosmological parameters of the ΛCDM model are precisely inferred from cosmological fits using three main cosmological probes: CMB anisotropy spectra, luminosity distances from SNe Ia and measurement of the BAO scale from the *galaxy clustering* (i.e. the spatial distribution of galaxies, Section 1.5.2). Beside the measure of the BAO scale, the galaxy clustering can probe the growth rate of structure f through redshift space distortions (RSD) which provide constraints on dark energy models and tests of potential deviations from GR. The latest results from galaxy clustering measurements come from the eBOSS collaboration (Alam et al., 2021a) and are given in Figure 1.19. This figure compares the eBOSS measurements of the BAO scale and growth rate of structure to the corresponding the best-fit ΛCDM

predictions from the Planck collaboration in Table 1.2, showing good agreement between the two. The combination of different cosmological probes allows us to give tighter constraints on cosmological parameters, as illustrated in Figure 1.20. However, despite the good agreement between cosmological measurements at high redshift from the three main probes, the value of H_0 directly measured from low redshift SNe Ia calibrated with Cepheids is in strong disagreement (more than 5σ) with the cosmological constraints derived from the CMB (Riess et al., 2022). The nature of this discrepancy remains unresolved today, and may be due to unknown systematic effects which may bias the measurement in either analysis, or may require new physics beyond the standard cosmological model.

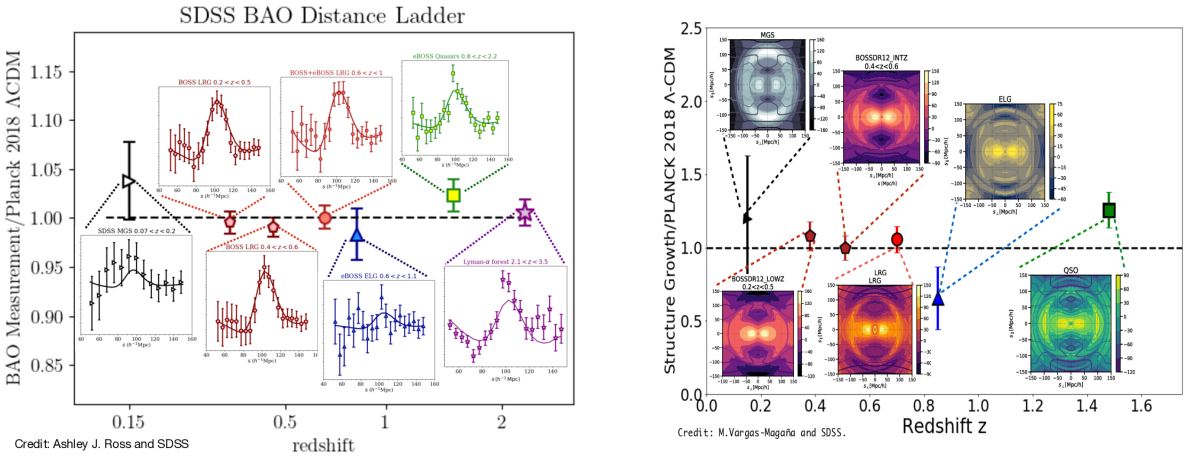

Figure 1.19: *Measurements of the (isotropic) BAO scale (left) and growth rate of structure (right) for different galaxy tracers and redshifts from eBOSS (Alam et al., 2021a) compared to ΛCDM predictions from the best-fit to Planck data (Planck Collaboration et al., 2020).*

Beyond galaxy clustering, it is important to note that LSS have been also used to constrain cosmological parameters using the weak lensing effect and its cross-correlation with galaxy clustering in photometric surveys. Numerous results from this type of study have been published in recent years (Collaboration et al., 2022, Dalal et al., 2023, Heymans et al., 2021) and weak lensing will be the main cosmological probe of new experiments such as the Vera Rubin observatory LSST (LSST Dark Energy Science Collaboration (2012), first light expected in 2024) or Euclid (Laureijs et al., 2011), which was successfully launched two days ago, as I write these lines. The wide variety of probes allows the same quantities to be measured with different physical processes and observation techniques, so that the systematic effects are different from one analysis to the other, which helps to confirm (or not) the cosmological parameter constraints. This is a promising avenue for future cosmological studies.

1.8 Outline of the thesis

During the last two decades, the emergence of galaxy surveys and especially spectroscopic galaxy surveys has provided an important probe of cosmology. We previously mentioned results from the eBOSS survey of the SDSS collaboration but many other spectroscopic surveys have also contributed to the precise determination of cosmological parameters (e.g. Blake et al. (2011),

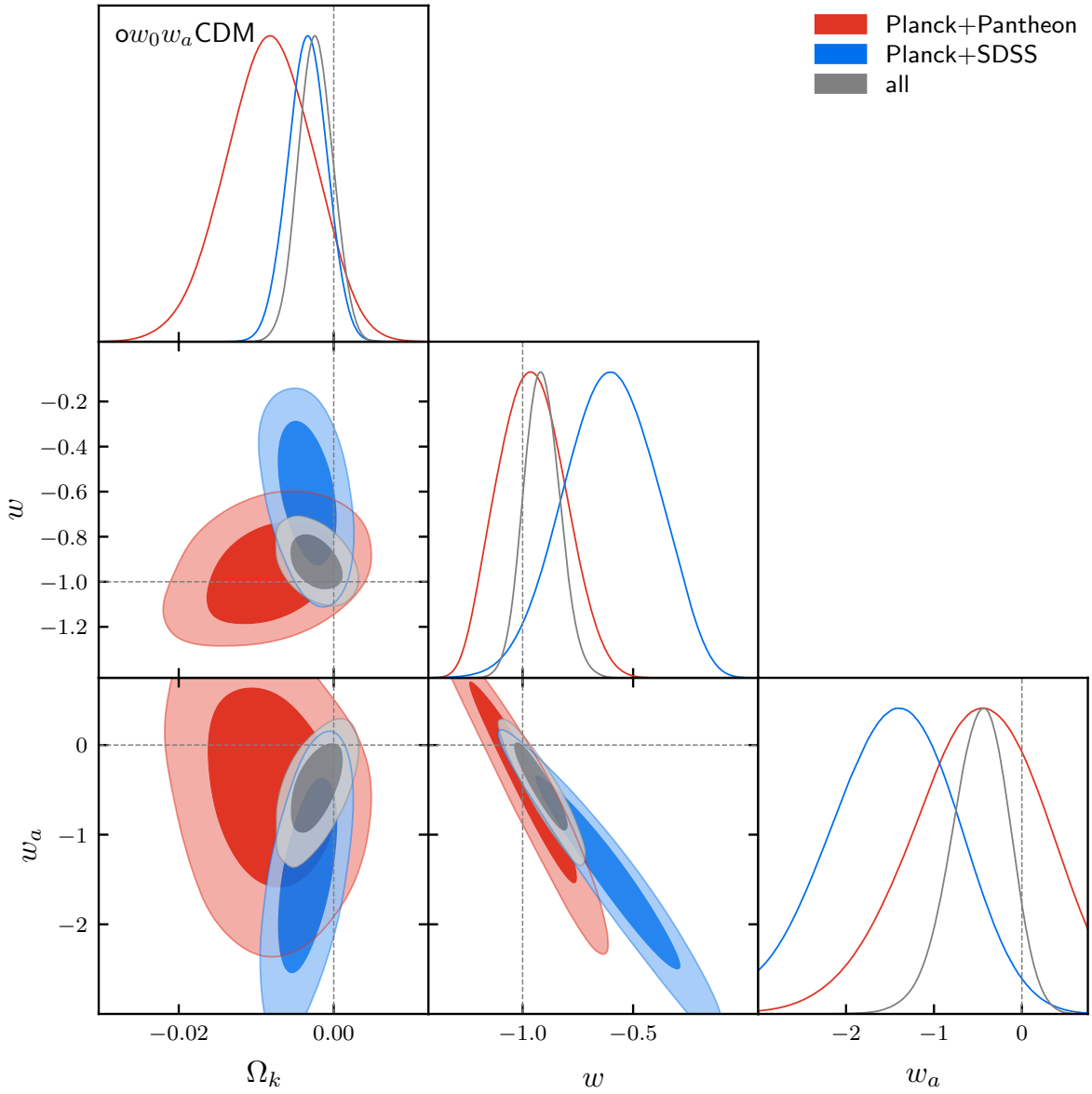

Figure 1.20: *Two dimensional contours on w_0, w_a, and Ω_k under the assumption of an open w_0w_a-CDM cosmological model. The one-dimensional constraints on each independent parameter are presented in the top panels. The red contours represent the 68% and 95% constraints when using the full Planck data (Temperature, Polarisation and lensing) and the Pantheon SNe Ia measurements. The blue contours represent the constraints from Planck and SDSS BAO+RSD, while the grey contours represent the combination of all measurements. Figure taken from Alam et al. (2021a)*

Guzzo et al. (2014), Aihara et al. (2018)). Stage-III spectroscopic surveys[1] ended in 2020 with the results from the eBOSS collaboration. Over 20 years, the various cosmological surveys from the SDSS collaboration measured more than 2 million of galaxy/quasar redshifts.

The Dark Energy Spectroscopic Instrument (DESI) (Collaboration et al., 2016) is the first Stage-IV spectroscopic survey on sky. Commissioning started in 2020 and the main survey of 5 years started in June 2021. DESI aims to measure over 40 million of redshifts using four different tracers: bright galaxies (BGS), luminous red galaxies (LRGs), emission line galaxies (ELGs) and quasars (QSOs). DESI will constrain cosmological parameters through the precise determination of the BAO scale and the estimation of the linear growth rate of structure through RSD measurements with a sub percent precision level. To reach that precision, clustering analyses have to be tested on simulated catalogues of galaxies in large interval of scales spanning the linear (> 50 Mpc/h) and quasi-linear regimes ($20 < s < 50$ Mpc/h). These tests rely heavily on N-body simulations coupled to prescriptions to describe the connection between dark matter halos and galaxies, the so called *galaxy-halo connection*. As was shown in Alam et al. (2021b) for a Stage-III spectroscopic survey, systematic uncertainties related to the galaxy-halo connection were found at that time to be negligible with respect to other sources of systematic errors. As those are expected to be reduced in DESI, galaxy-halo connection studies are becoming increasingly important to derive robust systematic error budgets for cosmological analyses. Galaxy-halo connection studies also provide invaluable information about the physics of galaxy formation, and their distribution within the cosmic web. To probe the galaxy-halo connection it is mandatory to have clustering measurements down to very small scales (< 1Mpc/h).

This is the framework of this thesis, whose aim is to study the galaxy-halo connection of ELGs by modelling the small scale clustering of DESI ELGs. We use the data from the last two months of observations of the survey validation phase before the start of the main survey. During this time, DESI observed 267k ELGs, which makes it the largest ELG spectroscopic sample to date ($\sim 173k$ in eBOSS). We present an overview of the DESI instrument and the survey strategy in Chapter 2. We describe the statistics we use to analyse the data, namely the two-point correlation function (2PCF). We also address the different (known) systematic effects that can affect the measurement and mention the different techniques to mitigate them. Finally, we describe the observational effects used to constrain cosmological parameters. The first part of Chapter 3 provides the theoretical framework of structure formation and evolution in the cold dark matter scenario and presents the different techniques to model it with numerical simulations. Then, we focus on galaxies and detail the theoretical and empirical prescriptions to study the galaxy-halo connection. In the second part, we focus on the particular case of ELGs and review previous theoretical and observational results about the ELG galaxy-connection. Among the prescriptions to describe the galaxy-halo connection, we use the halo occupation distribution (HOD), an empirical approach whose parameters are derived from clustering measurements at small scales. In Chapter 4, we present a novel and promising technique based on Gaussian processes (GP) to perform accurate and precise fits of HOD models, and test it using N-body simulations. This techniques is then applied to DESI ELG data in Chapter 5. As these

[1]The classification in stages was introduced in the Dark Energy Task Force report (Albrecht et al., 2006). Stage I refers to cosmological experiments that started before 2005, Stage II in the late 2000s (2005-2010), Stage-III in the 2010s (2010-2020) and Stage-IV refers to those that start in the 2020s. Today, the discussions and projects are starting to design Stage-V experiments that will be carried out in a couple of decades. At each stage, the precision on cosmological parameters increases by several orders of magnitude.

data allow clustering to be measured at very small scales, never probed before, we provide new insights on the galaxy-halo connection for ELGs. In particular, we demonstrate the need for galactic conformity to reproduce the observations and the first observational hint that a fraction of ELGs reside in the outskirts of the dark matter halos, as mentioned in theoretical studies but never observed so far. Finally, we conclude and discuss the implication of this thesis in Chapter 7 and give several prospects for future work.

Bibliography

Abazajian, K. N., Adshead, P., Ahmed, Z., et al. 2016, CMB-S4 Science Book, First Edition, arXiv, doi: 10.48550/arXiv.1610.02743

Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, Publications of the Astronomical Society of Japan, 70, S4, doi: 10.1093/pasj/psx066

Alam, S., Aubert, M., Avila, S., et al. 2021a, Physical Review D, 103, 083533, doi: 10.1103/PhysRevD.103.083533

Alam, S., de Mattia, A., Tamone, A., et al. 2021b, Monthly Notices of the Royal Astronomical Society, 504, 4667, doi: 10.1093/mnras/stab1150

Albrecht, A., Bernstein, G., Cahn, R., et al. 2006, Report of the Dark Energy Task Force, arXiv, doi: 10.48550/arXiv.astro-ph/0609591

Alpher, R. A., Bethe, H., & Gamow, G. 1948a, Physical Review, 73, 803, doi: 10.1103/PhysRev.73.803

Alpher, R. A., Herman, R., & Gamow, G. A. 1948b, Physical Review, 74, 1198, doi: 10.1103/PhysRev.74.1198.2

Alpher, R. A., & Herman, R. C. 1949, Physical Review, 75, 1089, doi: 10.1103/PhysRev.75.1089

Astier, P., Guy, J., Regnault, N., et al. 2006, Astronomy & Astrophysics, 447, 31, doi: 10.1051/0004-6361:20054185

Begeman, K. G., Broeils, A. H., & Sanders, R. H. 1991, Monthly Notices of the Royal Astronomical Society, 249, 523, doi: 10.1093/mnras/249.3.523

Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, The Astrophysical Journal Supplement Series, 208, 20, doi: 10.1088/0067-0049/208/2/20

Blake, C., Davis, T., Poole, G., et al. 2011, Monthly Notices of the Royal Astronomical Society, 415, 2892, doi: 10.1111/j.1365-2966.2011.19077.x

Chevallier, M., & Polarski, D. 2001, International Journal of Modern Physics D, 10, 213, doi: 10.1142/S0218271801000822

Collaboration, D., Aghamousa, A., Aguilar, J., et al. 2016, The DESI Experiment Part I: Science,Targeting, and Survey Design, arXiv. http://arxiv.org/abs/1611.00036

Collaboration, D., Abbott, T. M. C., Aguena, M., et al. 2022, Physical Review D, 105, 023520, doi: 10.1103/PhysRevD.105.023520

Collaboration LiteB I R D, Allys, E., Arnold, K., et al. 2023, Progress of Theoretical and Experimental Physics, 2023, 042F01, doi: 10.1093/ptep/ptac150

Conklin, E. K. 1969, Nature, 222, 971, doi: 10.1038/222971a0

Cooke, R. J., Pettini, M., & Steidel, C. C. 2018, The Astrophysical Journal, 855, 102, doi: 10.3847/1538-4357/aaab53

Dalal, R., Li, X., Nicola, A., et al. 2023, Hyper Suprime-Cam Year 3 Results: Cosmology from Cosmic Shear Power Spectra, arXiv, doi: 10.48550/arXiv.2304.00701

De Felice, A., & Tsujikawa, S. 2010, Living Reviews in Relativity, 13, 3, doi: 10.12942/lrr-2010-3

Dicke, R. H., Peebles, P. J. E., Roll, P. G., & Wilkinson, D. T. 1965, The Astrophysical Journal, 142, 414, doi: 10.1086/148306

Dodelson, S., & Schmidt, F. 2020, Modern Cosmology (Elsevier Science)

Efstathiou, G., Sutherland, W. J., & Maddox, S. J. 1990, Nature, 348, 705, doi: 10.1038/348705a0

Einstein, A. 1905, Annalen der Physik, 322, 891, doi: 10.1002/andp.19053221004

Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, The Astrophysical Journal, 633, 560, doi: 10.1086/466512

Fixsen, D. J. 2009, The Astrophysical Journal, 707, 916, doi: 10.1088/0004-637X/707/2/916

Fixsen, D. J., Cheng, E. S., Gales, J. M., et al. 1996, The Astrophysical Journal, 473, 576, doi: 10.1086/178173

Friedman, A. 1922, Zeitschrift für Physik, 10, 377, doi: 10.1007/BF01332580

Gamow, G. 1948, Nature, 162, 680, doi: 10.1038/162680a0

Guth, A. H. 1981, Physical Review D, 23, 347, doi: 10.1103/PhysRevD.23.347

Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, Astronomy & Astrophysics, 566, A108, doi: 10.1051/0004-6361/201321489

Henry, P. S. 1971, Nature, 231, 516, doi: 10.1038/231516a0

Heymans, C., Tröster, T., Asgari, M., et al. 2021, Astronomy & Astrophysics, 646, A140, doi: 10.1051/0004-6361/202039063

Hubble, E. 1929, Proceedings of the National Academy of Science, 15, 168, doi: 10.1073/pnas.15.3.168

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, Euclid Definition Study Report, Tech. rep. https://ui.adsabs.harvard.edu/abs/2011arXiv1110.3193L

Leavitt, H. S., & Pickering, E. C. 1912, Harvard College Observatory Circular, 173, 1. https://ui.adsabs.harvard.edu/abs/1912HarCi.173....1L

Lemaître, G. 1927, Annales de la Société Scientifique de Bruxelles, 47, 49. https://ui.adsabs.harvard.edu/abs/1927ASSB...47...49L

Linder, E. V. 2003, Physical Review Letters, 90, 091301, doi: 10.1103/PhysRevLett.90.091301

LSST Dark Energy Science Collaboration. 2012, Large Synoptic Survey Telescope: Dark Energy Science Collaboration, Tech. rep. https://ui.adsabs.harvard.edu/abs/2012arXiv1211.0310L

Markevitch, M., Gonzalez, A. H., Clowe, D., et al. 2004, The Astrophysical Journal, 606, 819, doi: 10.1086/383178

MICROSCOPE Collaboration, Touboul, P., Métris, G., et al. 2022, Physical Review Letters, 129, 121102, doi: 10.1103/PhysRevLett.129.121102

Nicolis, A., Rattazzi, R., & Trincherini, E. 2009, Physical Review D, 79, 064036, doi: 10.1103/PhysRevD.79.064036

Ostriker, J. P., & Peebles, P. J. E. 1973, The Astrophysical Journal, 186, 467, doi: 10.1086/152513

Particle Data Group, Workman, R. L., Burkert, V. D., et al. 2022, Progress of Theoretical and Experimental Physics, 2022, 083C01, doi: 10.1093/ptep/ptac097

Peccei, R. D., & Quinn, H. R. 1977, Physical Review D, 16, 1791, doi: 10.1103/PhysRevD.16.1791

Penzias, A. A., & Wilson, R. W. 1965, The Astrophysical Journal, 142, 419, doi: 10.1086/148307

Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, The Astrophysical Journal, 517, 565, doi: 10.1086/307221

Phillips, M. M. 1993, The Astrophysical Journal, 413, L105, doi: 10.1086/186970

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, Astronomy and Astrophysics, 641, A6, doi: 10.1051/0004-6361/201833910

Pospelov, M., & Pradler, J. 2010, Annual Review of Nuclear and Particle Science, 60, 539, doi: 10.1146/annurev.nucl.012809.104521

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, The Astronomical Journal, 116, 1009, doi: 10.1086/300499

Riess, A. G., Yuan, W., Macri, L. M., et al. 2022, The Astrophysical Journal Letters, 934, L7, doi: 10.3847/2041-8213/ac5c5b

Rubin, V. C., Burley, J., Kiasatpoor, A., et al. 1962, The Astronomical Journal, 67, 491, doi: 10.1086/108758

Rubin, V. C., & Ford, Jr., W. K. 1970, The Astrophysical Journal, 159, 379, doi: 10.1086/150317

Sami, M., & Myrzakulov, R. 2015, Late time cosmic acceleration: ABCD of dark energy and modified theories of gravity, arXiv, doi: 10.48550/arXiv.1309.4188

Schumann, M. 2019, Journal of Physics G: Nuclear and Particle Physics, 46, 103003, doi: 10.1088/1361-6471/ab2ea5

Slipher, V. M. 1917, Proceedings of the American Philosophical Society, 56, 403. https://ui.adsabs.harvard.edu/abs/1917PAPhS..56..403S

Smoot, G. F., Bennett, C. L., Kogut, A., et al. 1992, The Astrophysical Journal, 396, L1, doi: 10.1086/186504

Tsujikawa, S. 2010, in Lectures on Cosmology: Accelerated Expansion of the Universe, ed. G. Wolschin, Lecture Notes in Physics (Berlin, Heidelberg: Springer), 99–145, doi: 10.1007/978-3-642-10598-2_3

—. 2013, Classical and Quantum Gravity, 30, 214003, doi: 10.1088/0264-9381/30/21/214003

2

DESI: The Dark Energy Spectroscopic Instrument

Contents

In the previous chapter, we set out the cosmological context for this work. The whole theoretical framework constructed to give a mathematical and physical description of our Universe would be mere speculation if there were no observational confirmations of the theory. In astronomy, observations play a major role in our understanding of physical processes, and sometimes lead to unexpected discoveries. From Galileo Galilei's first astronomical observation in the 17th century to today's large ground-based and space-based telescopes, observation systems have been considerably improved.

Large-scale galaxy surveys have been used to study the structure and dynamics of the Universe by measuring large numbers of extragalactic objects. There are two types of galaxy survey: *photometric* and *spectroscopic*. The former observe objects on the sky with a given magnitude limit that depends on the *depth* of the survey. They use different filters to measure galaxy colours and morphologies, and are the basis of the next generation of large scale structure (LSS) surveys, mainly for weak lensing analysis (Laureijs et al., 2011, LSST Dark Energy Science Collaboration, 2012). On the other hand, in addition to object positions on the sky, spectroscopic surveys of galaxies aim to measure object distances from their redshift, deduced from their spectra. As already mentioned in the first chapter, spectroscopic galaxy surveys are important for cosmology, as they provide measurements of the BAO scale and the growth rate of structure. Over the past forty years, they have become an essential tool in cosmology.

This thesis work was carried out as part of the Dark Energy Spectroscopic Instrument (DESI) collaboration. I had the chance to work and used the first data of this complex instrument which results from many years of research and development conducted by women and men. This chapter is devoted to describe the DESI instrument (Section 2.3) and the science objectives (Section 2.2) of the DESI survey. We describe the data sample –the DESI One-Percent survey– (Section 2.6) and the different galaxy tracers (Section 2.4.2) that are used for the galaxy-halo connection analysis of this work. Then we present how the two-point correlation function in estimated from the data and describe the different (known) systematic effects affecting this measure and how to mitigated them (Section 2.7). Finally we present the clustering of the ELG sample from the One-Percent survey (Section 2.9). But first, we briefly review the previous galaxy spectroscopic surveys and their current status.

2.1 Brief history of galaxy spectroscopic surveys

The first major spectroscopic surveys of galaxies began in the 1980s with the Centre for Astrophysics (CfA) redshift survey, which measured 2,401 galaxy redshifts in the nearby Universe (Huchra et al., 1983), followed by the CfA 2 (CfA2) survey, which recorded 18,000 spectra of bright galaxies between 1985 and 1995 (Falco et al., 1999). The results of the CfA survey led to the discovery of the *Great Wall* shown in Figure 2.1, which was the largest single structure ever detected. It was one of the first evidence of the existence of a cosmic web due to the clustered nature of the galaxy distribution. It also provided strong indications that cold dark matter alone could not explain the observed distribution of galaxies (Vogeley et al., 1992).

In the 2000s, the Sloan Digital Spectroscopic Survey (SDSS) and the 2dFGRS (dFGRS standing for degree field galaxy redshift survey) were the first to detect the BAO signal in the galaxy distribution, with galaxy magnitude-limited samples at $z \sim 0.5$ (Cole et al., 2005, Eisenstein et al., 2005, Percival et al., 2007). Since then, several spectroscopic studies (e.g.

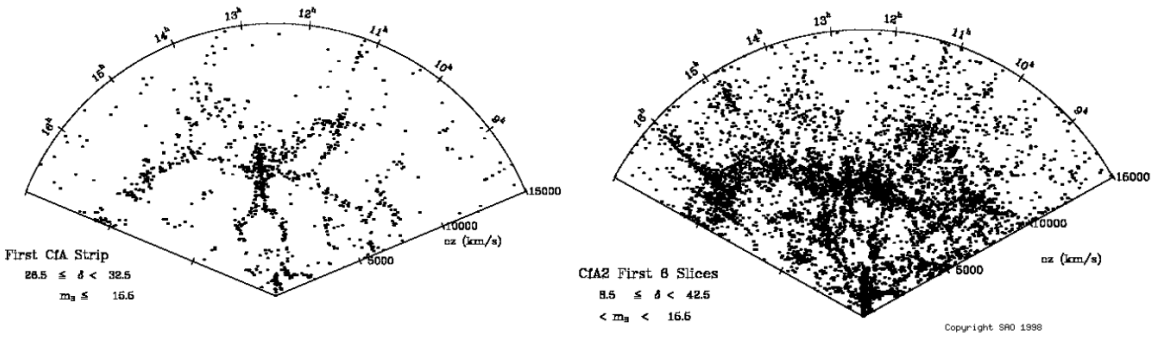

Figure 2.1: *Left panel: galaxy position map from the Centre for Astrophysics (CfA) redshift survey. Right panel: position map of 18,000 galaxies from the CfA2 survey, revealing the presence of a Great Wall in the structure of the local Universe. Credit: CfA redshift surveys, (Huchra et al., 1983, Vogeley et al., 1992).*

BOSS/eBOSS, 6dFGRS, WiggleZ, VIPERS...) have increased the number of observed galaxies, the target redshifts and the size of the footprint on the sky (Alam et al., 2021, Guzzo et al., 2014, Jones et al., 2009, Parkinson et al., 2012) as illustrated in Figure 2.2.

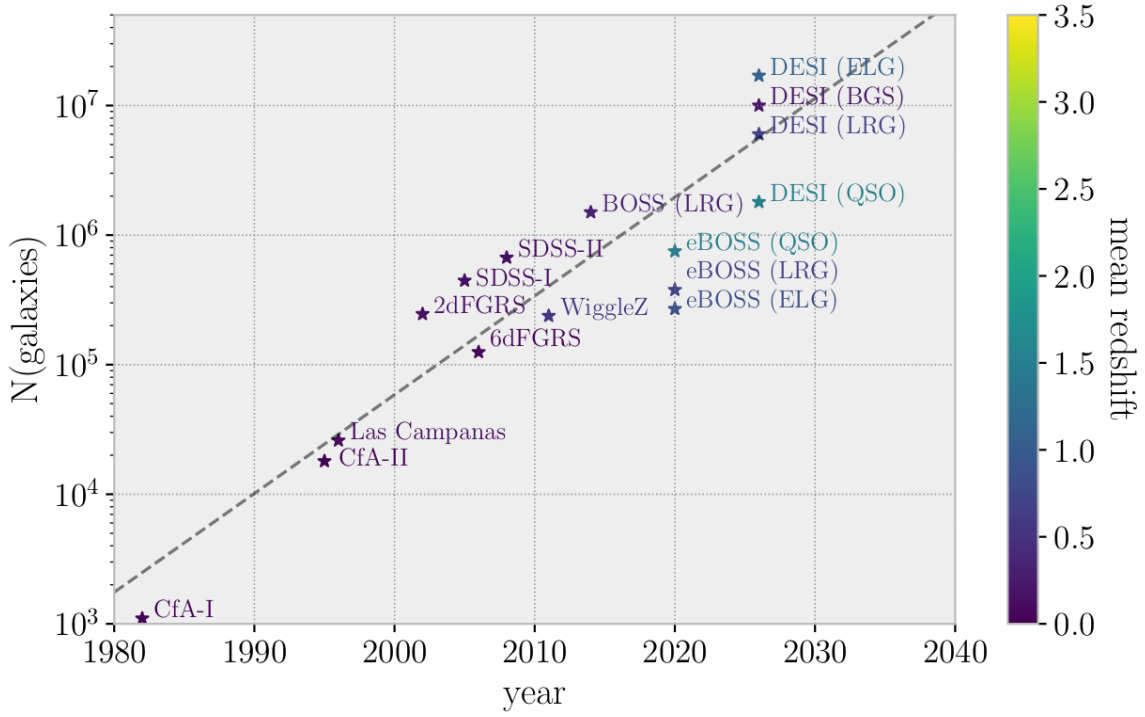

Figure 2.2: *Number of galaxies observed by the spectroscopic surveys over time. Credit: E. Chaussidon.*

Thanks to galaxy spectroscopic surveys and galaxy clustering studies, numerous discoveries have been made: detection of the BAO signal over cosmic time in the range $0 < z < 3.5$ using $\sim 2M$ galaxy redshifts (see Figure 1.11), highlight of the structure of the cosmic web with filaments, superclusters and voids, and increase in precision and accuracy of the cosmological parameter measurements. The results of 20 years of observations by the SDSS collaboration were published in 2020 (Alam et al., 2021), reviewing two decades of cosmological results for

Stage I (experiment before 2005), Stage II (2005-2010) and Stage III (2010-2020) surveys, using numerous probes. These results are summarised in Figure 2.3. Started in 2000, the SDSS (York et al., 2000) represents, so far, the largest survey of its kind. The latest public release of SDSS data, DR17, includes 5,580,057 optical and near-infrared spectra after quality cuts (Abdurro'uf et al., 2022). Data from SDSS have been used in more than 11,802 peer-reviewed publications.

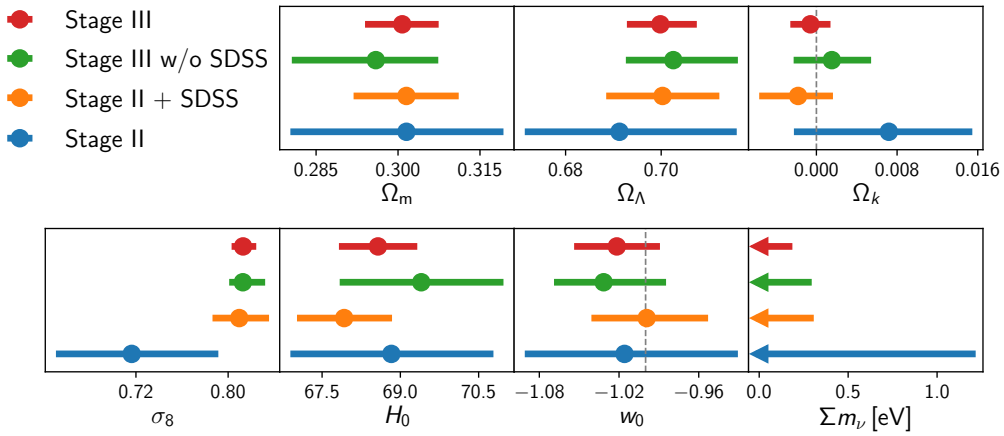

Figure 2.3: *Central values and 68% quantiles for each of the parameters describing the history of Universe expansion and growth of structure in an open, massive neutrino wCDM model. Results are shown for each data set combination, where Stage-II corresponds to a combination of the CMB (WMAP) (Bennett et al., 2013), Supernovae (JLA) (Betoule et al., 2014), and SDSS DR7 data (Abazajian et al., 2009) and Stage-III corresponds to a combination of the SDSS BAO+RSD results (Alam et al., 2021), Planck (Planck Collaboration et al., 2020), Pantheon SN Ia (Scolnic et al., 2018), and Dark Energy Survey (DES) data (Abbott et al., 2018).*

We are entering in the Stage IV era of cosmological experiments, which aim to provide high-precision measurements, up to sub-percent accuracy on cosmological parametres. DESI is the first Stage IV experiment on sky.

2.2 Overview of the DESI programme

The Dark Energy Spectroscopic Instrument (DESI) is a robotic, fibre-fed, highly multiplexed spectroscopic instrument that operates on the Mayall 4-meter telescope (equatorial mount) at Kitt Peak National Observatory (KPNO) on the Iolkam Du'ag mountain (Kitt peak) in Arizona, US (see Figure 2.4). This mountain is of particular importance to the Tohono O'odham Nation and DESI collaborators are honoured to be allowed to conduct scientific research there.

DESI is designed to measure simultaneously spectra of 5000 objects over a ~ 3 degree field and is currently conducting a five-year survey of 14 000 deg^2 (about a third of the sky), to obtain the spectra of about 40 million galaxies and quasars in a redshift range $0 < z < 3.5$. DESI aims to create a three-dimensional map of the distribution of matter covering an unprecedented volume, targeting different galaxy types.

At low redshift, $z < 0.5$, DESI carries out a *Bright Galaxy Survey* (BGS), creating a magnitude-limited sample of ultimately $\sim$ 13M galaxies to study cosmic structure in the dark energy-dominated epoch with high density sampling. At higher redshift, DESI will target in

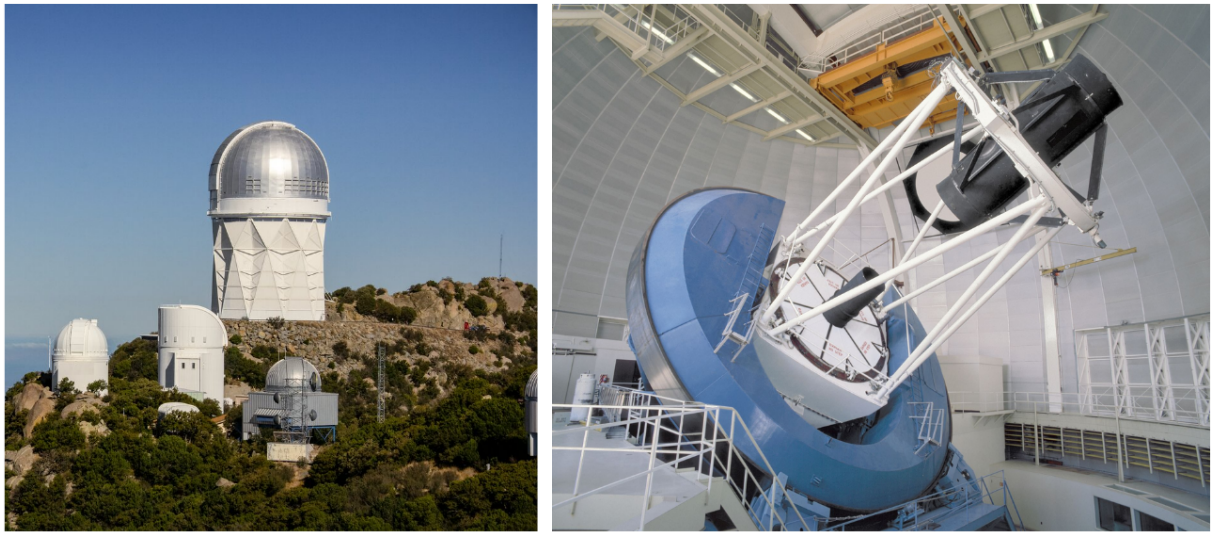

Figure 2.4: *Left: Picture of the dome of the Mayall telescope, the largest one at the Kitt Peak National Observatory (KPNO). Right: Picture of the Mayall telescope with the DESI instrument. Credit: DESI collaboration.*

total ∼8M *luminous red galaxies* (LRGs) between $0.4 < z < 1.1$, ∼17M *emission line galaxies* (ELGs) between $0.6 < z < 1.6$, and ∼3M *quasars* or *quasi stellar objects* (QSOs) between $0.8 < z < 3.5$, producing tight constraints on the large-scale clustering of the Universe to try and decipher the nature of cosmic acceleration.

The main science goal of DESI is to measure the baryon acoustic oscillation scale at different redshifts in order to precisely constrain the expansion history of the Universe. DESI will drastically reduce the errors on individual BAO measurements compared to previous experiments, with the aim of making sub-percent measurements on the BAO scales across a wide range of redshifts, as shown in Figure 2.5.

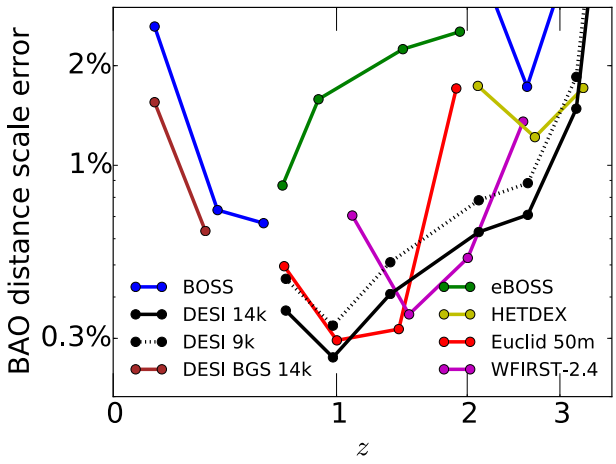

Figure 2.5: *Relative error on the (isotropic) BAO scale measurements from different past and future surveys. DESI points are initial forecasts from Collaboration et al. (2016).*

Clustering measurements will also make it possible to detect anisotropies in the galaxy distribution, known as *redshift space distortions* (RSD). This allows us to directly measure the properties of gravity through its effect on galaxy motions, by measuring the growth rate of

structure f (see Section 2.8.2). Figure 2.6 shows DESI forecasts on the expansion history of the Universe and the growth rate of structure, and presents a slice of the DESI target sample observed up to a redshift $z \sim 3.5$. These forecasts were built after the validation phase of the DESI survey (SV), a phase prior to the main survey that aimed to confirm that the survey design, instrument performance, and data quality would be sufficient to meet the scientific requirements. During SV, the DESI data and operation teams proved their ability to optimize operations (Collaboration et al., 2023a) and to efficiently process the spectra through the DESI spectroscopic pipeline (Guy et al., 2023).

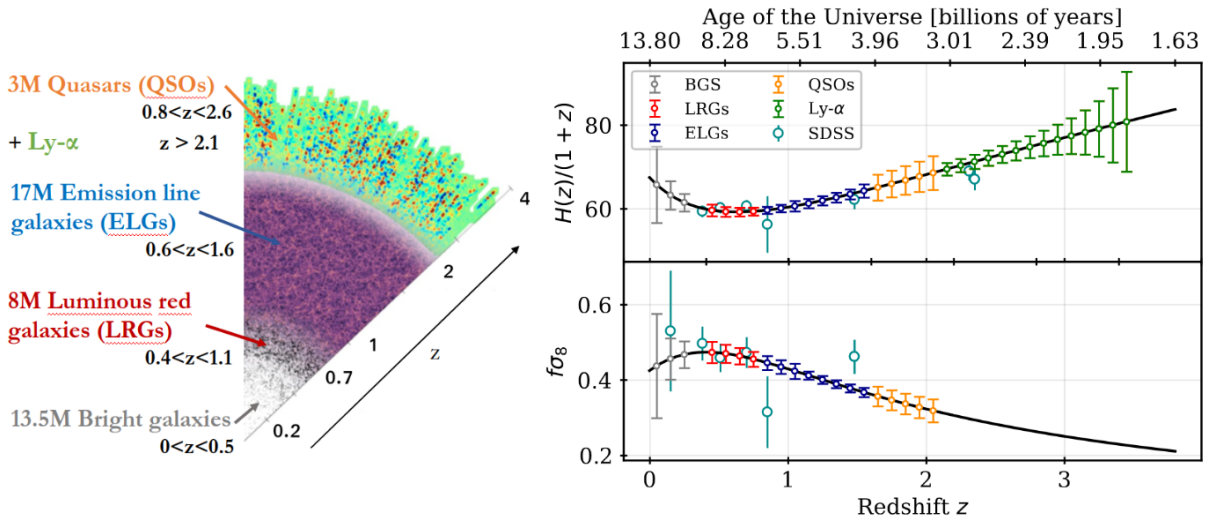

Figure 2.6: *Left: illustration of the different DESI targets and their redshift coverage. Right: 5-year DESI forecasts for BAO scale (top) and growth rate of structure (bottom) measurements, assuming central values to agree with the Planck ΛCDM best-fit model. The comparison with current results from SDSS highlights the gains expected from DESI in terms of the number of measurements (every 0.1 in z) and their accuracy.*

At high redshift, DESI quasar spectra are used to measure *Lyman-α* forests, which are absorption features in the spectrum of high redshift quasars due to neutral hydrogen clouds present between the quasar and the observer. This measurement probes the intergalactic medium and the power spectrum of matter at small scales, which contain information that can be used to constrain the sum of neutrino masses $\sum_\nu m_\nu$. This was done in eBOSS, leading to one of the best upper limits on the sum of neutrino masses to date, $\sum_\nu m_\nu < 0.115$ eV at 95% confidence (Alam et al., 2021).

In addition to constraints on dark energy, DESI will also constrain models of primordial inflation by measuring potential primordial non-gaussianities $f_{\rm NL}$ caused by inflation in the large-scale distribution of galaxies. Describing this measurement is beyond the scope of this thesis, and more detail on primordial non-gaussianities can be found in Desjacques & Seljak (2010) and Chen (2010).

Finally, DESI is also undertaking a Milky Way Survey (MWS), which will observe $\sim$ 7M stellar spectra over 5 years, in order to provide new constraints on the assembly history of the Milky Way and its dark matter distribution through measurements of chemical composition and radial velocity dispersions (Cooper et al., 2023).

2.3 Instrument design

The DESI instrument is complex, and its construction required a great deal of technology and effort. The technical details of the instrument are described in various articles: the focal plane system (Silber et al., 2023), the optical corrector (Miller et al., 2023), the spectrographs (Perruchot et al., 2020), the fibre systems (Poppett et al., 2020), and the instrument overview (Abareshi et al., 2022). In the following we give a brief overview of the instrument.

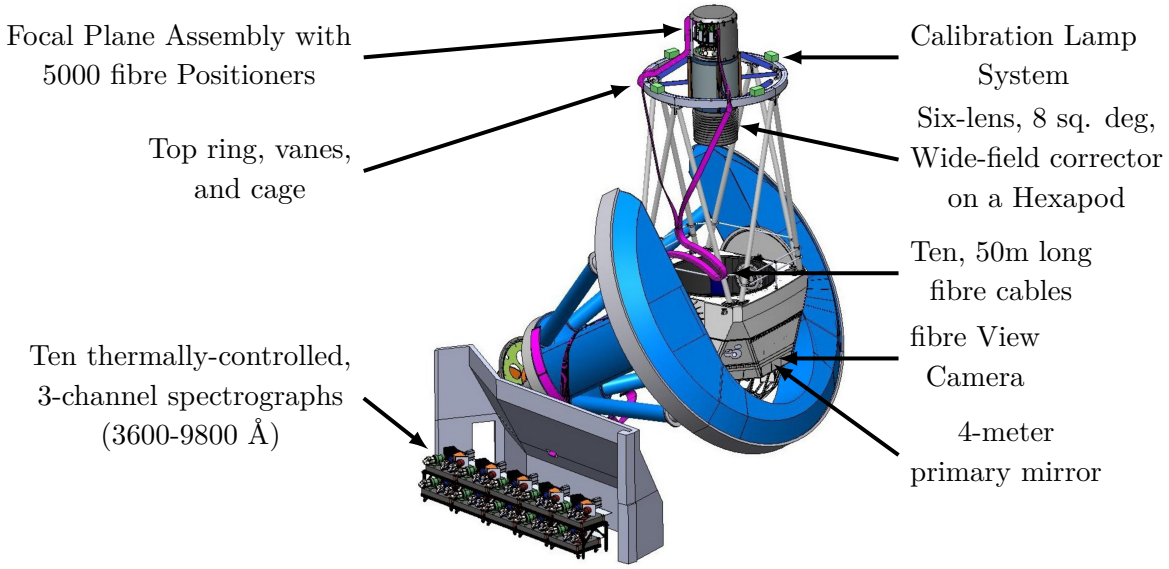

Figure 2.7: *3D model of DESI installed on the Mayall Telescope. The most relevant parts of the instrument are annotated. Credit: Image from DESI collaboration and labels from E. Chaussidon.*

DESI is mounted on the Mayall telescope, which has a 4-meter primary mirror. The instrument has been designed and built over a period of around ten years (2010-2020). Figure 2.7 shows a schematic picture of the instrument. The focal plane is the most important innovation of DESI. It has a diameter of 0.8 m and is composed of 5000 fibres, which are automatically positioned by robot positioners (see Figure 2.8).

By way of comparison, in SDSS the $\sim$ 1000 optical fibres were placed by hand on a plate and changing exposure in SDSS took $\sim$ dozens of minutes whereas in DESI it takes < 2 min. The focal plane is subdivided into 10 petals, each containing 500 robotic fibre positioners and a Guide Focus Alignment camera (GFA). Six of the GFAs are configured as guide cameras and four are used to maintain optical alignment between the optical corrector and the primary mirror.

The robotic positioners, which carry the fibers have a diameter of 4 mm while the fibre diameter is 107 μm. The positioners have two axes of rotation, the first axis θ is centered on the positioner and the second eccentric axis ϕ is centered along an arm located nominally 3 mm from the axis θ. Each positioner can place the fibre in a patrol region with a diameter of 12 mm, to an accuracy of $\sim$ 10 μm. The fibre positions on the focal plane are optimized to maximize focal plane coverage, and can be placed up to 10.4 mm from neighboring units. This optimization allows patrol regions between fibres to overlap, so a software has been developed to avoid collisions between neighboring positioners (Kent et al., 2023).

Each petal is linked to a single three-arm spectrograph, covering a wavelength range from 3600 to 9800 Å. A schematic representation of a spectrograph is given in Figure 2.9. Each

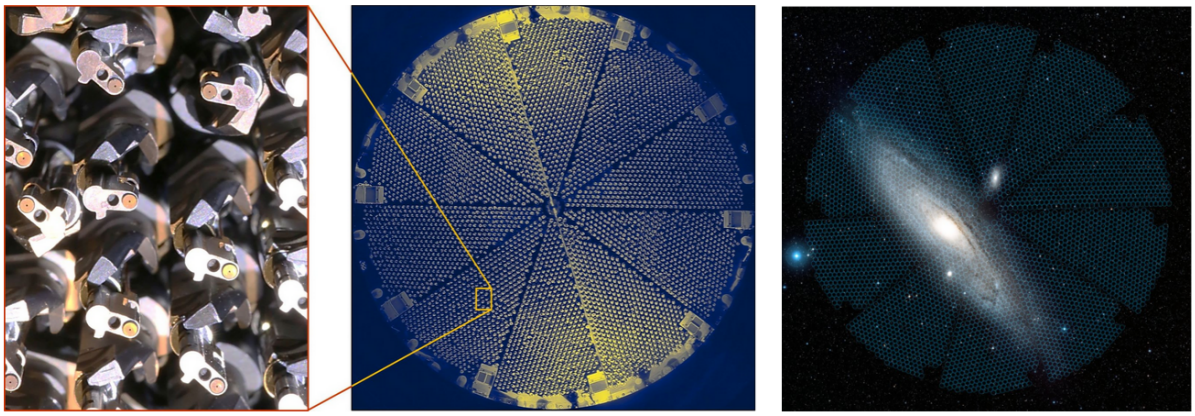

Figure 2.8: *Left: zoom on the fibre positioners in a small part of the focal plane (middle). Right: image of the focal plane on the sky, with the image of the galaxy Andromeda superimposed. The focal plane covers* ~ 8 *deg*^{2} *on the sky. Credit: DESI collaboration.*

spectrograph is equipped with two high-transmission dichroics (efficiency $> 95\%$) which divide the light into three wavelength channels: Blue (B), Red (R) and Near Infrared (Z) (see Table 2.1). This split optimises throughput, increases spectral coverage and gives each channel its own spectral resolution. The spectral resolution of the instrument is high enough to resolve the [OII] doublet of ELGs. Finally, the light is collected by CCD (Charge-coupled Device) sensors of 4096×4096 pixels. Each CCD is mounted in a vacuum cryostat ($< 3 \times 10^{-7}$ mBar) maintained at low temperature by a closed-cycle pulse tube cryocooler. The blue CCDs are at $\sim$163 K and the others at $\sim$140 K. The cryogenic machines ensure precise (± 1 K) and stable (± 0.1 K) temperature control, and were designed by our team at CEA Saclay. The team was also responsible of the cryostat mechanical mounting, CCD integration in the cryostats and CCD alignment w.r.t. the last optical lens of the spectrographs to achieve a parallelism within $\pm 15\mu$ with respect to the focal plane.

Channel	Spectral range (Å)	Spectral resolution
Blue (B)	$3600 - 5930$	$2000 - 3000$
Red (R)	$5600 - 7720$	$3500 - 4500$
Near Infrared (Z)	$7470 - 9800$	$4000 - 5500$

Table 2.1: *spectral range and resolution for each channel of the ten spectrographs of DESI (Abareshi et al., 2022).*

Another very important part of DESI is the *prime-focus corrector* (PFC). The corrector converts the light from the primary mirror and transfer it onto the focal surface of 0.8 m in diameter. The light collected from astronomical objects has to be focused into 107 μ diameter fibres on the focal plane. This requires high image quality with very little blurring (width of the optical point spread function (PSF)) to reduce the loss of light due to rays missing the core of the fibres. The wide-field corrector assembly comprises six lenses, the largest of which has a diameter of 1.1 m and the heaviest weighs 237 kg. The total mass of the lenses is 864 kg. The lenses are coated with a broadband anti-reflective coating that gives an average transmission $\geq 99.0\%$ over the wide passband (360-980 nm). Optical aberrations are corrected by a lens

assembly that delivers excellent images over the 3.2° field of view and wide passband (360-980 nm). Two lenses are dedicated to correcting atmospheric dispersion over a range of zenith angles from 0 to 60 degrees. These two lenses can be rotated independently to counter the effect of wavelength-dependent atmospheric dispersion, depending on the direction of observation of the telescope.

DESI is a complex instrument designed to record 5000 spectra per exposure, with high throughput and high-quality imaging to maintain excellent operational efficiency. To date, DESI is the largest multi-object spectrograph constructed and will measure ∼40 million spectra of galaxies and quasars over five years to probe dark energy and cosmological parametres sufficiently well to become the first completed Stage IV cosmological survey.

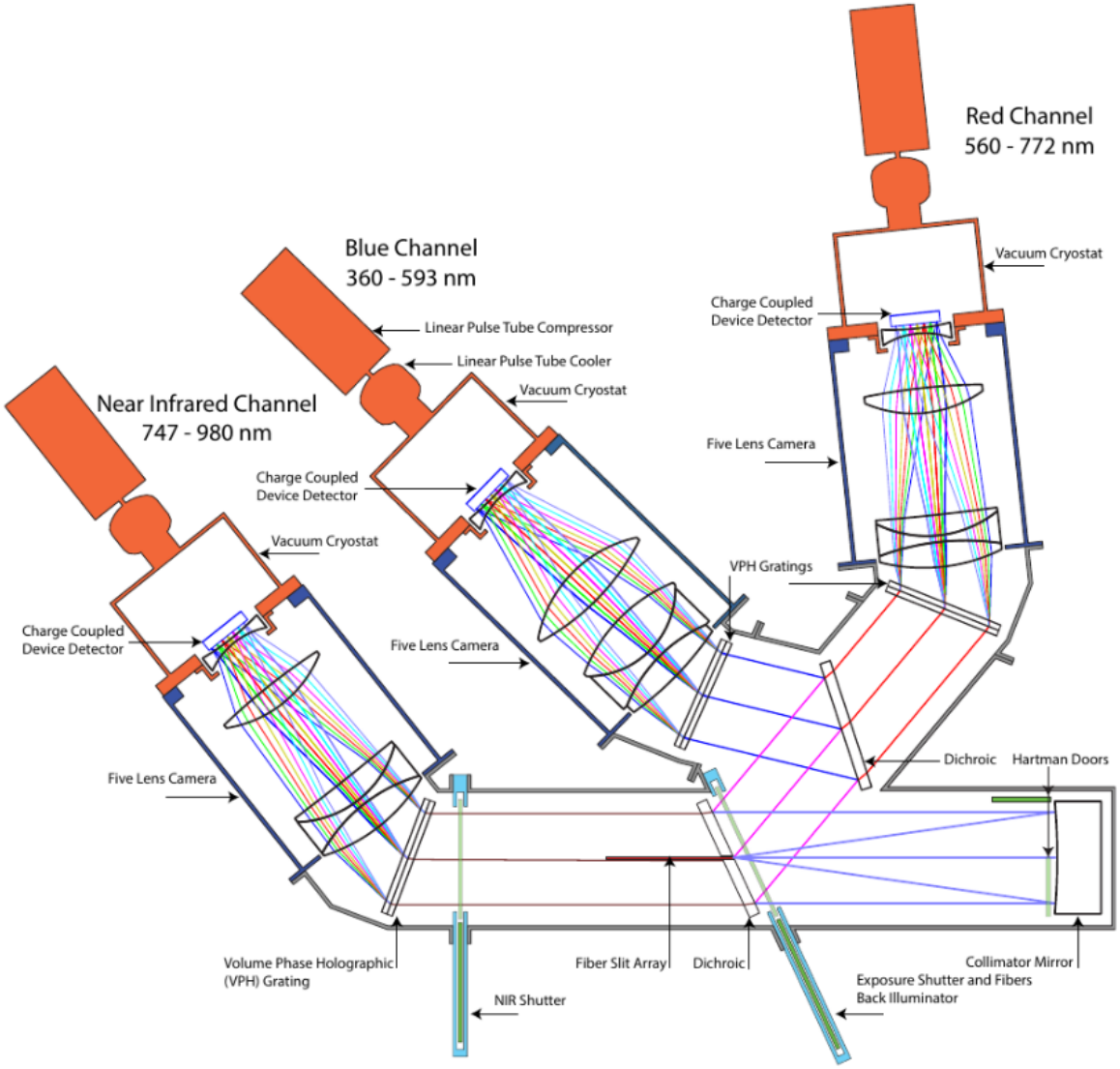

Figure 2.9: *Schematic representation of one spectrograph of the DESI instrument. This figure is taken from (Collaboration et al., 2016).*

2.4 Target selection

2.4.1 Photometric surveys

The first step of any spectroscopic surveys is to perform a target selection (TS). The selection of targets must be carried out before the start of spectroscopic operations and is based on a photometric survey.

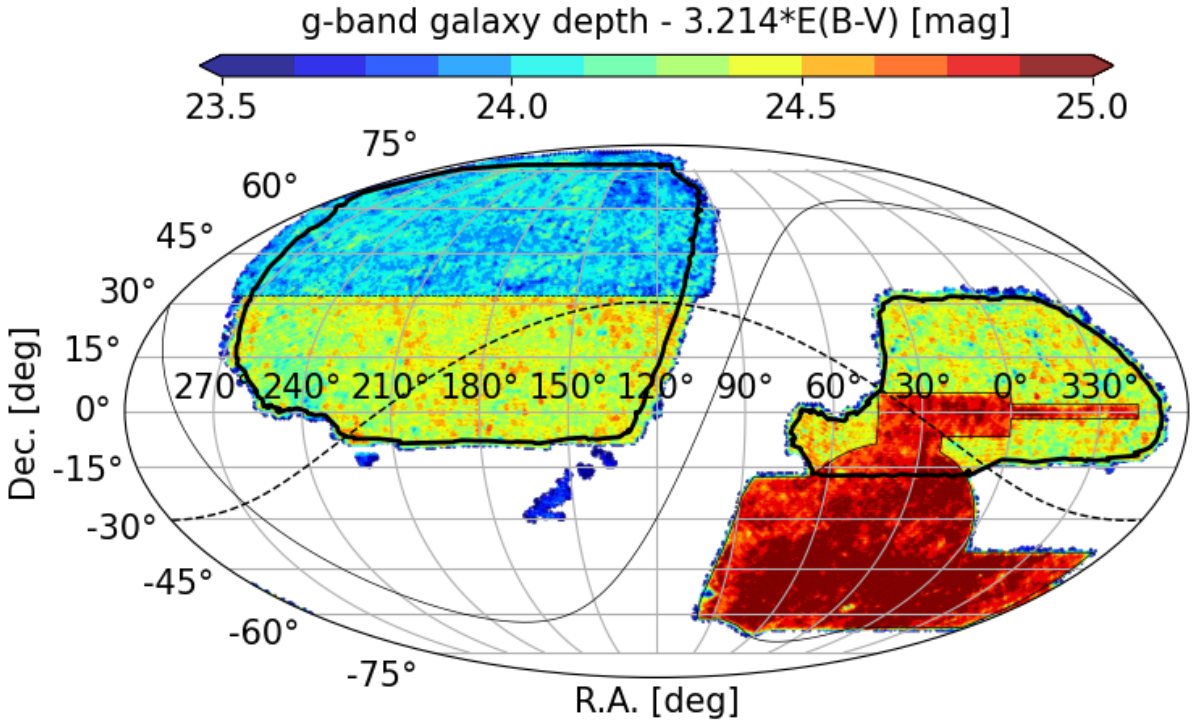

Figure 2.10: *Sky maps of the imaging g-band depth corrected from Galactic extinction (E(B-V)) which have been used to select the DESI ELG targets. The blue region corresponds to the BASS/MzlS photometric survey, the green region to DECaLS and the red region to DES. The thick black line represents the 14,000 deg^2 footprint covered by DESI. The Galactic plane is displayed as a solid black line and the Sagittarius plane is displayed as a dashed black line. This figure is taken from Raichoor et al. (2023).*

DESI targets have been selected using the Data Release 9 (DR9) of the Legacy Imaging Surveys programme. This survey covers $\sim 19,700$ deg^2 of the sky visible from the Northern hemisphere (see Figure 2.10), in three optical bands, g (centered at 470 nm), r (centered at 623 nm), and z (centered at 913 nm), covering the 14,000 deg^2 of the DESI footprint. A full description of the Legacy Imaging Surveys is available in Dey et al. (2019) and a sky viewer of the survey in available here. The optical bands were collected by different independent surveys:

➢ **BASS/MzLS:** using g and r band optical filters, the Beijing-Arizona Sky Survey (BASS) observed ∼5100 deg^2 of the North Galactic Cap (NGC) for a declination $dec > 32.375°$. The BASS survey was conducted on the Bok 2.3-m telescope and lasted four years, from 2015 to 2018 (Zou et al., 2017). The Mayall z-band Legacy Survey (MzLS) observed the same footprint as BASS in the z-band, using the Mayall telescope. This survey was conducted over 230 nights between 2016 and 2017. The region covered by these surveys is shown in blue in Figure 2.10.

➢ **DES/DECaLS:** The Dark Energy Survey (DES) observed 5000 deg^2 in the south galactic cap (SGC) with 8 optical bands and was originally designed for weak lensing studies using the Dark Energy Camera (DECam) on the 4-m Blanco telescope located at the Cerro Tololo Inter-American Observatory in Chile (Abbott et al., 2018, Flaugher et al., 2015). The Dark Energy Camera Legacy Survey (DECaLS) expanded the DES footprint in the three optical bands g, r, z, by observing $\sim$ 4000 deg^2 in the SGC and $\sim$ 5000 deg^2 in the lower region of the NGC ($dec < 32.375°$), without overlapping the DES footprint (Dey et al., 2019). The DES imaging is significantly deeper because it is covered by a greater number of exposures than the standard DECaLS observations. The DeCALS and DES regions in the sky are represented in green and red in Figure 2.10.

The DESI footprint is shown in black on Figure 2.10 and the TS used the photometry in the three optical bands g, r, z from BASS/MzLS, DECaLS and the DES region included in DECaLS ($dec > -20°$). In addition to the DESI Legacy Imaging Surveys other external sample were used for the TS of DESI objects:

➢ **WISE:** : The Wide-field Infrared Survey Explorer (WISE) satellite has provided all-sky infrared observations in four bands W1, W2, W3 and W4 centered at 3.4, 4.6, 12, and 22 μm (spanning wavelenghts between $\sim$ 1 to 30 μm) (Cutri & et al., 2012, Wright et al., 2010). DESI target selection uses the two shortest-wavelength bands W1 and W2 in the TS of LRGs and QSOs (Chaussidon et al., 2023, Zhou et al., 2023).

➢ **Gaia** : The Gaia satellite observed positions and proper motions of stars in the Milky Way (Gaia Collaboration et al., 2016). DESI uses the observations in the G band ($330 - 1050\mu$m) and the star catalogues provided by the Gaia DR2 release, which provides observations for 1.7 billion stars over the whole sky during 22 months of observation (Carrasco et al., 2016, Gaia Collaboration et al., 2018). It was used for the TS of the BGS sample (Hahn et al., 2023).

Using the photometric data described above, the TS can be performed by looking at the colours of the objects, which are magnitude differences between two photometric bands. Basically, each type of galaxy is different and exhibit different properties (colours). It is therefore possible to separate them from other objects in the sky by using different colour selections, the main contaminant of high redshift targets being the stars. The efficiency of the TS of DESI galaxies and QSOs was tested during the phase of survey validation (SV). The target selection of DESI was performed using a common pipeline described in Myers et al. (2023).

2.4.2 DESI targets

In this section we describe the selection of high redshift DESI targets. A summary of the expected target densities is given in Table 2.2.

➢ **BGS** In DESI, the bright galaxy sample (BGS) is a flux-limited and r-band selected sample of galaxies. Full details of the BGS selection procedure are described in Hahn et al. (2023). BGS targets two different samples: BGS Bright, a magnitude-limited sample $r < 19.5$ and BGS Faint, a fainter sample $19.5 < r < 20.175$ using colour selections to have high redshift efficiency. To

discriminate between stars and galaxies in the main BGS programme the TS procedure compares the G-band magnitude from Gaia with that in the r-band from the legacy survey ($G - r > 0.6$) as illustrated in Figure 2.11. It also examines the overlap of potential targets in the GAIA DR2 star catalogue. In total, BGS Bright has 864 targets/deg^2, BGS Faint has 533 targets/deg^2. Using the spectra from the SV, the BGS TS achieves a redshift success rate of $> 95\%$ (i.e. $> 95\%$ of BGS targets are spectroscopically confirmed to be a galaxy) for both the BGS Bright and Faint samples.

➢ **LRGs** Luminous Red Galaxies (LRGs) are massive, 'old' galaxies that have stopped forming stars and have a typical red spectral energy distribution. They have prominent characteristic "bump" in their spectra at 1.6 μm (rest frame) (John, 1988, Sawicki, 2002) which can be used to efficiently remove stars from the sample by looking at the optical/near-infrared (NIR) colour (see Figure 2.11). The DESI LRG sample is selected using a combination of the Legacy Survey g, r, z and WISE W1 bands. In addition, the LRG TS is optimised to select the most massive galaxies (in terms of stellar mass) with a high completeness defined as the ratio of selected LRGs to the expected total number of objects brighter than the LRG magnitude limit $z < 21.6$. The high stellar-mass completeness of the LRG sample allows a wide range of studies, including galaxy-galaxy lensing (e.g. Jullo et al. (2019)), galaxy-halo connection (e.g. Rodríguez-Torres et al. (2017)) and evolution of the most massive galaxies (e.g. Bundy et al. (2017)). In the end, the DESI LRG sample has a target density of 605/deg^2 with a redshift efficiency of 89.4% between $0.4 < z < 1.1$ and a high completeness for the most massive galaxies ($M_\star > 11.5$ [$M_\odot$]) in the range $0.4 < z < 1$. Full details of the TS for the DESI LRG sample are described in Zhou et al. (2023).

➢ **QSOs** Quasi-stellar objects or quasars are a type of *active galactic nuclei* (AGN). An AGN is the central nucleus of an active galaxy capable of producing long jets of gas (up to ~ 100 kpc). The luminosity emitted by the nucleus of a quasar exceeds the luminosity of the host galaxy. They are the brightest visible objects in the universe and apear as point sources in the sky (like stars). The spectrum of most quasars exhibits strong continuum emission in the visible, X-ray, and γ-ray regions with broad emission lines. Their high luminosity means that they can be detected at high redshifts. The photometric characteristics of QSOs mimic those of faint blue stars in optical wavelengths, making them difficult to select. As QSOs have a point-like morphology, the selection of objects was restricted to those with stellar morphology in the legacy survey and with a magnitude in the r-band such as $16.5 < r < 23.0$. Then, to discriminate QSOs from stars, the selection uses a random forest classifier based on a colour selection that combines optical-only and optical+IR colours, as shown in Figure 2.11. Details of the DESI QSO selection are described in Chaussidon et al. (2023). In the end, the density of selected QSO targets is ~ 310/deg^2. Using the spectra collected during SV, the main quasar selection has over 200 quasars/deg^2 confirmed spectroscopically, including 60 quasars/deg^2 with $z > 2.1$ that can be used for the Ly-α forest analysis.

➢ **ELGs** Emission Line Galaxies (ELGs) are the main tracer of the DESI survey and are the subject of this PhD work. Emission lines in galaxy spectra are correlated with star formation (Favole et al., 2023, Moustakas et al., 2006). ELGs are therefore typically star-forming, 'young' galaxies. Any galaxy actively forming stars at a sufficiently high rate will be considered as an

ELG. In DESI, the TS of ELGs is optimised to measure the [OII] doublet ($\lambda\lambda$ 3726,3729 Å) which provides an unambiguous signature and accurate redshift determination. ELGs are described in more detail in the next chapter.

Due to their vigorous star formation, they have a relatively blue continuum that allows ELG targets to be selected from optical photometry in the g, r, z bands. The selection of ELGs is described in details in Raichoor et al. (2023) and we present the main ideas in what follows. The main ELG selection consists of a magnitude cut in the g-band and a colour cut in $(g-r)$ relative to $(r-z)$, as illustrated in Figure 2.11. The main ELG sample is composed of two disjoint subsamples, the `ELG_LOP` and the `ELG_VLOP`. They have target densities of $\sim$ 1940 targets/deg^2 and $\sim$ 460 targets/deg^2, respectively. The TS of the `ELG_LOP` sample favours ELG targets in the redshift interval $1.1 < z < 1.6$, while the TS of the `ELG_VLOP` sample favours ELG targets between $0.6 < z < 1.1$. `LOP` and `VLOP` stands for low priority and very low priority in the fibre assignment process (see Section 2.7.1.1). After spectroscopy, the `ELG_LOP` and `ELG_VLOP` samples have $\sim$ 860 targets/deg^2 and $\sim$ 180 targets/deg^2 respectively, between redshift $0.6 < z < 1.6$ (see Table 2.2). Overall, $\sim$18.7M `ELG_LOP` ($\sim$2.7M `ELG_VLOP`) targets should be spectroscopically observed by DESI, and 12.M (2.4M) should provide a reliable redshift in the range $0.6 < z < 1.6$.

Object class	Targets /deg^2	Expected reliable redshifts /deg^2
BGS Bright	864	$\sim$ 820
BGS Faint	533	$\sim$ 506
LRG	615	$\sim$ 540
ELG_LOP	1940	$\sim$400 ($0.6 < z < 1.1$) & $\sim$460 ($1.1 < z < 1.6$)
ELG_VLOP	460	130 ($0.6 < z < 1.1$) & 50 ($1.1 < z < 1.6$)
QSO	310	$\sim$ 200

Table 2.2: *Number density of targets and expected number of reliable spectroscopic redshifts per square degree for each DESI target class (except for Milky Way stars).*

2.5 Observation strategy

The DESI observational strategy is described in (Schlafly et al., 2023). We will only describe the main ideas in the following. DESI has three observational programmes:

- **The dark programme** is the primary programme of DESI and will observed LRGs, ELGs, QSOs from $0.4 < z < 3.5$. This programme is observed whenever conditions are good (seeing, sky background, transparency, airmass).
- **The bright programme** aims to observed bright galaxies and Milky Way stars. This programme is observed when observing conditions are not good enough to conduct the dark programme, due to bright sky or poor seeing or transparency.
- **The backup programme** consists of observing brighter Milky Way stars. This programme is observed only when observational conditions are too poor to observe the bright programme.

The combination of the dark programme and the bright programme are called the *main survey*. The dark programme represents $\sim$ 90% of the effective observing time. This approach allows the

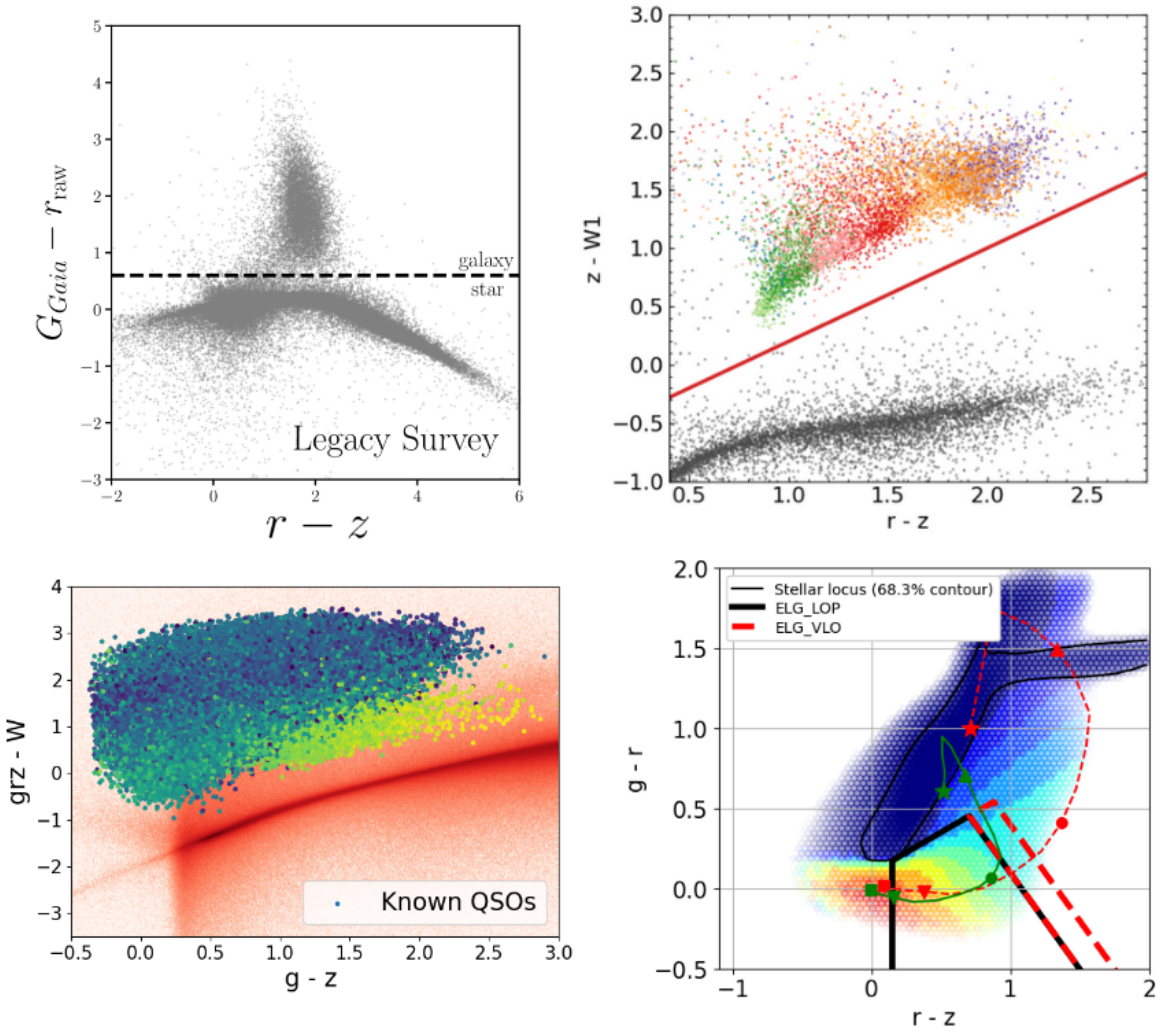

Figure 2.11: *Colour-colour diagram with the selection cuts uses to select DESI targets for BGS (top left), LRGs (top right), QSOs(bottom left) and ELG (bottom right). These figures are from the target selection papers Chaussidon et al. (2023), Hahn et al. (2023), Raichoor et al. (2023), Zhou et al. (2023).*

brightest targets to be observed in the worst conditions to limit systematic uncertainties. These programmes consist of observing a certain number of *tiles* across the footprint. A *tile* is a given telescope pointing combined with assignments of each fibre to a specific target for that telescope pointing. Each tile is associated with a single programme and each of these programmes have independent target lists. The main survey requires 7 passes for the dark programme, i.e. 7 tiles at the same sky location, with a slight offset (see Figure 2.17) to enable greater fibre assignment completeness for dense targets such as ELGs (so 9,929 tiles in total). The bright programme only requires 4 passes (2,657 tiles in total). The time for each exposure varies according to the observational conditions (seeing, sky background, transparency, airmass...), and is calculated using the Exposure Time Calculator (ETC), which is compared with an *effective time*. In DESI the observation must achieve a given pre accuracy or goal uncertainty when measuring the fluxes of distant galaxies. The *effective time* is the time required to reach the goal uncertainty for *nominal observing conditions*, defined as a seeing of 1.1", a sky background of 21.07 mag per square arcsecond in the r-band, photometric conditions, observations at zenith, through

zero Galactic dust reddening. The effective time per exposure for the dark programme is 1000 seconds, while with the bright programme it is 180 seconds (Schlafly et al., 2023). Each night, DESI observes around twenty tiles containing a total of ∼100,000 sources. The DESI survey strategy operates in a *depth-first* manner, meaning that it completes the observation tiles in a particular region first, rather than observing tiles in other parts of the sky. This allows many scientific programmes to proceed after the first year even if the sky coverage is lower. The other advantage is to minimise the negative impact of falling behind schedule, DESI "would prefer to end the survey with a complete 13,000 sq. deg. survey than an inhomogeneous 14,000 sq. deg. survey." (Schlafly et al., 2023). It also favours the observing of Lyman-α tracer (quasar at $z > 2.1$), which need to be identified in the sky from their initial observations, so that these quasars can be targeted for repeated observations (at least four times).

Before each night, the observation plans are defined to select the fields to be observed during the night. Then, during the night, the targets are assigned to each positioner on the fly immediately before the start of the exposure and the ETC determines the time needed to complete the observation. At the end of the night, the spectroscopic pipeline reduces, classifies, and measures redshifts for all targets (Guy et al., 2023), and visual (human) quality assurance is performed to see if any problems occurs overnight for each tile. For those tiles that have passed the quality assurance the Merged Target Ledger (MTL) is updated, updating the observation state and redshift of the observed targets.

2.5.1 Target priorities

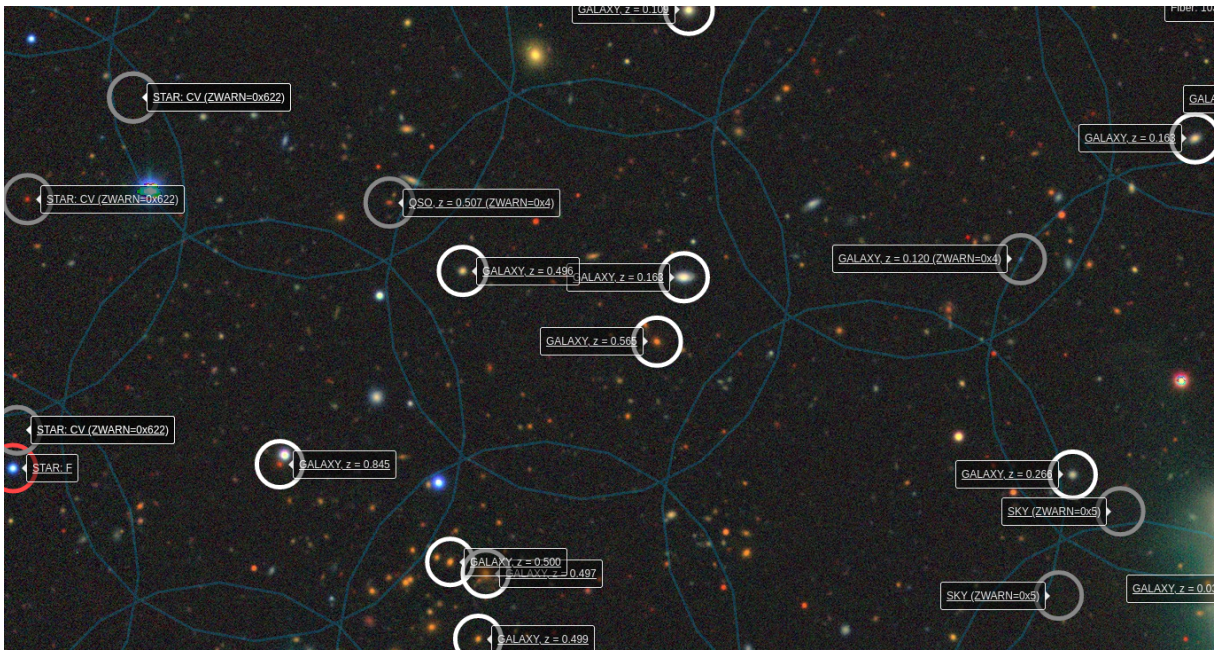

Figure 2.12: *Representation of the patrol radius of the fibres over a small region of the sky with the targets observed by DESI. The patrol region overlaps between the fibres. In this example, the telescope must observe this region at least 3 times to obtain the redshift of 3 galaxies in the middle of the image. This image comes from the Legacy Survey Sky Viewer.*

One of the main observational effects in DESI is the fibre assignment. The fibres have patrol regions 12 mm in diameter on the focal plane that correspond to a physical region on the sky of ~ 180 arcsec. If several targets are located within the same patrol region on the sky, only one is accessible by each fibre and other observations are required to reach all the targets, as illustrated in Figure 2.12. Consequently, priorities for fibre assignment are given to targets depending on the tracer. For dark time tracers, the highest priority is given to QSOs, then LRGs and finally to ELGs. As previously mentioned in Section 2.4.2, ELGs have two disjoint sub-samples, `ELG_LOP` and `ELG_VLOP` with the `ELG_LOP` sample having a higher priority than `VLOP`. In any case, ELGs will always be observed after all other tracers. The fibre assignment leads to incompleteness between the number of targets and the observed targets depending on the number of passes, as illustrated in Figure 2.13. The completeness is computed as the ratio between the observed targets $N_{\rm obs}$, and the initial targets, $N_{\rm targets}$ in a given region of the sky:

$$\text{Completeness} = \frac{N_{\rm obs}}{N_{\rm targets}} \tag{2.1}$$

ELGs are the most impacted by fibre assignment as they have the lowest priority, e.g. for $N_{\rm tile} = 3$, the completeness of QSO is $> 99\%$ compared to ~ 0.53 for ELGs. Missing objects due to fibre assignment can bias the galaxy clustering measurement and must be carefully corrected. We discuss on the effects that can bias this measurement, and how to mitigate them in Section 2.7.1.

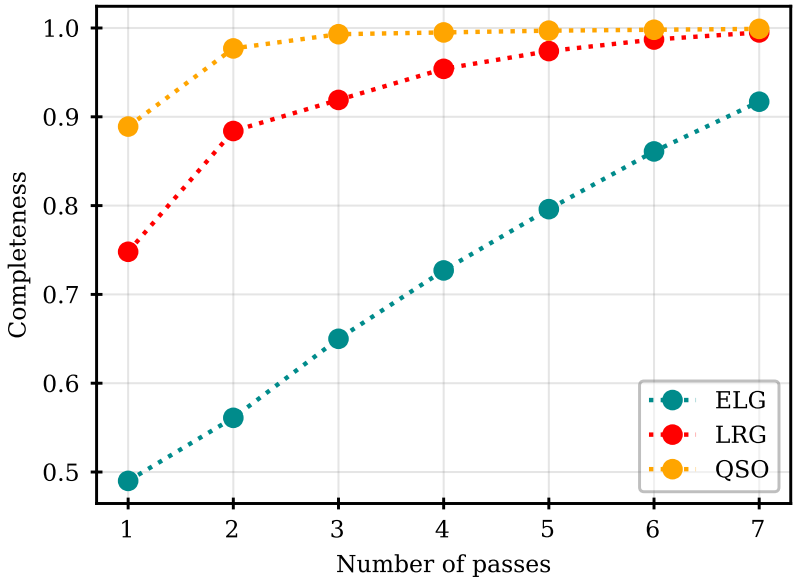

Figure 2.13: *Completeness of different tracers as a function of the number of tiles in the same region.*

2.5.2 Spectral classification and redshift determination

Once the observation has been made, the spectroscopic pipeline reduces, classifies, and measures redshifts for all targets the following morning. All the details of the spectroscopic pipeline are in

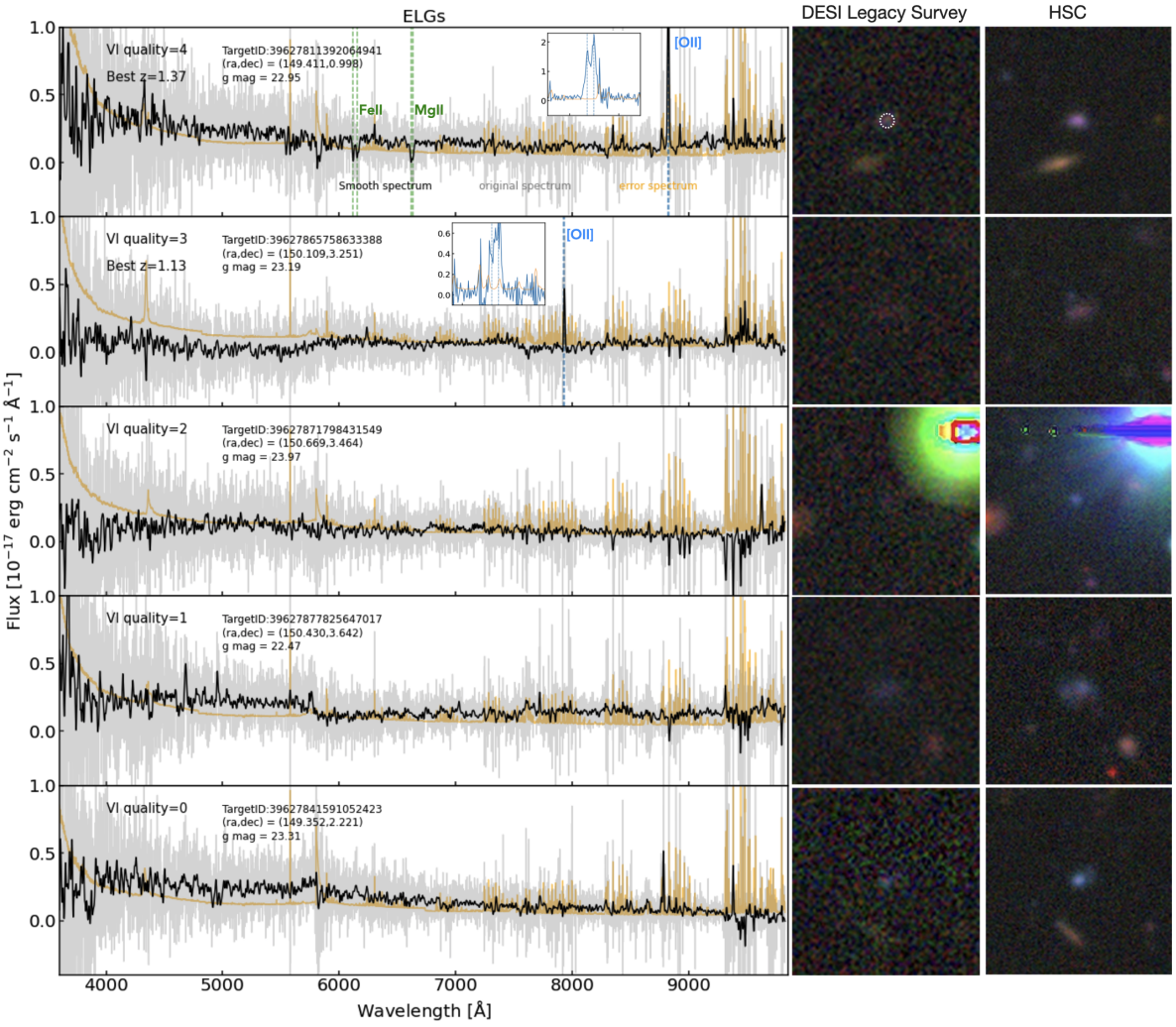

Figure 2.14: *Example of ELG spectra, ordered by their VI quality values from top with VI quality 4 to bottom with VI quality 0. The quality 4 spectrum shows a resolved [OII] doublet. The right two panels show the DESI Legacy Survey images (Dey et al., 2019) and the Hyper Supreme-Cam images (Aihara et al., 2019) of the ELG targets. The spectra in grey, black and orange colours are the original observed galaxy spectrum, the smooth spectrum with a median filter, and the error spectrum, respectively. Figure taken from Lan et al. (2022).*

this paper (Guy et al., 2023). Each spectrum is classified by a software called *Redrock*[1] (Bailey & DESI Collaboration, 2023). It is based on a template fitting method, i.e. χ^2 minimisation computed from a linear combination of spectral templates over the set of training templates. For each class of target, (stars, galaxy or quasar), templates are constructed from previous observations of each class of objects. The key parameters describing the best fit are the redshift, the redshift uncertainty, the spectral class (star, galaxy, or quasar), the coefficients to the spectral templates, the χ^2, and the value $\Delta\chi^2$, which is the difference in χ^2 between the best fitted redshift and the second best fit (the secondary minima in the χ^2 value). This value reflects the probability that the best-fit redshift is correct. Each spectrum is therefore associated with a redshift and a spectral type which is not necessarily the target type of the object (for example, a QSO target may be a star). In order to evaluate the performance of Redrock, *visual inspections* (VI) of galaxy and quasar spectra have been done during the SV (Alexander et al., 2023, Lan et al., 2022). I participated in the VI campaign for ELGs using the `prospect` software[1] to read the spectra together with the files generated by Redrock and facilitates the evaluation of the quality of the best fit. The key step in the VI procedure is to assess the quality of the VI redshift, according to different quality criteria:

- Quality 0: no signal, useless spectrum.
- Quality 1: unidentified feature in the spectrum, unlikely classification.
- Quality 2: possible classification, one strong spectral feature but unsure what it is.
- Quality 3: probable classification with at least one secure spectral feature, the redshift is likely to be correct.
- Quality 4: confident classification with two or more secure features in the spectra.

Each spectrum is checked by at least two inspectors whose quality assignments are averaged. The final VI redshift is robust if the overall VI quality ≥ 2.5, whereas an overall VI quality < 2.5 typically indicate a bad spectrum. Examples of ELG spectra of different quality are shown in Figure 2.14. In DESI, the target feature in the ELG spectra is the [OII] doublet $\lambda\lambda$3726, 3729 as shown in the top spectra of Figure 2.14. Among all the visually inspected ELG spectra (10315), $\sim 75\%$ have VI quality ≥ 2.5. In this sample of robust VI redshifts, the redshift recovery rates (in percent) by Redrock are 93.8 ± 0.2 and 97.5 ± 0.3 for `ELG_LOP` and `ELG_VLOP` respectively.

Most of the spectra obtained during the main survey will not have a VI counterpart, so we need to ensure that Redrock finds the right spectral type and redshift. Therefore, to increase the success rate from Redrock, an additional selection criterion is added to include most objects with redshifts successfully identified by Redrock and exclude most objects with incorrect redshifts. This criterion is a combination of the Redrock $\Delta\chi^2$ values and other parameters from DESI spectra, which is different for each tracer. To quantify the reliability of the best-fit redshift, the Redrock parameter $\Delta\chi^2$ is a good indicator, a large $\Delta\chi^2$ implying generally a reliable redshift measurement. However, ELGs spectra have in general a low SNR, so that the correct redshift could have a low $\Delta\chi^2$ because another solution due to a single emission line at a different redshift would still provide a comparable χ^2. To avoid ruling out a large proportion of good redshifts,

[1] https://github.com/desihub/redrock/releases/tag/0.15.4
[1] https://github.com/desihub

Raichoor et al. (2023) used a selection criterion that includes both $\Delta\chi^2$ and the signal-to-noise ratio (SNR) of the [OII] emission flux SNR([OII]), defined as follows:

$$\log_{10}(\mathrm{SNR}([\mathrm{OII}])) > 0.9 - 0.2 \times \log_{10}(\Delta\chi^2) \tag{2.2}$$

This criterion selects more than 95% of reliable redshifts (VI validated and recovered by Redrock Raichoor et al. (2023)), and corresponds to a redshift purity of 99.6% for ELGs at all redshifts Lan et al. (2022). Figure 2.15 shows the fraction of validated redshifts in the plane $\log_{10}(\Delta\chi^2)$-$\log_{10}(\mathrm{SNR}([\mathrm{OII}]))$ with the ELG criterion over all visually inspected spectra. Similarly to ELGs, other criteria are applied to other tracers, leading to a redshift purity $> 99\%$. During SV, the precision of the redshift measurements was tested using repeat observations. With independent exposures, spectra of the same targets were compared to determine the redshift errors from the Redrock pipeline. For BGS and ELGs, the redshift precision is ~ 10 km/s and that of LRGs is ~ 40 km/s and the redshift accuracy (compared to DEEP2 redshift measurements (Newman et al., 2013)) is around $\sim 6, -3, -1$ km/s for BGS, LRGs and ELGs, respectively (see Table 3 of Lan et al. (2022)). In the following, we consider the ELG sample which includes both `ELG_LOP` and `ELG_VLOP` samples.

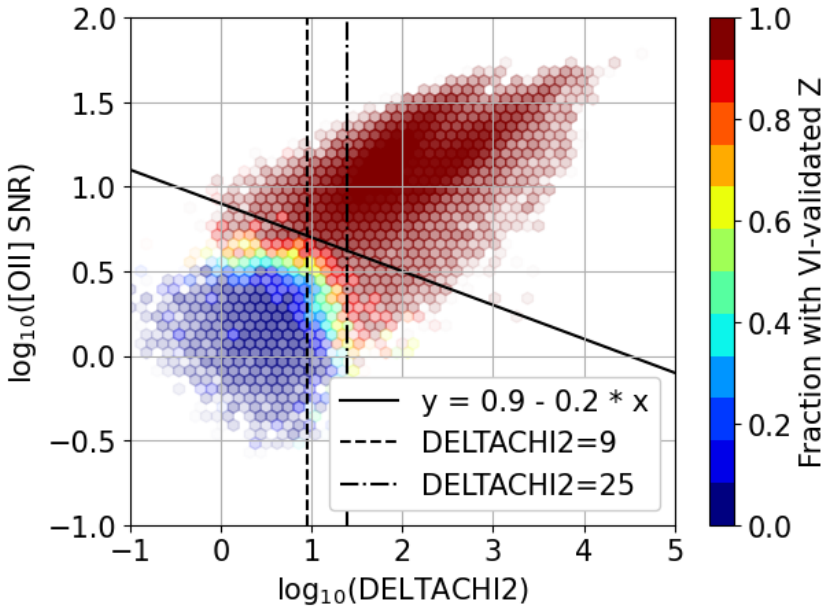

Figure 2.15: *Fraction of ELG redshifts validated by VI (VI quality $>$ 2.5) in the* $\log_{10}(\Delta\chi^2)$-$\log_{10}(\mathrm{SNR}([OII]))$ *plane. The slanted solid line is our criterion for selecting reliable redshift measurements, while the dashed and dotted–dashed vertical lines illustrate two threshold values for a lower cut in* $\log_{10}(\Delta\chi^2)$. *Figure taken from Raichoor et al. (2023).*

2.6 The DESI One-Percent survey

The DESI One-Percent survey is the third and final phase of the survey validation (SV3). It was conducted over 2 months (April and May 2021) prior to the start of the main survey in June 2021. As the name suggests, the One-Percent survey aims to mimic 1% of the main survey, covering $\sim$140 deg^2 with final target selection algorithms and depths similar to those of the

main survey. The footprint consists in twenty non-overlapping regions about the size of a focal plane, called *rosettes*, represented in red on Figure 2.16. Each region undergoes $\sim$ 13 visits to obtain high fibre assignment completeness (which is much higher than the main survey). For each visit, the centres are shifted slightly to increase completeness in regions of the focal plane that have no fibres (centre, petal edges).

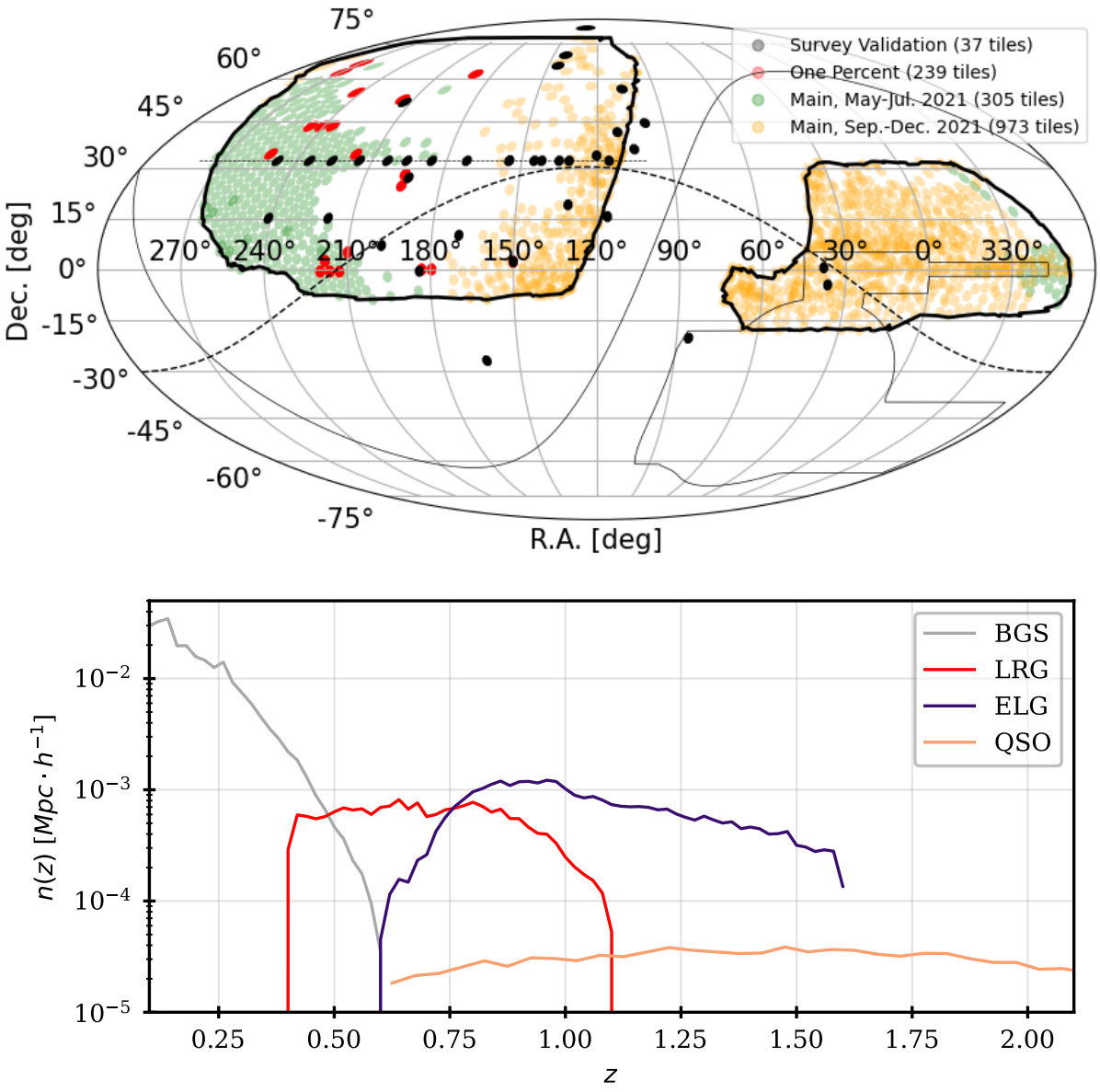

Figure 2.16: *Top: Sky distribution of the DESI-observed dark tiles during the One-Percent in red and previous phase of SV in black. Early coverage of the main survey is represented in green (2021 May–July) and orange (2021 September-December). This figure is taken from (Raichoor et al., 2023). Bottom: Redshift distrution of the BGS (grey), LRGs (red), ELGs (blue) and QSOs (orange) samples of the One-Percent survey.*

On the other hand, the edges of the rosette are less complete due to the reduced number of visits. The completeness of one rosette for ELGs is presented in Figure 2.17. Overall the completeness of ELGs is $\sim$ 86%, and $\sim$ 95% between 0.2 and 1.5 degrees from the centre of each

rosette. Thanks to its high completeness, the One-Percent survey provides precise measurements of the galaxy clustering down to very small scales. This sample is very appropriate to perform small-scale clustering studies, and therefore the study of the galaxy-halo connection. During the SV, DESI observed ~ 1.4M extragalactic redshifts including $\sim 730k$ during the One-Percent survey (2 months of observation) with 253,915 BGS, 312,790 ELGs, 137,317 LRGs and 34,173 QSOs (Collaboration et al., 2023b), and the number density of targets as function redshift is shown in Figure 2.16.

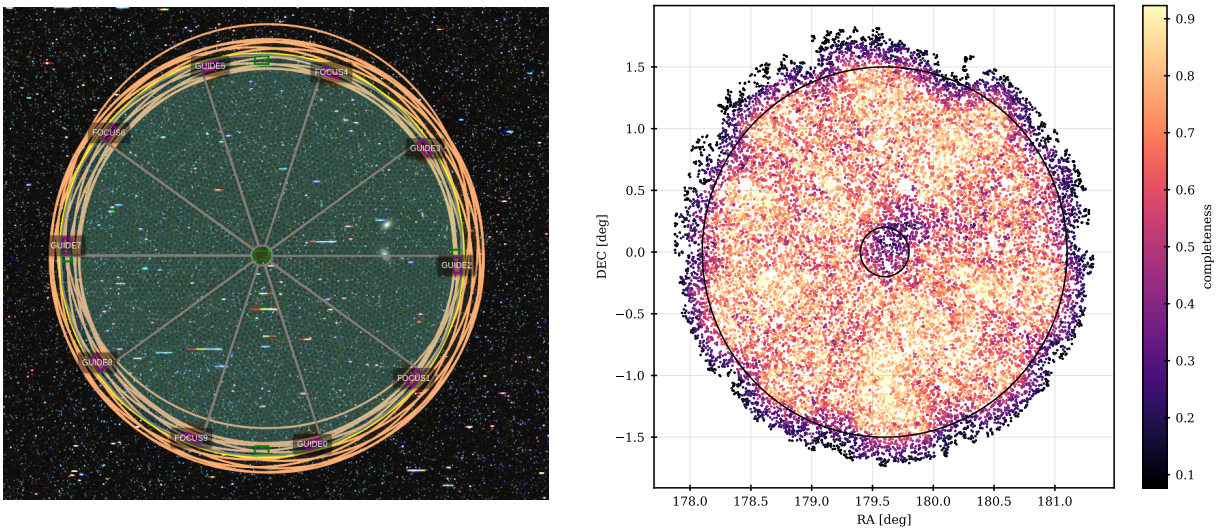

Figure 2.17: *Left: Focal plane on sky with the overlapping tiles used to observed one rosette in orange. Image from the legacy survey skyviewer. Right: The observational completeness of ELG targets on one rosette from the DESI One-Percent Survey. The centre and edge of the rosette has lower completeness due to the lower number of overlapping tiles. The two circle represent the rosette radius* $r = 0.2°$ *and* $r = 1.5°$*, which are the cut we used for our analysis.*

2.7 Estimator of the correlation function

The correlation function, and in particular the two-point correlation function (2PCF), $\xi(r)$, is the main statistic used in LSS analysis. The 2PCF measures the excess of pairs of galaxies separated by a distance r with respect to a random distribution. The estimator that makes the variance of the two-point correlation function nearly Poisson is the Landy-Szalay estimator (Landy & Szalay, 1993):

$$\hat{\xi}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} \tag{2.3}$$

where $DD(r)$, $DR(r)$, $RR(r)$ are the number of galaxy-galaxy, galaxy-random, random-random pairs separated by a distant r. Random refers to a random distribution of galaxies on the same geometry as the data. The advantage of this estimator is that the random distribution takes into account the survey geometry and potential masks, which minimise edge effects. Instead of calculating the 2PCF in separation r, we can decompose the distance r into two components, s and μ, where s is the observed distance between the pairs of galaxies and $\mu = \cos(\theta)$ with

θ being the angle between s and the line of sight (los). We can then expand the correlation function into Legendre polynomials $\mathcal{L}_\ell(\mu)$ to obtain estimates of the multipoles moments:

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1}\xi(s,\mu)\mathcal{L}_\ell(\mu)d\mu \tag{2.4}$$

The multipole moments provide a mechanism for compressing the anisotropy in the correlation function. The monopole $\xi_{\ell=0}$ is the isotropic component of the 2PCF, while the quadrupole $\xi_{\ell=2}$ (and higher even orders) contains information about the anisotropies in the correlation function. According to the cosmological principle, the galaxy distribution should be almost isotropic. However, *peculiar velocities* of galaxies induce anisotropies in the observed distribution of galaxies, known as the redshift space distortion (RSD) effect (see Section 2.8.2) that leads to non-zero even multipoles. As we will see in the following, in linear theory (and hence on large scales) Kaiser (1987) showed that the anisotropies in the 2PCF are proportional to the isotropic component of the 2PCF (in real space) by a factor $\propto \mu^4$ (see Equation (2.16)) which means that there is no contribution higher that μ^4. Thus, the cosmological signal is carried in the quadruopole and the hexadecapole ($\ell = 2, 4$).

To avoid the impact of galaxy peculiar velocities on small scales, the we can use the projected correlation function $w_p(r_p)$. Instead of decomposing the distance r between galaxies into (s, μ) we can decompose its components along and perpendicular to the line-of-sight π and r_p. The projected correlation function is obtained by integrating $\xi(r_p, \pi)$ along the line-of-sight:

$$w_p(r_p) = \int_{\pi_{min}}^{\pi_{max}}\xi(r_p,\pi)d\pi \tag{2.5}$$

The projected correlation function is widely used in galaxy-halo connection studies because it has the advantage of being almost insensitive to the peculiar velocity of galaxies on small scales (Bosch et al., 2013). We show the measurement of the projected correlation function of ELG data from the One-Percent survey in Section 2.9.

2.7.1 Systematics effects

The measurement of galaxy clustering can be bias due to systematic effects. One of the most important keys to obtaining reliable cosmological results is to study and correct these effects. We describe below some of the main effects that affect the clustering of galaxies in spectroscopic surveys. As always in the search for systematic studies, there are known systematic effects that can be corrected, and unknown systematic effects, for which null tests must be performed to try to avoid them.

2.7.1.1 Fibre assignment

As mentioned in Section 2.5.1, due to fibre assignment, some targets are missed inside the patrol region of the fiber, and we need other pointings of the telescope in the same region to observe all targets. In practice, we do not have enough passes to observe all the targets, so we need to correct for this effect in the measurements. The simplest way to correct for this effect is to up-weight galaxies in a given region of the sky according to the number of targets in that

region, i.e. 2 galaxies observed in a region that initially has 4 targets are up-weighted by 2. This correction provides a good recovery of the 2PCF on large scales but can not correct for scales smaller than the size of the patrol region. In DESI these weights are called *completeness weights*.

To recover the missing pair of objects at small scales, we rely on the pairwise-inverse-probability weighting scheme combined with angular correction (PIP+ANG) (Bianchi & Percival, 2017, Mohammad et al., 2020). The PIP weight accounts for incompleteness in the fibre assignment process. They are defined for each galaxy pairs by running a set of multiple realisations of the fibre assignment (FA) algorithm. Indeed, the FA process is only one random realisation of multiple FA, e.g. if a fibre has 2 targets with the same priority the observed target is chosen randomly. For each realisation, the output for a galaxy is 0 (unobserved) or 1 (observed), and it is stored as bitwise weight $w_i^{(b)}$ for each target (list of 0 or 1). Then, the PIP weight is estimated as the number of realisations N_{runs} in which a given pair could have been targeted divided by the number of times it was actually targeted:

$$w_{mn} = \frac{N_{runs}}{\mathrm{popcnt}[w_m^{(b)} \& w_n^{(b)}]} \tag{2.6}$$

where popcnt is the *population count* operator which returns the number of elements other than 0 and & is the logical operation `AND`. We can also compute the individual-inverse-probability (IIP) weights for individual targets, simply by replacing $m = n$ the equation above. The IIP weights are equivalent to the completeness weights described above. The PIP weighting scheme is only unbiased if there are no pairs with zero selection probability. However, for galaxy pairs within the same patrol region, regardless of the number of realisation, these pairs are never observed. To recover these missing pairs, we can use the angular up-weighting scheme proposed in Percival & Bianchi (2017) (originally used in the 2dFGRS analysis from Hawkins et al. (2003)). This up-weighting scheme is a (good) approximation compared to PIPs that are exact. The pairs DD and DR at a given separation angle θ are up-weighted according to:

$$\begin{aligned} w_{\mathrm{ang}}^{DD}(\theta) &= \frac{DD^{\mathrm{par}}(\theta)}{DD_{\mathrm{PIP}}^{\mathrm{fib}}(\theta)}, \\ w_{\mathrm{ang}}^{DR}(\theta) &= \frac{DR^{\mathrm{par}}(\theta)}{DR_{\mathrm{IIP}}^{\mathrm{fib}}(\theta)}. \end{aligned} \tag{2.7}$$

The superscripts par refers to the pair of targets in the reference parent sample (initial target sample) and fib to pair of targets that receive fibres. From this weighting scheme the corrected DD and DR pair counts are calculated by summing the weights of the galaxy pairs in the separation bin $\mathbf{s}$:

$$\begin{aligned} DD(\mathbf{s}) &= \sum_{\mathbf{s}=\mathbf{s}_m-\mathbf{s}_n} w_{mn} \times w'_{m,tot} w'_{n,tot} \times w_{\mathrm{ang}}^{DD}(\theta), \\ DR(\mathbf{s}) &= \sum_{\mathbf{s}=\mathbf{s}_m-\mathbf{s}_n} w_m \times w'_{m,tot} w_{n,tot} \times w_{\mathrm{ang}}^{DR}(\theta), \end{aligned} \tag{2.8}$$

where w_{mn}, w_m are the PIP and IPP weights, $w_{i,tot}$ are the other weights for each individual objects and the superscript $'$ refers to the weights for the randoms. PIP+ANG can successfully recover the 2PCF measurement at small scales as demonstrated using DESI BGS simulations (Smith et al., 2019). We rely on this weighting scheme in our analysis.

2.7.1.2 Imaging systematics

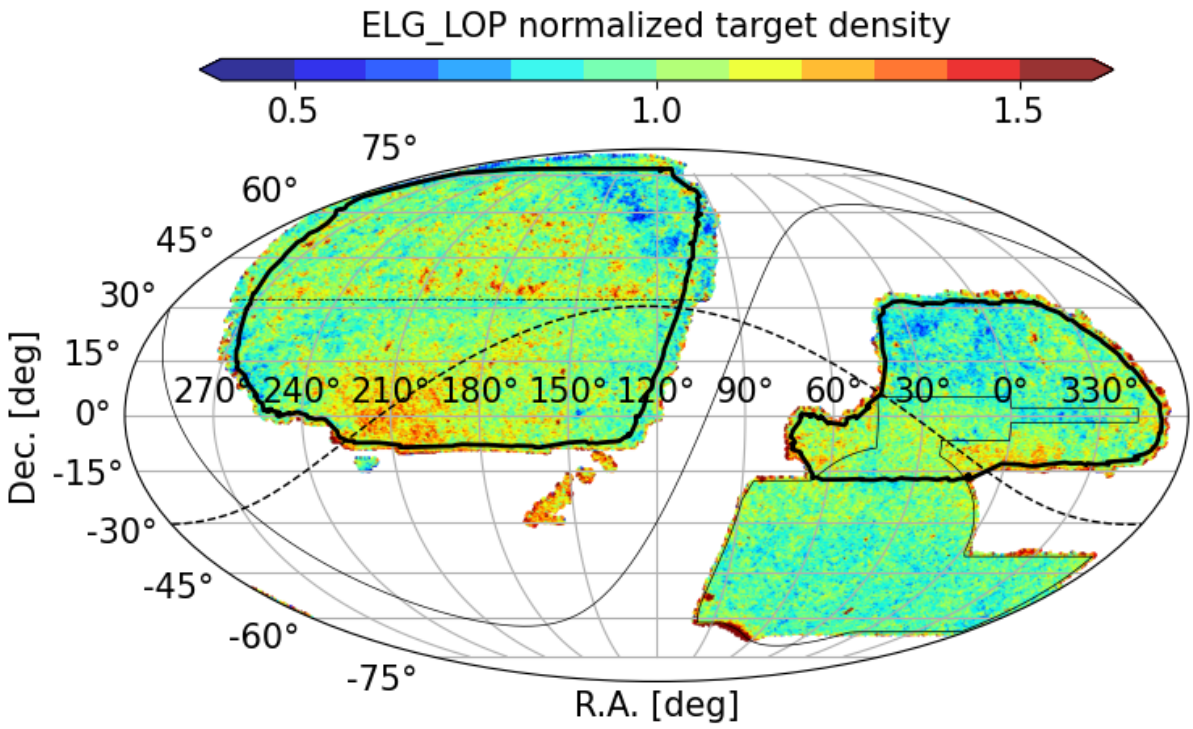

Figure 2.18: *ELG_LOP sample density sky map. The density is divided by the overall average value (1940 /deg^2) to display the fractional difference to the average. The thick black line represents the 14,000 deg^2 footprint covered by DESI. The Galactic plane is displayed as a solid line, while the Sagittarius plane is displayed as a dashed line. Figure taken from Raichoor et al. (2023).*

The target density field is expected to be uniform over the sky. However, the selection can be biased due to the quality, the depth of the photometry and more generally all photometric features (e.g. galactic extinction E(B-V)) used to performed the selection. In addtion, contaminants (e.g. stars) can introduce spurious fluctuations in the target density field and bias the galaxy clustering. In DESI, we can see from Figure 2.18 that the density of ELG targets is not uniform across the DESI footprint, certainly due to stellar contamination. We therefore need to correct for fluctuations in the target density field due to the target selection. In order to mitigate these effects the goal is to obtain a relative density that is independent of observational features such as galactic extinction (E(B-V)), stellar density, depth of the PSF... In DESI, we rely on a method that uses random forest (RF) regression (`regressis`[1]) based on observational feature templates to mitigate imaging systematics and derived the photometric weights w_p (Chaussidon et al., 2021). An example of the contaminations and mitigations for ELGs is illustrated in Figure 2.19. This Figure show that the density of ELGs is strongly dependent on the photometric features and the correction seems to mitigate all these effects.

[1] https://github.com/echaussidon/regressis/tree/main

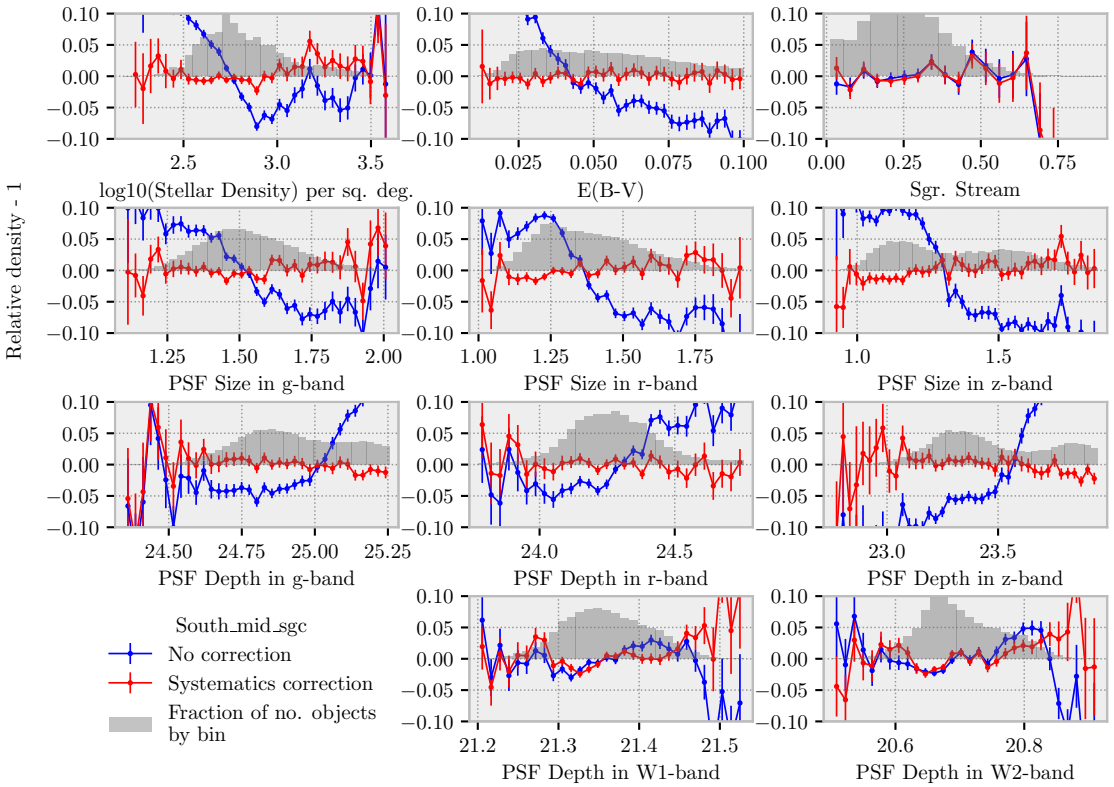

Figure 2.19: *Relative density of ELG targets as a function of different observational features (see Chaussidon et al. (2021)) for the definition of the features). The blue line is the relative density of ELG targets without any correction and the red line is after applying RF correction and the errors correspond to binomial errors. The grey histograms represent the number of object per bins. Credit: E. Chaussidon.*

2.7.1.3 Spectroscopic systematics

Similarly to imaging systematics, spurious contamination can arise from redshift or spectroscopic failures. To investigate potential spectroscopic systematics we use the spectroscopic success rate (SSR), defined as the ratio of the number of valid redshifts over the total number of ELG spectra. From the VI, we can define the average SSR for each target class. For the `ELG_LOP` sample the SRR is $\sim 72\%$. In principle we want SSR to be flat with respect to observational or instrumental features (e.g. across the focal plane, see Figure 2.20). Any variation of the SSR is quantified and if it is found to be significant compared to what one would expect randomly, it needs to be mitigated and corrected by applying a spectroscopic weight $w_{\rm spec}$.

During my thesis, I performed some tests to investigate the potential variations of SSR as a function of different observational and instrumental features using the data from the One-Percent survey. We did not find any significant trends, and the weights derived from this study do not have a significant impact on the clustering. So we decided not to include them. However, this effort is still under study for the Year 1 data sample. To give an example, the SSR should be independent of the position on the focal plane and, therefore, any trend on the focal plane should be mitigated. The variation in SSR over the focal plane for the `ELG_LOP` sample from the One-Percent survey is presented in Figure 2.20. For the present work we assume that the variations over the focal plane have a negligible impact. But this will be studied further for Y1 clustering analyses of DESI.

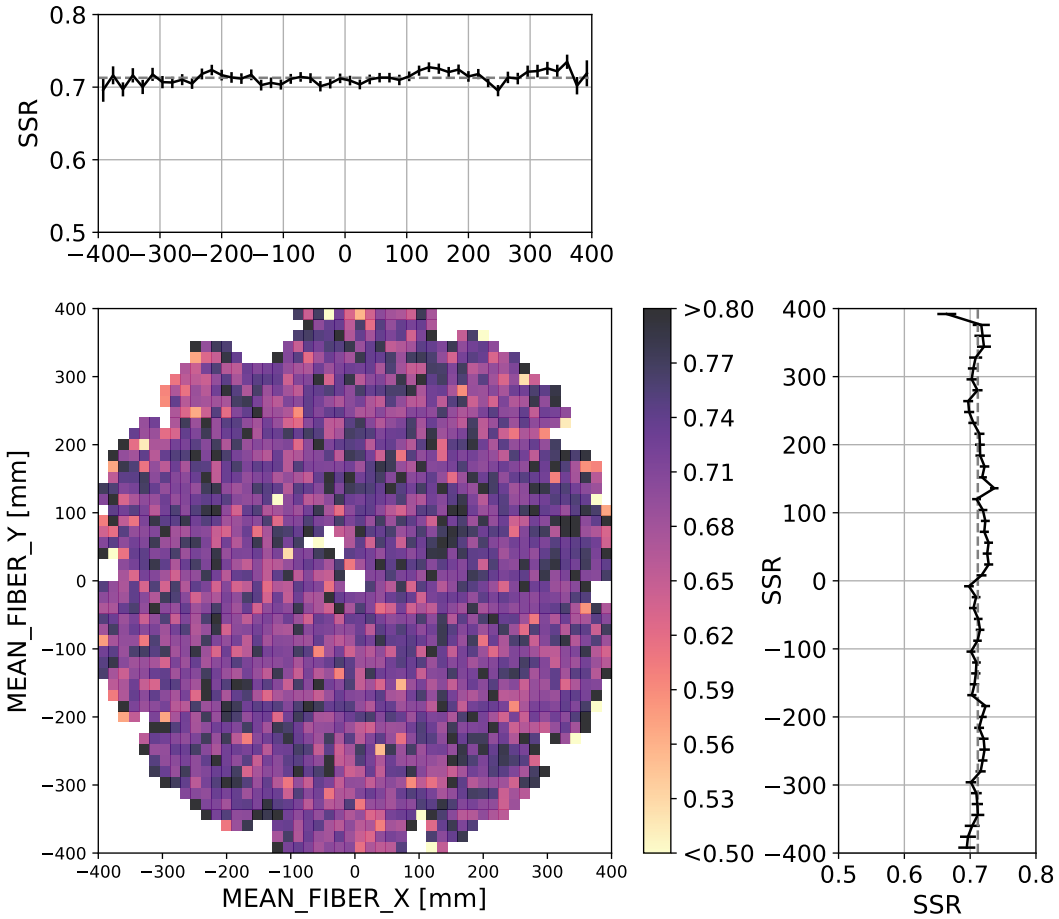

Figure 2.20: *Spectroscopic success rate (SSR) (in colour) on the 2D representation of the DESI focal plane. The SSR in the x and y axis is shown in the side panels and errors are binomial.*

2.7.1.4 FKP weights

The FKP weights (Feldman et al., 1994) are not a correction for systematic effects but are commonly used to improve the variance of measured two-point statistics in order to optimise the signal-to-noise of the galaxy field at a given scale k_0 (usually used for BAO scale around $k_0 = 0.14\ h/\mathrm{Mpc}$). Assuming that the galaxy field follows a Poisson distribution with a mean power spectrum $\bar{P}(k)$ the optimal FKP weights are:

$$w_{\mathrm{FKP}}(z) = \frac{1}{1 + \bar{n}(z)\bar{P}(k_0)} \tag{2.9}$$

where $\bar{n}(z)$ is the average number density of galaxies at a given redshift, and $\bar{P}(k_0)$ is the power-spectrum at a wavelength of interest k_0. These weights are used by default in our DESI analyses but do not affect our scales of interest ($< 30\mathrm{Mpc}$).

In summary, once all the individual weights have been calculated, the final individual weight $w_{\rm tot}$ for each galaxy is the multiplication of all the different weights.

2.8 Observational effects

In the following we describe the main observational effects that affect the galaxy clustering and show how they can be used to infer cosmological information.

2.8.1 Alcock-Paczynski effect

The two-point statistic is usually computed using Cartesian coordinate, whereas in galaxy surveys, catalogues are provided with angular positions and redshifts. Angular positions and redshifts can be converted to comoving distances in Cartesian space assuming a *fiducial cosmology*, which is likely to be different from the unknown cosmology in the data. The use of a wrong cosmological model to convert redshift to distance in the data, create detectable distortions in the galaxy clustering, known as the Alcock-Paczynski (AP) effect. The distortions occur in the radial and angular comoving distances, D_H and D_A (see Equation (1.25), Equation (1.26)) (at an effectively redshift $z_{\rm eff}$). We can define two scaling parameters, perpendicular and parallel to the line-of-sight:

$$\begin{aligned} q_{\parallel} &= \frac{D_H\left(z_{\rm eff}\right)}{D_H^{\rm fid}\left(z_{\rm eff}\right)} \\ q_{\perp} &= \frac{D_A\left(z_{\rm eff}\right)}{D_M^{\rm fid}\left(z_{\rm eff}\right)}. \end{aligned} \tag{2.10}$$

The superscript $^{\rm fid}$ refers to the fiducial cosmology used for cosmological distances. These distortions are used when measuring the BAO scale, r_d (see Equation (1.38)), in galaxy survey and are parameterized as follows:

$$\boxed{\begin{aligned} \alpha_{\parallel} &= \frac{D_H(z_{\rm eff})r_d^{\rm fid}}{D_H^{\rm fid}(z_{\rm eff})r_d} \\ \alpha_{\perp} &= \frac{D_A(z_{\rm eff})r_d^{\rm fid}}{D_A^{\rm fid}(z_{\rm eff})r_d} \end{aligned}} \tag{2.11}$$

These α-parameters can be varied in during cosmological inference, which allow the Hubble parameter, $H(z)$ and the comoving angular distance, $D_A(z)$, both divided by r_d, to be constrained.

2.8.2 Redshift-space distortions

As mentioned in the first chapter Section 1.4 the redshift have different contributions, the main one being the expansion of the Universe. The proper motions of galaxies, known as *peculiar velocities* also make a small contribution to the redshift. When measuring the redshift only the component along the line of sight (los) of peculiar velocities, $\mathbf{v}_{\parallel} = v(\mathbf{r}) \cdot \hat{los}$ (where $\hat{los}$ is a unit vector along the los), affects the measurement. Since velocities are driven by gravity, the RSD effect can be used to constrain gravity models on large-scale and dark energy. The position in real space $\mathbf{r}$ is mapped to *redshift space* (observed) $\mathbf{s}$ following:

$$\mathbf{s} = \mathbf{r} + \frac{v(\mathbf{r}) \cdot \hat{los}}{aH(a)} \tag{2.12}$$

In observations, we cannot disentangle the two contributions and we only have access to the position in redshift space. Here, we adopt the plane-parallel approximation, i.e. all lines of sight are considered to be parallel between the galaxies since they are far away from us. This approximation breaks down when considering a large separation between objects, and additional effects (such as wide-angle effects) have to be taken into account, which is not the case in this work. As the effect of velocities is only visible along the los, these distortions create anisotropies in the distribution of observed galaxies, and have two main signatures in the clustering of galaxies illustrated in Figure 2.21. As we will see in the next chapter (Section 3.1.3), structures grow continuously under the effect of gravity. On large scales, this growth is the main source of RSD. If we consider galaxies at the near or far end of an overdense region, as they are falling towards the overdensity. The one at the near end move away from us, increasing its redshift, and the one at the far end move towards us, decreasing its redshift. In this way, the observed clustering is *squashed* along the line of sight in large-scale redshift space, which is known as the *Kaiser* effect. On the other hand, at small scales, galaxies fall into collapsed objects with deep potential wells, and the velocity of the objects is dominated by random motions that introduce an apparent elongation along the line of sight. This effect is known as *Fingers of God* (see Figure 2.21).

The squashing effect on large scales causes an increase in the measured power of the 2PCF that can be easily modelled. Kaiser (1987) has derived how the change in power due to peculiar velocities of galaxies is related to the growth rate of structure on large/linear scale:

$$\xi^s(s, \mu) = (1 + f\mu^2)^2 \xi^r(r, \mu) \tag{2.13}$$

where μ is the cosine of the angle between the direction of the galaxy pair and the line of sight. f is the linear growth rate, which describes how fast the cosmic structure grow. We give more details on the linear growth rate f in the next chapter (see section Section 3.1.3). Finally, the superscripts r and s refer to the groupings in real space and in redshift space. From this simple model, we can model the 2PCF and compare it to the data to infer the value of f[1], which is directly related to the theory of gravity and dark energy (see Equation (3.37)).

Figure 2.22 shows the distortions induces by RSD and AP effetcs in the 3D distribution of galaxies.

2.8.3 Galaxy bias

In the discussion above, we do not differentiate between the distribution of galaxies and the distribution of matter. However, as we saw in the first chapter, baryons represent only a small fraction, $\sim 25\%$ of the total mass of the universe, the other part being invisible dark matter. As we will see in the next chapter, galaxies follow the distribution of dark matter, residing mainly at the centre of dark matter halos, which are overdense regions of the cosmic web. Consequently, galaxies are tracers of the matter distribution, which means that where there is a galaxy, there

[1] Actually, the constrained value is $f\sigma_8$, where σ_8 is the root-mean-square of the density fluctuations in a sphere of radius $R = 8$ Mpc$/h$, see Equation (3.13)

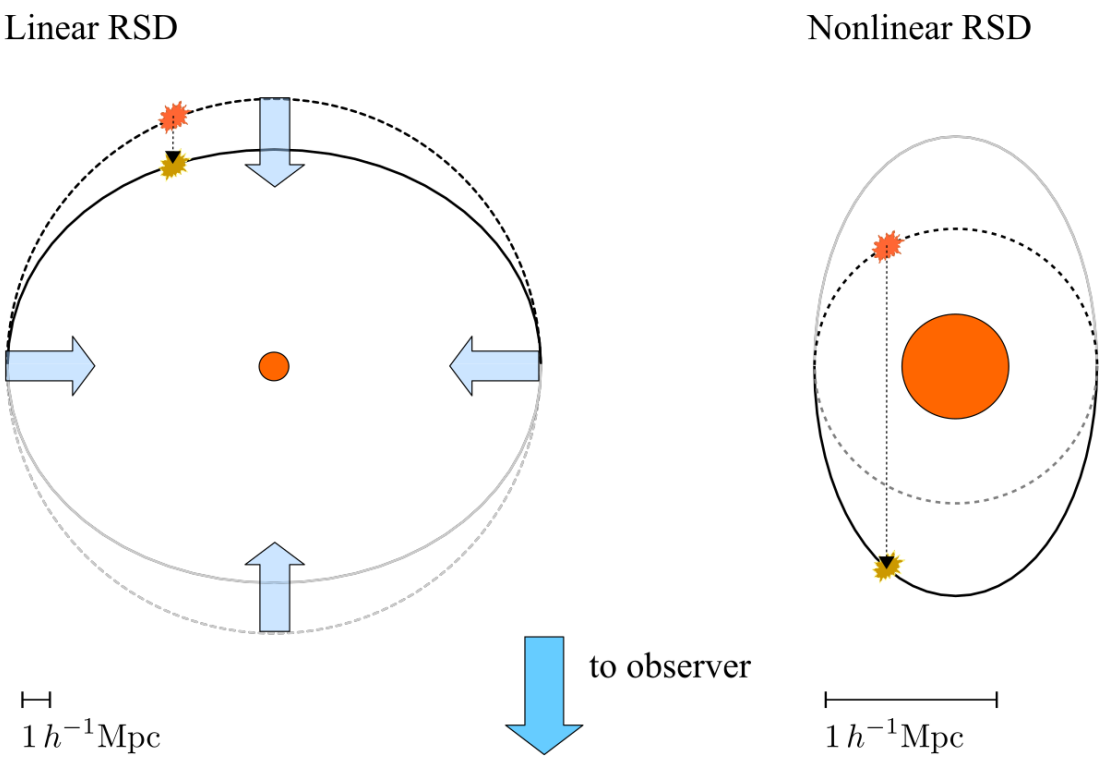

Figure 2.21: *Schematic illustration of the Redshift-space distortions effect. The observer is assumed to be far away below the figure, so the los in vertical. The central overdensity is represented by the filled orange circle. The wide blue arrows indicate the direction of the velocity flow and arrows with dashed lines indicate the velocity contribution along the line-of-sight of the object. Left panel: Linear/large-scale effect of RSD, a constant density contour circular in real space (dashed line) is squashed in redshift space (solid line), due to slow motion of object towards the overdensity. Right panel: Small-scale (non-linear) effect of RSD. The redshift space contour is elongate along the los. The velocity of galaxies are dominated by random motions and an object on the "far side" (top) of the overdensity in real space is displaced on the opposite side. Figure taken from (Dodelson & Schmidt, 2020).*

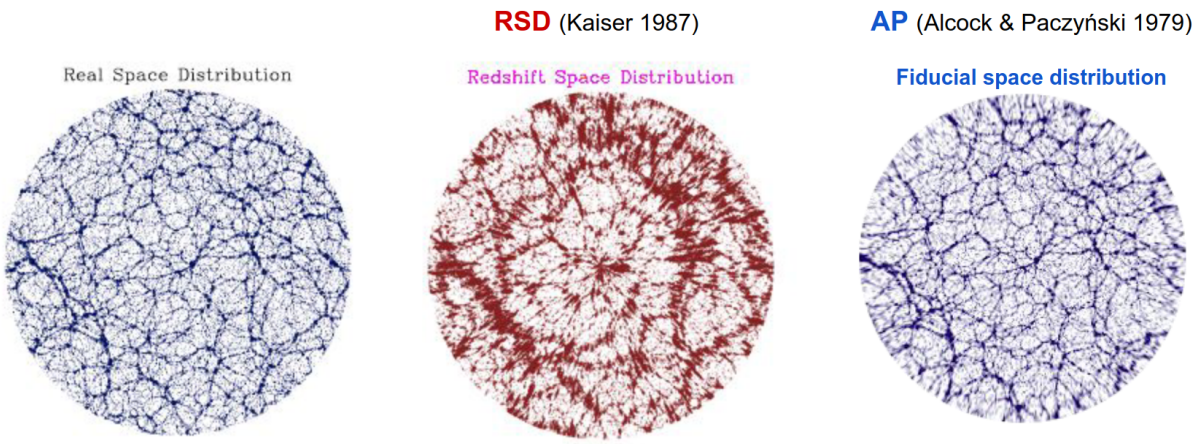

Figure 2.22: *The spatial distribution of galaxies in real space (left panel), redshift space (middle panel) and the distortion from AP effect in real space (different cosmology model to convert distances) computed using a cosmological simulation (right panel). In the observation we have the combination of the 2 effects. Credit: Slide from Carlos Mauricio Correa Fayn*

must also be dark matter. However, the fact that there are no galaxies does not mean that there is no dark matter. Thus, the field of galaxies δ_g is *biased* relative to the total matter field δ_m. The standard prescription to model that effect is:

$$\delta_g = b\delta_m \tag{2.14}$$

where b is called the galaxy linear bias. We can also write the relation between the two-point correlation functions of the matter field ξ_{mm} and the galaxy fields ξ_{gg} as follows:

$$\xi_{gg}(r) = b^2\xi_{mm}(r) \tag{2.15}$$

We can rewrite the Kaiser formula in Equation (2.16) adding the galaxy bias:

$$\xi^s_{gg}(s,\mu) = b^2(1+\frac{f}{b}\mu^2)^2\xi^r_{mm}(r,\mu) \tag{2.16}$$

The *bias* between these the galaxy fields and the matter fields depends on the type of galaxy, the physics of galaxy formation, and may therefore be a general function, (e.g. depending on the redshift, scale considered...). The galaxy bias is typically measured directly from the data and is, on large scales, a only a function of redshift for a given galaxy population (e.g. see Laurent et al. (2017) for quasars).

2.9 Small scale clustering of ELGs from the One-Percent survey

For the purposes of this thesis, and the study of the galaxy-dark matter halo connection, we need to measure the galaxy clustering on small scales. We dedicate, in the next chapter, an entire section on galaxy halo connection and particularly on ELGs (see Section 3.4). Thanks to the high completeness of the One-Percent survey, the clustering can be measured down to very small scales, 0.04 Mpc/h in r_p for the projected two-point correlation function w_p and 0.17 Mpc/h in galaxy pair separation s for the monopole and quadrupole of the 2PCF. Figure 2.23 shows the projected clustering (integrated between -40 and 40 Mpc/h), the monopole and the quadrupole of the 2PCF for the ELG sample of the One-Percent survey, and the impact of the completeness and the PIP+ANG weights, where we compute the clustering for the whole ELG sample ($\sim$ 86% completeness) with and without the weights and when we restrict the sample only to the radius of each rosette between 0.2° and 1.50° (see Figure 2.17) without weights ($\sim$ 95% completeness). The clustering measurement is restricted to the redshift range $0.8 < z < 1.6$. As shown by Figure 2.23, increasing the completeness of the sample reduced the observed systematic effect in the 2PCF measurements.

We can observe two features on small scales in the projected clustering. Firstly, there is a drop in clustering power on the smallest scales. This is expected due to blending effects. Blending occurs when two objects are closer than the size of the point spread function (PSF) of the image, so that the objects cannot be resolved, i.e. we see only one object in the sky (see Figure 2.25). The scale of the PSF corresponds roughly to the size of the seeing, so objects separated by an angle of less than twice the seeing cannot be resolved. On Figure 2.24, we can

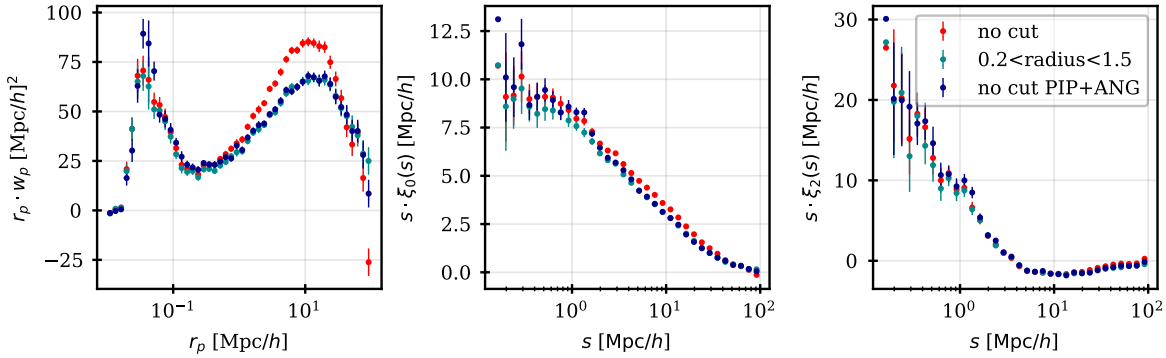

Figure 2.23: *ELG clustering measurement from the One-Percent survey in the redshift range* $0.8 < z < 1.6$. *Left panel: Projected correlation function* $w_p(r_p)$ *as a function of* r_p *(integrated between* $\pi_{min} = -40$ *and* $\pi_{max} = 40$ *Mpc/*h*). Middle: 2PCF monopole* $\xi_0(s)$ *times* s*. Right: Quadrupole of the 2PCF* $\xi_2(s)$ *times* s*. The red and blue dots correspond to the full ELG sample without and with the PIP+ANG weights respectively. The green dots correspond to the ELG sample restricted to the radius of each rosette between 0.2° and 1.50°. On each panel, the error bars are calculated using the delete-one Jackknife method for the One-Percent survey footprint divided into 128 independent regions.*

see that this scale (2*seeing, dotted line) corresponds to the drop in the projected clustering measurement. At Kitt Peak, on clear skies the seeing is ~ 1.3 arcsec (Dey et al., 2019).

The second feature is the increase in power at $r_p < 0.1$ Mpc/h, which means that a non-negligible part of the ELGs are very closely separated. This feature was not anticipated from previous studies of ELGs because, in previous experiments, scales < 0.1 Mpc/h were not well measured, and most studies of ELG clustering at small scales stopped at $r_p = 0.1$ Mpc/h (Avila et al., 2020, Lin et al., 2023, Okumura et al., 2021). We are therefore looking for possible systematic effects likely to generate this signal on these small scales. In particular, we are testing potential foreground effects in the images. To test this, we measure the projected clustering, w_p, in different π bins (along the line of sight), see Figure 2.24. We see that the strong clustering power at small scales is mainly due to separations along the line of sight below $\pi = 4$Mpc $/h$, which means that this feature is driven by objects that are closely separated in the transverse direction but also in the radial direction, and is therefore not due to foreground effects.

Finally, we examined the image and spectra of observed ELG targets closely separated on the sky with small redshift difference $\Delta z < 0.001$. Examples of these images are shown in Figure 2.25. Most of the pairs look real, where we can see 2 different targets on the sky. There are 267,345 ELGs in the sample from the One-Percent survey. Among them, there are 830 galaxy pairs that have a separation angle $\theta < 2.6$ arcsec on the sky corresponding to 828 unique targets. As these scales are subject to blending effect and fragmentation effect (several targets found for a single object), there are potentially ~ 414 spurious targets, which represent $\sim 0.1\%$ of the total sample. Removing these targets from the sample does not change the increase in clustering power on scales < 0.2Mpc/h. Looking at higher separation angle $2.6 < \theta < 25$ arcsecond, there are 67434 pairs of objects, corresponding to 62895 unique objects, representing $\sim 23\%$ of the ELG sample. After many checks, we did not find any significant systematic effects that could generate this increase in small-scale clustering power. Therefore, we tried to model this feature in our analysis in Chapter 5.

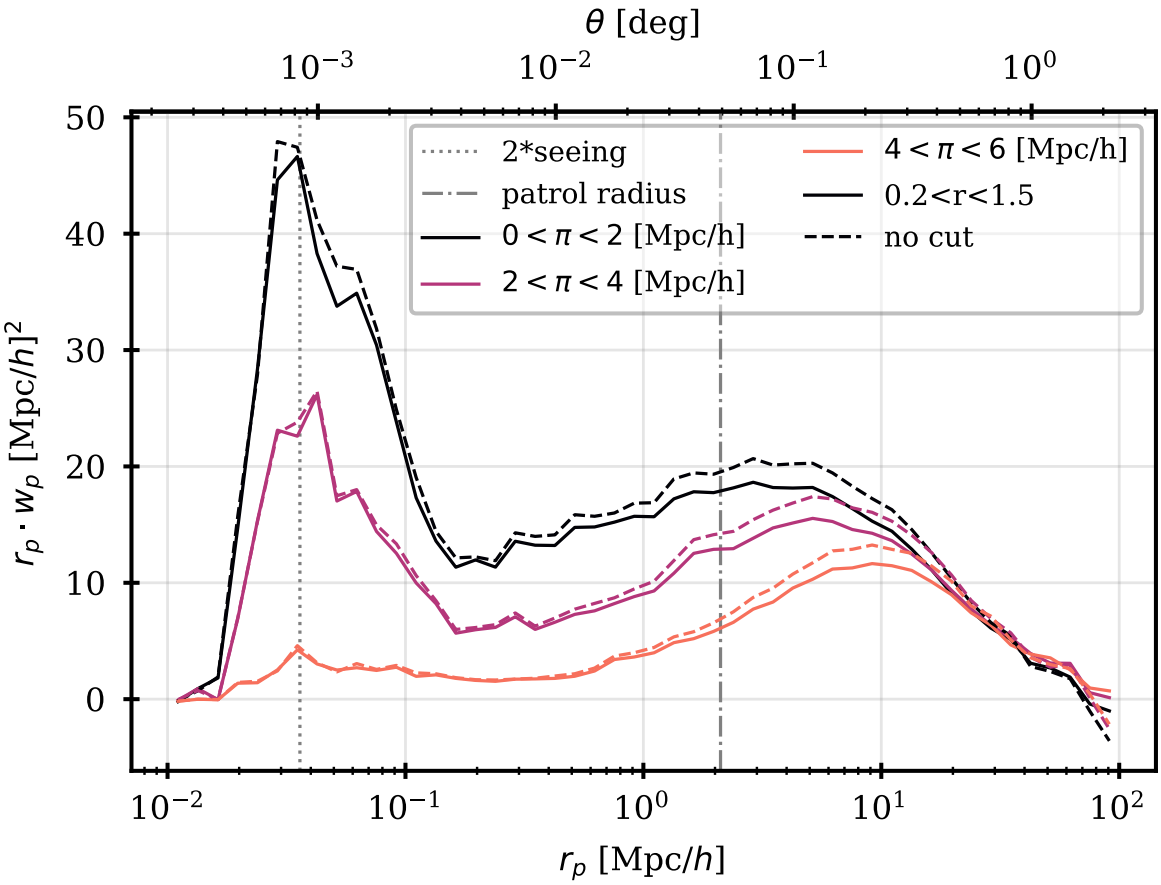

Figure 2.24: *DESI clustering measurements for the One-Percent survey ELG data sample restricted to the redshift range* $0.8 < z < 1.6$. *The 2D correlation function in successive bins of* $2\mathrm{Mpc}/h$ *in the galaxy-pair separation along the line-of-sight is shown as a function of the separation perpendicular to the line-of-sight,* r_p. *No correction weight has been applied. Measurements using the whole survey footprint (solid lines) are compared with measurements excluding the inner and outer regions of the rosettes where the survey was less incomplete (dashed lines). Also indicated are the separation corresponding to the fibre patrol region (dot-dashed grey line) and the limit corresponding to twice the mean survey seeing (dotted grey line). Below this limit, target blending cannot be resolved, leading to a loss of power. This plot demonstrates that the strong increase in power at small scales (below* $0.2\mathrm{Mpc}/h$*) is not due to the (slight) incompleteness of the One-Percent survey.*

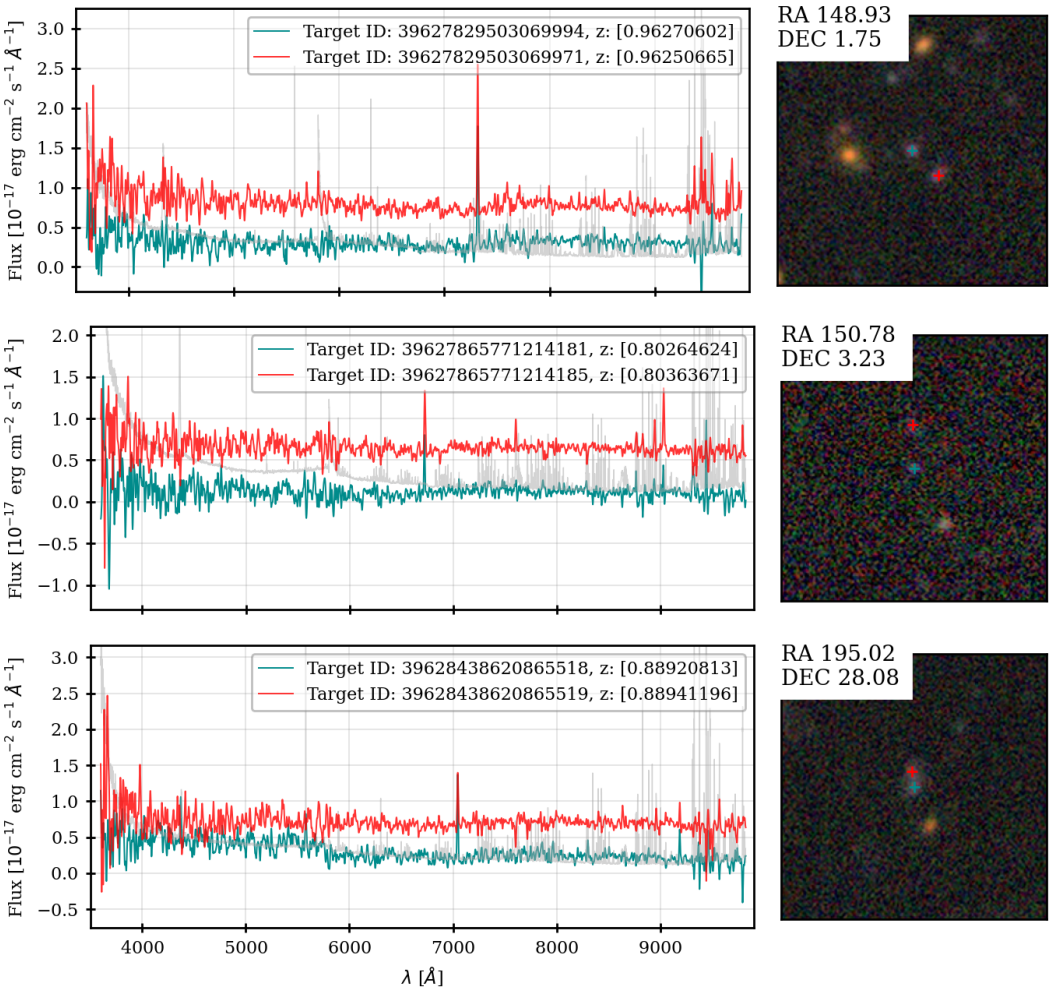

Figure 2.25: *Examples of pair of spectra for close galaxy pairs on sky and redshift from the ELG sample of the One-Percent survey. The two upper spectra are cleary 2 different objects on sky (the image have low quality but on the sky viwer we can clearly see the 2 objects), while the two targets on the bottom panel can potentially be the same object (fragmentation). The size of each image is* $\sim 21 \times 21$ *arcsec. Images url (from top to bottom):* https://www.legacysurvey.org/viewer/desi-edr-spectra/?ra=148.938&dec=1.758&zoom=20, https://www.legacysurvey.org/viewer/desi-edr-spectra/?ra=150.788&dec=3.238&zoom=20, https://www.legacysurvey.org/viewer/desi-edr-spectra/?ra=195.028&dec=28.088&zoom=20

Bibliography

Abareshi, B., Aguilar, J., Ahlen, S., et al. 2022, Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument, doi: 10.3847/1538-3881/ac882b

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, The Astrophysical Journal Supplement Series, 182, 543, doi: 10.1088/0067-0049/182/2/543

Abbott, T., Abdalla, F., Alarcon, A., et al. 2018, Physical Review D, 98, 043526, doi: 10.1103/PhysRevD.98.043526

Abdurro'uf, Accetta, K., Aerts, C., et al. 2022, The Astrophysical Journal Supplement Series, 259, 35, doi: 10.3847/1538-4365/ac4414

Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, Publications of the Astronomical Society of Japan, 71, 114, doi: 10.1093/pasj/psz103

Alam, S., Aubert, M., Avila, S., et al. 2021, Physical Review D, 103, 083533, doi: 10.1103/PhysRevD.103.083533

Alexander, D. M., Davis, T. M., Chaussidon, E., et al. 2023, The Astronomical Journal, 165, 124, doi: 10.3847/1538-3881/acacfc

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, Monthly Notices of the Royal Astronomical Society, 499, 5486, doi: 10.1093/mnras/staa2951

Bailey, S. J., & DESI Collaboration. 2023, in prep.

Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, The Astrophysical Journal Supplement Series, 208, 20, doi: 10.1088/0067-0049/208/2/20

Betoule, M., Kessler, R., Guy, J., et al. 2014, Astronomy & Astrophysics, 568, A22, doi: 10.1051/0004-6361/201423413

Bianchi, D., & Percival, W. J. 2017, Monthly Notices of the Royal Astronomical Society, 472, 1106, doi: 10.1093/mnras/stx2053

Bosch, F. v. d., More, S., Cacciato, M., Mo, H., & Yang, X. 2013, Monthly Notices of the Royal Astronomical Society, 430, 725, doi: 10.1093/mnras/sts006

Bundy, K., Leauthaud, A., Saito, S., et al. 2017, The Astrophysical Journal, 851, 34, doi: 10.3847/1538-4357/aa9896

Carrasco, J. M., Evans, D. W., Montegriffo, P., et al. 2016, Astronomy & Astrophysics, 595, A7, doi: 10.1051/0004-6361/201629235

Chaussidon, E., Yèche, C., Palanque-Delabrouille, N., et al. 2021, Monthly Notices of the Royal Astronomical Society, 509, 3904, doi: 10.1093/mnras/stab3252

—. 2023, The Astrophysical Journal, 944, 107, doi: 10.3847/1538-4357/acb3c2

Chen, X. 2010, Advances in Astronomy, 2010, 1, doi: 10.1155/2010/638979

Cole, S., Percival, W. J., Peacock, J. A., et al. 2005, Monthly Notices of the Royal Astronomical Society, 362, 505, doi: 10.1111/j.1365-2966.2005.09318.x

Collaboration, D., Aghamousa, A., Aguilar, J., et al. 2016, The DESI Experiment Part I: Science,Targeting, and Survey Design, arXiv. http://arxiv.org/abs/1611.00036

Collaboration, D., Adame, A. G., Aguilar, J., et al. 2023a, doi: 10.5281/zenodo.7858207

—. 2023b, doi: 10.5281/zenodo.7964161

Cooper, A. P., Koposov, S. E., Allende Prieto, C., et al. 2023, The Astrophysical Journal, 947, 37, doi: 10.3847/1538-4357/acb3c0

Cutri, R. M., & et al. 2012, VizieR Online Data Catalog, II/311. https://ui.adsabs.harvard.edu/abs/2012yCat.2311....0C

Desjacques, V., & Seljak, U. 2010, Advances in Astronomy, 2010, 1, doi: 10.1155/2010/908640

Dey, A., Schlegel, D. J., Lang, D., et al. 2019, The Astronomical Journal, 157, 168, doi: 10.3847/1538-3881/ab089d

Dodelson, S., & Schmidt, F. 2020, Modern Cosmology (Elsevier Science)

Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, The Astrophysical Journal, 633, 560, doi: 10.1086/466512

Falco, E. E., Kurtz, M. J., Geller, M. J., et al. 1999, Publications of the Astronomical Society of the Pacific, 111, 438, doi: 10.1086/316343

Favole, G., Gonzalez-Perez, V., Ascasibar, Y., et al. 2023, Characterizing the ELG luminosity functions in the nearby Universe, arXiv. http://arxiv.org/abs/2303.11031

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, The Astrophysical Journal, 426, 23, doi: 10.1086/174036

Flaugher, B., Diehl, H. T., Honscheid, K., et al. 2015, The Astronomical Journal, 150, 150, doi: 10.1088/0004-6256/150/5/150

Gaia Collaboration, Prusti, T., De Bruijne, J. H. J., et al. 2016, Astronomy & Astrophysics, 595, A1, doi: 10.1051/0004-6361/201629272

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, Astronomy & Astrophysics, 616, A1, doi: 10.1051/0004-6361/201833051

Guy, J., Bailey, S., Kremin, A., et al. 2023, The Astronomical Journal, 165, 144, doi: 10.3847/1538-3881/acb212

Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, Astronomy & Astrophysics, 566, A108, doi: 10.1051/0004-6361/201321489

Hahn, C., Wilson, M. J., Ruiz-Macias, O., et al. 2023, The Astronomical Journal, 165, 253, doi: 10.3847/1538-3881/accff8

Hawkins, E., Maddox, S., Cole, S., et al. 2003, Monthly Notices of the Royal Astronomical Society, 346, 78, doi: 10.1046/j.1365-2966.2003.07063.x

Huchra, J., Davis, M., Latham, D., & Tonry, J. 1983, The Astrophysical Journal Supplement Series, 52, 89, doi: 10.1086/190860

John, T. L. 1988, Astronomy and Astrophysics, 193, 189. https://ui.adsabs.harvard.edu/abs/1988A&A...193..189J

Jones, D. H., Read, M. A., Saunders, W., et al. 2009, Monthly Notices of the Royal Astronomical Society, 399, 683, doi: 10.1111/j.1365-2966.2009.15338.x

Jullo, E., Torre, S. d. l., Cousinou, M.-C., et al. 2019, Astronomy & Astrophysics, 627, A137, doi: 10.1051/0004-6361/201834629

Kaiser, N. 1987, Monthly Notices of the Royal Astronomical Society, 227, 1, doi: 10.1093/mnras/227.1.1

Kent, S., Neilsen, E., Honscheid, K., et al. 2023, Astrometric Calibration and Performance of the Dark Energy Spectroscopic Instrument Focal Plane, arXiv. http://arxiv.org/abs/2307.06238

Lan, T.-W., Tojeiro, R., Armengaud, E., et al. 2022, The DESI Survey Validation: Results from Visual Inspection of Bright Galaxies, Luminous Red Galaxies, and Emission Line Galaxies, arXiv, doi: 10.48550/arXiv.2208.08516

Landy, S. D., & Szalay, A. S. 1993, The Astrophysical Journal, 412, 64, doi: 10.1086/172900

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, Euclid Definition Study Report, Tech. rep. https://ui.adsabs.harvard.edu/abs/2011arXiv1110.3193L

Laurent, P., Eftekharzadeh, S., Goff, J.-M. L., et al. 2017, Journal of Cosmology and Astroparticle Physics, 2017, 017, doi: 10.1088/1475-7516/2017/07/017

Lin, S., Tinker, J. L., Blanton, M. R., et al. 2023, Monthly Notices of the Royal Astronomical Society, 519, 4253, doi: 10.1093/mnras/stac2793

LSST Dark Energy Science Collaboration. 2012, Large Synoptic Survey Telescope: Dark Energy Science Collaboration, Tech. rep. https://ui.adsabs.harvard.edu/abs/2012arXiv1211.0310L

Miller, T. N., Doel, P., Gutierrez, G., et al. 2023, The Optical Corrector for the Dark Energy Spectroscopic Instrument, arXiv. http://arxiv.org/abs/2306.06310

Mohammad, F. G., Percival, W. J., Seo, H.-J., et al. 2020, Monthly Notices of the Royal Astronomical Society, 498, 128, doi: 10.1093/mnras/staa2344

Moustakas, J., Kennicutt, Jr., R. C., & Tremonti, C. A. 2006, The Astrophysical Journal, 642, 775, doi: 10.1086/500964

Myers, A. D., Moustakas, J., Bailey, S., et al. 2023, The Astronomical Journal, 165, 50, doi: 10.3847/1538-3881/aca5f9

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, The Astrophysical Journal Supplement Series, 208, 5, doi: 10.1088/0067-0049/208/1/5

Okumura, T., Hayashi, M., Chiu, I.-N., et al. 2021, Publications of the Astronomical Society of Japan, 73, 1186, doi: 10.1093/pasj/psab068

Parkinson, D., Riemer-Sørensen, S., Blake, C., et al. 2012, Physical Review D, 86, 103518, doi: 10.1103/PhysRevD.86.103518

Percival, W. J., & Bianchi, D. 2017, Monthly Notices of the Royal Astronomical Society: Letters, 472, L40, doi: 10.1093/mnrasl/slx135

Percival, W. J., Cole, S., Eisenstein, D. J., et al. 2007, Monthly Notices of the Royal Astronomical Society, 381, 1053, doi: 10.1111/j.1365-2966.2007.12268.x

Perruchot, S., Blanc, P.-, Guy, J., et al. 2020, in Ground-based and Airborne Instrumentation for Astronomy VIII, ed. C. J. Evans, J. J. Bryant, & K. Motohara (Online Only, United States: SPIE), 179, doi: 10.1117/12.2561275

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, Astronomy and Astrophysics, 641, A6, doi: 10.1051/0004-6361/201833910

Poppett, C., Jelinsky, P., Guy, J., et al. 2020, in Ground-based and Airborne Instrumentation for Astronomy VIII, 161, doi: 10.1117/12.2562565

Raichoor, A., Moustakas, J., Newman, J. A., et al. 2023, The Astronomical Journal, 165, 126, doi: 10.3847/1538-3881/acb213

Rodríguez-Torres, S. A., Comparat, J., Prada, F., et al. 2017, Monthly Notices of the Royal Astronomical Society, 468, 728, doi: 10.1093/mnras/stx454

Sawicki, M. 2002, The Astronomical Journal, 124, 3050, doi: 10.1086/344682

Schlafly, E. F., Kirkby, D., Schlegel, D. J., et al. 2023, Survey Operations for the Dark Energy Spectroscopic Instrument, arXiv. http://arxiv.org/abs/2306.06309

Scolnic, D. M., Jones, D. O., Rest, A., et al. 2018, The Astrophysical Journal, 859, 101, doi: 10.3847/1538-4357/aab9bb

Silber, J. H., Fagrelius, P., Fanning, K., et al. 2023, The Astronomical Journal, 165, 9, doi: 10.3847/1538-3881/ac9ab1

Smith, A., He, J.-h., Cole, S., et al. 2019, Monthly Notices of the Royal Astronomical Society, 484, 1285, doi: 10.1093/mnras/stz059

Vogeley, M. S., Park, C., Geller, M. J., & Huchra, J. P. 1992, The Astrophysical Journal, 391, L5, doi: 10.1086/186385

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, The Astronomical Journal, 140, 1868, doi: 10.1088/0004-6256/140/6/1868

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, The Astronomical Journal, 120, 1579, doi: 10.1086/301513

Zhou, R., Dey, B., Newman, J. A., et al. 2023, The Astronomical Journal, 165, 58, doi: 10.3847/1538-3881/aca5fb

Zou, H., Zhou, X., Fan, X., et al. 2017, Publications of the Astronomical Society of the Pacific, 129, 064101, doi: 10.1088/1538-3873/aa65ba

3

The large scale structures of the Universe

Contents

In the first chapter we described the dynamics and the evolution of a homogeneous and isotropic universe on large scale. However, on smaller scales, where inhomogeneities appear, the Universe cannot be treated as homogeneous and isotropic anymore. This chapter aims to describe how the fluctuations laid out by inflation, which are imprinted in the temperature anisotropy power spectrum of the CMB ($\Delta T/T \sim 10^{-5}$), have been amplified by the influence of gravity and have led to the formation of large-scale structures. This chapter is inspired from a series of online lectures given by Franck van den Bosch on the theory of galaxy formation (https://campuspress.yale.edu/astro610/).

3.1 From overdensities to DM halos

3.1.1 Statistical properties of cosmic fields

Due to their quantum nature, fluctuations resulting from inflation cannot be predicted or directly measured. These fluctuations represent a single occurrence within an infinite ensemble of possible realisations that could have arisen from a random process during inflation. Consequently, to study these fluctuations it becomes necessary to employ a statistical perspective and utilise a probabilistic description. In this section, we provide some important tools and relevant properties to statistically describe cosmic fields.

The matter density field $\rho(\mathbf{x})$ can be translated into the *density perturbation field* $\delta_{\mathbf{x}}$ defined as:

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \langle\rho\rangle}{\langle\rho\rangle} \tag{3.1}$$

where $\rho(\mathbf{x})$ is the density at the position $\mathbf{x}$ and $\langle\rho\rangle$ is the mean density. In the following, we will apply the ergodic principle, which states that volume-averaged quantities are equal to their expectation values, i.e. $\bar{\rho} = \langle\rho\rangle$. One property of the density contrast is that its mean has to be zero $\langle\delta\rangle = 0$, to respect the cosmological principle (homogeneity on large scales). Because simple inflation models (e.g. single-field slow-roll) predict that initial density perturbations originate from numerous *independent quantum fluctuations*, the central limit theorem implies that the density perturbation field $\delta(\mathbf{x})$ at an early time, or on large scales, is very close to be Gaussian-distributed. In this case, all the statistical information about the density perturbation field can be completely characterised by its mean and variance. Since the mean of the distribution is zero according to the cosmology principle $\langle\delta\rangle = 0$, the variance, i.e. the *two-point correlation function (2PCF)*, fully describes the statistical distribution of the field:

$$\langle\delta(\mathbf{x})\rangle = 0 \tag{3.2}$$

$$\xi(\mathbf{x}, \mathbf{x}') = \langle\delta(\mathbf{x})\delta(\mathbf{x}')\rangle \tag{3.3}$$

Assuming statistical isotropy and homogeneity of the field, the 2PCF depends only on the distance $r = ||\mathbf{r}||$ between the two positions $\mathbf{x}$ and $\mathbf{x}'$:

$$\xi(r) = \langle\delta(\mathbf{x})\delta(\mathbf{x} + \mathbf{r})\rangle \tag{3.4}$$

The 2PCF measures the excess over the random probability that two fluctuations are separated by a distance r. If the correlation $\xi(r)$ is positive, there are more fluctuations separated by a distance r than if they were uniformly distributed. In cosmology, density perturbations are often described by modes in Fourier space. Throughout this manuscript, we adopt the following convention for the Fourier transform and inverse transform:

$$\begin{aligned}\delta(\mathbf{k}) &= \int \delta(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d^3\mathbf{x}, \\ \delta(\mathbf{x}) &= \int \delta(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}} \frac{d^3\mathbf{k}}{(2\pi)^3}.\end{aligned} \tag{3.5}$$

The *power spectrum* $P(k)$ characterises the correlation of Fourier modes and is defined as the correlation between two Fourier modes:

$$\boxed{\langle \delta(\mathbf{k})\delta(\mathbf{k}')\rangle = (2\pi)^3 \delta_D^{(3)}(\mathbf{k}+\mathbf{k}')P(\mathbf{k})} \tag{3.6}$$

where $\delta_D^{(3)}$ is the Dirac delta function. Similarly to the 2PCF, the power spectrum depends only on $k = ||\mathbf{k}||$. The power spectrum is the Fourier transform of the correlation function (Wiener–Khinchin theorem):

$$P(k) = \int \xi(r) e^{-i\mathbf{k}'\cdot\mathbf{x}} d^3\mathbf{x} = 4\pi \int_0^\infty r^2 \xi(r) \frac{\sin(kr)}{kr} dr, \tag{3.7}$$

and conversely,

$$\xi(r) = \int P(k) e^{i\mathbf{k}\cdot\mathbf{x}} \frac{d^3\mathbf{k}}{(2\pi)^3} = \int_0^\infty k^2 P(k) \frac{\sin(kr)}{kr} \frac{dk}{2\pi^2}. \tag{3.8}$$

The last part of both equations consider homogeneous isotropic fields (so one can apply spherical symmetry).

The power spectrum (in *Fourier space*) decomposes the probability of the correlation function (in *real space*) to find two fluctuations at a distance r, into characteristic lengths, $k = 2\pi/r$. As the correlation function and the power spectrum form a Fourier pair, they provide the same information.

3.1.2 The initial power spectrum

Inflation models predict the initial shape of the power spectrum of primordial fluctuations. If we consider scalar-field inflation models (e.g. slow-roll), the perturbations generated during inflation are *adiabatic*, and (almost) Gaussian-distributed. These perturbations of the inflationary field (δ_ϕ) lead to spatial curvature perturbations ($\mathcal{R}$). The dimensionless power spectrum of spatial curvature perturbations $\Delta^2_{\mathcal{R}}(k)$ is predicted by scalar inflation models to be *nearly scale-invariant*, with the following parametrisation:

$$\Delta^2_{\mathcal{R}}(k) \equiv \frac{k^3}{2\pi^2} P_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_p}\right)^{n_s - 1} \tag{3.9}$$

where A_s is the amplitude of the power spectrum at the pivot scale k_p (generally set at $k_p = 0.05$ Mpc^{-1}) and n_s is the spectral index. In the special case where $n_s = 1$, the dimensionless power spectrum becomes scale-independent (called the Harrison-Zel'dovich spectrum). The value of

the spectral index predicted in scalar-field inflation models is $n_s \sim 0.96$. The value n_s is well-constrained by the CMB observations and the results from (Planck Collaboration et al., 2020) reported a value of $n_s = 0.9649 \pm 0.00842$. Due to the exponential expansion during the inflationary period, perturbations at the end of inflation become super-horizon (or super-Hubble), meaning that their characteristic scales $k \ll aH$. In this case, curvature and density perturbations are related by the Poisson equation $k^2\Phi_B/a^2 = -4\pi G\delta_k$ (where the Bardeen potential Φ_B can be related to the comoving curvature perturbation $\mathcal{R}$). Thus, the initial power spectrum of the density perturbations can be written as:

$$P_\delta^i(k) = A_s \left(\frac{k}{k_p}\right)^{n_s} \tag{3.10}$$

In the linear regime $\delta \ll 1$, different Fourier modes evolve independently of each other. Consequently, the linear power spectrum at any redshift z is related to the initial power spectrum by a linear *transfer function*:

$$T(k,z) = \frac{\delta(k,z)\delta^i(k=0)}{\delta^i(k)\delta(k=0,z)} \tag{3.11}$$

where δ^i are given by the initial conditions of the inflation field. The transfer function depends on the cosmological parameters and is defined for each species (s) that composes the Universe. It can be calculated using numerical codes such as CLASS (Lesgourgues, 2011). The linear matter power spectrum can thus be expressed as a function of the initial power spectrum:

$$P_\delta(k,z) = T^2(k,z)P_\delta^i(k) \tag{3.12}$$

For simplicity, we will adopt P instead of P_δ throughout the thesis from now on. As we saw above, the shape of the linear power spectrum can be predicted from the initial conditions of the inflationary field. However, to obtain a complete description of the power spectrum, we need to define its amplitude A_s which is not predicted a priori by initial conditions but must instead be fixed by observations. Instead of A_s, the historical prescription for normalising the power spectrum in large scale structure analyses uses σ_8:

$$\sigma_8^2 = \int_0^\infty P(k)\widetilde{W}_R^2(k)k^2\frac{dk}{2\pi^2} \tag{3.13}$$

where σ_8 encodes the amount of matter fluctuations averaged over a sphere of radius $R = 8\mathrm{Mpc}/h$. $\widetilde{W}_R(k)$ is a *window function* that takes into account geometric effects in the way galaxies are selected. In this case, this function is the Fourier transform of a spherical top-hat function, where galaxies are selected only within a spherical volume V of radius $R = 8\mathrm{Mpc}/h$:

$$W_R(r) = \begin{cases} 3/\left[4\pi R^3\right] & \text{if } r \le R \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

$$\widetilde{W}_R(k) = \frac{3}{(kR)^2}\left(\sin(kR) - kR\cos(kR)\right) \tag{3.15}$$

Consequently, we can relate the power spectrum for any redshift to the power spectrum at $z_0 = 0$ by normalising it by the ratio of their amplitudes:

$$P(k,z) = \frac{\sigma_8^2(z)}{\sigma_8^2(z_0)}P(k,z_0) \tag{3.16}$$

Today, many observations have measured the σ_8 parameter. The value from (Planck Collaboration et al., 2020) measurement is $\sigma_8(z=0) = 0.8111 \pm 0.0060$.

Note: *Recently, the measurements of σ_8 from different probes have been in tension. Large scale structure measurements at redshift $z \sim 1$ (mainly through weak lensing effects) are in $\sim 2-3\sigma$ disagreement with CMB measurements (Abdalla et al., 2022). Baryonic effects, which occur at small scales (high k) and are non-linear, are a possible explanation for this mild tension. Future experiments such as Euclid or LSST will collect large data sets and attempt to resolve this tension.*

3.1.3 Linear growth of perturbations

The aim of the upcoming section is to describe the temporal evolution of *adiabatic* perturbations generated during the inflation period. The analysis of perturbation evolution will depend on the different regimes in which perturbations grow. There exist two distinct regimes within which perturbations can evolve:

- The *sub-Hubble* regime where comoving Fourier modes $k \gg aH$ and the Hubble expansion can be neglected.
- The *supper-Hubble* regime, where comoving Fourier modes $k \ll aH$ and the Hubble friction terms (relativistic effets) dominate the field dynamics.

The evolution of large scale perturbations ($\Rightarrow$ small k, supper-Hubble) is governed by general relativity and requires perturbation of the FLRW metric (Equation (7.1)).

In the following, we will concentrate solely on the linear theory (sub-Hubble), considering small perturbations $\delta \ll 1$. This implies that gravitational fields are weak so we can adopt a Newtonian approach. As long as density fluctuations remain sufficiently small, non-linearities can be treated using perturbation theory. This approach breaks down at small scales, where the field is highly non-linear and perturbation theory is no longer valid. The validity of linear regime corresponds to scales above $40-50\mathrm{Mpc}/h$. Shortly after recombination, the baryons are completely decoupled from photons and evolve with dark matter through gravitation as a single fluid of matter. We consider an ideal, single-flow fluid composed of dark matter and baryons with density $\rho_\mathrm{m} = \rho_\mathrm{cdm} + \rho_\mathrm{b}$, velocity v, and pressure p evolving under the influence of a gravitational field ϕ in an expanding universe. To discuss the temporal evolution of the perturbations in an expanding universe, it is best to replace physical positions $\mathbf{r}$ and velocities $\mathbf{v}$ by comoving ones $\mathbf{u}$ and $\mathbf{x}$:

$$\begin{aligned} \mathbf{r} &= a(t) \cdot \mathbf{x} \\ \mathbf{v} &= \dot{a}(t) \cdot \mathbf{x} + \mathbf{u} \\ \nabla_\mathbf{r} &= \nabla_\mathbf{x}/a \end{aligned} \tag{3.17}$$

We note that $\mathbf{u}$ is a peculiar velocity which comes on top of the Hubble flow $\dot{a}(t) \cdot \mathbf{x}$. The density, velocity and gravitational fields can be split into their homogeneous mean values and small perturbations, $\delta\rho_\mathrm{m}$, $\delta\mathbf{u} = \delta\mathbf{v}/a$ and $\delta\phi$. An equation of state links density and pressure fluctuations, $\delta p = c_s^2 \delta\rho_\mathrm{m}$ in the case of a barotropic fluid (i.e. pressure p only depends on the

density ρ). Density perturbations can be expressed in terms of the density contrast $\delta(\mathbf{x})$ using Equation (7.3). We define the following notation, which we will use in this section: $\delta_x = \delta(\mathbf{x})$. In this framework, equations governing the perturbation motion in comoving coordinates are (in terms of δ_x):

$$\begin{cases} \dot{\delta}_x + \dfrac{1}{a}\nabla_{\mathbf{x}} \cdot \delta\mathbf{u} = 0 & \text{(Continuity)} \\ \dot{\mathbf{u}} + \dfrac{\dot{a}}{a}\delta\mathbf{u} + \dfrac{c_s^2}{a}\nabla_{\mathbf{x}}\delta_x + \dfrac{1}{a}\nabla_{\mathbf{x}}\delta\phi = 0 & \text{(Euler)} \\ \nabla^2_{\mathbf{x}}\delta\phi = 4\pi G a^2 \bar{\rho}_{\mathrm{m}}\delta_x & \text{(Poisson)} \end{cases} \tag{3.18}$$

We recall that $\dot{f}$ denotes the time derivative of f. By combining the above system of equations (3.18), we can derive a single second-order differential equation:

$$\ddot{\delta}_x + \underbrace{2\frac{\dot{a}}{a}\dot{\delta}_x}_{\textit{Hubble drag}} - \frac{c_s^2}{a^2}\nabla^2_{\mathbf{x}}\delta_x - \underbrace{4\pi G\bar{\rho}_{\mathrm{m}}\delta_x}_{\textit{gravitation}} = 0 \tag{3.19}$$

The second term in the above equation is the *Hubble drag*, which tends to attenuate the growth of perturbations due to the expansion of the Universe. It competes with the gravitational term of the Poisson equation (last term in the equation above) which cause perturbations to grow through gravitational instability. We can translate the evolution equation Equation (3.19) in Fourier space using $\nabla^2_{\mathbf{x}} \to -k^2$ in the Fourier transformation:

$$\ddot{\delta}_k + 2\frac{\dot{a}}{a}\dot{\delta}_k + \left[\frac{k^2c_s^2}{a^2} - 4\pi G\bar{\rho}_{\mathrm{m}}\right]\delta_k = 0 \tag{3.20}$$

The above equation allows us to define the Jeans length λ_J (or its corresponding wavenumber k_J) which is the scale at which pressure and gravitational forces are equal (cancelling the last term):

$$k_J = \frac{\sqrt{4\pi a^2 G\bar{\rho}_{\mathrm{m}}}}{c_s}, \tag{3.21}$$

$$\lambda_J = \frac{2\pi a}{k_J} = c_s\sqrt{\frac{\pi}{G\bar{\rho}_{\mathrm{m}}}} \tag{3.22}$$

The Jeans length defines two domains of solutions:

- $k < k_J$: pressure cannot withstand gravity, perturbations can grow under gravitation.
- $k > k_J$: perturbations will not grow but oscillate as sound waves propagating at the sound speed.

For baryons, the Jeans length defines the scale at which pressure balances gravity. After recombination, only perturbations with $k < k_J$ can increase and we can approximate the sound speed of baryons to that of a non-relativistic mono-atomic gas:

$$c_s = \sqrt{\frac{5k_BT}{3m_p}} \tag{3.23}$$

where m_p is the mass of the proton and k_B the Boltzmann's constant. We can derive the corresponding Jeans length from Equation (3.23):

$$\lambda_J \approx 0.01(\Omega_{b,0}h^2)^{-1/2} \sim 0.67 \text{ Mpc} \tag{3.24}$$

and the corresponding Jeans mass, considering the mass in a sphere of radius $\lambda_J/2$:

$$M_J = \frac{4}{3}\pi\bar{\rho}_{(m,0)}\left(\frac{\lambda_J}{2}\right)^3 = 1.5 \cdot 10^5(\Omega_{b,0}h^2)^{-1/2}\ M_\odot \tag{3.25}$$

After recombination the Jeans mass is comparable to the mass of a globular cluster. For comparison, prior to recombination, photons are coupled with baryons and the corresponding sound speed is given in Equation (1.37). In this case the corresponding Jeans mass is:

$$M_J = 1.5 \cdot 10^{16}(\Omega_{b,0}h^2)^{-2}\ M_\odot \tag{3.26}$$

Altogether, before recombination the Jeans mass (or λ_J) is so high that perturbations cannot grow, which prevents any structure formation. At recombination, the photons decouple from baryons, which dramatically reduces the pressure, leading to a huge drop in the Jeans mass. Perturbation can grow and structure formation starts.

We consider only those perturbations that can grow with time in the linear regime ($k \ll k_J$) to eventually lead to gravitational collapse and hence large scale structure formation. The pressure component becomes negligible in Equation (3.19) and we end up with:

$$\ddot{\delta}_x + 2\frac{\dot{a}}{a}\dot{\delta}_x - 4\pi G\bar{\rho}_\text{m}\delta_x = 0 \tag{3.27}$$

The linear growth equation (3.27) can be solved by considering the different cosmological epochs – radiation, matter or dark energy dominance – seen in the first chapter in Section 1.6.

Radiation-dominated

During this period, the Hubble parameter is dominated by radiation, and $\bar{\rho}_\text{m}$ is negligible compared to H ($4\pi G\bar{\rho}_\text{m}\delta_x \ll \dot{a}/a$). Thus, Equation (3.27) can be simplified into:

$$\ddot{\delta}_x + 2\frac{\dot{a}}{a}\dot{\delta}_x = 0 \tag{3.28}$$

As the scale factor at this epoch evolves as $a \propto t^{1/2}$ (see Table 1.1) we find the following solution:

$$\delta_x(t) = A_1 + A_2\ln(t) \tag{3.29}$$

where A_1 and A_2 are integration constants. Perturbations increase only slowly (logarithmically) during the radiation-dominated period.

Matter-dominated

In a matter-dominated era, the scale factor evolves as $a(t) \propto t^{2/3}$ (see Table 1.1). In this case, the solution of Equation (3.27) is a linear combination of growing modes $D_+(t)$ and decaying modes $D_-(t)$:

$$\delta_x(t) = D_+(t)\delta_x^+(0) + D_-(t)\delta_x^-(0) \tag{3.30}$$

with the time dependence of D_+ and D_- as:

$$D_+(a) \propto a \propto t^{2/3}, \quad D_-(a) \propto a^{-3/2} \propto t^{-1}. \tag{3.31}$$

The growing mode D_+ evolves proportionally to the scale factor a, and the decaying mode is inversely proportional to time. Thus, from now on we consider only the growing mode D_+ which may lead to gravitational instability eventually. This growing mode is also known as the *linear growth factor* D. Since matter dominance lasts for 10 billion years (while the radiation-dominated period lasts only for $\sim 50,000$ years), thanks to the scaling of the growing mode $\propto t^{2/3}$, the structures of the Universe grow significantly during this period. The growth of structures is described generically by the *linear growth rate of structure* f defined by:

$$\boxed{f \equiv \frac{d\ln(D)}{d\ln(a)}} \tag{3.32}$$

From this definition, we can show that the growing solution $\delta_x(t) = D(t)\delta_x(0)$ satisfies:

$$\dot{\delta}_x(t) = fH\delta_x(t) \tag{3.33}$$

Using the continuity equation in Equation (3.18), we see that the linear regime is characterised by a linear coupling between the density and velocity fields proportional to the linear growth rate:

$$\boxed{\nabla_{\mathbf{x}} \cdot \delta_{\mathbf{u}} = -aHf\delta_x} \tag{3.34}$$

Note that in the simple case of matter-domination $D_+(a) \propto a$ and hence $f = 1$.

Dark energy domination

To describe dark energy, we consider the case of of a cosmological constant. Similarly to the case of radiation domination, $\bar{\rho}_{\rm m}$ is negligible compared to H, the latter being dominated by dark energy ($4\pi G\bar{\rho}_{\rm m}\delta_x \ll \dot{a}/a$). The evolution of perturbation is therefore given by Equation (3.27). With the scale factor increasing exponentially with time $a \propto e^{Ht}$ (see Table 1.1) we find the following solution:

$$\delta_x(t) = C_1 + C_2 e^{-2Ht} \underset{t\to\infty}{\longrightarrow} \text{constant} \tag{3.35}$$

where C_1 and C_2 are integration constants. The second term decreases exponentially and quickly becomes negligible and density perturbations stop growing.

In summary, we have seen different evolution of linear density perturbations at different epochs of the Universe:

$$\delta_x(t) \propto \begin{cases} \log(t) \propto \ln(a) & \text{(radiation dominated)} \\ t^{2/3} \propto a & \text{(matter dominated)} \\ \text{constant} & (\Lambda \text{ dominated}) \end{cases} \tag{3.36}$$

Large-scale structures were formed mainly during the era of matter domination, when the linear growth rate of structure f describes the effectiveness of gravitational attraction in comparison

to the expansion of the Universe. Linder (2003) shows that f can be parametrised as a function of $\Omega_{\rm m}$. For a flat Universe with matter and a cosmological constant ($\Omega_{\rm m} + \Omega_\Lambda = 1$), a good parametrisation is $f(\Omega_{\rm m}) = \Omega_{\rm m}^\gamma$ where γ is the growth index related to the equation of state of dark energy by:

$$\gamma = \frac{3(1 - w_{DE})}{5 - 6w_{DE}} \tag{3.37}$$

For the ΛCDM model, $w_{DE} = -1$ and we obtain $\gamma = 5/9 \approx 0.56$. Therefore, probing the linear growth rate of structures can directly constrain the nature of dark energy and deviations from general relativity.

3.1.4 Non-linear evolution of perturbations: the gravitational collapse

In the previous section, we saw the evolution of perturbations in the linear regime $\delta \ll 1$, where typical scales are larger than $\sim 4050{\rm Mpc}/h$. At smaller scales, we enter the non-linear regime $\delta \gg 1$ and perturbation theory is no longer valid. However, in the quasi-linear regime, where $\delta \sim 1$, perturbation theory remains valid to describe the evolution of perturbations. This corresponds to the intermediate scales $\sim 20 - 50{\rm Mpc}/h$. At these scales, the density field looses its Gaussian properties due to the mode-coupling in equations Equation (3.20). Consequently, higher-order moments can be used to fully describe the density field. In this manuscript, we will not go into the details of higher-order perturbation theory for the quasi-linear regime. Many textbooks and papers describe the subject in detail, such as Bernardeau et al. (2002), Peter et al. (2013), Scoccimarro (2004), Taruya et al. (2012). To complete our description of structure formation, we need to go beyond the linear and quasi-linear regimes and address the evolution of overdensities at small-scales $< 20{\rm Mpc}/h$ where perturbations are highly non-linear $\delta \gg 1$ and eventually collapse under the effect of gravity. In general, in the non-linear regime, there are no analytical solutions to the equation of motion, and we need to use computer simulations to track the evolution of gravitational dynamics. However, using simple assumptions, such as spherical symmetry of the system, it is possible to build analytical models. In the next section, we will review the spherical collapse model.

3.1.4.1 Spherical collapse

The spherical collapse model, first introduced by (Gunn & Gott, 1972), describes the evolution of an initial spherical perturbation under the effect of gravitation. This model makes several assumptions:

- the Universe is homogeneous, with the exception of a single, top-hat, spherical perturbation.
- the Universe is matter-dominated, following an Einstein de Sitter (EdS) cosmology, $\Omega_{\rm m} = 1$, $\Omega_\Lambda = 0$.
- we consider only a collisionless fluid of dark matter.

In this model, an overdensity is seen as a large number of individual, thin mass shells, like onion shells. We can apply the Birkhoff's theorem to describe the evolution of a single mass shell of radius r:

$$\frac{d^2r}{dt^2} = -\frac{GM}{r^2}. \tag{3.38}$$

M is the mass enclosed in the shell and can be expressed as follows:

$$M = \frac{4\pi}{3} r^3 \rho_{\rm m} = \frac{4\pi}{3} r^3 \Omega_{\rm m} \rho_{\rm c} \tag{3.39}$$

where $\rho_{\rm c}$ critical density is defined in Equation (1.19), and $\Omega_{\rm m} = 1$ for an EdS universe. If we consider that the enclosed mass is independent of time before shell crossing, we can integrate Equation (3.38) and obtain:

$$\frac{1}{2}\left(\frac{dr}{dt}\right)^2 - \frac{GM}{r} = E \tag{3.40}$$

where E is the total energy of the system. The system is gravitationally bound, which implies the collapse of the spherical perturbation, when $E < 0$. In this case, the motion of a mass shell can be parametrised as follows:

$$\begin{aligned} r &= R_m(1 - \cos(\tau)) \\ t &= t_{\rm m}(\tau - \sin(\tau)) \end{aligned} \tag{3.41}$$

with $R_m = GM/2|E|$, $t_{\rm m} = GM/(2|E|)^{3/2}$ and $\tau \in (0, 2\pi)$. R_m and $t_{\rm m}$ are linked by the simple relation $R_m^3 = GMt_{\rm m}^2$. We can expand the above solution to track the evolution of the perturbation. In the linear regime, we use the Maclaurin expansions for $\cos(\tau)$ and $\sin(\tau)$:

$$\begin{aligned} \lim_{\tau\to 0}(r(\tau)) &= R_m\left(\frac{\tau^2}{2} - \frac{\tau^4}{24}\right) \\ \lim_{\tau\to 0}(t(\tau)) &= t_{\rm m}\left(\frac{\tau^3}{6} - \frac{\tau^5}{120}\right) \end{aligned} \tag{3.42}$$

To first order $r = R_m\tau^2/2$ and $t = t_{\rm m}\tau^3/6$. We can express r as a function of t at next order by combining the two equations above (Equation (3.42)):

$$R_{lin}(t) = \frac{1}{2}(6t)^{2/3}(GM)^{1/3}\left[1 - \frac{1}{20}\left(\frac{6t}{t_{\rm m}}\right)^{2/3}\right]. \tag{3.43}$$

Solving the evolution equation for the background gives the following solution for the background scale factor:

$$R_{bg}(t) = \frac{1}{2}(6t)^{2/3}(GM)^{1/3} \tag{3.44}$$

Using the above equations and the conservation of mass ($\rho_{\rm m} R_{lin}^3 = \bar{\rho}_{\rm m} R_{bg}^3$), the density contrast $\delta_{\rm sc}$ of the spherical overdensity can be derived as a function of t:

$$1 + \delta_{\rm sc}(t) \equiv \frac{\rho_{\rm m}}{\bar{\rho}_{\rm m}} = \frac{R_{bg}^3}{R_{lin}^3} = \left[1 - \frac{1}{20}\left(\frac{6t}{t_{\rm m}}\right)^{2/3}\right]^{-3} \tag{3.45}$$

Initially, when $t \ll t_{\rm m}$ (i.e. $\delta_{\rm sc}$ is in the linear regime), we can expand the right-hand side to linear order ($(1+x)^n \underset{x\ll 1}{\longrightarrow} 1 + nx$) leading to:

$$\delta_{lin} \simeq \frac{3}{20}\left(\frac{6t}{t_{\rm m}}\right)^{2/3} \tag{3.46}$$

We simply find the evolution of perturbations in the linear regime, $\delta_{\rm sc} \propto t^{2/3}$, as we have already seen in the case of a matter-dominated Universe ($D \propto t^{2/3}$ see Equation (3.31)). The spherical overdensity extends until $t = \pi t_{\rm m}$, where the radius reaches its maximum $r_{\rm ta} = 2R_m$ (for $\tau = \pi$). It then turns around and collapses, reaching $r = 0$ at $t = 2\pi t_{\rm m}$ for $\tau = 2\pi$. From Equation (3.46) we can compute the value of the linear at turnaround ($t = \pi t_{\rm m}$):

$$\delta_{lin}^{ta} = \frac{3\pi}{20}(6\pi)^{2/3} \simeq 1.06 \tag{3.47}$$

and at collapse time (at linear order), $t = 2\pi t_{\rm m}$ ($\tau = 2\pi$):

$$\delta_c \equiv \delta_{lin}^{col} = \frac{3}{20}(12\pi)^{2/3} \simeq 1.686. \tag{3.48}$$

Here we introduce the *critical density contrast*, δ_c which is time independent. It corresponds to the linear density threshold above which a spherical overdensity in an EdS universe collapses into a single point becoming infinitely dense, as can be see, from Equation (3.41).

Beyond the linear regime, we can also compute the value of the non-linear density contrast at turnaround ($t = \pi t_{\rm m}$) using Equation (3.41) and Equation (3.44):

$$\delta_{\rm sc}^{ta} = \frac{R_{bg}^3(\pi t_{\rm m})}{r_{ta}^3} - 1 = \frac{9\pi^2}{16} - 1 \simeq 4.55 \tag{3.49}$$

In practice the collapse is never perfectly spherical and the overdensity does not reach infinite density. Instead, during collapse, shell crossings occur, leading to exchanges of angular momentum. The system relaxes toward an equilibrium when the gravitational energy equals twice the kinetic energy $E = 2K + U = 0$, the so-called *virial equilibrium*. This implies that the radius at virialisation is half the maximum radius reached by the overdensity $r_{vir} = r_{\rm ta}/2$. At this moment $t = 2\pi t_{\rm m}$, the density inside the sphere has increased by a factor 2^3. Since in a matter-dominated universe $a \propto t^{2/3}$ and $\bar{\rho}_{\rm m} \propto a^{-3}$, the background density of the Universe has decreased by a factor of 2^2. Therefore, the virialised overdensity corresponding to $t = 2\pi t_{\rm m}$ is:

$$\delta_{\rm vir} = \delta_{\rm sc}^{ta} \cdot 2^3 \cdot 2^2 \simeq 177 \tag{3.50}$$

After virialisation of the system, we finally obtain a virialised object of radius R_{vir} called a *halo*. Figure 3.1 shows a sketch of the evolution of the spherical collapse model.

This simple phenomenological model provides us with useful keys to understanding the non-linear evolution of spherical perturbations. When perturbations reach a density $\delta_{\rm vir} \simeq 177$ times higher than the mean density of the Universe, they form gravitationally-bounds halos composed of dark matter (as in this model we assume a collisionless dark matter fluid). For EdS cosmology the value of $\delta_{\rm vir}$ is exactly known. In the case of the standard ΛCDM cosmological model, due to the increase in the expansion rate, the background density is lower at the time of virialisation, implying a larger value of $\delta_{\rm vir}$. Other models including dark energy can be constructed, but there is no analytical solution to spherical collapse. Weinberg & Kamionkowski (2003) used numerical integration and find the following evolution of $\delta_{\rm vir}$ for flat cosmological models with a cosmological constant that has an equation of state of the form $w_{DE} = constant$:

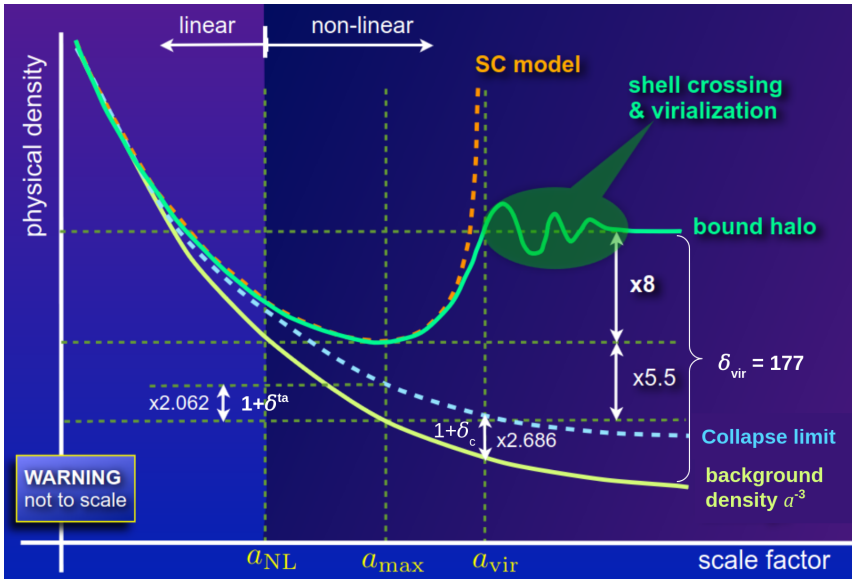

Figure 3.1: *Schematic representation of the spherical collapse model. Figure taken from Frank van den Bosch lectures.*

$$\begin{aligned} \delta_{\rm vir} &\simeq 177\left(1+b_1\theta^{b_2}(z)\right) \\ \theta &\equiv \frac{1-\Omega_{\rm m}(z)}{\Omega_{\rm m}(z)} \\ b_1 &= 0.399 - 1.309(|w_{DE}|^{0.426}-1) \\ b_2 &= 0.941 - 0.205(|w_{DE}|^{0.938}-1) \end{aligned} \tag{3.51}$$

The evolution of the critical density $\delta_{\rm vir}$ with different values of w_{DE} is shown in Figure 3.2 considering a fixed cosmological background from Table 1.2 to obtain the evolution of $\Omega_{\rm m}(z)$. We see that the evolution of $\delta_{\rm vir}$ in cosmological models with a cosmological constant converges to the EdS model at high redshift $z \geq 1$, and differs from $\delta_{\rm vir} = 177$ at lower redshift where most large scale structures are already formed. Consequently, the spherical collapse model in an EdS universe is a good approximation for the ΛCDM cosmology at $z > 1.5$. This is to be expected, since Λ is subdominant during the early stages of collapse when the Universe was dominated by matter. By the time dark energy has become non-negligible, the collapse regions are already much denser than the density background and have largely decoupled from the Hubble flow.

3.1.4.2 The mass function of collapsed objects

The spherical collapse model predicts that a region with volume $V = 4/3\pi R^3$ with a density contrast δ exceeding the spherical collapse threshold δ_c will collapse into a halo of mass $M = \bar{\rho}_{\rm m}(1+\delta_{\rm vir})V$. Press & Schechter (1974) derive the abundance of collapsed objects by considering that, for a Gaussian random field of mean density $\bar{\rho}$, we can statistically count the number of regions with an overdensity $\delta > \delta_c$. Suppose that, at a given redshift z, we smooth the random Gaussian field of density fluctuations over cells of radius R containing on average a mass of $M = 4/3\pi R^3(1+\delta)\bar{\rho}$ and that *all* cells with $\delta > \delta_c$ collapse into halos. Given the linear smoothing, the density field in the cells is also Gaussian with a variance σ_M^2 which can be given

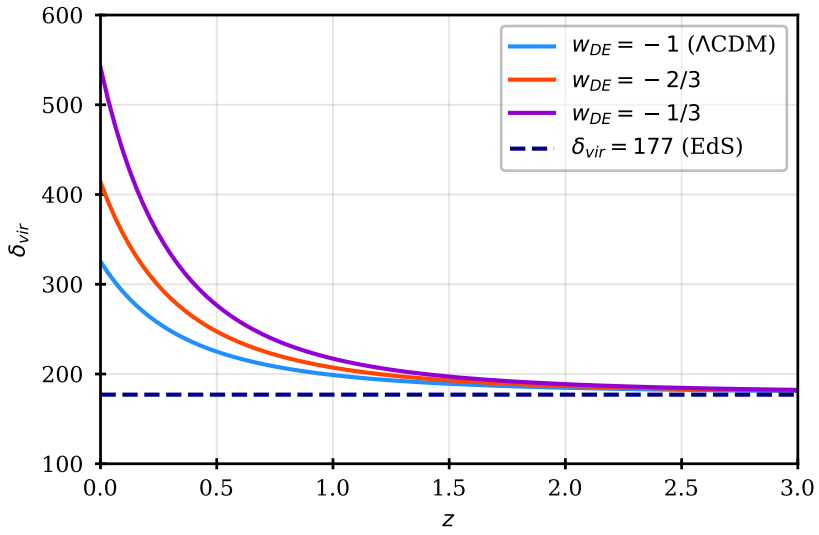

Figure 3.2: *Evolution of the non-linear critical density $\delta_{\rm vir}$ as a function of redshift for cosmological models including dark energy as a cosmological constant. Different values of w_{DE} are shown in solid lines for a fixed cosmological background from Table 1.2. The value for $\delta_{\rm vir}$ in an EdS universe is shown by the dashed-line. The line $w_{DE} = -1$ represents the ΛCDM model. This figure is inspired from Weinberg & Kamionkowski (2003).*

by Equation (3.13). Therefore, the probability for a cell to exceed the overdensity threshold δ_c is given by:

$$P(M|\delta > \delta_c) = \frac{1}{\sqrt{2\pi}\sigma_M} \int_{\delta_c}^{\infty} \exp\left(-\frac{\delta_M^2}{2\sigma_M^2}\right) d\delta_M = \frac{1}{2}\mathrm{erfc}\left(\frac{\delta_c}{\sqrt{2}\sigma_M}\right) \tag{3.52}$$

where $\mathrm{erfc}(x)$ is the complementary error function. $P(M|\delta > \delta_c)$ represents the fraction of collapsed regions, i.e. halos with a mass M greater than that corresponding to δ_c. We use the above formula to obtain the fraction of objects whose mass lies between $[M, M + dM]$:

$$dP(M) = \left|\frac{\partial P(M|\delta > \delta_c)}{\partial M}\right| dM \tag{3.53}$$

which leads formally to the halo mass function, i.e. the number density of halos per mass interval dM:

$$n_h(M)dM = \frac{\bar{\rho}}{M}\left|\frac{\partial P(M|\delta > \delta_c)}{\partial M}\right| dM \tag{3.54}$$

This mass function does not appear to be properly normalised, since integration over all the mass included in the halos only recovers half of the total mass (Press & Schechter, 1974). In their paper, Press & Schechter argued, without a proper demonstration, that matter in initially under-dense regions will eventually be accreted by the collapsed objects, doubling their masses without changing the shape of the mass function. Thus, they introduced a "fudge factor" 2 to ensure mass conservation. A rederivation based on excursion set theory revealed the true origin of this factor Bond et al. (1991). The Press-Schechter (PS) mass function is usually given in terms of logarithmic mass bins, hence $n_h(M) = \frac{n_h(M,z)}{M}\frac{dM}{d\ln(M)}$ and Equation (3.54) become

(with the fudge factor 2):

$$n_h(M)dM = \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}\delta_c}{M^2\sigma_M}\exp\left(-\frac{\delta_c^2}{2\sigma_M^2}\right)\left|\frac{d\ln(\sigma_M)}{d\ln(M)}\right|dM \tag{3.55}$$

We introduce the variable $\nu = \delta_c/\sigma_M$ and the multiplicity function:

$$f_{PS}(\nu) = \sqrt{\frac{2}{\pi}}\nu\exp\left(\frac{-\nu^2}{2}\right) \tag{3.56}$$

which gives the fraction of mass associated with halos in a unit range of $\ln(\nu)$. Thus, we can rewrite the PS mass function in a more compact form:

$$\boxed{n_h(M)dM = \frac{\bar{\rho}}{M^2}f_{PS}(\nu)\left|\frac{d\ln(\nu)}{d\ln(M)}\right|dM} \tag{3.57}$$

If we define a characteristic mass M^* such that $\sigma_{M^*} = \delta_c \Rightarrow \nu(M^*) = 1$ we can guess the behaviour of the mass function:

$$n_h(M)dM \propto \begin{cases} M^{\alpha-2} & \text{for } M \ll M^* \\ \exp(-\nu^2/2) & \text{for } M \gg M^* \end{cases} \tag{3.58}$$

where $\alpha = d\ln(\sigma_M)/d\ln(M)$ is a function of cosmology. For ΛCDM, $\alpha \to 0$, therefore at small mass the halo mass function decreases $\propto M^{-2}$, while at large mass the abundance of halo decreases exponentially. We have not mentioned the time dependence of the above mass function. The evolution of the mass function with time is related that of $\delta_c(t)$. As we mentioned in the previous section, in ΛCDM, $\delta_c \propto D(t^{-1})$ and thus decreases with time (as $D(t)$ increases, see Equation (3.31)) which means that the characteristic mass M^* grows as a function of time. As a result, more and more massive halos form over time.

The simplicity of the PS approach relies on rough approximations and hazardous extrapolation of the linear theory. Surprisingly, the PS approach remains fairly accurate over a range of masses and redshifts to reproduce the abundances of objects obtained with numerical simulations (see Figure 3.3). In a more general context, the multiplicity function $f_{PS}(\nu)$ for the PS theory that characterises the mass function in Equation (3.57) can be replaced with more general $f(\nu)$ functions. As an example, a popular and more accurate model for predicting the shape of the halo mass function is the Sheth-Tormen (ST) mass function, based on the ellipsoidal collapse model (EC) (Sheth et al., 2001, Sheth & Tormen, 1999). The derived form of the multiplicity function in the ST model is given by:

$$f_{EC}(\nu) = A\left(1 + \frac{1}{\tilde{\nu}^{2q}}\right)f_{PS}(\tilde{\nu}) \tag{3.59}$$

where $\tilde{\nu} = 0.84\nu$, $A \approx 0.322$, $q = 0.3$ are derived from numerical resolution. Figure 3.3 compares the halo mass function from numerical N-body simulations of dark matter particles in the Millenium simlulation (Springel et al., 2005) with those predicted by the PS and ST theories. The ST model provides a very good representation of the simulation results. Although the PS prediction is not very accurate, it provides a valuable first approximation of the mass function (the prediction being worse with increasing redshift).

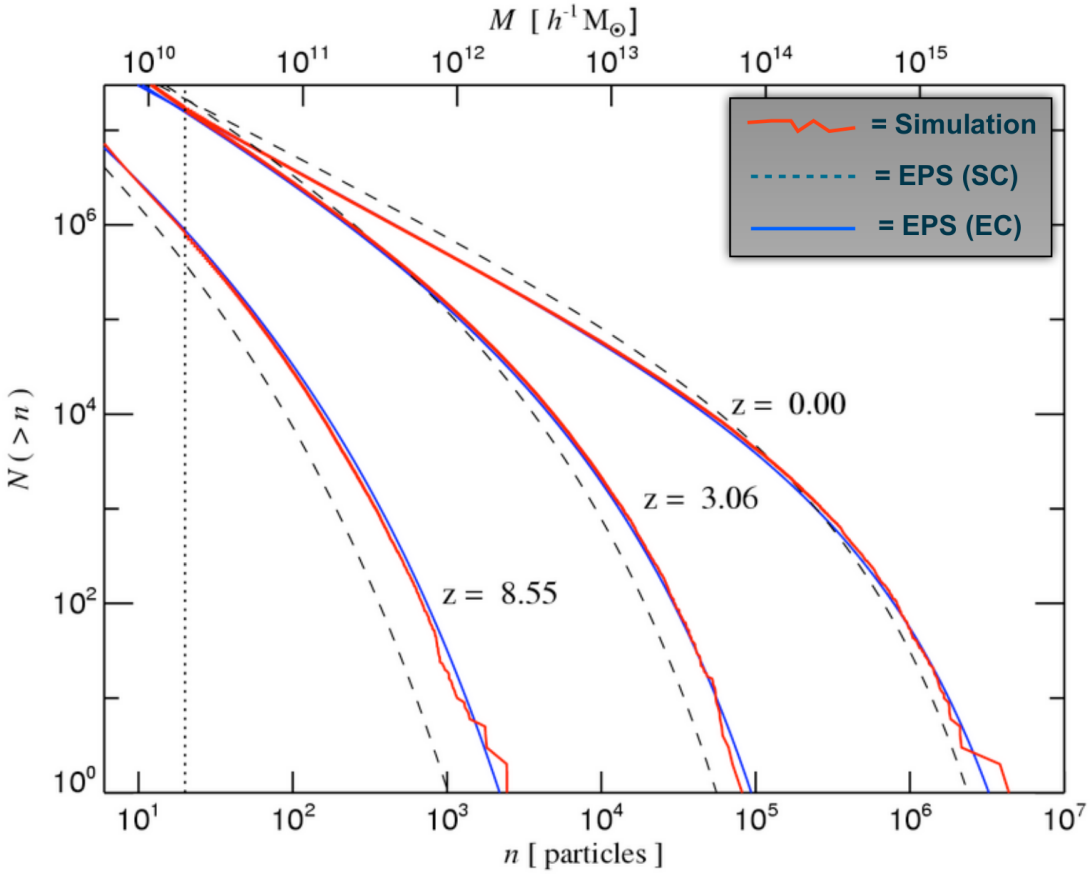

Figure 3.3: *Number of dark matter halos as function of mass at three different redshifts. The red line is the simulation output. The dashed-line is the prediction of the Press & Schechter theory. The blue line is the prediction of Sheth & Tormen's analytical fitting function. This figure is taken from Frank van der Bosch's lecture, originally from V. Springel.*

Note: *N-body simulations are particle based. Therefore, the identification of halos in the simulation box depends of the halo finder algorithm, which can have an impact on the shape of the halo mass function. The above results rely on a "Friends-Of-Friends" (FOF) algorithm (Davis et al., 1985, Huchra & Geller, 1982) .*

3.1.5 Internal structure of dark matter halos

So far, we discussed how primordial density fluctuations in the dark matter field evolve into virialised structures. We will now focus on the internal structure of dark matter halos.

Density profile

A dark matter (DM) halo can be modelled to a first approximation as a spherical object of radius R_h enclosing a mass M_h. We first focus on the density profile of DM halos. We have seen that viralized structures form as a function of their density environment around the initial overdensity. We might expect the density profile of DM halos to depend on their specific formation history.

However, contrary to this expectation, DM halos exhibit approximately universal spherically-averaged density profiles, as first demonstrated by Navarro Frenk and White (Navarro et al., 1997). Based on that work, a generalised density profile of dark matter halos can be given by a double power (Zhao, 1996):

$$\rho(r) = \frac{\rho_s}{(r/r_s)^\gamma \left[1 + (r/r_s)^\alpha\right]^{(\beta-\gamma)/\alpha}} \tag{3.60}$$

where (α, β, γ) are power law indexes, r_s is the *scale radius* of the density profile and ρ_s is the density at r_s. This density profile has two different behaviours in the outer or inner regions of the halo:

$$\rho(r) \propto \begin{cases} r^{-\gamma} & \text{if } r \ll r_s \\ r^{-\beta} & \text{if } r \gg r_s \end{cases} \tag{3.61}$$

The α parameter controls the sharpness of the break (see Figure 3.4). Several parametrisations have been studied in the literature. Among them, the Navarro-Frenk-White (NFW) profile (α, β, γ) = (1, 3, 1)) which provides a good description of the density around virialised halos in numerical simulations:

$$\boxed{\rho(r) = \frac{\rho_s}{(r/r_s)\left[1 + (r/r_s)\right]^2}} \tag{3.62}$$

The NFW profile is represented by the black line in Figure 3.4.

The above profile can be used to determine the density of any halo, ρ_h, at any moment of its evolution. At a given time, ρ_h is related to the halo matter overdensity, Δ_h and to cosmological parameters by $\rho_h = \Delta_h \bar{\rho}_\mathrm{m} = \Delta_h \rho_\mathrm{c} \Omega_\mathrm{m}$ with $\bar{\rho}_\mathrm{m}$ the mean density of matter in the Universe and ρ_c the critical density defined in Equation (1.19) at that time. Halo mass and radius are linked to halo density by:

$$\rho_h = \frac{3M(< R_{\Delta_h})}{4\pi R^3_{\Delta_h}} \tag{3.63}$$

where $M(< R_{\Delta_h}) \equiv M_h$ is the mass of the halo enclosed within the halo radius $R_{\Delta_h} \equiv R_h$. We further introduce the *concentration parameter*:

$$c \equiv \frac{R_h}{r_s} \tag{3.64}$$

which essentially describes how the mass is distributed in the halo profile. The value of the overdensity is $\Delta_h = 177$ in the case of the spherical collapse model at the time of halo formation. However, as the virialisation criterion is not strict, other definitions are also in use in the literature, one of them being $\Delta_h = 200$. The halo properties will depend on the redshift and on the background cosmology. The universal nature of the DM halo density profile is explained by the highly non-linear nature of DM halos, which have gone through a phase of gravitational collapse that has erased information pertaining to their individual formation history. Once a halo is formed, accretion material increases its mass and size, without adding much material to its inner region. The halo radius R_h increases while r_s remains unchanged. Consequently, the concentration parameter $c = R_h/r_s$ should decrease with increasing halo mass. This result

is confirmed by numerous studies using N-body simulations (Ludlow et al., 2014, Prada et al., 2012).

The halo properties can be reformulated in terms of concentration c. Integrating the NFW density profile to obtain the mass enclosed up to a given radius r gives:

$$M(<r) = \int_0^r 4\pi r^2 \rho(r')dr' = 4\pi\rho_s r_s^3 g(r/r_s) = 4\pi\rho_s r_s^3 g(cs) \tag{3.65}$$

where $s = r/R_h$ and:

$$g(x) = \ln(1+x) - \frac{x}{1+x}. \tag{3.66}$$

The mass enclosed within a given radius r can be expressed as function of the halo mass M_h and the concentration:

$$\frac{M(r)}{M_h} = \frac{g(cs)}{g(c)}. \tag{3.67}$$

with $M_h = 4\pi\rho_s R_h^3 g(c)$. A useful quantity derived from the NFW profile is the circular (i.e. Keplerian) velocity:

$$V_c(r) = \sqrt{\frac{GM(r)}{r}} \tag{3.68}$$

Similarly to Equation (3.67) we can express the circular velocity for any radius in terms of velocity at the halo radius:

$$\frac{V_c(r)}{V_c(R_h)} = \sqrt{\frac{g(cs)}{sg(c)}} \tag{3.69}$$

The circular velocity reaches a maximum at $V_{c,max} \simeq 0.465 V_h \sqrt{c/g(c)}$ corresponding to a radius $r_{v,max} \simeq 2.163 r_s$ (Navarro et al., 1997). This quantity is useful for deriving the concentration parameter in simulations. The NFW profile has been widely used to describe the density profile of DM halos in cosmology. Other density profile models also showed good agreement with simulation. Among them, the Einasto profile, which follows a decreasing exponential form:

$$\rho(r) = \rho_{-2} \exp\left(-2n\left[\left(\frac{r}{r_{-2}}\right)^{1/n} - 1\right]\right) \tag{3.70}$$

where r_{-2} corresponds to the radius for which $\rho = \rho_{-2}$ and is the equivalent of r_s for an NFW profile. n is the Einasto index defining the steepness of the power law and typical values from simulations span the range $4.54 < n < 8.33$ (Navarro et al., 2004). Figure 3.4 shows the Einasto profile in blue for $n = 5$. Its shape is quite comparable to that of the NFW profile.

Substructure of dark matter halos

In the course of their history, dark matter halos, once formed, evolve under the influence of gravity, accumulate mass and merge with each other. A small halo that merges with a much larger one will most likely become a sub-halo orbiting in the potential well of its host. Over time, the sub-halo is subjected to strong tidal forces, resulting in a loss of mass. It is also subject to dynamic friction, which causes it to lose energy and angular momentum. In other words, the longer the sub-halo remains in orbit, the greater the loss of mass. As a result, halos that assemble earlier will be more likely to destroy their sub-halos, while halos that assemble later will have more substructures. The survival of a sub-halo also depends on its mass and concentration

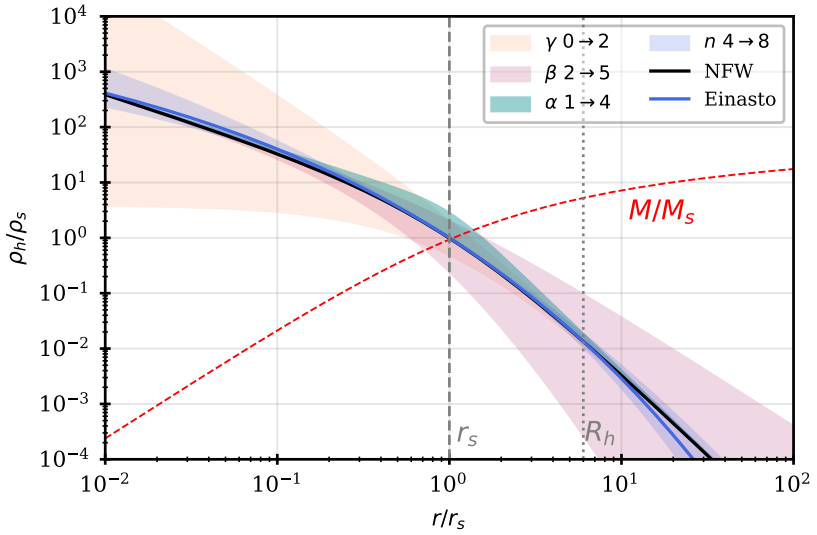

Figure 3.4: *The density profile of dark matter halos from the double power law prescription in Equation (3.60). The black line represents the NFW profile* $(\alpha, \beta, \gamma) = (1, 3, 1)$. *The shaded regions represent the following variations of* (α, β, γ) *values:* $1 \to 4$, $3 \to 5$ *and* $0 \to 2$ *in green, purple and yellow, respectively. The grey dashed vertical line is the scale radius* r_s *where the density is* ρ_s. *The grey dotted line is an arbitrary halo radius where we cut the halo profile, otherwise the enclosed mass (red dashed-line) would diverge. For comparison, the blue shaded region represents the Einasto profile for* n $4 \to 8$ *and the blue line corresponds to* $n = 5$.

relative to those of the host halo. Sub-halos with a mass greater than a few percent of that of the main halo should merge rapidly, while objects with significantly less mass will survive for long periods (Diemand et al., 2007). In the ΛCDM framework, massive halos assemble later and should therefore host a larger fraction of sub-halos than less massive halos (Giocoli et al., 2008). These results are mainly based on high-resolution numerical simulations, which allow us to trace the history of DM matter halos over time and identify substructures. These numerical simulations are the subject of the next section.

3.2 The Universe in boxes

Numerical simulations in cosmology and astrophysics are essential and widespread for understanding the formation and evolution of structures on all scales, and in particular for exploring dynamics in the non-linear regime where perturbation theory breaks down. They are based on two main ingredients: the initial conditions given by the shape of the initial power spectrum, and the physics governing the evolution of the initial conditions over time. In general, two types of simulations are used in cosmology: *N-body simulations*, which consider cold, collisionless dark matter particles evolving under gravity in boxes with periodic boundary conditions to mimic large-scale homogeneity and isotropy, and *hydrodynamic simulations* (full-physics) that take complex baryonic physics into account.

3.2.1 N-body simulations

In N-body cosmological simulations, the dark matter field is represented by a set of non-relativistic particles (points) of mass m interacting only by gravitational forces, and collisions between particles are not taken into account. Their advantage is that they can be carried out on large volumes, $\sim$ a few Gpc^3, as they only resolve Newtonian gravity and ignore baryonic interactions on small scales (a few Mpc/h). Indeed, the impact of baryonic effects is weak on large scales and baryons simply follow the dynamics of dark matter, so these effects can be neglected when studying large volumes as in the analysis of large-scale structures. The evolution of these particles can be described by considering their phase-space distribution function $f(\mathbf{x}, \mathbf{p}, t)$, which operates in a 6+1 dimensional space (3 comoving coordinates $\mathbf{x}$, 3 comoving momentum components $\mathbf{p}$, physical time t). The evolution of this distribution function is governed by the *Vlasov-Poisson* equations:

$$\begin{aligned} \frac{\mathrm{d}f}{\mathrm{d}t} &= \frac{\partial f}{\partial t} + \frac{\mathbf{p}}{ma^2} \cdot \frac{\partial f}{\partial \mathbf{x}} - \frac{m}{a} \nabla_{\mathbf{x}} \Phi(\mathbf{x}) \cdot \frac{\partial f}{\partial \mathbf{p}} = 0 \\ \triangle \Phi(\mathbf{x}) &= \frac{4\pi G m}{a} \left(\int d^3\mathbf{p} f(\mathbf{x}, \mathbf{p}, t) - \frac{1}{V} \int \int d^3\mathbf{x}\, d^3\mathbf{p} f(\mathbf{x}, \mathbf{p}, t) \right) \end{aligned} \tag{3.71}$$

where $\triangle = \nabla^2$ is the Laplacian, V is the comoving volume and comoving coordinates $\mathbf{x}$ are defined in Equation (3.17). Note that $\mathbf{p} = ma\mathbf{v}$ is the comoving momentum related to the peculiar velocity $\mathbf{v} = a\frac{d\mathbf{x}}{dt}$. The number of particles defines the resolution of the simulation: the more particles there are, the better the resolution of the simulation. The first N-body simulations in astrophysics appeared in the 1960s. In the 1970s, (Peebles, 1970, White, 1976) carried out simulations of galaxy clusters with a few dozen of particles. Shortly afterwards, cosmological-scale simulations with a few thousand particles emerged to theoretically model the cosmic web (Aarseth et al., 1979, Centrella & Melott, 1983, Doroshkevich et al., 1980, Efstathiou & Eastwood, 1981, Frenk et al., 1983, Press & Schechter, 1974). Nowadays, state-of-the-art numerical simulations can handle dynamics of $\sim 10^{10} - 10^{12}$ particles, with a particle mass of $\sim 10^8 - 10^{10} M_\odot/h$. Simulations are essential for large-scale structure analyses, as they can be used to estimate measurement errors, generate covariance matrices, test cosmological pipelines, mitigate systematic errors, study the galaxy-halo connection or training cosmological inference emulators. Next-generation surveys require high-volume, high-resolution simulations so that the survey volume can be simulated several times with great accuracy. In DESI, we use the large simulation suite called AbacusSummit that has been specifically designed to meet the scientific requirements of the survey. Obviously, increasing the number of particles also increases computing time. A compromise must therefore be made between resolution and volume in order to generate a simulation that meets the scientific requirements of the next generation survey. At first-order the computation time scales with the number of particles N_p as $\mathcal{O}(N_p^2)$. There are many techniques for numerically evolving particles through gravitational interaction and we describe some of them below.

➢ **Initial conditions** Before solving the equations of gravity, the initial conditions must be defined. Once the cosmological model has been defined, the basic idea is to generate a smooth background, sample it with particles, then add fluctuations given by the initial power spectrum in Equation (3.10). The first challenge is to generate the initial particle distribution. Uniform random sampling generates undesirable structure formation due to sampling noise, even if no

perturbation is imposed. Another solution is to place the particles on a regular grid, but this leads to preferential directions and scales. To solve these issues, Baugh et al. (1995), White (1994) suggested to use a glass-like distribution. Starting from a random uniform distribution, particles are displaced by the gravity solver according to the inverse of the gravitational force, so that particles tend to repel each other until they freeze in comoving coordinates. The final particle distribution shows no preferred direction or scale. Then a Gaussian perturbation field is generated based on the initial linear power spectrum (provided by Boltzmann codes such as CLASS and CAMB (Lesgourgues, 2011, Lewis et al., 2000)). The particles are displaced (slightly) and are assigned initial peculiar velocities according to the input power spectrum. This step is usually performed using Lagrangian perturbation theory at first order (Zel'dovich approximation, (Zel'dovich, 1970)) or higher orders (Hahn et al., 2021). Typically, N-body simulations start around $z \sim 100$, when the perturbations are still linear. The initial power spectrum is shifted to this redshift with the appropriate transfer function $T(k)$ (Equation (3.11)), which depends on the cosmological background.

➢ **The particle-particle (PP) method** is the simplest, but most time-consuming, way to solving the N-body problem. In this case, at each time step, the exact Newtonian gravitational forces between two particles of mass m_i and m_j separated by r_{ij} are given by:

$$F_{ij} = \frac{G m_i m_j}{r_{ij}^2} \tag{3.72}$$

and the peculiar velocity of each particle i is calculated using:

$$\dot{\mathbf{v}}_i + H\mathbf{v}_i = \frac{1}{m_i} \sum_{j \neq i} \mathbf{F}_{ij} \tag{3.73}$$

This is the familiar momentum equation for gravitating systems of particles, written in comoving coordinates, with an additional drag term due to the Hubble expansion. Gravitational forces diverge if two particles are at null separation. In practice, a softening parameter ϵ is introduced to avoid this divergence by changing the denominator r_{ij}^2 by $(r_{ij}^2 + \epsilon^2)^{3/2}$ (Plummer, 1911). This softening corresponds to a smoothing on small scales, i.e. it reduces shot noise. However, ϵ introduces a bias with respect to pure Newtonian dynamics and affects the growth of structure on scales many times larger than ϵ (Garrison et al., 2019). An optimum must therefore be found between shot noise and bias. The PP method is (apart from the softening parameter) highly accurate, with the accuracy depending on the time step. The main disadvantage of this approach is that it scales as $\mathcal{O}(N_p^2)$, making it very difficult to use with a very large numbers of particles. Typically, it can be used with a maximum of 10^6 particles, whereas the largest N-body simulations have up to 10^{12} particles.

➢ **The particle-mesh (PM) method** evaluates the density of particles on a grid (a mesh) using an interpolation kernel (e.g. cell counting, triangular-shaped-cell (TSC) or nearest-grid-point). The gravitational potential is then calculated for each cell using the Poisson equation in Fourier space with Fast Fourier Transforms (FFT). The motion of particles is computed from this gravitational potential with the same interpolation kernel. FFT algorithms scales as $\mathcal{O}(N_c \log(N_c))$, where N_c is the number of cells in the grid, and the interpolation scales as the number of particles $\mathcal{O}(N_p)$ (see Feng et al. (2016) for an optimised use of the PM method,

and Chuang et al. (2019) for its application to the UNIT cosmological simulation). Although this method is very fast, its accuracy deteriorates at scales several times larger than the size of the mesh. Methods have been developed to increase the precision on small scales, such as the adaptive mesh refinement (AMR) which dynamically increases the mesh resolution in high density regions (e.g. RAMSES code (Teyssier, 2002)).

➢ **Tree codes** (Appel, 1985, Barnes & Hut, 1986) are nowadays the most widely used methods for solving N-body problems in cosmology (e.g. GADGET code (Springel et al., 2005) used (in its thrid version) for the MultiDark simulation suite (Klypin et al., 2016)). The tree algorithm organizes the matter distribution along a tree. Particles are gathered in a large cubic cell, which is then subdivided into smaller and smaller cells, until each cell contains one particle. The tree is fixed, and does not need to be recalculated at each time step. Once the tree has been defined, the gravitational potential is calculated by descending the tree cells and performing a hierarchical multipole expansion for sufficiently distant cells. This means that a single force is computed for the centre of mass of the cells, instead of for each particle. The descent stops when the opening angle (cell size over distance) is less than a fixed acceptation angle θ. The force for nearby particles is then computed individually. In this way, the force resolution can be as high as the PP method in very dense regions. This method can be as efficient as $\mathcal{O}(N_p \log(N_p))$ depending on the acceptation angle.

➢ **Hybrid methods** mix a PM method on large scales with a PP method on small-scales (also called particle-particle/particle-mesh (P^3M) (Hockney, 1988)). The idea is to use the efficiency and speed of FFT methods on a mesh without losing accuracy on small scales. The field is decomposed into 2 components, a *far-field* for the long-range force which is calculated with a PM method and a *near-field* for short-range contribution obtained by direct computation of individual interactions between nearby particles (PP method). The TreePM code is an example of these hybrid methods which mixes a tree-code method for dense regions and a PM for under-dense regions. These codes are highly efficient, enabling large simulations to be carried out with very good resolution.

By combining these new algorithms with the exponential growth in computing power during the last decades, it has been possible to increase both volume and mass resolution in simulations. For instance, Euclid Flagship (Potter et al., 2016) simulations evolve $\sim$ 2 trillion particles in a cubic box of 4 Gpc/h length size, with a mass resolution $\sim 10^9 M_\odot$. In DESI we use the AbacusSummit simulations based on the Abacus code described hereafter. As an alternative, DESI analyses also use the UCHUU simulation Ishiyama et al. (2021), which is a 2 Gpc/h length size and very well-resolved, 2.1 trillion particles (12800^3) with a particle mass of $M_{\rm part} = 3.27 \cdot 10^8 M_\odot/h$. This simulation was run produced with the GreeM code (TreePM code Ishiyama et al. (2009, 2012)) and use the ROCKSTAR halo finder (Behroozi et al., 2013b).

AbacusSummit simulations

The AbacusSummit simulations[1] (Maksimova et al., 2021) have been designed to reach the scientific requirements of the DESI survey. It is a large suite of high-accuracy cosmological N-body simulations produced with the Abacus N-body run on the Summit supercomputer at the Oak

[1] https://abacussummit.readthedocs.io/en/latest/index.html

Ridge Leadership Computing Facility. This suite is composed of 150 simulation boxes, covering 97 cosmological models. The `base` simulations have 6912^3 particles with mass $2 \cdot 10^9 M_\odot/h$ in a 2 Gpc/h cubic box. The fiducial cosmology `c000` corresponds to the Planck 2018 results (Planck Collaboration et al., 2020) based on the mean estimates of the TT,TE,EE+lowE+lensing likelihood chains (see Table 1.2). There are 25 `base` mass-resolution boxes with different initial conditions for the fiducial cosmology, a visualisation of which is shown in Figure 3.5. A series of 1883 `small` cubic boxes of 500 Mpc/h length size with the same mass resolution, and additional boxes at both lower and higher resolutions, are also available (see Table 3.1). Table 3.2 shows the cosmological parameters for the fiducial cosmology and four secondary cosmologies. For each of the secondary cosmologies 6 `base` boxes and one `fixedbase` box with fixed initial conditions are available. Other cosmologies can be found on the AbacusSummit website or in (Maksimova et al., 2021). Finally, a set of 52 `base` boxes at different cosmologies covering a wide range of the 8-dimensional parameter space has been created as a basis for training cosmlogical emulators.

Name	PPD	Size (Mpc/h)	Particle mass ($M_\odot/h$)
`base`	6912^3	2000	2×10^9
`highbase`	3456^3	1000	2×10^9
`high`	6300^3	1000	3×10^8
`hudge`	8640^3	7500	5×10^{10}
`fixedbase`	4096^3	1185	2×10^9
`small`	1728^3	500	2×10^9

Table 3.1: *AbacusSummit simulation characteristics. PPD stands for particles per dimension. Particle mass may vary slightly with different cosmologies. The size is the length of the cubic box.*

Description	Ω_b	ω_c	h	$10^9 A_s$	n_s	α_s	N_{ur}	N_ncdm	$10^4\omega_{ncdm}$	$w_{0,fld}$	$w_{a,fld}$	$\sigma_{8,m}$	$\sigma_{8,cb}$
`c000` Baseline ΛCDM	0.02237	0.1200	0.6736	2.0830	0.9649	0.0	2.0328	1	6.4420	-1.0	0.0	0.807952	0.811355
`c001` ow ω_c ΛCDM	0.02242	0.1134	0.7030	2.0376	0.9638	0.0	2.0328	1	6.4420	-1.0	0.0	0.776779	0.780222
`c002` Thawing dark energy	0.02237	0.1200	0.6278	2.3140	0.9649	0.0	2.0328	1	6.4420	-0.7	-0.5	0.808189	0.811577
`c003` $N_\mathrm{eff} = 3.70$	0.02260	0.1291	0.7160	2.2438	0.9876	0.0	2.6868	1	6.4420	-1.0	0.0	0.855190	0.858583
`c004` low $\sigma_{8,m}$ ΛCDM	0.02237	0.1200	0.6736	1.7949	0.9649	0.0	2.0328	1	6.4420	-1.0	0.0	0.749999	0.753159

Table 3.2: *Cosmological parameters for the fiducial `c000` and 4 secondary cosmologies of the AbacusSummit suite of simulations.*

The Abacus code

The Abacus code is a high-performance code designed to perform large-scale cosmological N-body simulations with high precision in the calculation of gravitational forces (Garrison et al., 2021). It is based on an hybrid method whose near field and far field decomposition has a few specific features compared to standard hybrid P^{3}M methods. First, the decomposition between the far and near fields is strict, meaning that every pairwise interaction is given by either the near field or the far field, but never by both. The far-field calculation relies on the concept of multipole expansion rather than calculating the gravitational force on a particle mesh. The gravitational potential is expanded in Taylor series around a location (up to a given polynomial order p). The coefficients of this expansion are related to the multipole moments of the density field in another distant region.

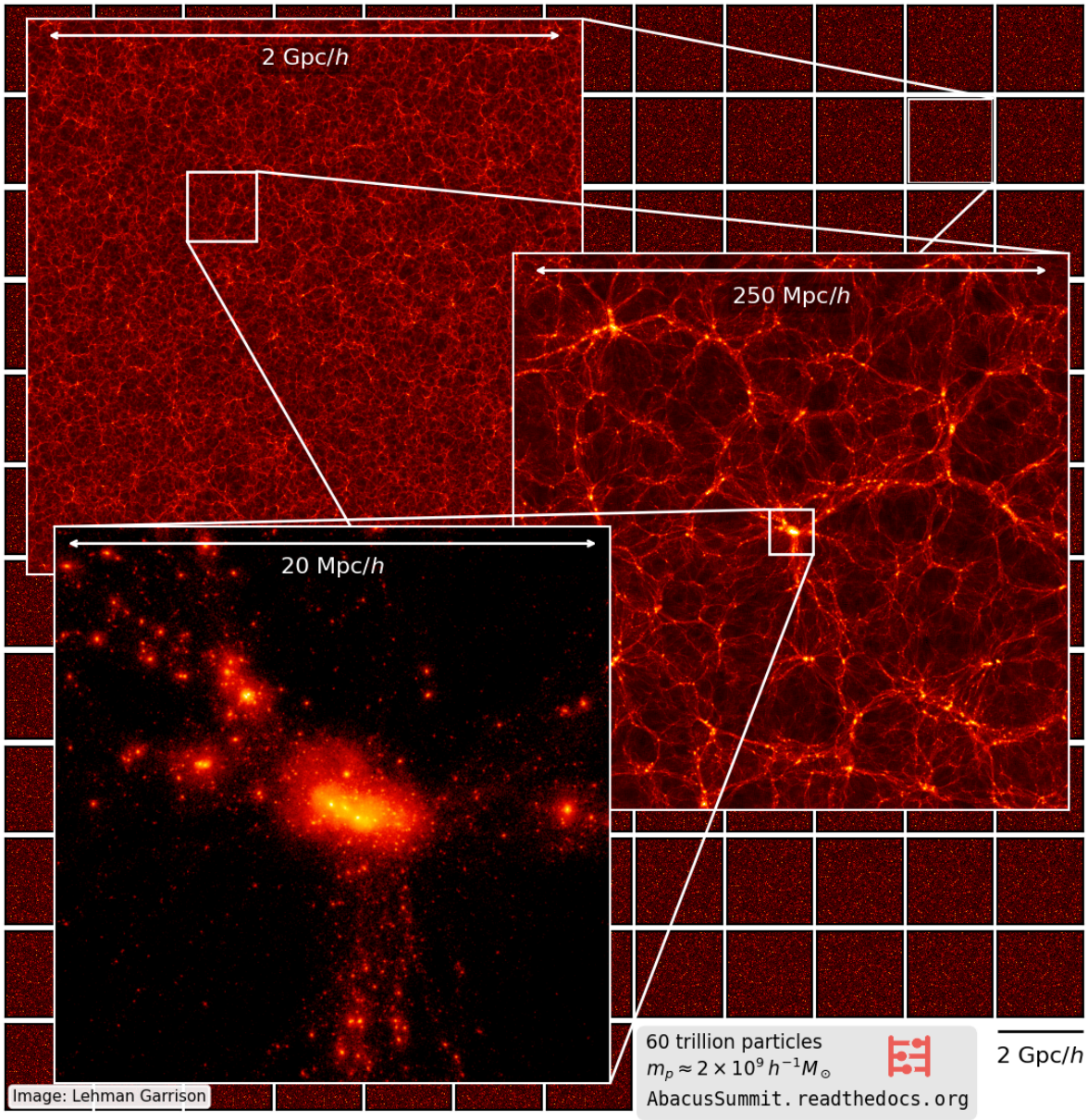

Figure 3.5: *Visualisation of the* AbacusSummit `base` *resolution boxes with progressive zoom-ins from the full box down to the cluster scale. The* `AbacusSummit_base_c000_ph000` *simulation at* $z = 0.1$ *is displayed in the zooms. Projections are 10 Mpc/h deep.*

This method works in the same way as the tree method, except that the simulation is decomposed into a 3D Cartesian grid of cells and the multipole moments are calculated for each cell. The three-dimensional Cartesian grid allows the Taylor series of the gravitational potential to be computed as a convolution over the cells, rather than performing numerous individual interactions between paired cells, which is very efficient in terms of computation time. The near field force is computed with Newtonian gravity for particles in all near-field cells, using a softening force from a spline method (Hernquist & Katz, 1989) instead of the standard softening method discussed above. Typically, the number of neighbouring cells in the near field, 5^3, the softening range of 7 kpc/h and an order $p = 8$ for the Taylor expansion give excellent accuracy of gravitational force computations for cosmological simulations. Figure 3.6 shows the decomposition of the far and near fields.

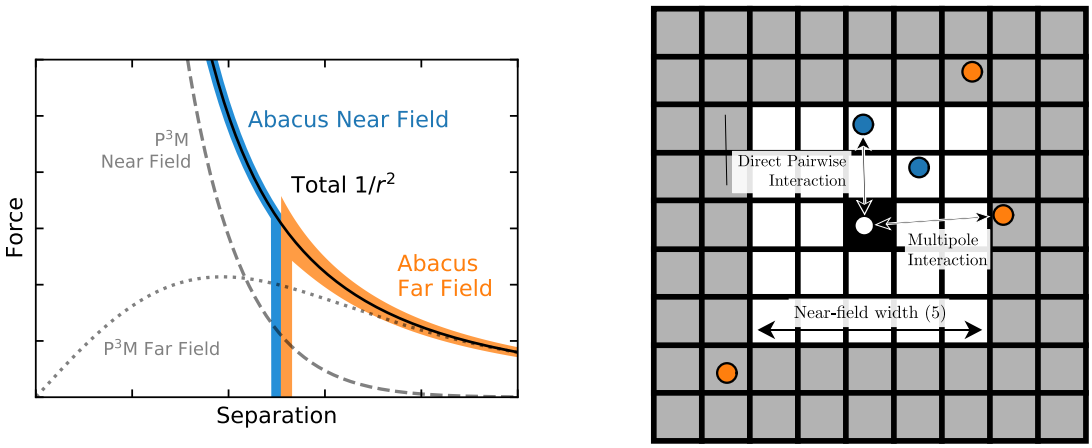

Figure 3.6: *Left: Schematic illustration of the near-field/far-field decomposition in the* Abacus *simulation. The grey lines represent another scheme for the near-field (dashed line)/far-field (dotted line) decomposition. Right: Schematic view of the decomposition on a slab. We consider here the particle in the black cell (in the centre). The near-field is represented in white and the far-field in grey. These two figures are taken from Garrison et al. (2021).*

Another notable feature of the Abacus code is the organisation of the data which is designed to minimise memory consumption. Cells are organised into slices of one cell wide in a chosen direction called *slabs*. The positions and velocities of each particle (for the near field) and the cell-based Taylor series coefficients (for the far field) are stored in grid order indexed by slabs and cells. At each time step, slabs are introduced into the pipeline until the entire volume has been processed. For each slab, particle data and cell-based multipole moments are updated and stored on disk. The near-field calculation requires to have at least 5 slabs in memory. After scanning the entire volume, the far-field operation converts the multipole moments in all cells into Taylor series coefficients via a convolution, preparing for the next scan. Memory is only required to process $\sim$ 7 slabs in the pipeline (5 loaded slabs + a little more space for calculations), so the simulation can be run even if the particle data is larger than the available memory – a so-called *out-of-core algorithm*. Decomposition into slabs is also an advantage for parallel implementation, where each node can manage a range of slabs. The AbacusSummit simulations typically take 1100 time steps.

3.2.2 Hydrodynamical simulations

Dark matter accounts for $\sim 95\%$ of the mass in the Universe. Consequently, N-body simulations enable us to understand the formation and evolution of structures down to small scales in the non-linear regime. However, at cluster scales and below $\sim 1-2$ Mpc baryonic physics has an impact on the dynamics of non-linear evolution of density perturbations and must be taken into account. The use of hydrodynamical simulations is therefore necessary to model these complex baryonic phenomena. By simultaneously resolving the evolution of dark matter and baryons for gravity and hydrodynamics in a cosmological context, hydrodynamical simulations are able to model galaxy formation and evolution. Many baryonic processes can be incorporated, such as radiative gas cooling, stellar feedback, star formation, active galactic nuclei (AGN) and SN feedback, radiative transfer... I will not go into the details of all baryonic processes, but give the reference to a very nice review on this topic (Somerville & Davé, 2015).

Typically two main approaches are used to solve hydrodynamical equations: the smoothed particle hydrodynamics method (SPH) or adaptive grid codes (e.g. adaptive mesh refinement (AMR)). The SPH method (also called Lagrangian method) discretizes the gas into fluid particles that carry physical information. Local gas properties (i.e. temperature, density...) are obtained by a convolution with a smoothing kernel over neighbouring particles within a given smoothing length. The other approach, called Eulerian method, discretizes the fluid into grid cells, and the physical properties of the fluid are computed at the level of the cell. To increase grid resolution, hydrodynamic codes are generally based on adaptive mesh refinement (AMR). To precisely simulate the scale of galaxy formation the grid must have fine resolution. However, cosmological simulations have to be carried out on large volumes, which makes it difficult to achieve a sufficiently fine resolution to fully model the range of scales required for galaxy formation. Some parametrisations must be introduced to accurately simulate scales below the resolution scale, generally referred to as *subgrid physics*. These parametrisations need to be adjusted, either by direct tests with observations or by comparison to other empirical models that connect galaxies to dark matter halos.

Although there are some approximations below the resolution scale, hydrodynamical simulations provide our best understanding of the physical processes of galaxy formation. Running such a simulation on cosmological scales with fine resolution requires a lot of computing resources, which is one of the main limitations. Considerable efforts have been made to produce realistic galaxy populations in cosmological-scale hydrodynamical simulations, thanks to increased computing power and technical improvements. One example is the IllustrisTNG project, a suite of hydrodynamical simulations in cubic boxes of up to 300 Mpc/h length size (Nelson et al., 2021). Figure 3.7 shows a composite image of IllustrisTNG simulations for box sizes of 100 and 300 Mpc/h (TNG100, TNG300). One can appreciate on this image the matter distribution on large scales and the fine resolution on galaxy scales. The volume of the TNG300 simulation is considered a huge volume for hydrodynamical simulations, but it is smaller by a factor of 10^3 than that of the largest N-body simulations. One of the latest hydrodynamical simulation, MilleniumTNG, operates on a box size of 740 Mpc/h (Hernández-Aguayo et al., 2023).

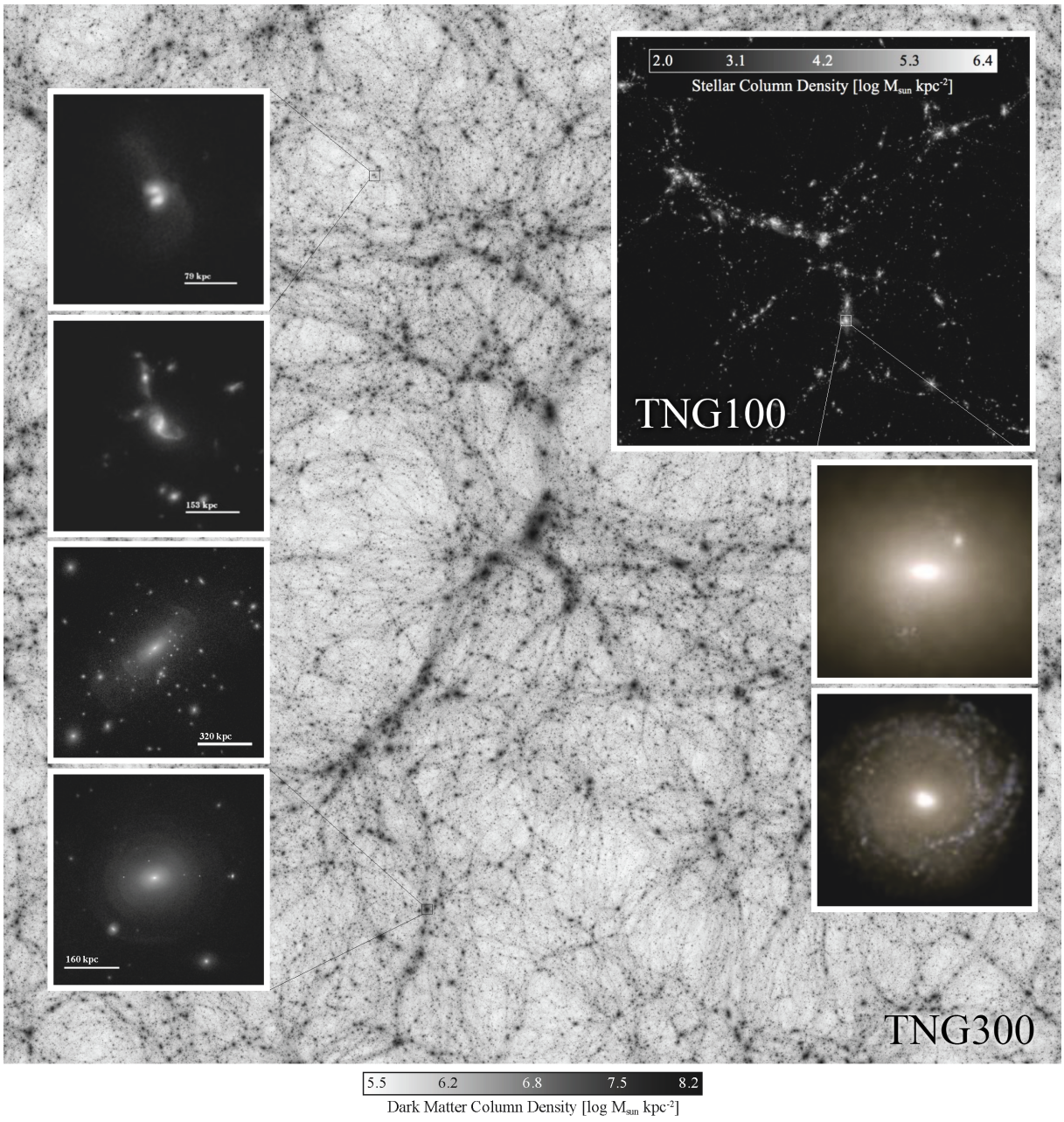

Figure 3.7: *Composite image of the TNG300 and TNG100 simulations: in the background, the full dark matter map of the TNG300 volume. In the upper right image, the distribution of stellar mass across the TNG100 volume. Panels on the left show galaxy-galaxy interactions and the fine resolution of structure on galactic scales. Panels on the right show stellar light projections from two massive central galaxies at* $z = 0$. *Credit: IllustrisTNG collaboration.*

3.2.3 Halo-finders

Once N-body simulations have been run, one major challenge is to define halos and subhalos from DM particles using *halo-finder* codes. The following section aims to explain how halo finders work and to demonstrate the complexity of this procedure. A halo-finder analyses the particles generated by a simulation and identifies dense regions where dark matter halos are likely to be located. Once a halo is identified, the halo-finder provides information about its properties, such as its mass, radius, shape, concentration... A wide variety of methods and codes have been developed to identify DM halos in simulation. Knebe et al. (2011) gives a nice review on this topic and provides a quantitative comparison of different halo-finder methods and codes. Most of these are based on two very popular methods:

➢ **The Spherical Overdensity (SO)** first mentioned by Press & Schechter (1974) and further developed by Lacey & Cole (1994), Warren et al. (1992). The SO method uses integrated densities over a spherical volume to identify dark matter halos as volumes which enclose a mean overdensity that corresponds to the value for virialised halos. A local density is first computed for each particle from their N nearest neighbours and the particle with the largest local density is selected as the candidate centre of a halo. A sphere is grown around this particle until the specified overdensity is reached (shifts in the halo centre are also accounted for in this process, see Lacey & Cole (1994)). All particles enclosed inside the final radius are members of the halo. Codes based on the SO method are numerous and encompass Bound Density Maxima (BDM Klypin et al. (1999)), Amiga's Halo Finder (AHF, Knollmann & Knebe (2009)), Adaptive Spherical Overdensity Halo Finder (ASOHF, Planelles & Quilis (2010)), parallel SO (pSO, Sutter & Ricker (2010)).

➢ **The Friend-of-Friends (FoF)** algorithm introduced in astrophysics by (Davis et al., 1985, Huchra & Geller, 1982). In the FoF algorithm, particles are linked to each other if their distances are lower than a characteristic linking length l_{FoF}. The resulting group is consider as a halo. The main differences compared to the SO method is that FoF halos are unstructured, coordinate-free and are defined by only one parameter, l_{FoF}. The typical linking length is $l_{FoF} = 0.2 \cdot l_{mean}$ (where l_{mean} is the mean interparticle separation) which roughly translates into an overdensity of ~ 180 times the background density (More et al., 2011). Examples of the many codes based on the FoF method are SUBFIND (Springel et al., 2001), LANL (Habib et al., 2009), parallel FOF (pFOF, (Rasera et al., 2010)).

After the initial definition of halos, most methods apply a pruning phase during which particles not bound by gravity are removed from the halo. Further studies extended these two approaches by including information in phase-space (positions + velocities) such as the Six-Dimensional Friends-of-Friends (6DFOF, Diemand et al. (2006)) or ROCKSTAR (Behroozi et al., 2013b) halo-finders. Halo-finder algorithms are challenging to develop. A lot of computing resources are needed to process the large number of dark matter particles in the simulations. The algorithms also face several issues. The traditional problem in halo finding is the *identification of halo mergers and sub-halos* especially when these are close to the centres of their host halos. In these cases the use of 6D space helps. Each of the 2 methods described above have their own problems. For example, in the FoF algorithm, particles are uniquely assigned to a halo,

avoiding the intersection between FoF groups. However if two structures are close enough, they can be connected by a *bridge*, resulting in a weirdly shaped structure. Another limitation is the *resolution* of the simulations. In simulations, small halos may not be well-resolved (because of their low number of particles) and can be missed by the halo-finder. In addition, *numerical noise* (Poisson) can be problematic when dealing with low-mass (sub)halos and can introduce spurious halos or affect the halo properties. These examples show the difficulty to construct a reliable DM halo catalogue in N-body simulations and explain why using different halo-finder codes on the same simulation can lead to different results, especially at low halo masses (see Knebe et al. (2011), for a halo-finder code comparison).

3.2.3.1 CompaSO halo-finder

In this thesis work, we rely on AbacusSummit simulations. The COMPetitive Assigment to Spherical Overdensities (CompaSO) algorithm (Hadzhiyska et al., 2022a) was developed as a group-finding tool for the Abacus N-body code. It runs "on-the-fly", and was written to meet the requirements of massive N-body simulations and high speed calculations. The algorithm achieved a rate of ~ 30 million particles/second/node on the Summit supercomputer. A brief summary of this algorithm is described in the following and the reader is referred to Hadzhiyska et al. (2022a) for more details.

CompaSO is a hybrid algorithm, composed of three levels of group finding, based on both FoF and SO methods. It first estimates the local density for each particle, Δ, using a weighting kernel $W(r, b_k) = 1 - r^2/b_k^2$ where r denotes the interparticle separation and $b_k = 0.4$ is the kernel radius. Since the squared distances are already computed for the near-field forces, obtaining the local density on-the-fly is (almost) "free". The local density helps to identify substructures and the core of a halo, as depicted in Figure 3.8.

Level 0 (L0) groups particles into halos using a modified FoF procedure, with a linking length $l_{FoF} = 0.25 \cdot l_{mean}$ but only for particles with a local density value $\Delta > 60$. From the L0 halos, Level 1 (L1) and Level 2 (L2) halos are then constructed.

In each L0 group, the particle with the highest kernel density Δ is selected and becomes the first halo nucleus. Then, the L1 halo radius R_{L1} is defined as the innermost radius that encloses a density below the L1 density threshold Δ_{L1}. All particles within this radius are preliminary assigned to this L1 halo. Once this process is done, the 20% of particles furthest from the centre of the halo are considered "eligible" to potentially be the nucleus of a halo that could be orbiting at the periphery. Among these "eligible" particles, the algorithm searches that with the highest kernel density that meets the minimum local density criterion, i.e. the particle must have the highest density within the kernel radius (b_k), including particles 'eligible' or not. The search for new halo nuclei ends when the density of particles becomes too low to create a new halo (see fig:compaso).

Finally, particles that fall within the radius of two halos are assigned to one of them by competitive assignment. This means that a particle is attributed to the new halo only if it is estimated to have an enclosed density with respect to this new halo that is at least twice larger than that of the enclosed density with respect to its currently assigned halo.

Then, for each L1 halo, the competitive SO algorithm steps is run to find L2 subhalos with density threshold Δ_{L2} enclosed in R_{L2}. The largest L2 subhalo is used as the centre-of-mass and defines a centre for the output of the L1 statistics. The L1 and L2 density thresholds are defined

for an Einstein-de Sitter cosmology to be $\Delta_{L1} = 200$, $\Delta_{L2} = 800$ and vary with redshift as $\Delta_{L1} = (200/18\pi^2)\Delta_{\rm base}(z)$ and $\Delta_{L2} = (800/18\pi^2)\Delta_{\rm base}(z)$ where $\Delta_{\rm base}(z)$ is the fitting function provided by Bryan & Norman (1998), which defines the density with respect to the critical density:

$$\Delta_{\rm base}(z) = 18\pi^2 + 82(\Omega_{\rm m}(z) - 1) + 39(\Omega_{\rm m}(z) - 1)^2. \tag{3.74}$$

Figure 3.8 illustrates a schematic picture of the CompaSO algorithm.

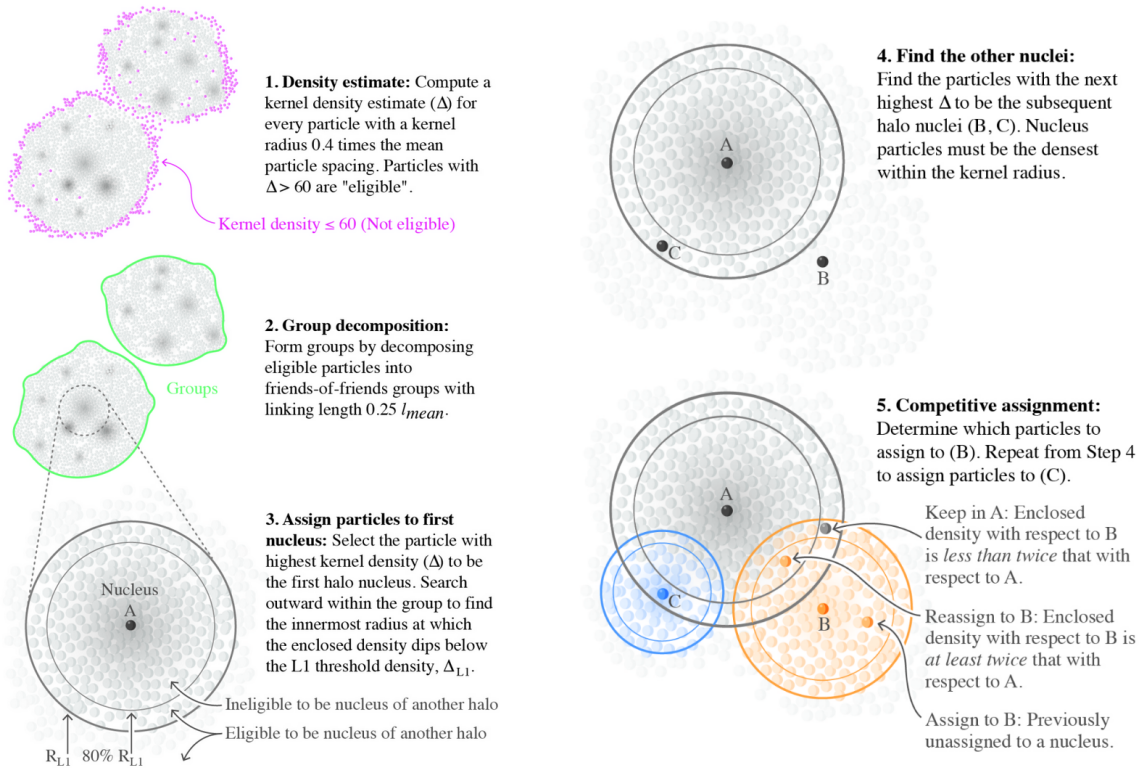

Figure 3.8: *Visualization from Hadzhiyska et al. (2022a) of the CompaSO algorithm.*

The initial version of the CompaSO algorithm suffered from fragmented or re-collapsed structures at different epochs. A post-processing cleaning procedure was developed in Bose et al. (2022), based on merger-tree information. It checks the fraction of particles in a halo at a time t_i that comes from a main progenitor at a previous time t_{i-1}. If this fraction is too large, the newer halo is marked a "potential split" and merged into the larger halo. In addition, halos for which the peak mass (maximal mass during halo history) exceeds twice the present day mass are declared unphysical and merged into a more massive neighbour, from whom it had presumably split off (see Bose et al. (2022) for details).

3.3 From darkness to light: illuminating dark matter halos

Until now, we have mainly focused on the dynamics of dark matter collapsing into halos, and on how to simulate part of the Universe in boxes. Hydrodynamical simulations can model the

evolution of dark matter and baryons to study galaxy formation and evolution. Although these simulations give many insights into the processes of galaxy formation, they are highly dependent on the choice of physical effects to take into account and how these should be implemented. They are also very computationally expensive, so that not all the physical prescriptions can be tested. Therefore, other alternatives are used to implement (paint) galaxies in N-body simulations. The basic assumption of our current view of galaxy formation is that galaxies form in dark matter halos. Consequently, the growth, internal properties and spatial distribution of dark matter halos can be related to those of galaxies. N-body codes provide the backbone of galaxy formation models, and various techniques are then applied to connect galaxies and dark matter halos in simulations, the so-called *galaxy-halo connection*. These techniques are then used to constrain the galaxy-halo connection from data, providing invaluable information about the physics of galaxy formation. But these constraints are also essential for guaranteeing the robustness of the cosmological results from galaxy surveys, as they allow us to produce reliable mocks (i.e. catalogues of simulated galaxies) used to test clustering analyses and to derive their systematic uncertainty budget (Alam et al., 2021).

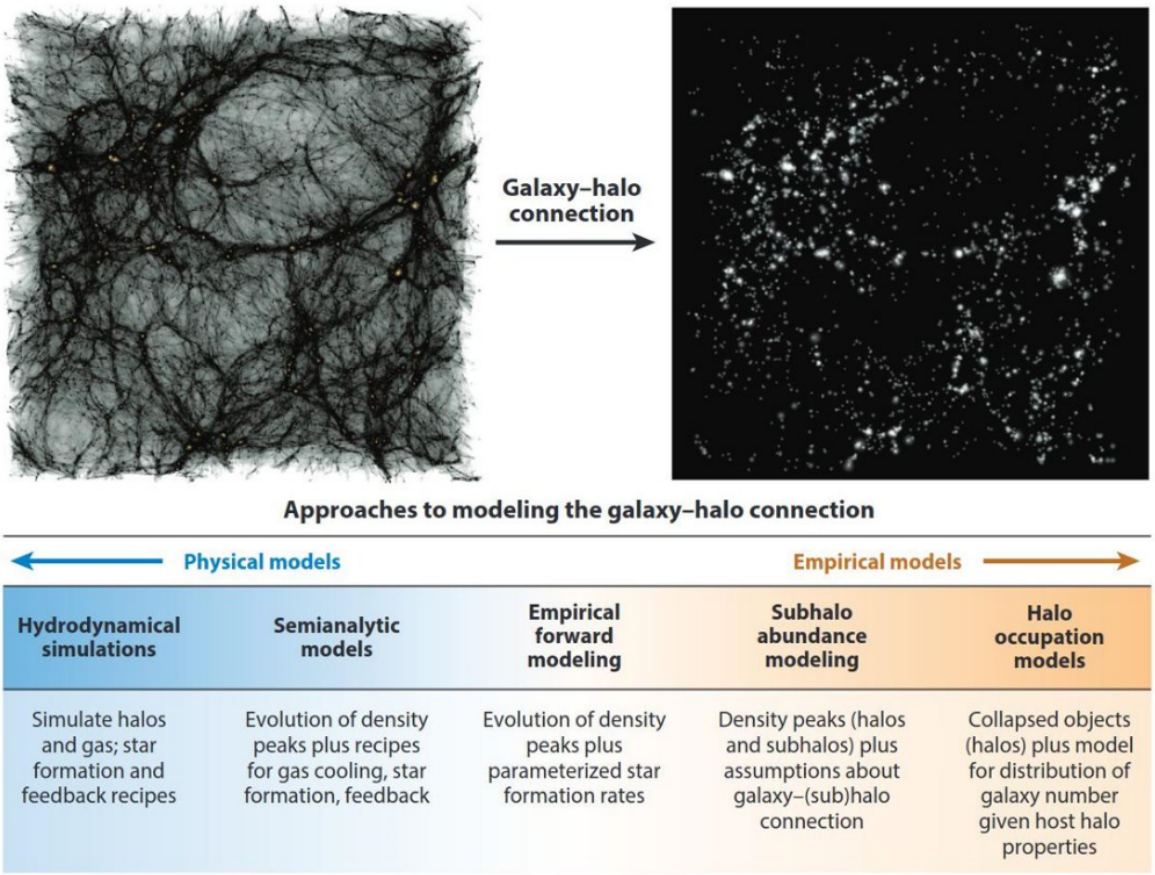

Figure 3.9: *Different approaches to model the galaxy-halo connection, from the most physical (on the left) to the most empirical model (on the right). This figure is taken from Wechsler & Tinker (2018).*

In this section, we first give a quick overview of galaxy evolution in a cosmological context. We then review the various techniques for connecting galaxies and dark matter halos in simulations, used in this thesis or in other DESI analyses which I compare my results to. These tech-

niques encompass empirical approaches as the halo occupancy distribution (HOD) or sub-halo abundance matching (SHAM) as well as more physical models such as semi-analytical models (SAM) of galaxy formation. A schematic picture of the galaxy-halo connection is presented in Figure 3.9. This section is inspired by the very nice review of the galaxy-halo connection from (Wechsler & Tinker, 2018).

3.3.1 A foreword about galaxies

We first start with a quick reminder about galaxy evolution to introduce the vocabulary that will be necessary in the following. A galaxy is a system of stars, stellar remnants, interstellar gas, dust, and dark matter, bound together by gravity. Historically, galaxies have been classified by many properties, i.e. morphology (spiral, elliptical, irregular...), colour (blue or red), star formation rate... In cosmology, we use galaxies as tracers of the dark matter field to study the structure of the Universe. Tracer selections are based on magnitudes and colours, but these selections cover various types of galaxies that are important to know since they trace different regions of the cosmic web.

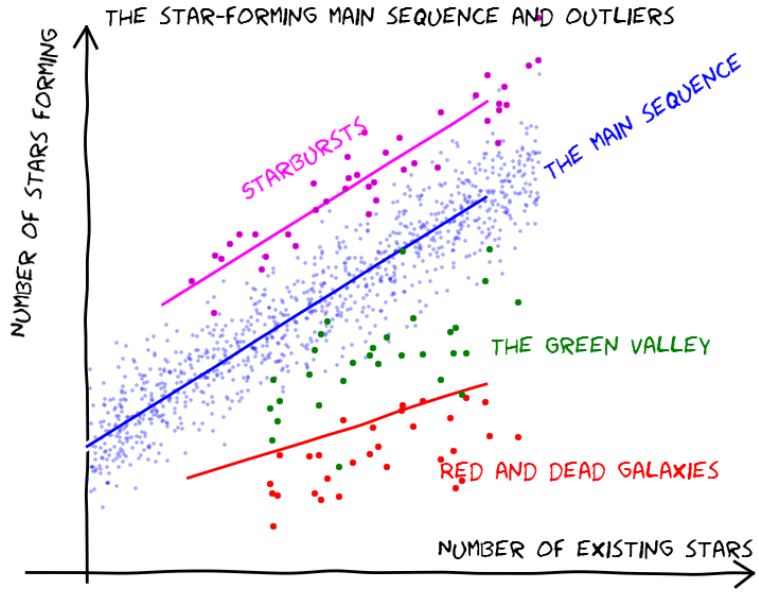

Figure 3.10: *Cartoon plot of the main sequence of star-forming galaxies. The y-axis, "number of stars forming" refers to the star formation rate and the x-axis, "number of existing stars" refers to the stellar mass. This figure comes from the CANDELS collaboration.*

We can roughly classify galaxies into two populations: blue, star-forming galaxies and red, quiescent or "dead" galaxies which no longer create stars, also known as *quenched galaxies* (Strateva et al., 2001). Figure 3.10 presents a pedagogical view of galaxy evolution, with a *main sequence of star-forming blue galaxies* shown in blue and quiescent galaxies in red. The main sequence of star-forming galaxies is a linear relation between the galaxy star formation rate (SFR) and stellar mass ($M_\star$) that has been measured in redshift bins over the range $0 < z < 6$ in both data and hydrodynamical simulations (Popesso et al., 2022). The main sequence slope

does not vary with redshift while the normalisation is a decreasing power law of the Universe age.

The standard picture about star formation in a galaxy is that stars form out of molecular gas at $T < 10^2$ K that cooled from warm and hot gas ($T = 10^3 - 10^6$ K) previously accreted in the galaxy halo from cosmological filaments. The cooled gas then collapsed locally, turning into stars. In the history of galaxy evolution, blue galaxies are considered to be "young" and less massive than red galaxies. Once blue galaxies have "completed" their star formation period, they leave the main sequence and become red galaxies. There are two pathways for galaxies to go out of the main sequence. At some point of their evolution, galaxies may form stars at a much higher rate than on the main sequence, becoming *starburst galaxies*, which rapidly consume their baryonic reservoir, resulting in a rapid transition through the so-called green valley towards the red sequence (see Figure 3.10). Starburst galaxies are usually interpreted as being driven by a merging event that boosts their galaxy star formation. They account only for a minor fraction of the cosmic star formation rate density e.g. 10% at $z \sim 2$ (Rodighiero et al., 2011) but represent one pathway to leave the main sequence. In contrast, galaxies on the second pathway show a slow decline in star formation and depart gradually from the main sequence. Those two pathways are illustrated in Figure 3.11.

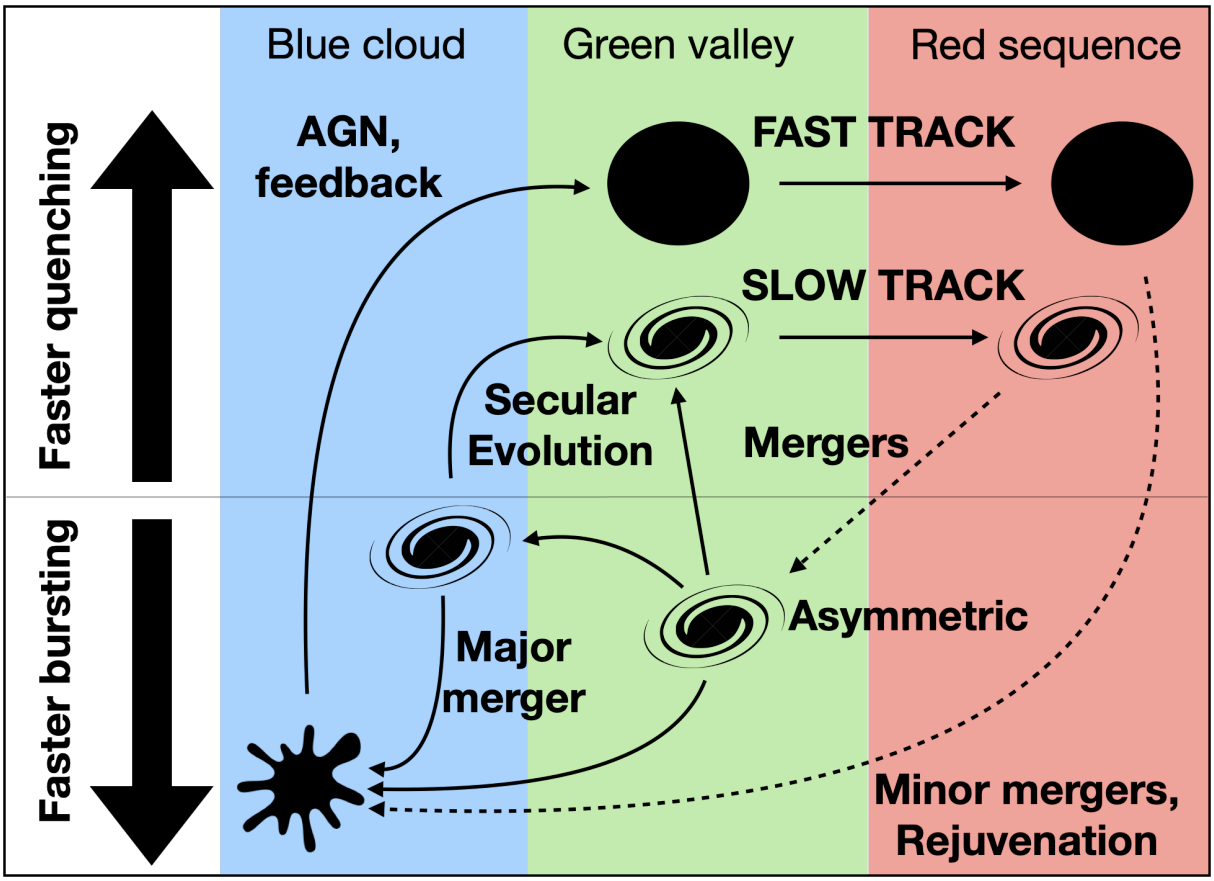

Figure 3.11: *Schematic presentation of the evolutionary pathways of blue galaxies towards the red sequence. This figure is taken from de Sá-Freitas et al. (2021).*

The cessation of star formation is called *quenching*. There are many mechanisms for this, the effects of which fall into five broad classes: preventing gas from accreting, cooling or form-

ing stars, or leading to the gas consumption or removal (Man & Belli, 2018). Examples of mechanisms are the so-called stellar or AGN feedback, which will be described later, shocks that heat up the interstellar medium, galaxy mergers... The details of these processes are not yet fully understood, nor is their interplay, since several of these mechanisms may be involved simultaneously but on different timescales.

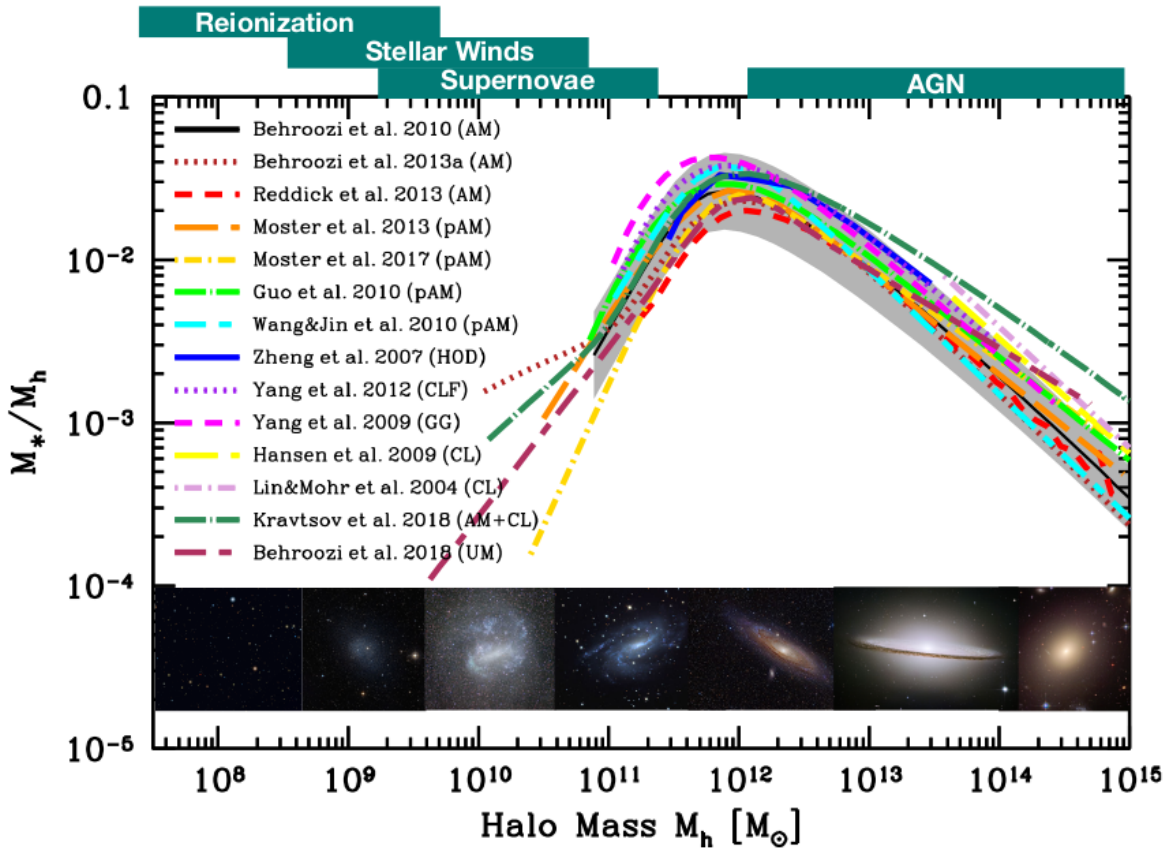

Figure 3.12: *The galaxy stellar mass-to-halo mass ratio of central galaxies at $z = 0$. This figures compiles a wide range of models from different galaxy-halo connection methods compared to various data sets. The band of galaxy images shows example galaxies that are hosted by DM halos in the specified mass range. Indicated on the top of the figure are key physical processes that may eject gas, heat gas or suppress star formation at those mass scales. This figure is taken from Wechsler & Tinker (2018).*

An interesting relation between halo and stellar mass, called the stellar-to-halo mass ratio (SHMR) has been measured in data and studied in simulations (Behroozi et al., 2013a, Leauthaud et al., 2012). This relation, illustrated in Figure 3.12, measures the efficiency of star formation as a function of the mass of the host halo. The ratio peaks at halo masses around $\sim 10^{12} M_\odot$. At higher and lower halo masses, star formation is less efficient due to different processes, mainly stellar and AGN feedbacks, as indicated at the top of the figure. In particular, the SHMR decrease above $\sim 10^{12} M_\odot$ is consistent with quenching effects, as in this region of halo masses, central galaxies are observed to be predominantly quiescent. Moreover, numerical simulations show that shocks are formed only when halos are more massive than this same threshold, which suggests that virial shock heating may play a role in the onset of star formation quenching (Man & Belli, 2018).

Note: *The efficiency of galaxies to convert baryons into stars is low, barely reaching a few percent. Assuming that all halos contain the universal baryon fraction* $\Omega_b/\Omega_m \simeq 0.17$*, with a star formation efficiency of 100% (all baryons turn into stars) we would expect to have the same fraction of stellar-to-halo mass. As the peak observed in the SHMR is at the level of* $\sim 4\%$ *we can infer that only ∼20-30% of baryons have turned into stars.*

The Emission Line Galaxies (ELGs) and Luminous Red Galaxies (LRGs) targeted in DESI (see Section 2.4.2) can be roughly associated with the two galaxy populations previously discussed, red galaxies for LRGs and blue galaxies for ELGs. According to what was explained above, LRGs are expected to reside in hot and massive halos, while ELGs should reside in halos where star formation is high, around $\sim 10^{12} M_\odot$. In DESI, the LRG target selection is optimised to select the most massive galaxies with a high degree of completeness. Note that the completeness mentioned here is different from the completeness in the observations, discussed in Section 2.5.1. Here, completeness refers to the number of target LRGs relative to the "true" number (i.e. the expected number in the Universe) of massive galaxies (Zhou et al., 2023). Such an optimisation was also implemented in SDSS for the CMASS (Complete-MASS) galaxy sample (Dawson et al., 2013). This type of sample is called *(stellar) mass-complete sample*, which means that all objects above a given luminosity threshold (and therefore a given stellar mass) are selected. In contrast, the ELG target selection in DESI (or eBOSS) has been optimised to select [OII] emitters, so mostly star-forming galaxies (Raichoor et al., 2023). Such ELG samples are not meant to be complete in terms of luminosity or stellar mass, but their selection is instead equivalent to a selection by their star formation rate, or even their specific star formation rate (i.e. the SFR per stellar mass) as shown in Hadzhiyska et al. (2021). Therefore, ELGs and LRGs are expected to reside in halos of different masses and in different environments of the cosmic web.

In the following, we present the various techniques used to connect galaxies and dark matter halos in simulations.

3.3.2 Semi-analytical models

Semi-analytic models of galaxy formation (SAMs) (Cole et al., 2000, Guo et al., 2013, Kauffmann et al., 1993, Lacey & Cole, 1993, Somerville & Primack, 1999, White & Frenk, 1991) aim to predict the properties of galaxies, such as luminosity, morphology, metallicity, star formation history... It combines analytical calculations with N-body simulations to model galaxy formation processes in a computationally efficient way. These models are based on the *hierarchical* growth of dark matter halos that drives galaxy formation. Various physical processes associated with galaxy formation are treated using approximate analytical prescriptions that are traced through *merger trees* extracted from N-body simulations. A merger tree gives the "family tree" of DM halos in N-body simulations. By identifying the evolution of dark matter halos, it traces the evolution of DM halo masses with redshift and the times when progenitor halos merge together to form a larger halo. The complete merger history of any dark matter halo is a complex structure containing a wealth of information. An example of a merger tree is shown in Figure 3.13. From a high redshift, the mass that ends up in the halo at $z = 0$ originates from many smaller branches that merge into larger halos over time. It is worth noting that extracting merger trees from

N-body simulations is not straightforward, and that the results can be sensitive to the method used to identify the halos.

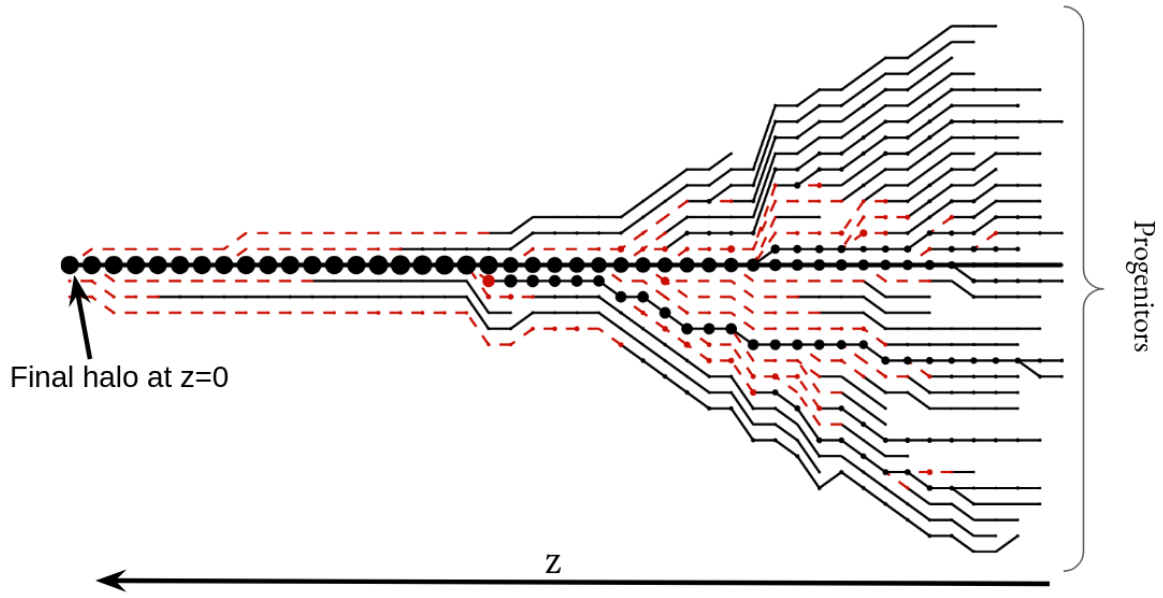

Figure 3.13: *Example of merger tree for a dark matter halo of mass* $M \simeq 3 \times 10^{12} M_{\odot}/h$ *at* $z = 0$*. The time increases from right to left. Figure adapted from Stewart et al. (2008).*

A wide range of physical effects can be treated within the SAMs, we list below some of the main processes that impact the galaxy formation and evolution:

Gas cooling: gas cooling is a necessary and fundamental ingredient of the galaxy formation and evolution processes. When a newly halo is formed (at the top level of the merger tree) or after a galaxy merger, the gas is heated due to shocks during the virialisation or merger process. Then, the gas cools down through radiative processes, and the cold gas then collapses and forms stars.

Stellar feedback: stars influence the surrounding gas in their host galaxy by injecting energy and momentum. This creates a feedback loop that regulates the star formation process. Different feedback channels are at play, such as supernova events that lead to both gas ejection and gas heating in the interstellar medium (Ciotti et al., 1991, Hou et al., 2016). Other channels are energy and momentum injection from stellar winds (e.g. from evolved massive galaxies), photoionisation, and radiation pressure resulting from radiation emitted by young, massive stars (Cattaneo, 2019, Hopkins et al., 2012). Stellar feedback is mostly efficient in low mass galaxies.

AGN feedback: AGN feedback from supermassive black holes provides an effective star formation regulation mechanism for high-mass galaxies. A large quantity of gas flows towards the black holes, generating a release of energy capable of driving powerful outflow jets that heat up the interstellar medium, and regulate star formation and the baryonic content of galaxies. This mechanism may even lead to quenching by removing the galaxy supply of gas (Bower et al., 2006, Pontzen et al., 2017).

Galaxy mergers: galaxy (or halo/sub-halo) mergers can be traced by the merger tree. During a merger, the gas from a galaxy in the smaller (sub-)halos is added to that of the main galaxy. The former can either become a satellite galaxy or completely merge with the main galaxy.

Dust extinction: dust in galaxies (mainly at high-mass) is heated by high-energy photons from stellar radiation. Dust attenuates stellar light and modifies galaxy fluxes and colours, resulting in a reddening of the observed galaxy colours.

Chemical enrichment of the interstellar medium: as stars evolve inside galaxies, they feed the interstellar medium with heavier elements and cold gas, notably through SN events.

Similar to hydrodynamical simulations, SAMs have several degrees of approximations, depending on the complexity of the underlying physics being addressed. Consequently, SAM assumptions need to be tested against hydrodynamical simulations and data. Although these models are considerably less CPU/GPU expensive than hydrodynamical simulations, their large number of parameters (up to 30) makes it difficult to explore the parameter space completely. However, recent studies have used Monte Carlo Markov chain techniques to directly constrain the parameter space of SAMs against data (Bower et al., 2010, Henriques et al., 2009, 2015, Lu et al., 2011, 2014). In the literature, several SAMs of galaxy formation have been developed by different groups (I list only some examples): the Munich model (Kauffmann et al., 1999), the Santa Cruz model (Somerville & Primack, 1999), `MORGANA` (Monaco et al., 2007), `MITAKA` (Nagashima & Yoshii, 2004), `GALICS` (Hatton et al., 2003), and the Durham model `GALFORM` (Cole et al., 2000).

The latter, initially developed by Cole et al. (2000) and improved over the years by adding more and more specificities and complexity to the description of galaxy formation processes has been used to study the evolution and clustering of emission-line galaxies and in particular of [OII] (Gonzalez-Perez et al., 2014, 2018, 2020) emitters. These specific results on the galaxy-halo connection of ELGs are covered in a dedicated section (see Section 3.4).

3.3.3 Sub-halo abundance matching

The Sub-halo abundance matching (SHAM) is an intuitive empirical method to model the non-linear relation between galaxies and halos including the substructure of DM halos (shortly introduced in Section 3.1.5). The idea behind abundance matching (AM) is that the most massive galaxies live in the most massive DM halos. In this framework, each halo and subhalo hosts a galaxy, whose properties (such as stellar mass and luminosity) are matched by abundance according to the mass or the velocity of the host DM (sub-)halo (Kravtsov et al., 2004, Tasitsiomi et al., 2004, Vale & Ostriker, 2006). This approach is non-parametric as it assumes a monotonic relation between galaxies and DM structures. The key question in a SHAM analysis is to find which halo property best matches which galaxy property, and what is the scatter, σ, between the galaxy property and the halo property. Different halo/sub-halo properties can be taken into account in a SHAM analysis:

- M_h: halo mass at the time considered,

- M_{acc}: mass at the time of accretion (for sub-halos),
- M_{peak}: highest mass achieved in the entire history of a halo,
- V_{max}: maximum circular velocity of the halo at the time considered,
- V_{peak}: highest circular velocity achieved in the entire history of a halo,
- V_{acc}: V_{max} at the time of accretion (for sub-halos).

In the literature, these properties are often associated with stellar mass or galaxy luminosity. A first success of this technique was to match a luminosity-selected galaxy sample using the property V_{max}/V_{acc} of dark matter halos/subhalos (Conroy et al., 2006). SHAM predictions have been shown to be in remarkable agreement with observations, as illustrated in Figure 3.14, which compares the projected clustering of galaxies, w_p, as predicted by SHAM models using different halo properties compared to that of a local galaxy sample ($z < 0.3$) from the New-York University value added catalogue based on SDSS DR7 (Blanton et al., 2005).

Abundance matching can also be parametrised to determine the galaxy stellar-to-halo mass relation (SHMR), which was presented in Figure 3.12 (Moster et al., 2010). As previously explained, unlike luminosity-selected samples, star-forming galaxy samples are expected to be incomplete. Therefore, extended implementations of SHAM models have been developed to take incompleteness effects into account (Favole et al., 2016, Rodríguez-Torres et al., 2017, Yu et al., 2022).)

3.3.4 Halo occupation distribution

The Halo Occupation Distribution (HOD) is an empirical formalism that describes the relation between a typical class of galaxies and dark matter halos, as the probability that a halo with mass M_h contains N such galaxies. HOD models have contributions from two galaxy populations, namely centrals and satellites, with $\langle N_{cent}(M_h)\rangle$ and $\langle N_{sat}(M_h)\rangle$ their respective mean numbers hosted per halo of a given halo mass. The most common mean HOD functional uses a step function for centrals, a power law for satellites and assumes generally that satellites can only be found in halos which already host a central galaxy (Zheng et al., 2007):

$$\langle N_{cent}(M_h)\rangle = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{M_h - M_{min}}{\sigma_M}\right)\right] \tag{3.75}$$

$$\langle N_{sat}(M_h)\rangle = \begin{cases} \langle N_{cent}(M_h)\rangle \left(\frac{M_h - M_0}{M_1}\right)^{\alpha} & \text{if } M_h > M_0 \\ 0 \quad \text{otherwise} \end{cases} \tag{3.76}$$

$$\tag{3.77}$$

Once the mean number of galaxies per halo is computed, a probability distribution function is used to assign central and satellite galaxies to a halo. Standard choices are a Bernoulli distribution for centrals and a Poisson distribution for satellites. Central galaxies are typically placed at the halo centre with a velocity given by the halo peculiar velocity, while satellites are placed assuming a halo profile (mainly NFW) or assigned to a random particle of the halo. This model, represented in Figure 3.15, has been proven to describe well the clustering of different galaxy populations, e.g luminosity selected (Zehavi et al., 2011) or stellar mass limited

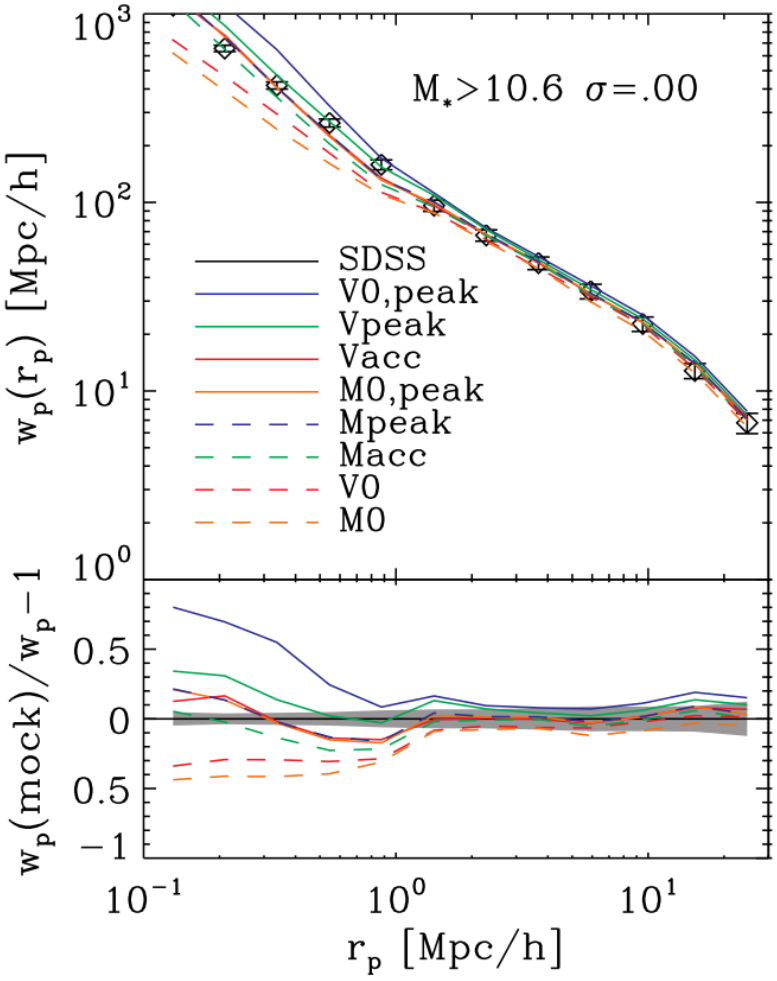

Figure 3.14: *Top panel: Projected galaxy clustering w_p as a function of transverse direction. Dots with error bars correspond to the measurements from a luminosity-selected sample from SDSS galaxies at mean redshift $z \sim 0.05$. Lines correspond to SHAM models with different halo properties considered to match the data, with no scatter applied. Lower panel: Ratio between models and data. This Figure is from Reddick et al. (2013).*

(Contreras et al., 2013) samples, like LRGs (Zheng et al., 2009) or QSOs (Smith et al., 2020). This standard HOD model considers that above a certain mass, all halos are populated by a central galaxy. However, as we mentioned above, ELGs are not complete at high mass, thus the standard HOD is not appropriate and other HOD shapes have been considered, such as a Gaussian or asymmetric Gaussian distribution (Avila et al., 2020). Again, we will discuss this further in Section 3.4.

Conditional luminosity function

Similar to the HOD model, the conditional luminosity function (CLF) $\Phi(L|M)$ links the galaxy luminosity function $\Phi(L)$ and the halo mass function $n(M)$ to parametrise the halo occupation (Yang et al., 2003). The CLF gives the average number of galaxies with luminosities between $[L, L+dL]$ hosted by a halo of mass M. Each halo is defined by a central galaxy whose luminosity

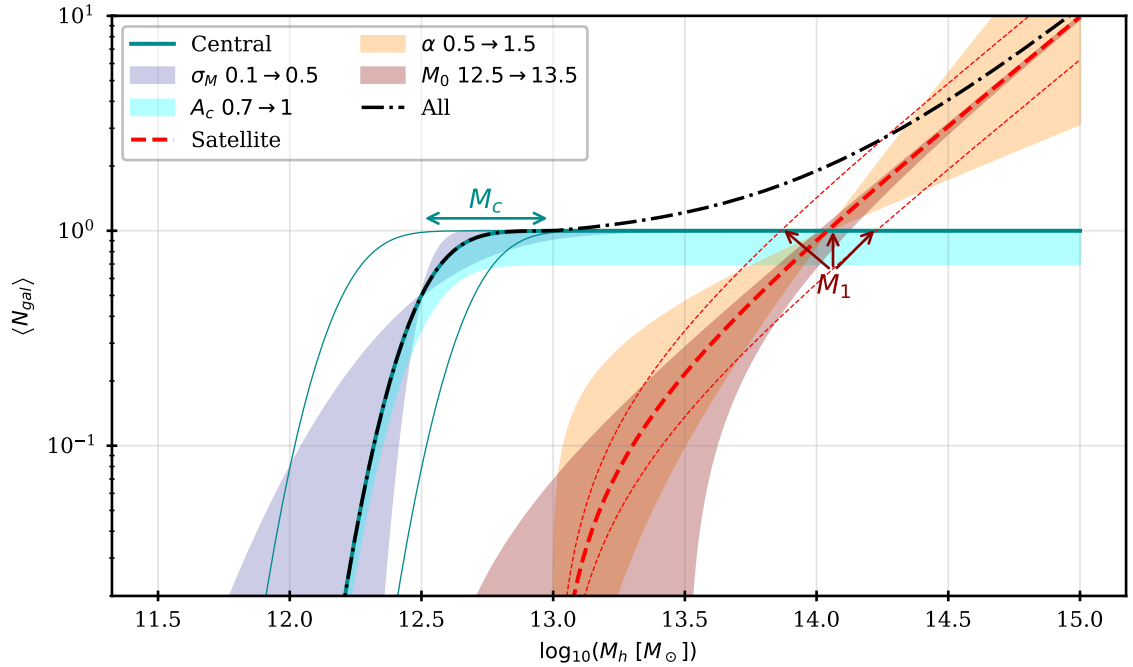

Figure 3.15: *Mean number of galaxies as a function of halo mass for the standard HOD model. The central contribution is represented by the solid blue line for* $M_c = 1.5$, $A_c = 1$, $\sigma_M = 0.2$. *Variations for* $M_c = [12.2, 12.8]$ *are shown in solid thin blue lines, those for* A_c *by the shaded cyan region and those for* σ_M *by the shaded purple region. The satellite contribution is represented by the red dashed line for* $M_0 = 13$, $M_1 = 14$ *and* $\alpha = 1$. *Variations for* $M_1 = [13.8, 14.2]$ *are shown in dotted thin red lines, those for* α *by the shaded orange region and those for* M_0 *by the shaded reddish region. The black dash-dot line is the contribution from both satellite and central distributions. All masses are expressed in* $\log_{10}([M_\odot/h])$.

L_c is assumed to be the brightest in the halo, and by satellite galaxies (if any) with luminosity L_s. The luminosity function for central galaxies is assumed to follow a lognormal distribution:

$$\Phi_c(L|M)dL = \frac{L_c}{\sqrt{2\pi}\sigma_c} \exp\left[-\left(\frac{\ln(L/L_c)}{\sqrt{2}\sigma_c}\right)^2\right]\frac{dL}{L} \tag{3.78}$$

and that for satellite galaxies is the Schechter luminosity function (Schechter, 1976):

$$\Phi_s(L|M)dL = \frac{\Phi_s}{L_s}\left(\frac{L}{L_s}\right)^\alpha \exp\left[-\left(\frac{L}{L_s}\right)\right]dL \tag{3.79}$$

with the parameters L_c, L_s, σ_c, Φ_s, α depending on the halo mass M. This model has also shown excellent agreement with data from a luminosity-selected galaxy sample at $z < 0.25$ from SDSS (Yang et al., 2009).

3.3.5 Beyond the standard HOD

Although HOD models are based only on halo mass, they have provided good modelling of data down to small scales. However, analytical models and hydrodynamic simulations show that features other than halo mass have an impact on galaxy formation. HOD models can be modified to take these properties into account.

Assembly bias

Each halo and galaxy is unique and has its own history. In the course of their evolution, galaxies and halos experience a wide variety of histories in their assembly pathway that can influence properties other than halo mass. Semi-analytical models and hydrodynamic simulations predict correlations between the spatial distribution of galaxies in halos of the same mass and halo secondary properties. This phenomenon is known as assembly bias (Croton et al., 2007, Gao & White, 2007, Wechsler et al., 2002, 2006). Recent studies (e.g. Hadzhiyska et al. (2022b), Hearin et al. (2016), Yuan et al. (2018)) have developed extended HOD models that take into account secondary properties of halos as seen in simulation to modify the average number of halos at a given mass as a function of these properties. In the literature, several secondary properties have been studied, such as halo concentration, density environment, tidal environment (shear), spin parameter, maximum accretion rate... Mao et al. (2018) present a summary of the correlations between several proxies of assembly history and secondary halo biases.

Velocity bias

In HOD models, the velocities of galaxies are defined by that of their host halos. Centrals take the velocities of the halos and satellites can take velocities derived from, for example, the NFW halo profile or from the halo dark matter particle velocities. However, due to baryonic effects, galaxy velocities can differ from the velocities of dark matter particles. Therefore, to accurately model small-scale clustering in redshift space, parameters can be used to shift galaxy velocities relative to their original assignment. This is known in the literature as *velocity bias* (Berlind & Weinberg, 2002, Skibba et al., 2011, Van Den Bosch et al., 2005, Yuan et al., 2018). The importance and nature of velocity biases differ from tracer to tracer.

Satellite occupation properties

Satellite distribution: Standard HOD models assume that the probability distribution function for satellite galaxies follows a Poisson distribution. This is quite true for luminosity-selected or complete mass samples such as LRGs, although slight deviations from the Poisson distribution have been seen in simulations for those galaxies (Hadzhiyska et al., 2022b). However, for ELGs, deviations from the Poisson distribution have been observed in hydrodynamical simulations and sub/super-Poissonian distributions have been considered in HOD models but no observational evidence has been reported (Avila et al., 2020, Hadzhiyska et al., 2022b, Jiménez et al., 2019).

Satellite profile: To recover small-scale clustering on scales below $\sim 1\mathrm{Mpc}/h$ (the one-halo term), the spatial distribution of satellites within their host halos needs to be determine. Typically, satellite positions assume a dark matter profile for the host halo (mainly NFW) or are assigned to randomly-selected dark matter particles. As with velocities, galaxies are affected by baryonic effects occurring at small scales, and their positions can differ from the dark matter profile. Although the density profile of dark matter halos has been extensively tested for high-mass halos, it may be slightly different from an NFW profile for low-mass halos. As a result, other density profiles can be studied in addition to the standard ones in HOD models. For example, Yuan et al. (2018) investigated a radial profile for satellite positions.

3.4 Galaxy-halo connection of ELGs

We have seen the different techniques for modelling galaxies in simulations, from the most empirical to the most physical models (Figure 3.9). In this section, we look specifically at the connection between dark matter halos and ELGs (in particular [OII] emitters as targeted in DESI), and describe what are the characteristics of DM halos that host ELGs and how they cluster, particularly on small-scales.

3.4.1 The halo occupation of ELGs

Strong emission lines in galaxy spectra are strongly correlated with the galaxy star formation rate. For galaxy samples selected by luminosity or stellar mass, such as luminous red galaxies (LRGs), the average number of central galaxies $\langle N_{cent} \rangle$ is well described by a a smooth step function eventually reaching 1 for large-mass halos (Figure 3.15). For ELGs that are selected by their star formation rates, the halo occupation for central ELGs is different from a step function. It is closer to an asymmetric Gaussian distribution that does not reach unity, as shown in Figure 3.16 (Cowley et al., 2016, Geach et al., 2012, Gonzalez-Perez et al., 2018). The latter point implies that not all dark matter halos are expected to host an ELG as a central galaxy, contrary to what happens for LRGs which are assumed to be complete above a given mass (i.e. above this mass, all halos contain one central galaxy). These results for ELGs are based on the semi-analytic model of galaxy formation `GALFORM`. Hydrodynamic simulations (Hadzhiyska et al., 2021, Osato & Okumura, 2022, Yuan et al., 2022b) give a similar shape for the ELG central occupation distribution. The halo occupation for satellite ELGs is well represented by a power law, similar to that in standard HOD models.

Thus, different HOD models have been developed to reproduce the Gaussian shape of the central distribution. In eBOSS, Avila et al. (2020) and Alam et al. (2020) used different HOD models for central ELGs to reproduce the HOD shape obtained in SAMs:

➢ **Gaussian HOD (GHOD):**

$$\langle N_{cent}(M) \rangle = \frac{A_c}{\sqrt{2\pi}\sigma_{\rm M}} \cdot e^{-\frac{(\log_{10} M - \log_{10} M_c)^2}{2\sigma_{\rm M}^2}} \equiv \left\langle N_{cent}^{GHOD}(M) \right\rangle \tag{3.80}$$

In this model, $\langle N_{cent} \rangle$ is simply a Gaussian function with mean M_c, width σ_M and A_c defines the amplitude of the distribution. This model is compared with SAM predictions in the left panel of Figure 3.17 (labelled as HOD-2).

➢ **Star-Forming HOD (SFHOD):**

$$\langle N_{cent}(M) \rangle = \begin{cases} \left\langle N_{cent}^{GHOD}(M) \right\rangle & M \leq M_c \\ \frac{A_c}{\sqrt{2\pi}\sigma_{\rm M}} \cdot \left(\frac{M}{M_c}\right)^{\gamma} & M > M_c \end{cases} \tag{3.81}$$

This model is a combination of a Gaussian distribution for low-mass halos $< M_c$ and a decreasing power law for high-mass halos $> M_c$. The result is an asymmetric shape (see left panel of Figure 3.17, labelled HOD-3) where the asymmetry is controlled by the γ parameter. This function describes the SAM predictions well, but has the disadvantage of being discontinuous at $M = M_c$.

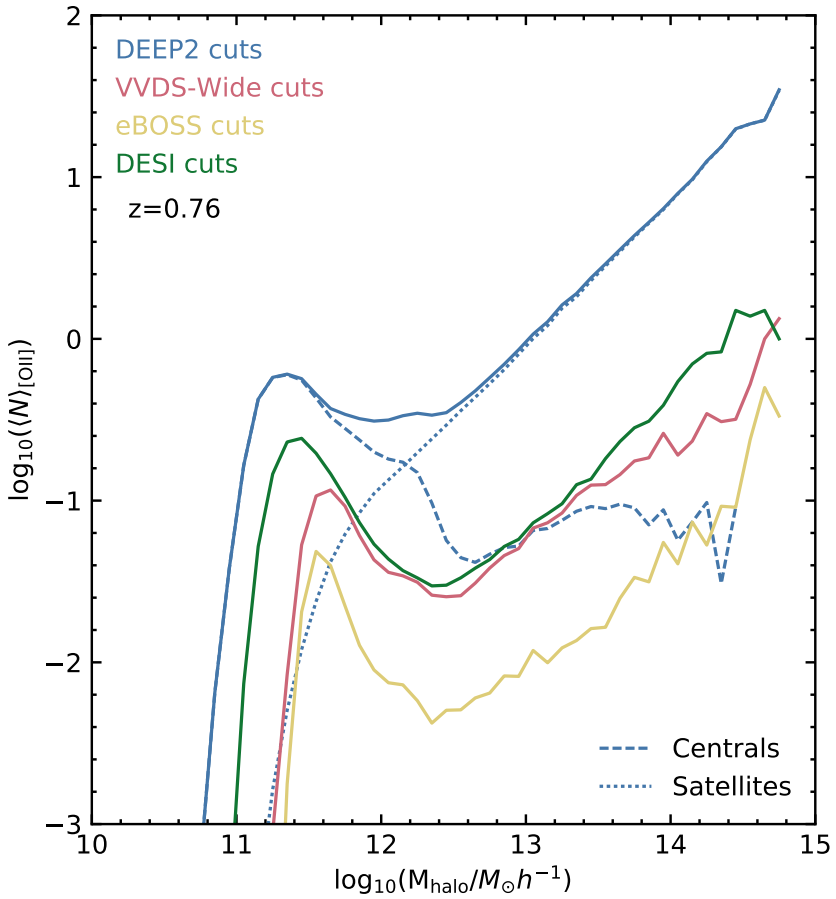

Figure 3.16: *The mean halo occupation distribution of model [OII] emitters (solid lines) for different photometric cuts (target selection) corresponding to different surveys including DESI prospects in green at $z = 0.76$. The contribution of centrals $\langle N_{cent} \rangle$ is represented in dashed lines and that of satellites in dotted lines (for DEEP2 cuts only). Figure taken from Gonzalez-Perez et al. (2018).*

➢ High mass quenched (HMQ):

$$\begin{aligned} \langle N_{cent}\left(M_h\right)\rangle &= 2A\phi\left(M_{\rm h}\right)\Phi\left(\gamma M_{\rm h}\right) + \frac{1}{2Q}\left[1+\mathrm{erf}\left(\frac{\log_{10}M_h - \log_{10}M_c}{0.01}\right)\right], \\ \phi(x) &= \mathcal{N}\left(\log_{10}M_c, \sigma_M\right), \\ \Phi(x) &= \int_{-\infty}^{x}\phi(t)dt = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \\ A &= \frac{p_{\rm max} - 1/Q}{\max(2\phi(x)\Phi(\gamma x))}. \end{aligned} \tag{3.82}$$

This model is a combination of a Gaussian function and an error function. p_{max} controls the amplitude of the low-mass Gaussian part relative to the high-mass plateau, whose level is set by Q that represents the quenching efficiency at high halo masses. The asymmetry of the Gaussian distribution is controlled by the parameter γ. The effect of the various parameters on the HMQ occupation function is illustrated in Figure 3.17 where the black line shows the fiducial model and each coloured line illustrates the impact of parameter variation.

Both SAM predictions and hydrodynamical simulations (Gonzalez-Perez et al., 2018, Hadzhiyska et al., 2021, Orsi & Angulo, 2018, Osato & Okumura, 2022, Yuan et al., 2022b) reveal that ELG

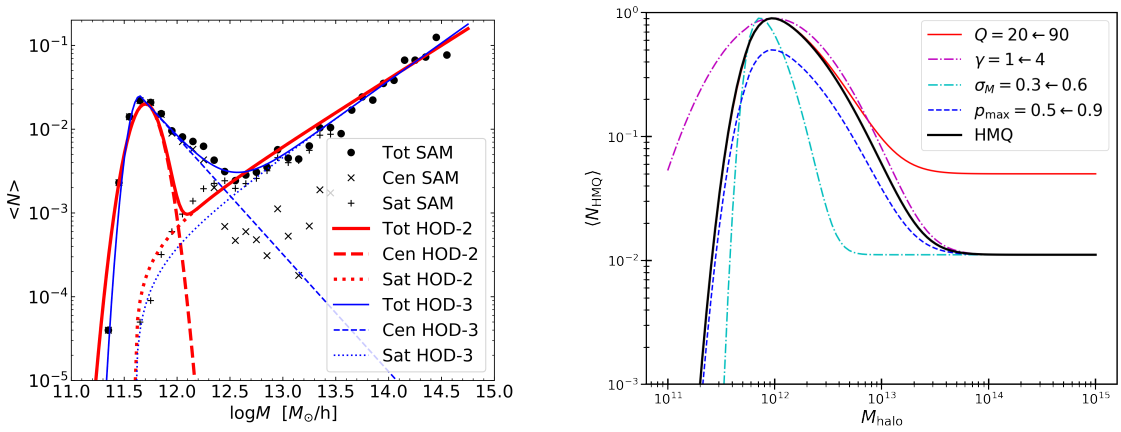

Figure 3.17: *Left: Mean number of ELG galaxies as a function of halo mass. The dots are the SAM results presented in Figure 3.16 from Gonzalez-Perez et al. (2018) with the central and satellite contributions as labelled. The red lines correspond to the Gaussian HOD model (labelled HOD-2) from Equation (7.9) and the blue lines are the star-forming HOD model (labelled HOD-3) from Equation (7.10). For both HOD models, the contribution of centrals (resp. satellites) is represented in dashed (resp. dotted) lines. Right: HOD of the high mass quenched model from (Alam et al., 2020). The effect of varying individual parameters is illustrated by coloured lines, while the solid black line represents the fiducial model. Solid red, dotted magenta, dotted cyan and dotted blue lines show the impact of parameters Q, γ, σ_M and p_{max} respectively, when varied from the fiducial values given in the legend.*

halo occupation peaks at around $\sim 10^{12} M_\odot$. This peak is robust to redshift, i.e. star formation occurs at roughly the same halo mass whatever the redshift, as shown in the left-hand panel of Figure 3.18 (Behroozi et al., 2013a). In terms of star formation history (right panel of Figure 3.18), the bulk of star formation occurs at redshift $z \sim 1.5 - 2$, when the Universe is predominantly matter-dominated. According to these results, a high density of ELGs is expected around this redshift, and they should mainly be hosted by halos of mass $\sim 10^{12} M_\odot$. SAM predictions (Gonzalez-Perez et al., 2018) and IllustrisTNG hydrodynamical simulations (Hadzhiyska et al., 2021) also provide the redshift evolution of HODs (see Figure 3.19). The shape of the distribution remains the same for all redshifts, but the peak slightly shifts towards higher halo masses with increasing redshift. HOD results on data (Alam et al., 2020, Avila et al., 2020, Favole et al., 2016, Guo et al., 2019, Lin et al., 2023, Okumura et al., 2021, Yuan et al., 2022a) report similar findings with a mean halo mass for ELG hosts around $\sim 10^{12} M_\odot$ with little redshift evolution.

Note: *In Figure 3.19 we note that the number of ELGs decreases with redshifts while we expect an increasing number of ELGs at $z \sim 1.5/2$. This is due to the photometric selection cuts (which do not evolve with redshift), as high redshift objects are fainter and therefore do not fulfill the selection criteria.*

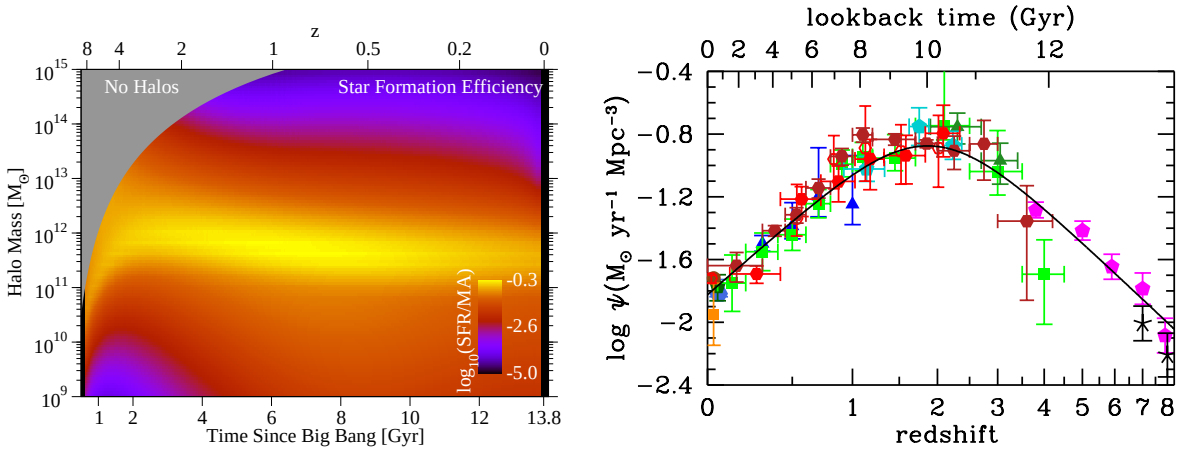

Figure 3.18: *Left: Star formation efficiency (the star formation rate divided by the halo mass accretion rate) as a function of redshift and halo mass. Figure taken from Wechsler & Tinker (2018) and originally from Behroozi et al. (2013a). Right: The history of cosmic star formation (star formation Ψ as a function of redshift). Data points with symbols are given in Table 1 from Madau & Dickinson (2014) where this figure is originally from. The peak of star formation in the history of the Universe is around redshift 2.*

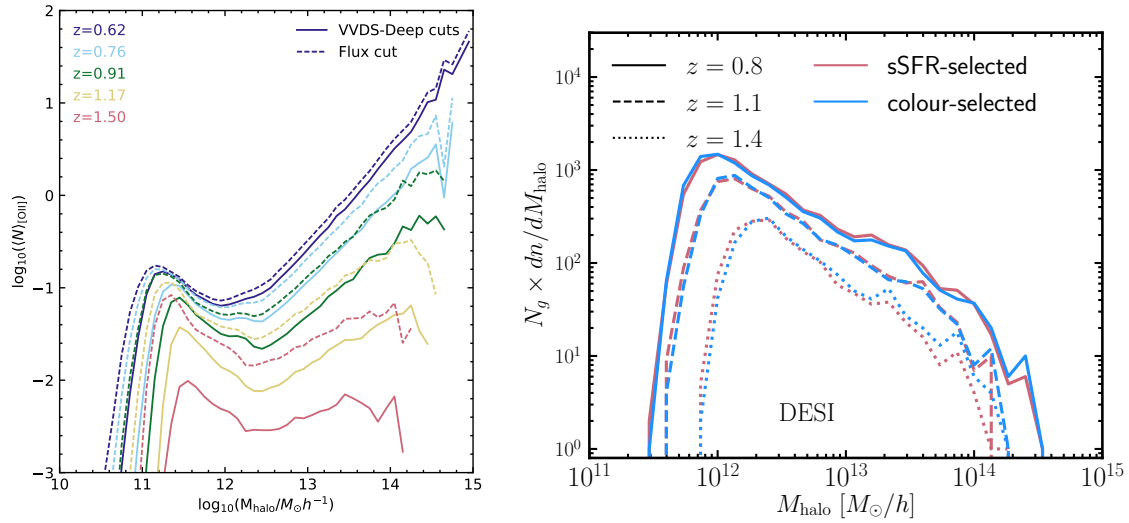

Figure 3.19: *Left: Model ELGs halo occupation distribution at different redshifts using two different sets of photometric selection cuts: VVDS-Deep (solid lines) and flux limited $F_{[OII]} > 1.9\times10^{-17} erg\cdot s^{-1}\cdot cm^{-2}$ (dashed-lines) from Gonzalez-Perez et al. (2018). Right: Total number of model galaxies per halo mass bin for 3 different redshift samples: $z = 0.8$ (solid line), $z = 1.1$ (dashed-line) and $z = 1.4$ (dotted-line) for 2 different selections, based on colour and sSFR (specific star-formation rate), from Hadzhiyska et al. (2021).*

3.4.2 Where are ELGs to be found ?

➢ The distribution within the cosmic web

The cosmic web is made up of knots, filaments, sheets and voids. These different environments exhibit different properties and are not clustered in the same way. In this section we review how ELGs and [OII] emitters trace the distribution of dark matter. A large number of observational studies such as GAMA (Kraljic et al., 2018), VIPERS (Malavasi et al., 2017) or COSMOS (Laigle et al., 2018) have found that star-forming and less massive galaxies are more likely to reside in filaments compared to quiescent and more massive galaxies that are found in denser regions (knots). Hydrodynamical simulations and SAM predictions for ELGs are in agreement with these findings (Gonzalez-Perez et al., 2018, Hadzhiyska et al., 2021, Osato & Okumura, 2022). Figure 3.20 shows a slice of the simulation box used in the SAM analysis where [OII] emitters are highlighted in blue circles and are found mostly in the filamentary structures of the cosmic web.

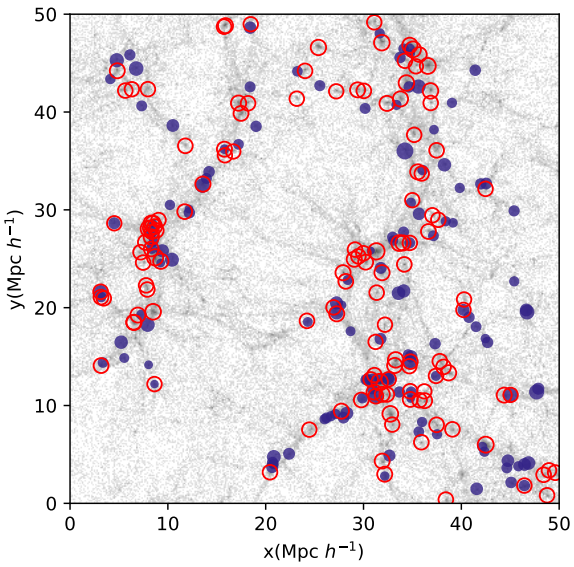

Figure 3.20: *Slice of a simulation box of volume* $50 \times 50 \times 10[Mpc/h]^3$ *at redshift* $z = 1$. *The cosmic web of the dark matter is represented in grey. The locations of [OII] emitters are indicated by the filled circles and the dark matter halos above* $10^{11.8} M_{\odot}/h$ *by the open circles. This figure is taken from Gonzalez-Perez et al. (2018).*

➢ Influence from assembly history of DM halos?

The aim is to see whether the assembly history of DM halos has an impact on the presence of ELGs, i.e. whether the assembly history triggered star formation. One way of looking at the impact of assembly history is to compare the clustering of ELGs from SAMs or hydrodynamical simulations with that of the same sample where ELGs are shuffled between halos of the same mass to erase any assembly history. Results from the literature have shown that the ELG clustering is influenced by the assembly history and secondary properties of DM halos (Contreras et al.,

2019, Jiménez et al., 2021, Xu et al., 2021, Zehavi et al., 2019). Figure 3.21 of Jiménez et al. (2021) shows the impact of assembly history for sample ELGs with different selections (including [OII]) using the Semi Analytical Galaxy (SAG) model of galaxy formation (Cora et al., 2018). Clustering of the [OII]-selected sample is $\sim$ 10 to 20% lower on large scales compared to the shuffled clustering, depending on the galaxy number density. Note that with increasing density, the difference is reduced, which is expected as there is a limited number of DM halos, given by the initial halo mass function. Therefore, if more halos are populated by galaxies, the impact of the assembly history is reduced. We note however that the impact reduction is weak in the case of samples selected by their [OII] emission.

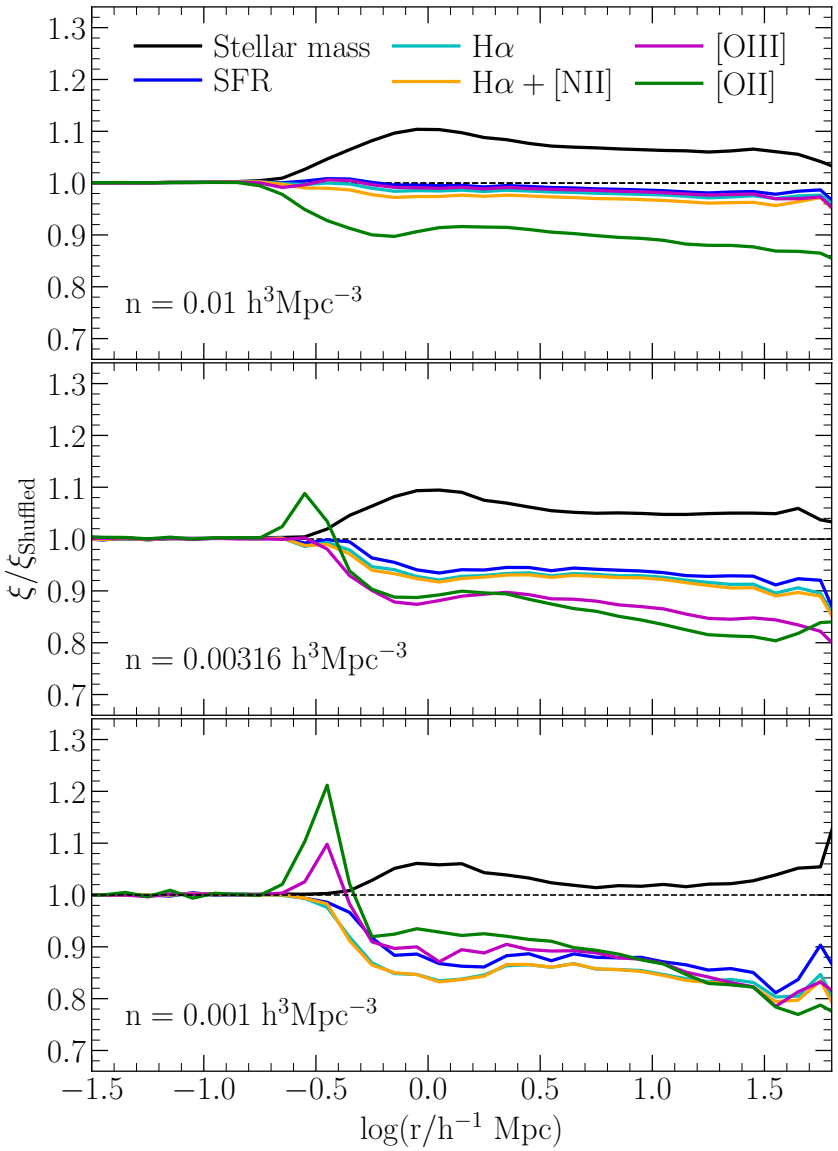

Figure 3.21: *Ratio of the predicted clustering for ELG samples selected by different properties to that of their corresponding shuffled sample. Each panel shows a different number density, as labelled.*

Furthermore, the assembly history tends to reduce the clustering power on large scales in a scale-dependent manner, meaning that the ELG bias depends on the scales considered (Jiménez et al., 2021). This is problematic for cosmological studies because the full shape modelling of the

2PCF assumes a linear bias on linear scales, so that a scale-dependent bias of the ELG samples could bias cosmological results.

Recent results using hydrodynamical simulations have led to a better understanding of the small-scale clustering of ELGs and the impact of secondary properties of DM halos hosting ELGs (Hadzhiyska et al., 2021, 2022b, Yuan et al., 2022b). Hadzhiyska et al. (2022b) use MilleniumTNG, a large-volume hydrodynamical simulation (box size of 500 Mpc/h), to extract a "true" sample of ELG-like galaxies and compare it to HOD predictions taking into account the mass only HOD and improved HOD models that take into account secondary properties of DM halos as follows:

$$\begin{aligned} \langle N'_{cent}(M)\rangle &= [1 + (a_{cent} f_a + b_{cent} f_b)(1 - \langle N_{cent}(M)\rangle)]\, \langle N_{cent}(M)\rangle \\ \langle N'_{sat}(M)\rangle &= [1 + (a_{sat} f_a + b_{sat} f_b)]\, \langle N_{sat}(M)\rangle \end{aligned} \tag{3.83}$$

The mean numbers of the HOD are modified according to the secondary halo parameters a and b. f_a and f_b are the normalised rank-ordered halo properties, a_{cen} (a_{sat}) and b_{cen} (b_{sat}) are free parameters for the entire central (satellite) sample, $\langle N_{cent}(M)\rangle$ ($\langle N_{sat}(M)\rangle$) are the mean number of centrals (satellites) in the mass bin of the halo under consideration. To compute f_a or f_b, halos in the mass bin are first ranked by decreasing values of the halo property and each halo is attributed a different value of f, which is a user-defined function of the considered property (e.g. a linear function decreasing between 0.5 and -0.5 when going from the top ranked halo to the last one). Figure Figure 3.22 shows the comparison between the clustering of "true" ELG samples and HOD models considering the mass-only HOD model and the improved HOD model with different secondary properties of DM halos:

- `conc`: halo concentration (see Section 3.1.5),
- `Mass peak`: highest mass achieved in the entire halo history,
- `VelAni`: velocity anisotropy, measured by the ratio of the tangential and radial velocity dispersions of the halo particles,
- `R Splash`: splashback radius[1], the radius where particles reach the apocentre of their first orbit which is a physically motivated definition of the halo boundary (Diemer & Kravtsov, 2014),
- `EnvAdapt` : local environment, i.e. local density around the halo,
- `ShearAdapt` : local *shear*, i.e. tidal environment (amount of anisotropic pulling due to gravity) around the halo.

Figure 3.22 displays results for ELG and LRG samples at redshift 0 and 1. The quantity indicated in Figure 3.22 is the difference between the 2PCF quadrupole of the true ELG sample and the HOD models divided by the monopole of the true sample $(\xi_{2,pred} - \xi_{2,true})/\xi_{0,true}$. Solid lines with error bars are the results from the improved HOD model which takes into account a given secondary parameter. Dotted lines with shaded areas are the results of the mass only HOD. Focusing on ELGs only, we see that the clustering is slightly influenced by secondary parameters, the parameter with the greatest impact being concentration, which provides better

[1] A quick introduction with lots of reference on the splashback radius can be found here

agreement on average compared to the "true" clustering on all scales ($r > 10\text{Mpc}/h$). The splashback radius also improves the agreement at small scales. However, none of the tested secondary parameters solves completely the disagreement at small scales ($r < 1\text{Mpc}/h$). In their studies Hadzhiyska et al. (2022b) also reported similar results in real space, combining the secondary properties of DM halos in pairs and reporting that shear and environment show slightly better agreement with the truth than each property separately. On the data side, only a few studies have examined the impact of assembly bias on the ELG sample. Among them, Lin et al. (2023) used a secondary halo property related to halo assembly history and found that the projected clustering of eBOSS ELGs matched better with assembly bias included.

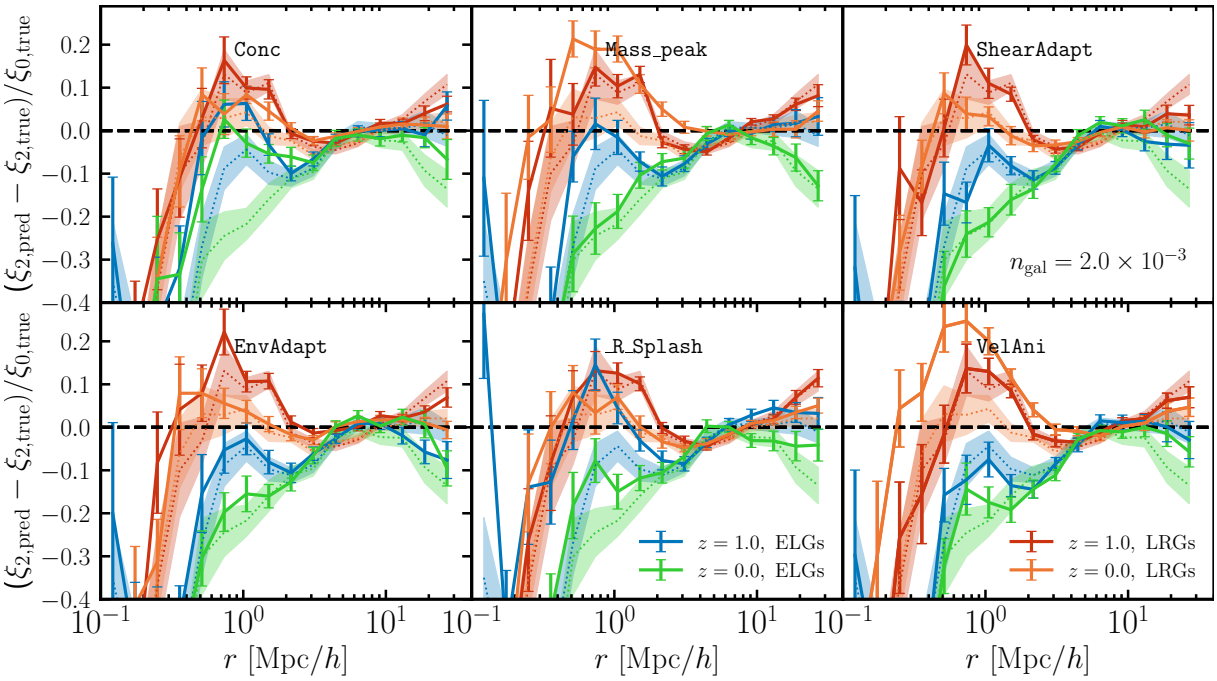

Figure 3.22: *Difference in clustering quadrupole between predicted and "true" ELG and LRG samples at $z = 1$ and $z = 0$. Predicted samples are obtained with HOD models either mass-dependent only (dotted lines) or considering secondary bias as in Equation (3.83) (solid line with error bars). Each panel adopts a different halo property as a secondary assembly bias proxy.*

➢ Positions and velocities within their host halos

After focusing on the environment of halos hosting ELGs, we turn our attention to their positions and velocities relative to their host. Based on `LGALAXIES`, a SAM of galaxy formation (Munich group), run on the MXXL (Angulo et al., 2012) N-body simulation, Orsi & Angulo (2018) studied the satellite kinematics of ELGs and argue that the quenching of star formation rate induced by gas stripping processes decreases the fraction of satellite ELGs in the inner part of DM halos. They show that satellite ELGs are made up of two populations with different properties: the first one corresponds to objects affected by gas-stripping processes but still forming stars, which occupy a wide range of radial positions within their parent halo, with infall velocities well described by a zero-centred Gaussian. The second one corresponds to recently accreted satellites that populate the outskirts of their parent halo, with a dominant infall velocity component. This can be seen in Figure 3.23, which displays the intra-halo radial distribution (left panel)

and tangential and radial velocities (right panel) of LRGs and ELGs, specifically highlighting the contribution of recently accreted galaxies ($t_{\rm infall} < 680$Myr). The infall contribution of these recently accreted galaxies is much higher in the case of ELGs. For LRGs, the contribution of recently accreted galaxies is very low. LRGs are more uniformly distributed in the halos, with a much less pronounced infall velocity.

Recent results from hydrodynamical simulations with IllustrisTNG (Yuan et al., 2022b) and MilleniumTNG (Hadzhiyska et al., 2022b) also indicate that satellites are more likely to be found at the periphery of DM halos, with a small proportion of satellites at a distance up to 3 times greater than the halo virial radius. But Yuan et al. (2022b) associate this finding to a possible side-effect of the halo finding algorithm. This publication also reports a slight velocity bias for satellite model ELGs and a modest one for central model ELGs, which exhibit velocity dispersions lower by a few percent than their host halos.

The above results are based on hydrodynamical simulations or SAMs, which rely on prescriptions to describe galaxy formation processes. Confirmation with data is still missing but would be essential to confirm the validity of these prescriptions. The clustering of ELGs at sub-halo scales is difficult to probe with data because we need to resolve very small separation scales that are strongly affected by fibre collisions in spectroscopic samples (see Section 2.7.1.1). Avila et al. (2020) carried out an HOD study on the ELG eBOSS sample and tested the dependence of the ELG clustering on the DM halo density profile (using either particles or the NFW profile), with measurements of the projected clustering w_p at small scales $0.2 < r_p < 4$Mpc$/h$. More dispersed satellite profiles w.r.t NFW were found to be preferred by data. Moreover, they report a positive velocity bias for satellites, i.e. greater velocity dispersions w.r.t. the halo velocity, possibly due to large infall velocities.

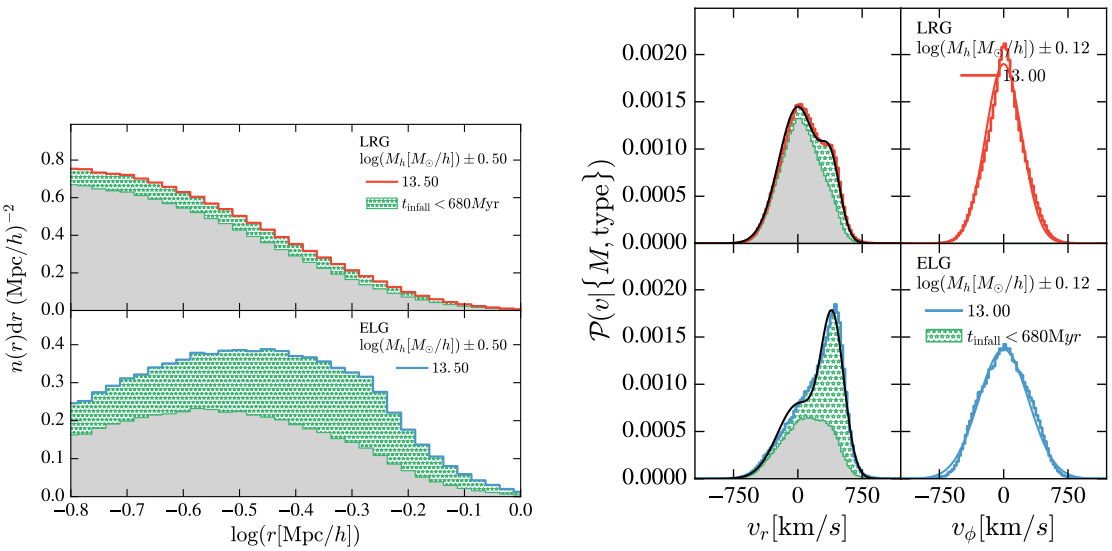

Figure 3.23: *The intra-halo radial distribution (left panel) and the intra-halo velocity distribution (right panel) of satellites from model LRG samples (top) and ELG samples (bottom) in host halos of mass* $\log_{10}(M_h\ [M_\odot/h]) = 13.5 \pm 0.5$. *The green dotted region indicates the contribution of satellites accreted into the host halo within the last* ~ 680 *Myr. Figures taken from Orsi & Angulo (2018).*

➢ Satellite fraction and distribution

The satellite fraction of ELGs can be determined in SAMs or hydrodynamical simulations, but also measured in data. First, let's define what we mean by a satellite galaxy. A satellite galaxy is considered to be enclosed within the radius of the halo and to orbit a central galaxy. This idea was introduced for mass-limited samples such as LRGs, meaning that LRG satellites are found associated with a central LRG. For ELGs, the story is a little different, as ELG satellites are not necessarily associated with a central ELG (they may be associated with a central LRG for example). The fraction of satellites is the number of galaxies of a given galaxy type that are considered to be satellite over the total number of galaxies of this type in the sample:

$$f_{sat} = \frac{N_{sat}}{N_{cen} + N_{sat}} \tag{3.84}$$

where N_{cen} and N_{sat} are the number of centrals and satellites of the given type. *For ELGs, we cannot infer the number of ELG central-satellite pairs from the ELG satellite fraction because satellite ELGs are not necessarily associated to a central ELG, contrary to LRGs.*

In theoretical models, we can determine whether an ELG is central or satellite. HOD models determine the satellite fraction by counting the number of objects drawn from the satellite HOD, and SHAM methods count the number of populated sub-halos. In the literature, a large number of studies using SAMs or hydrodynamical simulations have reported a wide range of satellite fractions for ELGs from $\sim 2\%$ to $\sim 40\%$. All models agree that ELGs are most likely central in low mass halos $< 10^{12.5} M_\odot/h$ and satellites in higher-mass halos (Gonzalez-Perez et al., 2018, Jiménez et al., 2019, Orsi & Angulo, 2018, Osato & Okumura, 2022, Yuan et al., 2022b). Osato & Okumura (2022) fitted their truth sample (based on IllustrisTNG hydrodynamical simulations) with a satellite fraction of 28% with different HOD models and found the resulting satellite fraction to range from 25 to 50% depending on the models. Thus, the satellite fraction of ELGs is poorly constrained by the physical models. The data results also show a wide range of satellite fractions depending on the method used (HOD or SHAM-SHMR) to fit the data. For example, for the eBOSS ELG sample, Guo et al. (2019) use the SHAM plus SHMR method and report a satellite fraction of $\sim 13\% - 17\%$ depending on the redshift in $0.7 < z < 1.2$. Lin et al. (2023) report $\sim 19\%$ with a SHAM analysis. Two HOD multi-tracer analyses on eBOSS data used the HMQ model (Equation (3.82)) and report ELG satellite fractions of ~ 7 and $\sim 17\%$, respectively (Alam et al., 2020, Yuan et al., 2022a). The former also report $\sim 12\%$ with the standard HOD model (Equation (3.75)). Once again, these results highlight the difficulty of constraining the satellite fraction of ELGs. Most results report satellite fractions in the range $\sim 10 - 30\%$.

The last point we address on satellite ELGs is the Poissonian shape of the satellite distribution. Recent results from SAMs and hydrodynamical simulations have shown that the ELG satellite population more likely exhibits a super-Poissonian behaviour rather than a Poissonian one in high-mass halos $\geq 10^{13} M_\odot/h$ (Hadzhiyska et al., 2022b, Jiménez et al., 2019), i.e. the mean number of satellites for a given halo mass remains the same but the variance increases.

3.4.2.1 ELG central-satellite conformity

Galactic conformity was introduced in Weinmann et al. (2006), which reports that the properties of satellite galaxies in SDSS data are strongly correlated with those of the central galaxy in their halo. They found that this correlation is even more important for early-type galaxies (such as

ELGs): "In particular, the early type fraction of satellites is significantly higher in a halo with an early type central galaxy than in a halo of the same mass but with a late type central galaxy". Since then, other studies have found a significant trend in favour of galactic conformity (Kauffmann et al., 2013, Knobel et al., 2015, Phillips et al., 2014, Robotham et al., 2013, Wang & White, 2012).

The results from Hadzhiyska et al. (2022b) based on hydrodynamical simulations show that the probability of a halo having a central ELG if it has at least one ELG satellite is about twice as large as the probability of a halo having a central galaxy whatever the number of satellites, in the mass interval $10^{12}-10^{13}M_{\odot}/h$, as illustrated in Figure 3.24. This may suggest that the presence of ELG satellites is more likely if the halo already hosts a central ELG. Several implementations of conformity in HOD models have been tested in the literature, conditioning the presence of satellite galaxies on prior information if the halo hosts a central galaxy or not (Alam et al., 2020, Hadzhiyska et al., 2022b, Jiménez et al., 2019). The latter studied the HOD of the eBOSS ELG and LRG samples (auto and cross correlations) and reported a signature of *1-halo galactic conformity* (central-satellite conformity) at more that 3σ of statistical significance.

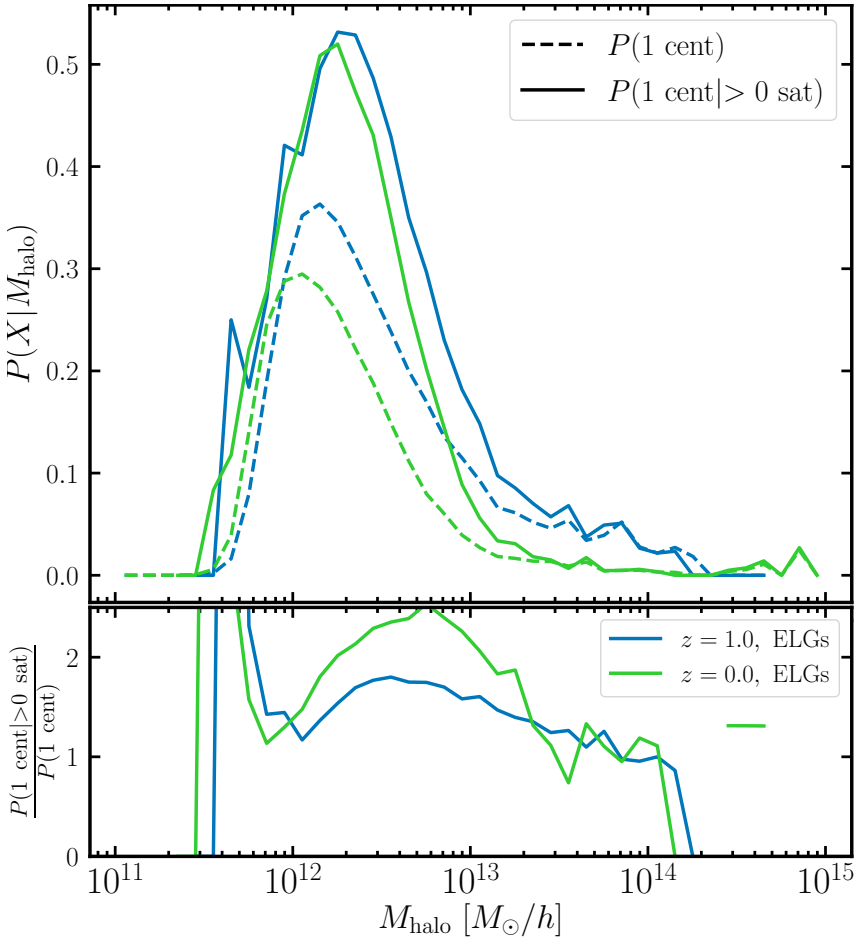

Figure 3.24: *Probability distribution of the ELG centrals at $z = 1$ (blue) and $z = 0$ (green) as a function of halo mass. The dashed lines indicate the probability that a halo of a given mass contains a central, and the solid lines correspond to the conditional probability that a halo contains a central given that it hosts one or more satellites. The lower panel shows the ratio of the solid to the dashed lines for each sample. Figure taken from Hadzhiyska et al. (2022b).*

In the literature, the presence of *2-halo conformity* is also observed, i.e. conformity extends to scales beyond the virial radius of the halos (up to 4Mpc/h for projected separations (Kauffmann et al., 2013)). One possible explanation for this effect is that halo pairs at these distances reside in the same large-scale tidal environment and that the mass accretion rates of dark matter halos are strongly correlated at large scales, leading to conformity between galaxies pairs at larger scales (up to thirty times the virial radius of either halo (Hearin et al., 2016)). However, this result is still debated as possibly due to selection bias (Lacerna et al., 2018, Sin et al., 2017, Tinker et al., 2018).

3.4.3 Global picture of ELG-dark matter connection

The previous sections describe what was known (or expected from models and simulations) about ELGs before DESI started. In DESI we select only a sub-sample of ELGs, those that are [OII] emitters, because we rely on the [OII] doublet ($3726 - 3729$ Å) to assess the redshift. To summarise, ELG is a generic term that refers to a population of (predominantly) star-forming galaxies and, at most $\sim 10\%$ of those are [OII] emitters in the redshift range $0.6 < z < 1.5$. This is consistent with the overall picture of comic history of star formation that peaks around redshift $\sim 1.5 - 2$ with the highest fraction of SFR in halos of mass $\sim 10^{12} M_\odot/h$. ELGs mainly reside in the filamentary structure of the cosmic web i.e. in less dense environments than LRGs. The mean mass of the halos hosting ELGs is around $\sim 10^{12} M_\odot/h$. In low mass halos $< 10^{12.5} M_\odot/h$, ELGs are mainly centrals, while they are mostly satellites in higher mass halos. ELGs are sensitive to assembly history at all scales, with larger dependence regarding shear, environment and concentration as halo secondary parameters. The satellite fraction of ELGs is found to be around $\sim 10 - 30\%$ depending on the model considered, but remains poorly constrained. ELG satellites are located mainly in the outskirts of the halos with a large component of infall velocity.

The physical picture: Most ELGs are active star-forming galaxies. Star formation generates strong emission lines in the intergalactic medium. One of the characteristic indicators of star formation is the [OII] doublet ($3726-3729$ Å), which is the strongest feature after the Hα Balmer line (6562.8 Å) (Moustakas et al., 2006). The [OII] doublet has the advantage of being easily identifiable and can be seen in optical spectra with moderate resolution up to high redshifts ($z \sim 1.6$), whereas the Hα line becomes inaccessible from ground observation at lower redshift. ELGs are mainly found in halos of mass $\sim 10^{12} M_\odot/h$. Star formation of a galaxy is modulated by feedback and environmental effects. It is expected that massive galaxies located in large halos are more often subject to strong star formation regulation mechanisms (e.g. AGN feedback) and therefore exhibit low star formation rates. In addition, low star formation rates are also expected in low-mass halos $< 10^{11} M_\odot$ due to a combination of other extinction mechanisms (e.g. supernova feedback) and the small amount of baryonic gas available. To be detected, satellite ELGs in high-mass halos need to be recently accreted by the halos to keep a significant SFR. Indeed, if an ELG enters a high-mass halo, its star formation rate drops rapidly, turning it into a red galaxy in ~ 1 Gyr. For this reason, ELGs exhibit high infall velocities towards the halo centres and are most often located in the outskirts of massive halos.

Bibliography

Aarseth, S. J., Gott, III, J. R., & Turner, E. L. 1979, The Astrophysical Journal, 228, 664, doi: 10.1086/156892

Abdalla, E., Abellán, G. F., Aboubrahim, A., et al. 2022, Journal of High Energy Astrophysics, 34, 49, doi: 10.1016/j.jheap.2022.04.002

Alam, S., Peacock, J. A., Kraljic, K., Ross, A. J., & Comparat, J. 2020, Monthly Notices of the Royal Astronomical Society, 497, 581, doi: 10.1093/mnras/staa1956

Alam, S., de Mattia, A., Tamone, A., et al. 2021, Monthly Notices of the Royal Astronomical Society, 504, 4667, doi: 10.1093/mnras/stab1150

Angulo, R. E., Springel, V., White, S. D. M., et al. 2012, Monthly Notices of the Royal Astronomical Society, 426, 2046, doi: 10.1111/j.1365-2966.2012.21830.x

Appel, A. W. 1985, SIAM Journal on Scientific and Statistical Computing, 6, 85, doi: 10.1137/0906008

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, Monthly Notices of the Royal Astronomical Society, 499, 5486, doi: 10.1093/mnras/staa2951

Barnes, J., & Hut, P. 1986, Nature, 324, 446, doi: 10.1038/324446a0

Baugh, C. M., Gaztañaga, E., & Efstathiou, G. 1995, Monthly Notices of the Royal Astronomical Society, 274, 1049, doi: 10.1093/mnras/274.4.1049

Behroozi, P. S., Wechsler, R. H., & Conroy, C. 2013a, The Astrophysical Journal, 770, 57, doi: 10.1088/0004-637X/770/1/57

Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013b, The Astrophysical Journal, 762, 109, doi: 10.1088/0004-637X/762/2/109

Berlind, A. A., & Weinberg, D. H. 2002, The Astrophysical Journal, 575, 587, doi: 10.1086/341469

Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. 2002, Physics Reports, 367, 1, doi: 10.1016/S0370-1573(02)00135-7

Blanton, M. R., Schlegel, D. J., Strauss, M. A., et al. 2005, The Astronomical Journal, 129, 2562, doi: 10.1086/429803

Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, The Astrophysical Journal, 379, 440, doi: 10.1086/170520

Bose, S., Eisenstein, D. J., Hadzhiyska, B., Garrison, L. H., & Yuan, S. 2022, Monthly Notices of the Royal Astronomical Society, 512, 837, doi: 10.1093/mnras/stac555

Bower, R. G., Benson, A. J., Malbon, R., et al. 2006, Monthly Notices of the Royal Astronomical Society, 370, 645, doi: 10.1111/j.1365-2966.2006.10519.x

Bower, R. G., Vernon, I., Goldstein, M., et al. 2010, Monthly Notices of the Royal Astronomical Society, 407, 2017, doi: 10.1111/j.1365-2966.2010.16991.x

Bryan, G. L., & Norman, M. L. 1998, The Astrophysical Journal, 495, 80, doi: 10.1086/305262

Cattaneo, A. 2019, Nature Astronomy, 3, 896, doi: 10.1038/s41550-019-0904-y

Centrella, J., & Melott, A. L. 1983, Nature, 305, 196, doi: 10.1038/305196a0

Chuang, C.-H., Yepes, G., Kitaura, F.-S., et al. 2019, Monthly Notices of the Royal Astronomical Society, 487, 48, doi: 10.1093/mnras/stz1233

Ciotti, L., D'Ercole, A., Pellegrini, S., & Renzini, A. 1991, The Astrophysical Journal, 376, 380, doi: 10.1086/170289

Cole, S., Lacey, C., Baugh, C., & Frenk, C. 2000, Monthly Notices of the Royal Astronomical Society, 319, 168, doi: 10.1046/j.1365-8711.2000.03879.x

Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2006, The Astrophysical Journal, 647, 201, doi: 10.1086/503602

Contreras, S., Baugh, C., Norberg, P., & Padilla, N. 2013, Monthly Notices of the Royal Astronomical Society, 432, 2717, doi: 10.1093/mnras/stt629

Contreras, S., Zehavi, I., Padilla, N., et al. 2019, Monthly Notices of the Royal Astronomical Society, 484, 1133, doi: 10.1093/mnras/stz018

Cora, S. A., Vega-Martínez, C. A., Hough, T., et al. 2018, Monthly Notices of the Royal Astronomical Society, 479, 2, doi: 10.1093/mnras/sty1131

Cowley, W. I., Lacey, C. G., Baugh, C. M., & Cole, S. 2016, Monthly Notices of the Royal Astronomical Society, 461, 1621, doi: 10.1093/mnras/stw1069

Croton, D. J., Gao, L., & White, S. D. M. 2007, Monthly Notices of the Royal Astronomical Society, 374, 1303, doi: 10.1111/j.1365-2966.2006.11230.x

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, The Astrophysical Journal, 292, 371, doi: 10.1086/163168

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, The Astronomical Journal, 145, 10, doi: 10.1088/0004-6256/145/1/10

de Sá-Freitas, C., Gonçalves, T. S., de Carvalho, R. R., et al. 2021, Monthly Notices of the Royal Astronomical Society, 509, 3889, doi: 10.1093/mnras/stab3230

Diemand, J., Kuhlen, M., & Madau, P. 2006, The Astrophysical Journal, 649, 1, doi: 10.1086/506377

—. 2007, The Astrophysical Journal, 657, 262, doi: 10.1086/510736

Diemer, B., & Kravtsov, A. V. 2014, The Astrophysical Journal, 789, 1, doi: 10.1088/0004-637X/789/1/1

Doroshkevich, A. G., Kotok, E. V., Novikov, I. D., et al. 1980, Monthly Notices of the Royal Astronomical Society, 192, 321, doi: 10.1093/mnras/192.2.321

Efstathiou, G., & Eastwood, J. W. 1981, Monthly Notices of the Royal Astronomical Society, 194, 503, doi: 10.1093/mnras/194.3.503

Favole, G., Comparat, J., Prada, F., et al. 2016, Monthly Notices of the Royal Astronomical Society, 461, 3421, doi: 10.1093/mnras/stw1483

Feng, Y., Chu, M.-Y., Seljak, U., & McDonald, P. 2016, Monthly Notices of the Royal Astronomical Society, 463, 2273, doi: 10.1093/mnras/stw2123

Frenk, C. S., White, S. D. M., & Davis, M. 1983, The Astrophysical Journal, 271, 417, doi: 10.1086/161209

Gao, L., & White, S. D. M. 2007, Monthly Notices of the Royal Astronomical Society: Letters, 377, L5, doi: 10.1111/j.1745-3933.2007.00292.x

Garrison, L. H., Eisenstein, D. J., Ferrer, D., Maksimova, N. A., & Pinto, P. A. 2021, Monthly Notices of the Royal Astronomical Society, 508, 575, doi: 10.1093/mnras/stab2482

Garrison, L. H., Eisenstein, D. J., & Pinto, P. A. 2019, Monthly Notices of the Royal Astronomical Society, 485, 3370, doi: 10.1093/mnras/stz634

Geach, J. E., Sobral, D., Hickox, R. C., et al. 2012, Monthly Notices of the Royal Astronomical Society, 426, 679, doi: 10.1111/j.1365-2966.2012.21725.x

Giocoli, C., Pieri, L., & Tormen, G. 2008, Monthly Notices of the Royal Astronomical Society, 387, 689, doi: 10.1111/j.1365-2966.2008.13283.x

Gonzalez-Perez, V., Lacey, C. G., Baugh, C. M., et al. 2014, Monthly Notices of the Royal Astronomical Society, 439, 264, doi: 10.1093/mnras/stt2410

Gonzalez-Perez, V., Comparat, J., Norberg, P., et al. 2018, Monthly Notices of the Royal Astronomical Society, 474, 4024, doi: 10.1093/mnras/stx2807

Gonzalez-Perez, V., Cui, W., Contreras, S., et al. 2020, Monthly Notices of the Royal Astronomical Society, 498, 1852, doi: 10.1093/mnras/staa2504

Gunn, J. E., & Gott, III, J. R. 1972, The Astrophysical Journal, 176, 1, doi: 10.1086/151605

Guo, H., Yang, X., Raichoor, A., et al. 2019, The Astrophysical Journal, 871, 147, doi: 10.3847/1538-4357/aaf9ad

Guo, Q., White, S., Angulo, R. E., et al. 2013, Monthly Notices of the Royal Astronomical Society, 428, 1351, doi: 10.1093/mnras/sts115

Habib, S., Pope, A., Lukić, Z., et al. 2009, Journal of Physics: Conference Series, 180, 012019, doi: 10.1088/1742-6596/180/1/012019

Hadzhiyska, B., Eisenstein, D., Bose, S., Garrison, L. H., & Maksimova, N. 2022a, Monthly Notices of the Royal Astronomical Society, 509, 501, doi: 10.1093/mnras/stab2980

Hadzhiyska, B., Tacchella, S., Bose, S., & Eisenstein, D. J. 2021, Monthly Notices of the Royal Astronomical Society, 502, 3599, doi: 10.1093/mnras/stab243

Hadzhiyska, B., Hernquist, L., Eisenstein, D., et al. 2022b, The MillenniumTNG Project: Refining the one-halo model of red and blue galaxies at different redshifts, arXiv. http://arxiv.org/abs/2210.10068

Hahn, O., Rampf, C., & Uhlemann, C. 2021, Monthly Notices of the Royal Astronomical Society, 503, 426, doi: 10.1093/mnras/staa3773

Hatton, S., Devriendt, J. E. G., Ninin, S., et al. 2003, Monthly Notices of the Royal Astronomical Society, 343, 75, doi: 10.1046/j.1365-8711.2003.05589.x

Hearin, A. P., Zentner, A. R., Bosch, F. C. v. d., Campbell, D., & Tollerud, E. 2016, Monthly Notices of the Royal Astronomical Society, 460, 2552, doi: 10.1093/mnras/stw840

Henriques, B. M. B., Thomas, P. A., Oliver, S., & Roseboom, I. 2009, Monthly Notices of the Royal Astronomical Society, 396, 535, doi: 10.1111/j.1365-2966.2009.14730.x

Henriques, B. M. B., White, S. D. M., Thomas, P. A., et al. 2015, Monthly Notices of the Royal Astronomical Society, 451, 2663, doi: 10.1093/mnras/stv705

Hernquist, L., & Katz, N. 1989, The Astrophysical Journal Supplement Series, 70, 419, doi: 10.1086/191344

Hernández-Aguayo, C., Springel, V., Pakmor, R., et al. 2023, The MillenniumTNG Project: High-precision predictions for matter clustering and halo statistics, arXiv, doi: 10.48550/arXiv.2210.10059

Hockney, R. W. 1988, Computer simulation using particles (Bristol [England] ; Philadelphia : A. Hilger). http://archive.org/details/computersimulati0000hock

Hopkins, P. F., Quataert, E., & Murray, N. 2012, Monthly Notices of the Royal Astronomical Society, 421, 3522, doi: 10.1111/j.1365-2966.2012.20593.x

Hou, J., Frenk, C. S., Lacey, C. G., & Bose, S. 2016, Monthly Notices of the Royal Astronomical Society, 463, 1224, doi: 10.1093/mnras/stw2033

Huchra, J. P., & Geller, M. J. 1982, The Astrophysical Journal, 257, 423, doi: 10.1086/160000

Ishiyama, T., Fukushige, T., & Makino, J. 2009, Publications of the Astronomical Society of Japan, 61, 1319, doi: 10.1093/pasj/61.6.1319

Ishiyama, T., Nitadori, K., & Makino, J. 2012, 4.45 Pflops Astrophysical N-Body Simulation on K computer – The Gravitational Trillion-Body Problem, doi: 10.48550/arXiv.1211.4406

Ishiyama, T., Prada, F., Klypin, A. A., et al. 2021, Monthly Notices of the Royal Astronomical Society, 506, 4210, doi: 10.1093/mnras/stab1755

Jiménez, E., Contreras, S., Padilla, N., et al. 2019, Monthly Notices of the Royal Astronomical Society, 490, 3532, doi: 10.1093/mnras/stz2790

Jiménez, E., Padilla, N., Contreras, S., et al. 2021, Monthly Notices of the Royal Astronomical Society, 506, 3155, doi: 10.1093/mnras/stab1819

Kauffmann, G., Colberg, J. M., Diaferio, A., & White, S. D. M. 1999, Monthly Notices of the Royal Astronomical Society, 303, 188, doi: 10.1046/j.1365-8711.1999.02202.x

Kauffmann, G., Li, C., Zhang, W., & Weinmann, S. 2013, Monthly Notices of the Royal Astronomical Society, 430, 1447, doi: 10.1093/mnras/stt007

Kauffmann, G., White, S. D. M., & Guiderdoni, B. 1993, Monthly Notices of the Royal Astronomical Society, 264, 201, doi: 10.1093/mnras/264.1.201

Klypin, A., Kravtsov, A. V., Valenzuela, O., & Prada, F. 1999, The Astrophysical Journal, 522, 82, doi: 10.1086/307643

Klypin, A., Yepes, G., Gottlöber, S., Prada, F., & Heß, S. 2016, Monthly Notices of the Royal Astronomical Society, 457, 4340, doi: 10.1093/mnras/stw248

Knebe, A., Knollmann, S. R., Muldrew, S. I., et al. 2011, Monthly Notices of the Royal Astronomical Society, 415, 2293, doi: 10.1111/j.1365-2966.2011.18858.x

Knobel, C., Lilly, S. J., Woo, J., & Kovac, K. 2015, The Astrophysical Journal, 800, 24, doi: 10.1088/0004-637X/800/1/24

Knollmann, S. R., & Knebe, A. 2009, The Astrophysical Journal Supplement Series, 182, 608, doi: 10.1088/0067-0049/182/2/608

Kraljic, K., Arnouts, S., Pichon, C., et al. 2018, Monthly Notices of the Royal Astronomical Society, 474, 547, doi: 10.1093/mnras/stx2638

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, The Astrophysical Journal, 609, 35, doi: 10.1086/420959

Lacerna, I., Contreras, S., González, R. E., Padilla, N., & Gonzalez-Perez, V. 2018, Monthly Notices of the Royal Astronomical Society, 475, 1177, doi: 10.1093/mnras/stx3253

Lacey, C., & Cole, S. 1993, Monthly Notices of the Royal Astronomical Society, 262, 627, doi: 10.1093/mnras/262.3.627

—. 1994, Monthly Notices of the Royal Astronomical Society, 271, 676, doi: 10.1093/mnras/271.3.676

Laigle, C., Pichon, C., Arnouts, S., et al. 2018, Monthly Notices of the Royal Astronomical Society, 474, 5437, doi: 10.1093/mnras/stx3055

Leauthaud, A., Tinker, J., Bundy, K., et al. 2012, The Astrophysical Journal, 744, 159, doi: 10.1088/0004-637X/744/2/159

Lesgourgues, J. 2011, The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview, arXiv. http://arxiv.org/abs/1104.2932

Lewis, A., Challinor, A., & Lasenby, A. 2000, The Astrophysical Journal, 538, 473, doi: 10.1086/309179

Lin, S., Tinker, J. L., Blanton, M. R., et al. 2023, Monthly Notices of the Royal Astronomical Society, 519, 4253, doi: 10.1093/mnras/stac2793

Linder, E. V. 2003, Physical Review Letters, 90, 091301, doi: 10.1103/PhysRevLett.90.091301

Lu, Y., Mo, H. J., Weinberg, M. D., & Katz, N. 2011, Monthly Notices of the Royal Astronomical Society, 416, 1949, doi: 10.1111/j.1365-2966.2011.19170.x

Lu, Y., Wechsler, R. H., Somerville, R. S., et al. 2014, The Astrophysical Journal, 795, 123, doi: 10.1088/0004-637X/795/2/123

Ludlow, A. D., Navarro, J. F., Angulo, R. E., et al. 2014, Monthly Notices of the Royal Astronomical Society, 441, 378, doi: 10.1093/mnras/stu483

Madau, P., & Dickinson, M. 2014, Annual Review of Astronomy and Astrophysics, 52, 415, doi: 10.1146/annurev-astro-081811-125615

Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., et al. 2021, Monthly Notices of the Royal Astronomical Society, 508, 4017, doi: 10.1093/mnras/stab2484

Malavasi, N., Arnouts, S., Vibert, D., et al. 2017, Monthly Notices of the Royal Astronomical Society, 465, 3817, doi: 10.1093/mnras/stw2864

Man, A., & Belli, S. 2018, Nature Astronomy, 2, 695, doi: 10.1038/s41550-018-0558-1

Mao, Y.-Y., Zentner, A. R., & Wechsler, R. H. 2018, Monthly Notices of the Royal Astronomical Society, 474, 5143, doi: 10.1093/mnras/stx3111

Monaco, P., Fontanot, F., & Taffoni, G. 2007, Monthly Notices of the Royal Astronomical Society, 375, 1189, doi: 10.1111/j.1365-2966.2006.11253.x

More, S., Kravtsov, A. V., Dalal, N., & Gottlöber, S. 2011, The Astrophysical Journal Supplement Series, 195, 4, doi: 10.1088/0067-0049/195/1/4

Moster, B. P., Somerville, R. S., Maulbetsch, C., et al. 2010, The Astrophysical Journal, 710, 903, doi: 10.1088/0004-637X/710/2/903

Moustakas, J., Kennicutt, Jr., R. C., & Tremonti, C. A. 2006, The Astrophysical Journal, 642, 775, doi: 10.1086/500964

Nagashima, M., & Yoshii, Y. 2004, The Astrophysical Journal, 610, 23, doi: 10.1086/421484

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, The Astrophysical Journal, 490, 493, doi: 10.1086/304888

Navarro, J. F., Hayashi, E., Power, C., et al. 2004, Monthly Notices of the Royal Astronomical Society, 349, 1039, doi: 10.1111/j.1365-2966.2004.07586.x

Nelson, D., Springel, V., Pillepich, A., et al. 2021, The IllustrisTNG Simulations: Public Data Release, arXiv. http://arxiv.org/abs/1812.05609

Okumura, T., Hayashi, M., Chiu, I.-N., et al. 2021, Publications of the Astronomical Society of Japan, 73, 1186, doi: 10.1093/pasj/psab068

Orsi, A., & Angulo, R. E. 2018, Monthly Notices of the Royal Astronomical Society, 475, 2530, doi: 10.1093/mnras/stx3349

Osato, K., & Okumura, T. 2022, Monthly Notices of the Royal Astronomical Society, 519, 1771, doi: 10.1093/mnras/stac3582

Peebles, P. J. E. 1970, The Astronomical Journal, 75, 13, doi: 10.1086/110933

Peter, P., Uzan, J.-P., Peter, P., & Uzan, J.-P. 2013, Primordial Cosmology, Oxford Graduate Texts (Oxford, New York: Oxford University Press)

Phillips, J. I., Wheeler, C., Boylan-Kolchin, M., et al. 2014, Monthly Notices of the Royal Astronomical Society, 437, 1930, doi: 10.1093/mnras/stt2023

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, Astronomy and Astrophysics, 641, A6, doi: 10.1051/0004-6361/201833910

Planelles, S., & Quilis, V. 2010, in Highlights of Spanish Astrophysics V, ed. J. M. Diego, L. J. Goicoechea, J. I. González-Serrano, & J. Gorgas, Astrophysics and Space Science Proceedings (Berlin, Heidelberg: Springer), 341–341, doi: 10.1007/978-3-642-11250-8_70

Plummer, H. C. 1911, Monthly Notices of the Royal Astronomical Society, 71, 460, doi: 10.1093/mnras/71.5.460

Pontzen, A., Tremmel, M., Roth, N., et al. 2017, Monthly Notices of the Royal Astronomical Society, 465, 547, doi: 10.1093/mnras/stw2627

Popesso, P., Concas, A., Cresci, G., et al. 2022, Monthly Notices of the Royal Astronomical Society, 519, 1526, doi: 10.1093/mnras/stac3214

Potter, D., Stadel, J., & Teyssier, R. 2016, PKDGRAV3: Beyond Trillion Particle Cosmological Simulations for the Next Era of Galaxy Surveys, arXiv, doi: 10.48550/arXiv.1609.08621

Prada, F., Klypin, A. A., Cuesta, A. J., Betancort-Rijo, J. E., & Primack, J. 2012, Monthly Notices of the Royal Astronomical Society, 423, 3018, doi: 10.1111/j.1365-2966.2012.21007.x

Press, W. H., & Schechter, P. 1974, The Astrophysical Journal, 187, 425, doi: 10.1086/152650

Raichoor, A., Moustakas, J., Newman, J. A., et al. 2023, The Astronomical Journal, 165, 126, doi: 10.3847/1538-3881/acb213

Rasera, Y., Alimi, J. M., Courtin, J., et al. 2010, 1241, 1134, doi: 10.1063/1.3462610

Reddick, R. M., Wechsler, R. H., Tinker, J. L., & Behroozi, P. S. 2013, The Astrophysical Journal, 771, 30, doi: 10.1088/0004-637X/771/1/30

Robotham, A. S. G., Liske, J., Driver, S. P., et al. 2013, Monthly Notices of the Royal Astronomical Society, 431, 167, doi: 10.1093/mnras/stt156

Rodighiero, G., Daddi, E., Baronchelli, I., et al. 2011, The Astrophysical Journal, 739, L40, doi: 10.1088/2041-8205/739/2/L40

Rodríguez-Torres, S. A., Comparat, J., Prada, F., et al. 2017, Monthly Notices of the Royal Astronomical Society, 468, 728, doi: 10.1093/mnras/stx454

Schechter, P. 1976, The Astrophysical Journal, 203, 297, doi: 10.1086/154079

Scoccimarro, R. 2004, Physical Review D, 70, 083007, doi: 10.1103/PhysRevD.70.083007

Sheth, R. K., Mo, H. J., & Tormen, G. 2001, Monthly Notices of the Royal Astronomical Society, 323, 1, doi: 10.1046/j.1365-8711.2001.04006.x

Sheth, R. K., & Tormen, G. 1999, Monthly Notices of the Royal Astronomical Society, 308, 119, doi: 10.1046/j.1365-8711.1999.02692.x

Sin, L. P. T., Lilly, S. J., & Henriques, B. M. B. 2017, Monthly Notices of the Royal Astronomical Society, 471, 1192, doi: 10.1093/mnras/stx1674

Skibba, R. A., Bosch, F. C. v. d., Yang, X., et al. 2011, Monthly Notices of the Royal Astronomical Society, 410, 417, doi: 10.1111/j.1365-2966.2010.17452.x

Smith, A., Burtin, E., Hou, J., et al. 2020, Monthly Notices of the Royal Astronomical Society, 499, 269, doi: 10.1093/mnras/staa2825

Somerville, R. S., & Davé, R. 2015, Annual Review of Astronomy and Astrophysics, 53, 51, doi: 10.1146/annurev-astro-082812-140951

Somerville, R. S., & Primack, J. R. 1999, Monthly Notices of the Royal Astronomical Society, 310, 1087, doi: 10.1046/j.1365-8711.1999.03032.x

Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, Monthly Notices of the Royal Astronomical Society, 328, 726, doi: 10.1046/j.1365-8711.2001.04912.x

Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Nature, 435, 629, doi: 10.1038/nature03597

Stewart, K. R., Bullock, J. S., Wechsler, R. H., Maller, A. H., & Zentner, A. R. 2008, The Astrophysical Journal, 683, 597, doi: 10.1086/588579

Strateva, I., Ivezić, , Knapp, G. R., et al. 2001, The Astronomical Journal, 122, 1861, doi: 10.1086/323301

Sutter, P. M., & Ricker, P. M. 2010, The Astrophysical Journal, 723, 1308, doi: 10.1088/0004-637X/723/2/1308

Taruya, A., Bernardeau, F., Nishimichi, T., & Codis, S. 2012, Physical Review D, 86, 103528, doi: 10.1103/PhysRevD.86.103528

Tasitsiomi, A., Kravtsov, A. V., Wechsler, R. H., & Primack, J. R. 2004, The Astrophysical Journal, 614, 533, doi: 10.1086/423784

Teyssier, R. 2002, Astronomy & Astrophysics, 385, 337, doi: 10.1051/0004-6361:20011817

Tinker, J. L., Hahn, C., Mao, Y.-Y., Wetzel, A. R., & Conroy, C. 2018, Monthly Notices of the Royal Astronomical Society, 477, 935, doi: 10.1093/mnras/sty666

Vale, A., & Ostriker, J. P. 2006, Monthly Notices of the Royal Astronomical Society, 371, 1173, doi: 10.1111/j.1365-2966.2006.10605.x

Van Den Bosch, F. C., Weinmann, S. M., Yang, X., et al. 2005, Monthly Notices of the Royal Astronomical Society, 361, 1203, doi: 10.1111/j.1365-2966.2005.09260.x

Wang, W., & White, S. D. M. 2012, Monthly Notices of the Royal Astronomical Society, 424, 2574, doi: 10.1111/j.1365-2966.2012.21256.x

Warren, M. S., Quinn, P. J., Salmon, J. K., & Zurek, W. H. 1992, The Astrophysical Journal, 399, 405, doi: 10.1086/171937

Wechsler, R. H., Bullock, J. S., Primack, J. R., Kravtsov, A. V., & Dekel, A. 2002, The Astrophysical Journal, 568, 52, doi: 10.1086/338765

Wechsler, R. H., & Tinker, J. L. 2018, Annual Review of Astronomy and Astrophysics, 56, 435, doi: 10.1146/annurev-astro-081817-051756

Wechsler, R. H., Zentner, A. R., Bullock, J. S., Kravtsov, A. V., & Allgood, B. 2006, The Astrophysical Journal, 652, 71, doi: 10.1086/507120

Weinberg, N. N., & Kamionkowski, M. 2003, Monthly Notices of the Royal Astronomical Society, 341, 251, doi: 10.1046/j.1365-8711.2003.06421.x

Weinmann, S. M., Bosch, F. C. v. d., Yang, X., & Mo, H. J. 2006, Monthly Notices of the Royal Astronomical Society, 366, 2, doi: 10.1111/j.1365-2966.2005.09865.x

White, S. D. M. 1976, Monthly Notices of the Royal Astronomical Society, 177, 717, doi: 10.1093/mnras/177.3.717

—. 1994, arXiv e-prints, astro, doi: 10.48550/arXiv.astro-ph/9410043

White, S. D. M., & Frenk, C. S. 1991, The Astrophysical Journal, 379, 52, doi: 10.1086/170483

Xu, X., Zehavi, I., & Contreras, S. 2021, Monthly Notices of the Royal Astronomical Society, 502, 3242, doi: 10.1093/mnras/stab100

Yang, X., Mo, H. J., & Bosch, F. C. v. d. 2003, Monthly Notices of the Royal Astronomical Society, 339, 1057, doi: 10.1046/j.1365-8711.2003.06254.x

—. 2009, The Astrophysical Journal, 695, 900, doi: 10.1088/0004-637X/695/2/900

Yu, J., Zhao, C., Chuang, C.-H., et al. 2022, Monthly Notices of the Royal Astronomical Society, 516, 57, doi: 10.1093/mnras/stac2176

Yuan, S., Eisenstein, D. J., & Garrison, L. H. 2018, Monthly Notices of the Royal Astronomical Society, 478, 2019, doi: 10.1093/mnras/sty1089

Yuan, S., Garrison, L. H., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022a, Monthly Notices of the Royal Astronomical Society, 510, 3301, doi: 10.1093/mnras/stab3355

Yuan, S., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022b, Monthly Notices of the Royal Astronomical Society, 512, 5793, doi: 10.1093/mnras/stac830

Zehavi, I., Kerby, S. E., Contreras, S., et al. 2019, The Astrophysical Journal, 887, 17, doi: 10.3847/1538-4357/ab4d4d

Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, The Astrophysical Journal, 736, 59, doi: 10.1088/0004-637X/736/1/59

Zel'dovich, Y. B. 1970, Astronomy and Astrophysics, 5, 84. https://ui.adsabs.harvard.edu/abs/1970A&A.....5...84Z

Zhao, H. 1996, Monthly Notices of the Royal Astronomical Society, 278, 488, doi: 10.1093/mnras/278.2.488

Zheng, Z., Coil, A. L., & Zehavi, I. 2007, The Astrophysical Journal, 667, 760, doi: 10.1086/521074

Zheng, Z., Zehavi, I., Eisenstein, D. J., Weinberg, D. H., & Jing, Y. P. 2009, The Astrophysical Journal, 707, 554, doi: 10.1088/0004-637X/707/1/554

Zhou, R., Dey, B., Newman, J. A., et al. 2023, The Astronomical Journal, 165, 58, doi: 10.3847/1538-3881/aca5fb

4

ELG HOD fitting with Gaussian processes

Contents

This chapter is devoted to the HOD fitting method for DESI ELGs that I developed during my thesis and which was published in (Rocher et al., 2023)

4.1 Introduction on HOD fitting methods

Building clustering predictions from HOD models using simulations is cheaper than running SAMs but requires a non-negligible amount of computational resources. First, one needs to load N-body simulation boxes, which can be memory expensive depending on the box size. Typically, for AbacusSummit simulations, the phase space, mass and properties have to be loaded for each of the ~ 40 (resp. 300) million halos present in simulation boxes of size 1 Gpc/h (resp. 2 Gpc/h). Next, the expected mean numbers of central and satellite galaxies from the HOD model are computed for each halo and used to randomly draw the actual numbers of such galaxies to be assigned to the halo, using a Bernoulli (resp. Poisson) distribution for central (resp. satellite) galaxies. Satellite positions and velocities are then assigned using a halo density profile or a randomly selected dark matter particle of the halo. The latter solution is more costly, as particle positions and velocities have to be loaded for each halo. The obtained galaxy catalogue is then compared with data using clustering statistics, which is again costly in terms of CPU resources. Consequently, full inference of HOD parameters can be CPU expensive. In addition, HOD models suffer from stochasticity induced by random draws, making minimisation difficult.

In the literature several approaches have been proposed to perform HOD fits more efficiently, in particular to limit the stochasticity of the procedure. One popular technique is the tabulated HOD method that pre-compute halo and particle clustering and convolve it with halo occupation distribution (Zheng & Guo, 2016). Other techniques used optimized and parallelized code to code HOD models and clustering statistics. For instance, the recent AbacusHOD (Yuan et al., 2022) pipeline first initialises random numbers for every halo and particle, down-samples halos and particles from N-body simulations, and then run inference on HOD parameters from a sampler to derive best-fit parameters.

In the following, I describe the new method I developed to fit HOD parameters on small scale clustering measurements using Gaussian Processes (Rasmussen & Williams, 2005). The methodology presented hereafter aims at performing accurate fitting of HOD model parameters while minimising CPU time consumption. To this purpose, inspired by Efficient Global Optimization algorithms (Jones et al., 1998), I developed a two-step procedure using Gaussian Processes (GP) to create a surrogate model of the likelihood posterior $\mathcal{L}$. In a first step, we sample the likelihood posterior to provide initial training to the GP. This initial GP model is further improved by successive iterations, each iteration adding one point until the predicted map becomes stable enough so that marginalised parameter values and posterior contours can be reliably derived. After a brief introduction to Gaussian processes, we describe the different steps of the fitting procedure and give its performance.

4.2 Gaussian Processes

Gaussian Processes (GP) (Rasmussen & Williams, 2005) offer an alternative route to build a model function that fits a set of observational data. In cosmology, GP have been recently used

to build emulators to perform cosmological inference, whether iterative (El Gammal et al., 2022, Neveux et al., 2022, Pellejero-Ibañez et al., 2020) or not (Angulo et al., 2021, Nishimichi et al., 2019, Sáez-Casares et al., 2023). In particular, El Gammal et al. (2022) is very instructive on the challenges to be faced to set up such a methodology. GP are also very efficient to perform global optimisation of expensive and stochastic functions (Garnett et al., 2008), as often encountered in HOD modelling.

Gaussian processes provide a way to predict the value of a function $f(\boldsymbol{x})$ for any set of parameter values $\boldsymbol{x} = (p_0, p_1, \dots, p_N)$, from an initial, restricted set of parameter values $\boldsymbol{X} = (\boldsymbol{x}_0, \dots, \boldsymbol{x}_n)$ and their corresponding function values $\boldsymbol{y} = f(\boldsymbol{X})$ called training sample (or data) in the following, where $n+1$ is the number of evaluations in the initial training sample and $N+1$ the number of parameters. $\boldsymbol{X}$ is a matrix of dimension $(N+1)\times(n+1)$. GP assume that the function $f(\boldsymbol{x})$ is drawn from a collection of random functions that are Gaussian-distributed (hence the name) around a mean function $m(\boldsymbol{x})$ with a covariance function, called kernel, $k(\boldsymbol{x}, \boldsymbol{x})$ which completely define the GP. The random functions are conditioned by the values from the training sample, i.e. they have prior information on the function f given by the initial training sample, $\boldsymbol{y} = f(\boldsymbol{X})$:

$$\begin{bmatrix} f(\boldsymbol{x}_0) \\ f(\boldsymbol{x}_1) \\ \vdots \\ f(\boldsymbol{x}_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{x}_0) \\ m(\boldsymbol{x}_1) \\ \vdots \\ m(\boldsymbol{x}_n) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{x}_0, \boldsymbol{x}_0) & k(\boldsymbol{x}_0, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_0, \boldsymbol{x}_n) \\ k(\boldsymbol{x}_1, \boldsymbol{x}_0) & k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}) & k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \right) \tag{4.1}$$

In this notation $\sim$ stands for "distributed according to...". As an illustration, Figure 4.1 shows random functions generated by GP with and without training sample.

Importantly, while the correlation of the function value at two points is assumed to be Gaussian, this neither means that the function is itself assumed to be Gaussian, nor that the mean of the family of functions is presumed to be Gaussian. The training set $(\boldsymbol{X}, \boldsymbol{y})$ can be used to make predictions of $f(\boldsymbol{X}_*)$ for any set of unobserved parameters $\boldsymbol{X}_*$. The unobserved values of $f(\boldsymbol{X}_*)$ and the observed values $\boldsymbol{y}$ are jointly distributed Gaussian variables:

$$p(f(\boldsymbol{X}_*), \boldsymbol{y}) = \begin{bmatrix} f(\boldsymbol{X}) \\ f(\boldsymbol{X}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{X}) \\ m(\boldsymbol{X}_*) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{X}, \boldsymbol{X}) & k(\boldsymbol{X}, \boldsymbol{X}_*) \\ k(\boldsymbol{X}_*, \boldsymbol{X}) & k(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix} \right) \tag{4.2}$$

The joint distribution $p(f(\boldsymbol{X}_*), \boldsymbol{y})$ can be used to compute the *conditional distribution* (also called the *GP posterior*) $p(f(\boldsymbol{X}_*)|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*)$[1], which also follows a Gaussian distribution:

$$p(f(\boldsymbol{X}_*)|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*) \sim \mathcal{N}(\mu(\boldsymbol{X}_*), \Sigma(\boldsymbol{X}_*)) \tag{4.3}$$

where $\mu(\boldsymbol{X}_*)$ is the mean vector, and $\Sigma(\boldsymbol{X}_*)$ the covariance matrix:

$$\begin{aligned} \mu(\boldsymbol{X}_*) &= m(\boldsymbol{X}_*) + k(\boldsymbol{X}_*, \boldsymbol{X}) \left(k(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} [\boldsymbol{y} - m(\boldsymbol{X})] \\ \Sigma(\boldsymbol{X}_*) &= k(\boldsymbol{X}_*, \boldsymbol{X}_*) - k(\boldsymbol{X}_*, \boldsymbol{X}) \left(k(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} k(\boldsymbol{X}, \boldsymbol{X}_*) \end{aligned} \tag{4.4}$$

Note that $diag(\Sigma(\boldsymbol{X}_*))$ provides the variances around the mean predictions of the GP, $\mu(\boldsymbol{X}_*)$ (the corresponding uncertainties are shown as shaded regions in Figure 4.1).

[1] It takes ~ 4 pages of matrix algebra to derive the conditional distribution $p(f(\boldsymbol{X}_*)|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*)$ from the joint distribution $p(\boldsymbol{y}, f(\boldsymbol{X}_*))$, so we skip the computation details in the manuscript. See (Rasmussen & Williams, 2005) for the derivation details.

GP can also describe noisy observations $f(\boldsymbol{X}) = \boldsymbol{y} + \boldsymbol{\epsilon}$ where the elements of the noise vector, $\boldsymbol{\epsilon} = \{\epsilon_0, \epsilon_1, ..., \epsilon_n\}$ are parametrised by zero-mean Gaussians with variances given by $\boldsymbol{\sigma}_{\boldsymbol{\epsilon}}^2 = \{\sigma_{\epsilon_0}^2, \sigma_{\epsilon_1}^2, ..., \sigma_{\epsilon_n}^2\}$, $\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon_i}^2)$. The prior kernel for one noisy observation $\boldsymbol{x_i}$ becomes:

$$k(\boldsymbol{x_i}, \boldsymbol{x_i}) \rightarrow k(\boldsymbol{x_i}, \boldsymbol{x_i}) + \sigma_{\boldsymbol{\epsilon_i}}^2 \tag{4.5}$$

If the noise is uncorrelated between observations $\boldsymbol{y}$, a diagonal matrix $\boldsymbol{D}$ whose elements are $\boldsymbol{\sigma}_{\boldsymbol{\epsilon}}^{\boldsymbol{2}}$ is added to the kernel. Similarly, the posterior distribution of predicted noisy observations will have additive noise $f(\boldsymbol{X}_*) = \boldsymbol{y}_* + \boldsymbol{\epsilon}_*$ and one should add to the posterior kernel $k(\boldsymbol{X}_*, \boldsymbol{X}_*)$ a diagonal matrix $\boldsymbol{D}_*$ with elements $\boldsymbol{\sigma}_{\boldsymbol{\epsilon}_*}^{\boldsymbol{2}}$. Therefore, in presence of noise, we rewrite Equation (4.4) as:

$$\begin{aligned} \mu(\boldsymbol{X}_*) &= m(\boldsymbol{X}_*) + k\left(\boldsymbol{X}_*, \boldsymbol{X}\right)\left(k(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{D}\right)^{-1}\left[\boldsymbol{y} - m(\boldsymbol{X})\right] \\ \Sigma(\boldsymbol{X}_*) &= k\left(\boldsymbol{X}_*, \boldsymbol{X}_*\right) + \boldsymbol{D}_* - k\left(\boldsymbol{X}_*, \boldsymbol{X}\right)\left(k(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{D}\right)^{-1} k\left(\boldsymbol{X}, \boldsymbol{X}_*\right) \end{aligned} \tag{4.6}$$

In the literature, a zero-mean function $m(\boldsymbol{x}) = 0$ is often assumed since the flexibility provided by the kernel is enough to model $f(\boldsymbol{x}_*)$ arbitrarily well. There is a wide range of possible kernel functions. In our procedure we performed tests with a squared exponential kernel (also known as Radial Basis Function, RBF) and a Matérn kernel of index $\nu = 5/2$ which is equivalent to the product of an exponential and a polynomial of order 5. We adopt the latter as our baseline for this method. The Matérn kernel function in one dimension has the following form:

$$k(x_p, x_q) = \sigma_k^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d(x_p, x_q)}{\ell}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{d(x_p, x_q)}{\ell}\right) \left(+\sigma_{\epsilon_p}^2 \delta_{pq}\right) \tag{4.7}$$

where Γ is the gamma function, K_ν is the modified Bessel function, ℓ is the length scale, $d(x_p, x_q)$ is the Euclidean distance between the two points and σ_k^2 is the variance of the kernel. δ_{pq} is the Kronecker delta function. The kernel is parametrised by the length scale and the kernel variance. The length scale can be viewed as a characteristic correlation length between points along the parameter axis and high (resp. low) values of ℓ generate smooth (resp. wiggly) random functions, as can be seen in Figure 4.1. In N-dimensions, there is one length scale for each dimension of the parameter space. The length scale and kernel variance values are found by marginalising the following likelihood over $\boldsymbol{\ell}$ and σ_k^2:

$$-\log p(\boldsymbol{y}|\boldsymbol{X}, (\boldsymbol{\ell}, \sigma_k^2)) = \frac{1}{2}\boldsymbol{y}^T \left(k(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{D}\right)^{-1} \boldsymbol{y} + \frac{1}{2}\log\left|k(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{D}\right| - \frac{n+1}{2}\log 2\pi \tag{4.8}$$

➢ Gaussian Process procedure

Altogether, a procedure based on Gaussian Processes can be summarised as follows:

- generate a training sample $(\boldsymbol{X}, \boldsymbol{y})$ for a given function f,
- choose a kernel k (here Matern $\nu = 5/2$) and a mean μ (we consider $\mu = 0$),
- choose a set of parameters $\boldsymbol{X}_*$ where we want to estimate the function i.e. $f(\boldsymbol{X}_*)$,
- maximize Equation (4.8) to obtain the kernel length scales,
- compute the expected mean value $\mu(\boldsymbol{X}_*)$ and its variance $\Sigma(\boldsymbol{X}_*)$ of $f(\boldsymbol{X}_*)$ using Equation (4.6).

Many documents on GP can be found in the literature. A nice introduction and visual exploration of Gaussian processes can be found at visual-exploration-gaussian-processes. In our inference method, we iteratively add a new set of parameter values to the training sample of the GP (see Section 4.3.6). So the GP prediction is updated by adding new prior on the parameter space.

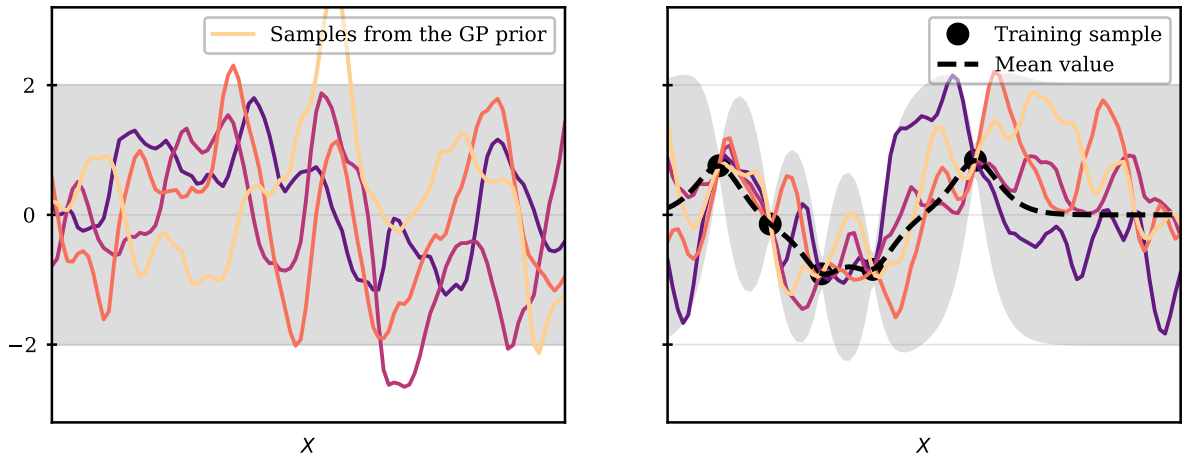

Figure 4.1: *Schematic view of Gaussian processes. Left: Sample of random functions drawn from a Gaussian process using the Matern($\nu = 5/2$) kernel. The shaded region represents the 2σ region from the mean $\mu = 0$. Right: Training points (black) are added to the procedure. The sample of random functions are now conditioned by the data points. The mean function μ is in black dashed-lines and the shaded grey band is the 2σ region from the mean given by Equation (4.6).*

4.3 HOD modelling framework

To develop the fitting procedure we rely on HOD-based fiducial mock catalogues that aim to reproduce the population of ELGs that we expect to find in DESI. HOD fitting for ELGs is challenging since these galaxies are expected to reside in low mass halos $\sim 10^{12} M_{\odot}$, which requires handling high resolution N-body simulations where dark matter halos are well-resolved down to (at least) $\sim 10^{11} M_{\odot}$, making the evaluation of the HOD model, and the estimation of clustering statistics, CPU and memory expensive (due to the fast increase of the halo mass function at low mass). We hereafter briefly recall the HOD framework introduced in Section 3.4.1.

4.3.1 HOD model

The HOD formalism describes the relation between a specific class of galaxies and dark matter halos, as the probability that a halo with mass M contains N such galaxies. It also specifies how galaxy positions and velocities are distributed within halos. HOD models have contributions from two galaxy populations, namely centrals and satellites, with $\langle N_{cent}(M) \rangle$ and $\langle N_{sat}(M) \rangle$ their respective mean numbers hosted per halo of a given halo mass.

As discussed in Section 3.4.1, the mean HOD predicted by ELG semi-analytical models can be fit reasonably well by a simple Gaussian or slightly asymmetric Gaussian for centrals, together with a power law for satellites. Accordingly, the baseline model followed in this chapter is the

Gaussian HOD model (GHOD) defined in Equation (7.9), which we recall here:

$$\begin{aligned}\langle N_{cent}(M)\rangle &= \frac{A_c}{\sqrt{2\pi}\sigma_{\rm M}}\cdot e^{-\frac{(\log_{10}(M)-\mu)^2}{2\sigma_{\rm M}^2}}\\ \langle N_{sat}(M)\rangle &= A_s\left(\frac{M-M_0}{M_1}\right)^{\alpha}\end{aligned} \tag{4.9}$$

A_c defines the amplitude of the central galaxy HOD, $\mu = \log_{10} M_c$ where M_c is the characteristic halo mass with maximal probability to host a central galaxy and $\sigma_{\rm M}$ is the width of the distribution. A_s defines the amplitude of the satellite galaxy HOD, M_0 is a cut-off halo mass below which no satellite can be present, α is a slope-parameter that controls the variation in satellite richness with increasing halo mass, and M_1 is the mass at which 1 satellite is expected per halo if $A_s = 1$ and M_0 is much lower than M_1. The total number density of the galaxy sample can be calculated as follows:

$$\bar{n}_{\rm gal} = \int \frac{{\rm d}n(M)}{{\rm d}M}\left[\langle N_{cent}(M)\rangle + \langle N_{sat}(M)\rangle\right]{\rm d}M \tag{4.10}$$

The total galaxy sample size is governed by both A_c and A_s and the fraction of satellites is controlled by their ratio. Moreover, all other conditions being equal, the same clustering is obtained whatever A_c and A_s values, provided their ratio is fixed and $\langle N_{cent}(M)\rangle$ remains lower than 1 (which is the case for all the fits performed in the following). We rely on this property to impose a fixed density in our fitting procedure.

The constraints adopted in our fits are the following. First, as $\log_{10}(M_1)$ cannot be constrained due to degeneracies with A_s and α, this parameter is kept fixed for the tests. To choose a sensible value for $\log_{10}(M_1)$, we follow Avila et al. (2020) and set:

$$\log_{10}(M_1) = \log_{10}(M_c)^{ref} + 0.3 = 11.93 \tag{4.11}$$

taking for $\log_{10}(M_c)^{ref}$ the value used to generate our pseudo-data catalogues (see Table 4.2). Second, in order to apply a density constraint to our fits to match that in DESI data, we treat the A_c and A_s parameters in the following way. At each point in the HOD parameter space, A_c is set to an initial value of 0.05, while A_s is sampled from a flat prior range (reported in Table 4.2). The total number density for these initial values of A_c and A_s is computed according to Equation (4.10) and the values of A_c and A_s are rescaled by the same factor (to preserve the clustering) in order to provide a fixed galaxy density of $10^{-3}{\rm Mpc}/h^3$ close to that expected for the DESI ELG sample. In the following, all our results are expressed as a function of the initial (i.e. unrescaled) value of A_s, corresponding to $A_c = 0.05$.

In the following, we complement the above functions for the mean numbers of central and satellite galaxies by the following assumptions. The actual number of central (resp. satellite) galaxies per halo of mass M follows a Bernoulli (resp. Poisson) distribution with mean equal to $\langle N_{\rm cent}(M)\rangle$ (resp. $\langle N_{sat}(M)\rangle$). Central galaxies are positioned at the center of their halos while satellite galaxy positions sample a Navarro-Frenk-White profile described in Section 3.1.5. We assume that satellite velocities are normally distributed around their mean halo velocity v_h, with a dispersion equal to that of the halo dark matter particles σ_{v_h}, rescaled by an extra free parameter denoted f_{σ_v} as follows:

$$\vec{v}_{\rm sat} \sim \mathcal{N}(\vec{v}_h, f_{\sigma_v}\cdot\sigma_{v_h}) \tag{4.12}$$

4.3.2 Simulation tests

name	cosmology	box size	resolution	realisations
baseline	Planck 2018 ΛCDM	1 Gpc/h	3456^3	1
cosmic variance	Planck 2018 ΛCDM	2 Gpc/h	6912^3	25

Table 4.1: *Cosmology, box size and mass resolution of the AbacusSummit simulations used in this chapter. The mass resolution is given as the number of particles in the box.*

parameter	$\log_{10}(M_c)$	α	A_s	$\log_{10}(M_0)$	$\log_{10}(M_1)$	$\sigma_{\rm M}$	f_{σ_v}
input	11.63	0.6	0.11	11.63	11.93	0.12	1.
priors	11.4-11.8	0.5-0.7	0.05-0.2	11.4-11.8	11.93	0.01-0.3	0.75-1.25

Table 4.2: *Top row: GHOD parameter input values used for pseudo-data catalogues. Bottom row: GHOD parameter flat prior ranges used in all fits performed in this chapter. In both rows, the indicated values of A_s are initial values and the initial value for A_c is 0.05. We recall that, at each point in the parameter space, these individual values are rescaled by the same factor to provide clustering modelling with a fixed density of $10^{-3}(Mpc/h)^3$ (see text).*

Our HOD fitting method uses the above GHOD model to populate simulated dark matter halos and produce mock galaxy catalogues for which clustering statistics are calculated and compared to data. The method is tested with simulated data (dubbed as pseudo-data in the following) that are themselves galaxy mock catalogues produced in the same way. For both purposes, we rely on the AbacusSummit suite and the corresponding cleaned halo catalogues obtained with the CompaSO algorithm, as described in Section 3.2.1.

We report in Table 5.1 the subset of simulations used in this chapter. They all use the base resolution, 6912^3 particles in a box of 2 Gpc/h length, which corresponds to a particle mass of about $2{\times}10^9 M_\odot/h$. With this particle mass, halos are well resolved down to $10^{11} M_\odot/h$ which provides ∼50 particles/halo (Maksimova et al., 2021). Moreover, the halos selected in this work have a mass larger than $3 \times 10^{11} M_\odot/h$ which corresponds to 150 particles/halo. Simulations in Table 5.1 are used to create mock catalogues for both pseudo-data and model predictions, as will be detailed in the next sections. For pseudo-data mocks, the GHOD parameters are fixed at values listed in Table 4.2 (top row). These values provide a clustering close to that expected for the ELG sample collected during the survey validation phase of DESI.

4.3.3 Clustering statistics

As clustering statistics, we adopt the projected correlation function, $w_p(r_p)$, which is robust against redshift-space distortions at small scales, and the two-point correlation function monopole, $\xi_0(s)$ and quadrupole $\xi_2(s)$. We first compute the galaxy two-point correlation function, $\xi(r_p, \pi)$, as a function of the galaxy pair separation components along (π) and perpendicular to the line-of-sight (r_p). Integration over the line-of-sight provides the projected correlation function:

$$w_p(r_p) = 2 \int_0^{\pi_{max}} \xi(r_p, \pi) d\pi. \tag{4.13}$$

Computing the two-point correlation function $\xi(s, \mu)$, as a function of the galaxy pair separation, s, and the cosine of the angle between the line-of-sight and separation vector, μ, provides the

two multipoles we use:

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1}\xi(s,\mu)\mathcal{L}_\ell(\mu)d\mu \tag{4.14}$$

with $\ell \in \{0,2\}$ and where $\mathcal{L}_\ell(\mu)$ denotes the Legendre polynomial of order ℓ. We rely on the DESI wrapper (see Section 4.5) around the CORRFUNC package (Sinha & Garrison, 2020) to compute the above two-point correlation functions $\xi(r_p,\pi)$ and $\xi(s,\mu)$ with the natural estimator, which compares galaxy pair counts to the expected pair count for a uniform distribution. For $w_p(r_p)$, we use 25 logarithmic bins in r_p between 0.03 and 30 Mpc/h, setting $\pi_{max} = 40\,\mathrm{Mpc}/h$. For the multipoles, we use 25 logarithmic bins between 0.8 and 30 Mpc/h in s and 100 linear bins in μ. In the galaxy pair count computation, the galaxy redshift to distance conversion uses the simulation cosmology as the fiducial cosmology and the z axis is chosen as a line-of-sight for the application of redshift space distortions.

4.3.4 χ^2 definition

At each point of the HOD parameter space, 20 model realisations are drawn. For each realisation, the model clustering is compared to that of the pseudo-data with the following χ^2 definition:

$$\chi^2 = (\xi_{data}-\xi_{model})^\top\left[\mathbf{C}_{data}/(1-D_{data}) + \mathbf{C}_{model}/(1-D_{model})\right]^{-1}(\xi_{data}-\xi_{model}) \tag{4.15}$$

where ξ denotes a vector of clustering measurements, $\mathbf{C}$ a component of the covariance matrix and D the Hartlap correction factor (Hartlap et al., 2007) applied to the inverse of the covariance matrix component:

$$D = \frac{n_b+1}{n_m-1} \tag{4.16}$$

with n_b the number of bins of the data vector and n_m the number of mocks used to build the covariance matrix component. These 20 measured χ^2 values are then averaged and the dispersion of the χ^2 values divided by $\sqrt{20}$ is used as an estimate of the uncertainty on the mean χ^2. For HOD input values in Table 4.2 (top row), this uncertainty is of order 2.3 for a mean χ^2 around 63 (for $75-6=69$ degrees of freedom). As the dynamical range of χ^2 variations is large over the HOD parameter space, which can make it hard to model the likelihood posterior, we use the natural logarithm of the mean χ^2 values and the corresponding errors as inputs to the GP. With the notations of Section 4.2, we provide the GP with $\boldsymbol{x} = \{\log_{10}(M_c), \alpha, A_s, \log_{10}(M_0), \log_{10}(M_1), \sigma_\mathrm{M}, f_{\sigma_v}\}$, $y = \ln(\chi^2)$, $\epsilon = \delta(\ln(\chi^2))$.

The computation of the covariance matrix for the pseudo-data depends on the test to be performed and is described in the next section. To build the model covariance we assume that correlations have small variations over the HOD parameter space and compute a fixed correlation matrix from 1000 realisations of the HOD model in table 4.2 (top row), drawn from the simulation box used for the model. At each point of the parameter space, the model covariance matrix is then obtained by normalising the previous correlation matrix using the variances of the clustering measurements over the 20 realisations drawn to compute the χ^2 at that point. This is the baseline for the computation of the model covariance matrix, which we changed slightly for specific tests, as detailed in Section 4.4.

4.3.5 GP training sample

To define the GP training sample, the HOD parameter space must be sampled efficiently. Two sampling methods were tested. The first one, commonly used in GP applications is the Latin Hypercube Sampling (LHS McKay et al. (1979)). The second method, which we adopt as our default, is the Hammersley sampling (Wong et al., 1997) which generates a more uniformly distributed sampling pattern.

➢ Latin hypercube sampling

The Latin hypercube sampling is an efficient technique to sample high-dimensional parameter spaces into bins of equal probability so as to provide a more even distribution of sample points than would be possible with pure random sampling. LHS is based on the *latin square*, a square equally divided in N samples along each axis. Each row and column contains a single point of the sample (see Figure 4.2). The *latin hypercube* is the extension of the latin square to higher dimensions. For a n-dimensional parameter space, each axis is divided equally in N samples and each sample point has to be the only one in each axis-aligned hyperplane containing it.

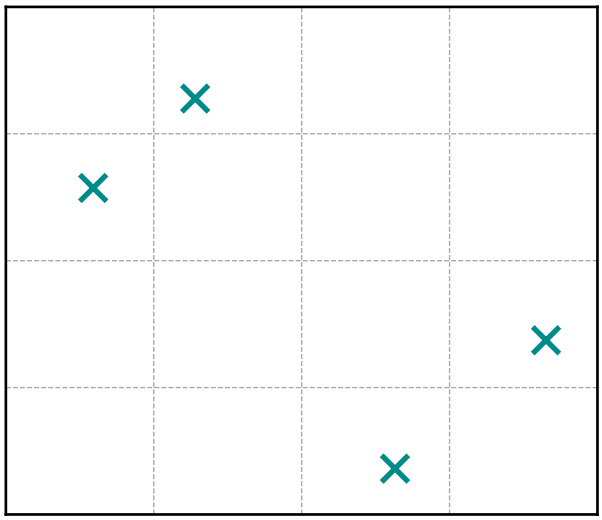

Figure 4.2: *Schematic view of a latin square sampling in 2D.*

➢ Hammersley sampling

Hammersley sampling (Wong et al., 1997) is part of the *low-discrepancy* sampling methods (or quasi-Monte Carlo methods) based on the Hammersley sequence. The discrepancy of a sequence refers to a quantitative measure of how much the distribution of samples deviates from an ideal uniform distribution (hence a low-discrepancy makes the distribution quasi-uniform). The Hammersley sequence is an approach that employs a deterministic algorithm to generate samples in an n-dimensional space as close as possible to a uniform sampling. This algorithm

generates N points using the radix-R (or prime base) notation of an integer[1]. An non-negative integer, k, can be represented in radix-R notation as an expansion over a prime base:

$$k = k_0 + k_1 p + k_2 p^2 + \cdots + k_r p^r \tag{4.17}$$

where each k_i is an integer in $[0, p-1]$. The inverse radix number function of k is then defined as:

$$\Phi_p(k) = \frac{k_0}{p} + \frac{k_1}{p^2} + \frac{k_2}{p^3} + \cdots + \frac{k_r}{p^{r+1}} \tag{4.18}$$

The n-dimensional Hammersley set of N points is defined by:

$$x(k) = \left(\frac{k}{N}, \Phi_{p_1}(k), \Phi_{p_2}(k), \ldots, \Phi_{p_{n-1}}(k)\right) \tag{4.19}$$

for $k = 0, 1, 2, \ldots, N-1$ and the values of $p_1 < p_2 < \cdots < p_{n-1}$ are the first $(n-1)$ prime numbers $(2, 3, 5, 7, 11, \ldots)$. This algorithm generates a set of N points in the n-dimensional parameter space $[0,1]^n$. In an Hammersley sequence, points in the first dimension (k/N) are located equidistant from each other. The Hammersley sequence (or set) is a particular case of the *Halton* sequence: $(\Phi_{p_1}(k), \Phi_{p_2}(k), \ldots, \Phi_{p_{n-1}}(k), \Phi_{p_n}(k))$. A 2D representation of the Hammersley sampling is compared to the LHS in Figure 4.3. By construction the Hammersley sampling exhibits a better uniformity than that of the LHS. This algorithm is reliable and efficient for low-dimensional problems only (less than 10 parameters) which is the case of this work.

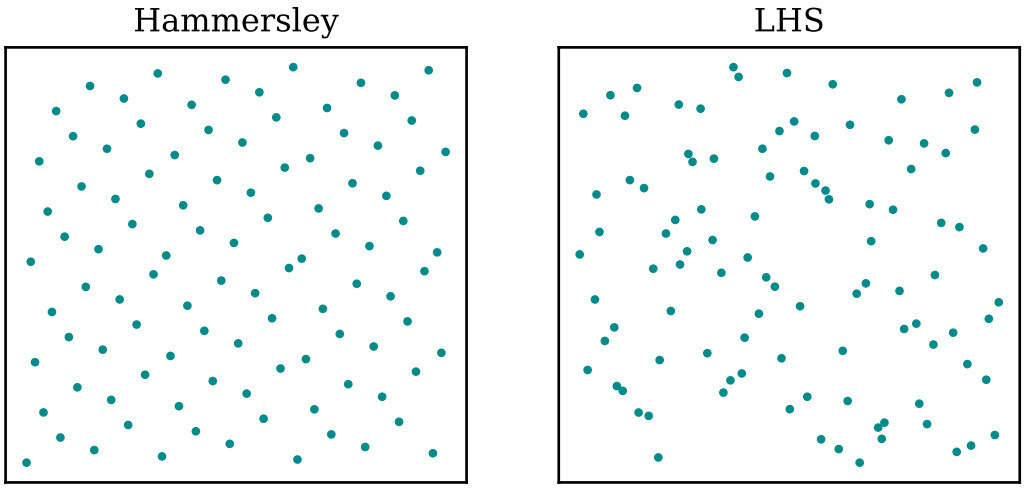

Figure 4.3: *Comparison between a Hammersley sampling (left) and a LHS (right) in 2D for a sample of 10 points per dimension (100 points in total).*

We performed tests using both sampling methods and noticed that LHS can be too sparse for HOD fitting, resulting in biased contours or even missed best fits. Thus, we adopt the Hammersley sampling as our default sampling. To define the training sample we draw $N = 600$ points in the HOD parameter space defined by Hammersley sampling and, in each point, compute the previously defined χ^2 and its error. The GP is then provided with the natural logarithm of

[1]In a positional numeral system, the radix or base is the number of unique digits, including the digit zero, used to represent numbers. For example, for the decimal system which has ten digits (from 0 to 9), the radix is ten

the N computed χ^2 values, together with the corresponding errors. The parameter values are drawn uniformly in ranges summarised in Table 4.2 (bottom row). Unless otherwise stated, we choose A_s as the parameter space dimension with equidistant points. Since the prior ranges do not change in the following tests, the Hammersley sampling for a given number of points and a given choice for the dimension with equidistant points is uniquely defined. We study the impact of changing these conditions in section 4.4.

4.3.6 Iterations and fit stability criterion

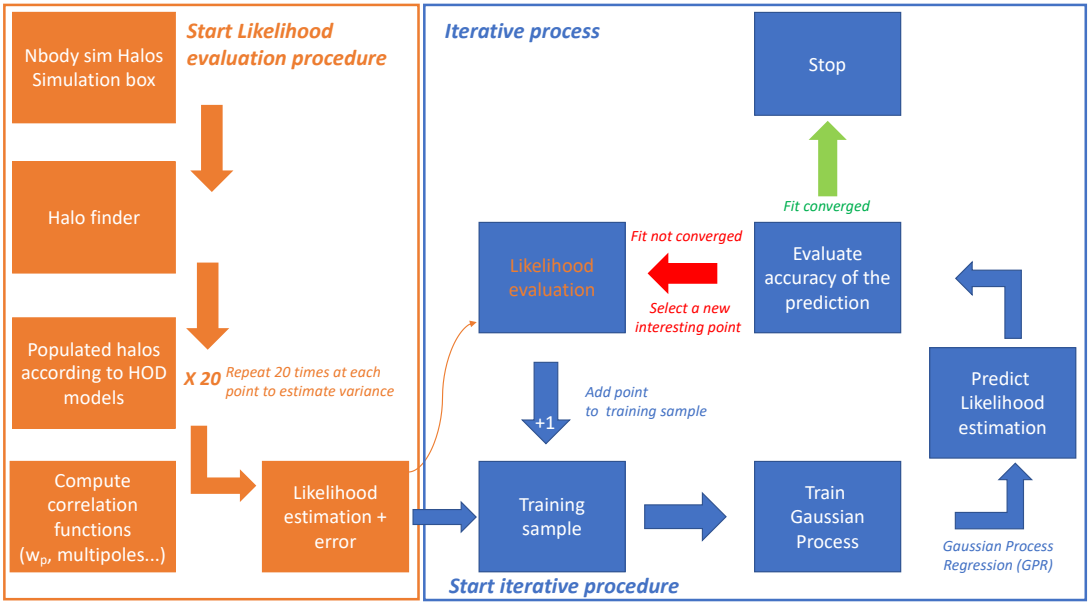

Figure 4.4: *Sketch of the fitting procedure. Left: mock catalogue creation, clustering measurements and likelihood computation. Right: fitting iterative procedure based on likelihood predictions driven by Gaussian processes.*

The surrogate model of the likelihood posterior provided by the GP from the initial training sample is iteratively improved by adding one point to the training sample at each iteration (see Figure 4.4 for a schematic view). Choosing the next point to add, x_{next}, that is choosing the GP acquisition function, can be done in several ways. The most popular one uses the Expected Improvement (EI) information acquisition function (Jones et al., 1998, Mockus et al., 1978). The latter defines how much the likelihood value at a given point is expected to improve over the current maximum and the point that gives the greatest expected improvement is taken as x_{next}. Applied to our case, this method proved to be efficient at finding the maximum of the likelihood function but did not provide accurate error contours. This illustrates the difficulties in defining an acquisition function that reaches a good compromise between exploring the full parameter space and focusing on high probability areas, as discussed in El Gammal et al. (2022).

In order to have both an accurate determination of the likelihood maximum and reliable error contours, we use the following method to determine x_{next}, based on previous works in cosmology (Neveux et al., 2022, Pellejero-Ibañez et al., 2020). At each iteration, the GP prediction is sampled by a Monte Carlo Markov chain (MCMC) algorithm and x_{next} is randomly selected in the MCMC chains. Its χ^2 value and error are computed, and the point is added to the training

sample to reiterate the procedure. This method has the advantage that the points inside the 3 σ contour around the maximum likelihood are more likely to be selected as points to be added to the training sample.

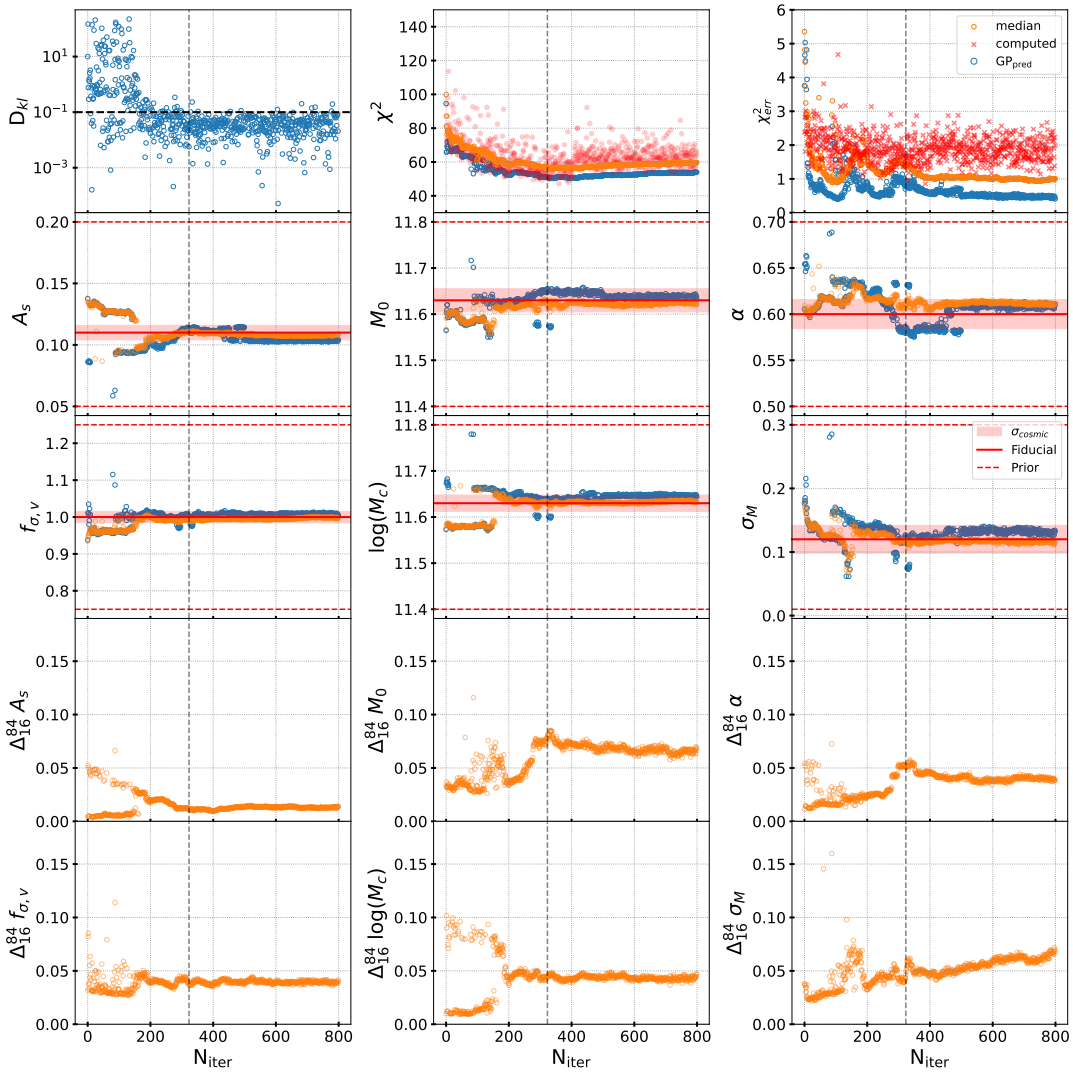

Figure 4.5: *Evolution of several fit result indicators with iteration number, for one fit from the accuracy test with cosmic variance included (see section 4.4). From left to right and top to bottom, indicators are the Kullback-Leibler divergence between MCMC chains sampling the GP prediction, the value of χ^2_{min} (blue) and that of χ^2_{med} (orange) from the chains, their errors, the values of the six HOD parameters at χ^2_{min} (blue) and their marginalised median values (orange), the $[16-84]$ percentile range of the six HOD parameters. Red dots in the middle (resp. right) top panel are χ^2 values (resp. errors) computed at the selected point added to the GP training sample at the next iteration. Red solid (resp. dashed) lines are fiducial (resp. prior boundary) values. The band indicates the dispersion from all fits with cosmic variance included. The vertical dashed line indicates the iteration at which the KL criterion is met.*

To stop the iterative procedure, we require stability of both the MCMC chains and fit results, as explained in the following. To characterise the chain stability, we rely on the *Kullback-Leibler* (KL) divergence (Kullback & Leibler, 1951) between the MCMC chains, computed at each iteration. The KL divergence (also called the relative entropy) quantifies the difference between

two distributions p and q for a set of points X and is defined as follows:

$$D_{KL}(p|q) = \sum_{x \in X} p(x) \log\left(\frac{p(x)}{q(x)}\right) \tag{4.20}$$

For two similar distributions $D_{KL} \to 0$. We consider our KL criterion as fulfilled when the KL divergence is below 0.1 in a set of 20 consecutive iterations before the current one. This threshold was determined empirically from our fits, but is rather common in iterative GP emulators (Neveux et al., 2022, Pellejero-Ibañez et al., 2020). To illustrate the stability of the chains and fit results, Figure 4.5 shows an example of the evolution of several indicators of the fit results as a function of the iteration number.

The fit results on HOD parameters at each iteration are characterized both by values corresponding to the minimum χ^2 value of the MCMC chains run on the GP prediction (hereafter called χ^2_{min}), and by marginalised values defined as the median values of the posterior distributions run from the same MCMC chains. Errors on these marginalised values are defined by the 16% and 84% percentiles of the parameter posterior distributions. Besides the KL divergence, the fit results reported in Figure 4.5 are thus the values of the six HOD parameters at χ^2_{min} and their marginalised values, as well as the value of χ^2_{min} and the χ^2 value for the GP prediction at the marginalised HOD parameter values (hereafter dubbed as χ^2_{med}), together with their corresponding errors. We also show the evolution of the size of the $[16-84]$ percentile range of the posterior distribution for the six HOD parameters. Finally, in the sub-plots related to χ^2 values and errors, we added the computed χ^2 value and its error for the point added at each iteration.

As can be seen from the figure, the KL divergence drops and remains below the threshold of 0.1 after iteration 300. The values of χ^2_{min} and χ^2_{med} reach a plateau after that iteration but the learning phase of the GP continues, as shown by the excursions in the χ^2_{min} error. The explored range in the parameter values, indicated by the excursions of their values at χ^2_{min} are limited and induce small variations of the marginalised HOD parameter values and their percentile ranges, which are all well stabilized at iteration 800, although the percentile range of σ_M may still be evolving slightly. For completeness, Figure 4.6 presents the contour plot of the fit at iteration 800 (see also Section 4.4.3) and compares the pseudo-data clustering to that from the HOD model defined by the marginalised parameter values at that iteration. The modelling of the pseudo-data clustering is good, well within errors due to model stochasticity and cosmic variance, the latter being the dominant effect at large scales. The χ^2_{med} value at the final iteration is 56.8 ± 2.2 for a number of degrees of freedom of $75 - 6 = 69$ (p-value of 85%).

Although the fit in the above example reaches stable and reliable results after 200 iterations are added once the KL criterion is met, nothing guarantees that this is generally the case (El Gammal et al., 2022). The stability of the results from the iterative method will thus be investigated in a systematic way in section 4.4, by running the procedure on a large set of simulated mocks, taking into account cosmic variance. All fits will be run up to a maximal number of iterations $N_{max} = 800$ and the fit precision and accuracy of the method will be explored with that stopping criterion, which we discuss further in section 4.4. The fit results will be defined by the marginalised HOD parameter values at iteration N_{max}, with statistical uncertainties on these given by the 16% and 84% percentiles of the parameter posteriors at that same iteration.

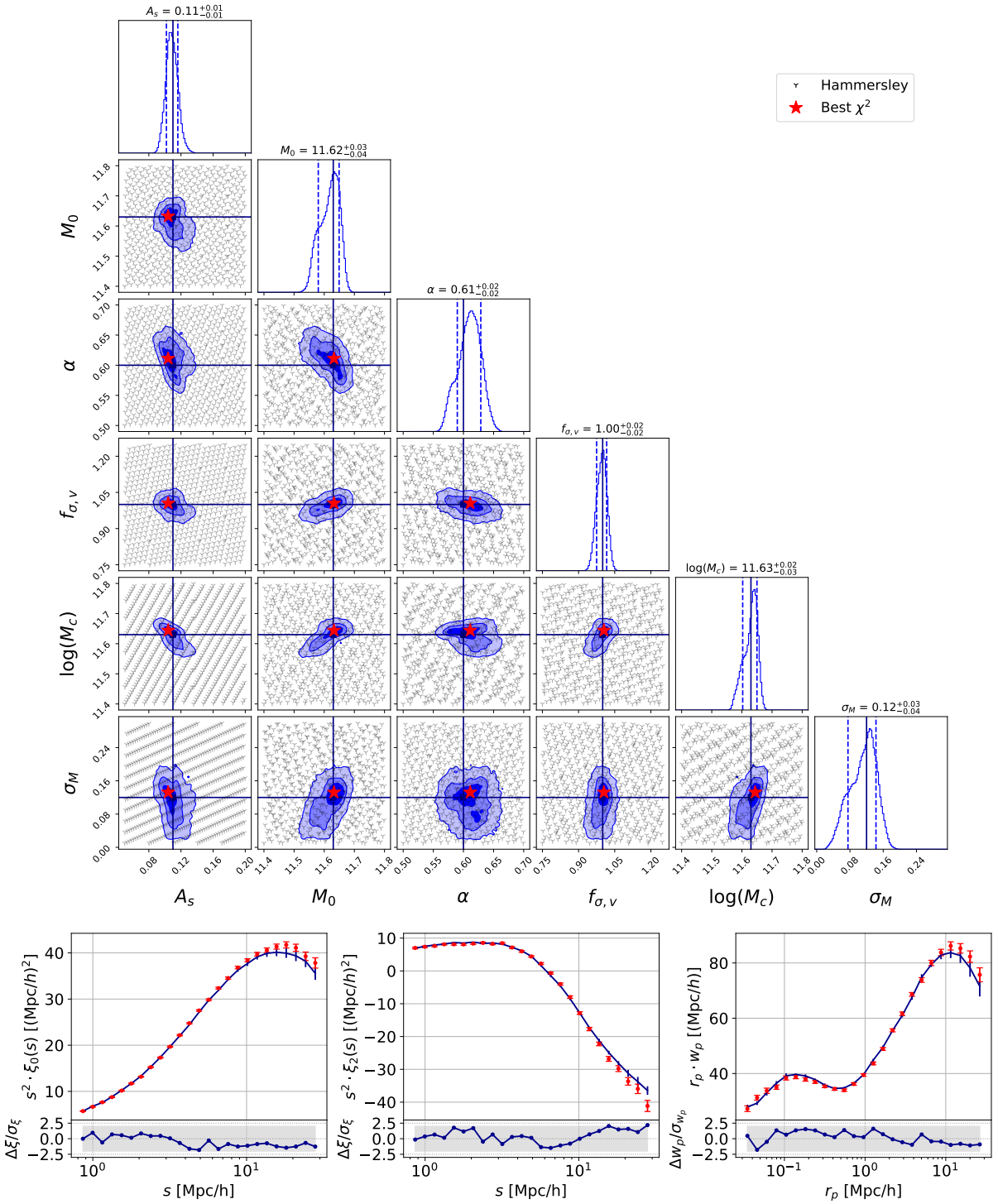

Figure 4.6: *Top: Contours and marginalised 1D-posteriors at iteration 800 for the fit in figure 4.5. Grey points are the Hammersley training sample. The red star corresponds to the minimal χ^2 of the GP prediction. The contours are obtained from MCMC chains run on the GP predictions, after burning phase. Solid lines indicate the parameter input values. Bottom: Clustering measurements predicted by the HOD model from the fit in the top plot. The shaded band corresponds to $\pm 2\sigma$ residuals, with errors on pseudo-data (red bars) and model (blue bars) added in quadrature.*

4.4 Tests of the method

We first test to what extent the procedure can be considered as reproducible and then include cosmic variance to test the method accuracy. We then study how the results evolve when the ingredients of the method are changed.

4.4.1 Reproducibility

To test the reproducibility of the method, we use the 1 Gpc/h length cubic box in the base cosmology (see Table 5.1) to create one pseudo-data mock with the HOD parameter input values in Table 4.2 (top row). The data covariance matrix is computed from 1,000 realisations of the same HOD model on the same box. This covariance matrix includes stochastic noise in the process of populating halos with galaxies and the statistical noise induced by the density of the resulting mock catalogues, the two irreducible sources of noise of the procedure. Data and model variances were compared (for the input HOD model) and found to be comparable in all separation bins of the clustering measurements.

We perform 24 independent fits to the pseudo-data mock, all with the same initial sampling of the parameter space (see section 4.3.5). Results are presented in Figure 4.7. The results of the 24 fits agree with each other within ± 0.002 for A_s, ± 0.005 for α, ± 0.004 for f_{σ_v}, ± 0.013 for $\sigma_{\rm M}$ and within ± 0.016 and ± 0.010 for M_0 and $\log_{10}(M_c)$ respectively. These numbers quantify the reproducibility limits at the 68% confidence level of our procedure, due to its stochastic nature. Also included in the plot are the expected dispersions when cosmic variance is also taken into account (see next section), showing that, except for M_0 and $\sigma_{\rm M}$, the intrinsic dispersion due to stochasticity is a sub-dominant component.

Spurious instabilities in the GP predicted likelihood posterior were observed for these fits. Indeed, in reproducibility fits, due to uncertainties on pseudo-data being very small, the surrogate model of the likelihood surface can present spikes, likely due to the stochasticity of the HOD modelling. Spikes occur in about 50% of the iterations and their locations in the HOD parameter space vary in the course of the iterative procedure. Besides, iterations showing spikes are generally associated with large χ^2 errors in the MCMC chains, showing that these spikes are most probably spurious and mostly due to regions with not enough points, so the GP can predict huge variations in the χ^2 value. In order to obtain reliable contours and errors, we remove points with large GP predicted χ^2 errors in the MCMC chains. This is illustrated in Figure 4.8 which shows that spurious spikes do indeed disappear, leaving the main component of the contours unchanged. This procedure is applied to reproducibility fits only, other fits presented below have smooth predicted likelihood surface at the final iteration.

4.4.2 Accuracy with cosmic variance

To test the method accuracy in more realistic conditions, cosmic variance must be included. We thus cut each of the 25 large boxes of 2 Gpc/h length in the base cosmology (see Table 5.1) into 1 Gpc/h length cubes, which allows us to create 200 independent mocks corresponding to different realisations of the same cosmology.

We take 25 out of these 200 mocks as pseudo-data (one per large box) and perform four series of fits, each series using a different 1 Gpc/h sub-cube for the model. These sub-cubes

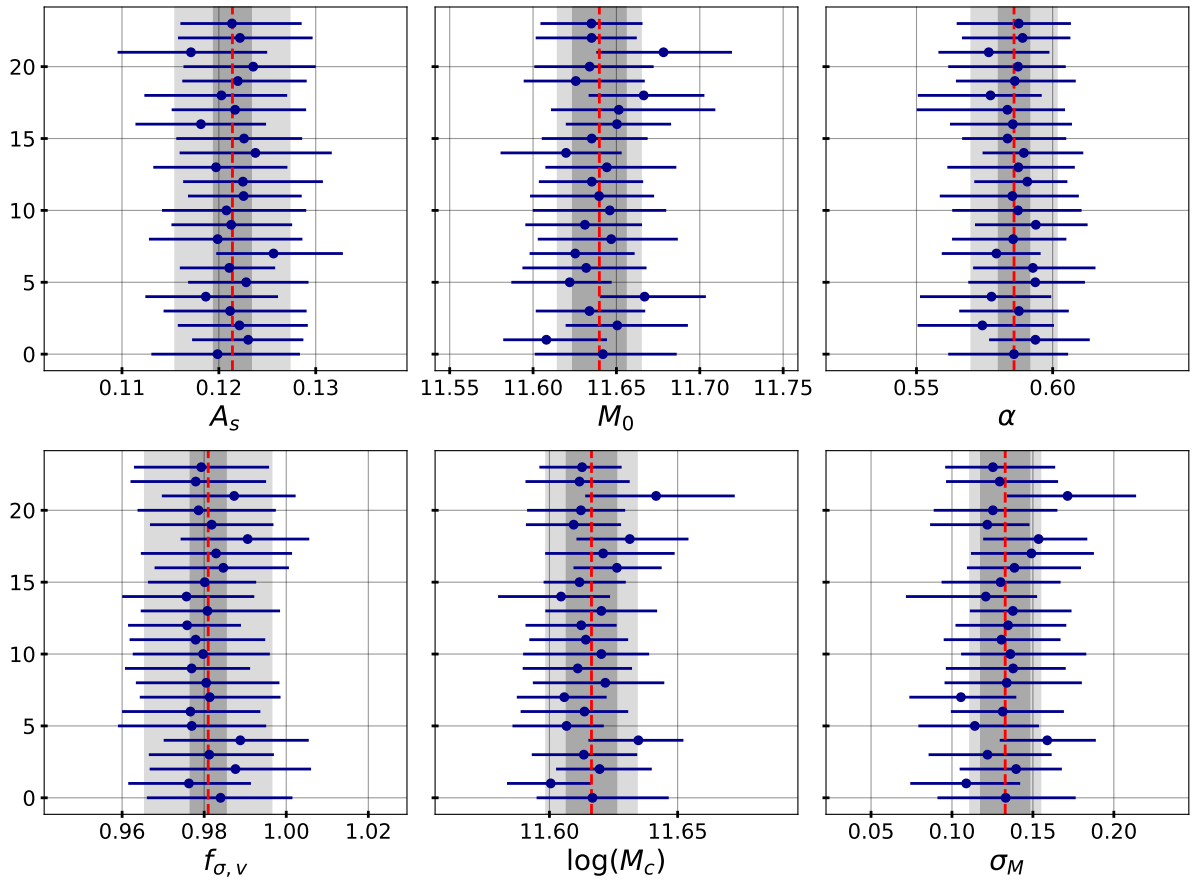

Figure 4.7: *Reproducibility test from 25 fits to the same pseudo-data under the same initial conditions. Blue dots are marginalised HOD parameter values with errors defined by the* 16% *and* 84% *percentiles of the fit posteriors. The dark grey band spans the* $\pm 1\sigma$ *interval around the average of the marginalised values given by the vertical red dashed line. The light grey band includes also cosmic variance (from the accuracy test of section 4.4.2).*

differ from those used to create the 25 pseudo-data mocks and three of them belong to the same large box. The training sample is recomputed for each of the 4 sub-cubes.

The data covariance matrix is built from the entire set of 200 mocks. This matrix includes stochastic noise in the process of populating halos with galaxies, statistical noise induced by mock density, and cosmic variance. The model covariance matrix is built as described in Section 4.3.4, but the variances used to compute the normalisation factor applied at each HOD point to the fixed correlation matrix are modified to account also for cosmic variance in the choice of a given sub-cube for the model. To the variances used in Section 4.3.4 to define the normalisation factor, we thus add the clustering measurement variances over the 8 sub-cubes of the box used for the model, computed for the input HOD, with one realisation per sub-cube. Data and model variances were compared and their difference was found to be within ± 0.6 times the pseudo-data variances, with no marked scale dependence.

The distribution of the first iteration meeting the KL criterion as defined in section 4.3.6, is presented in Figure 4.9 (left-hand plot, blue histogram). 86% of the fits pass the KL criterion before iteration 600 and only 4% did not reach stability at the last iteration, $N_{max} = 800$. Marginalised fit results are compared to the input HOD parameter values in Figure 4.10. The

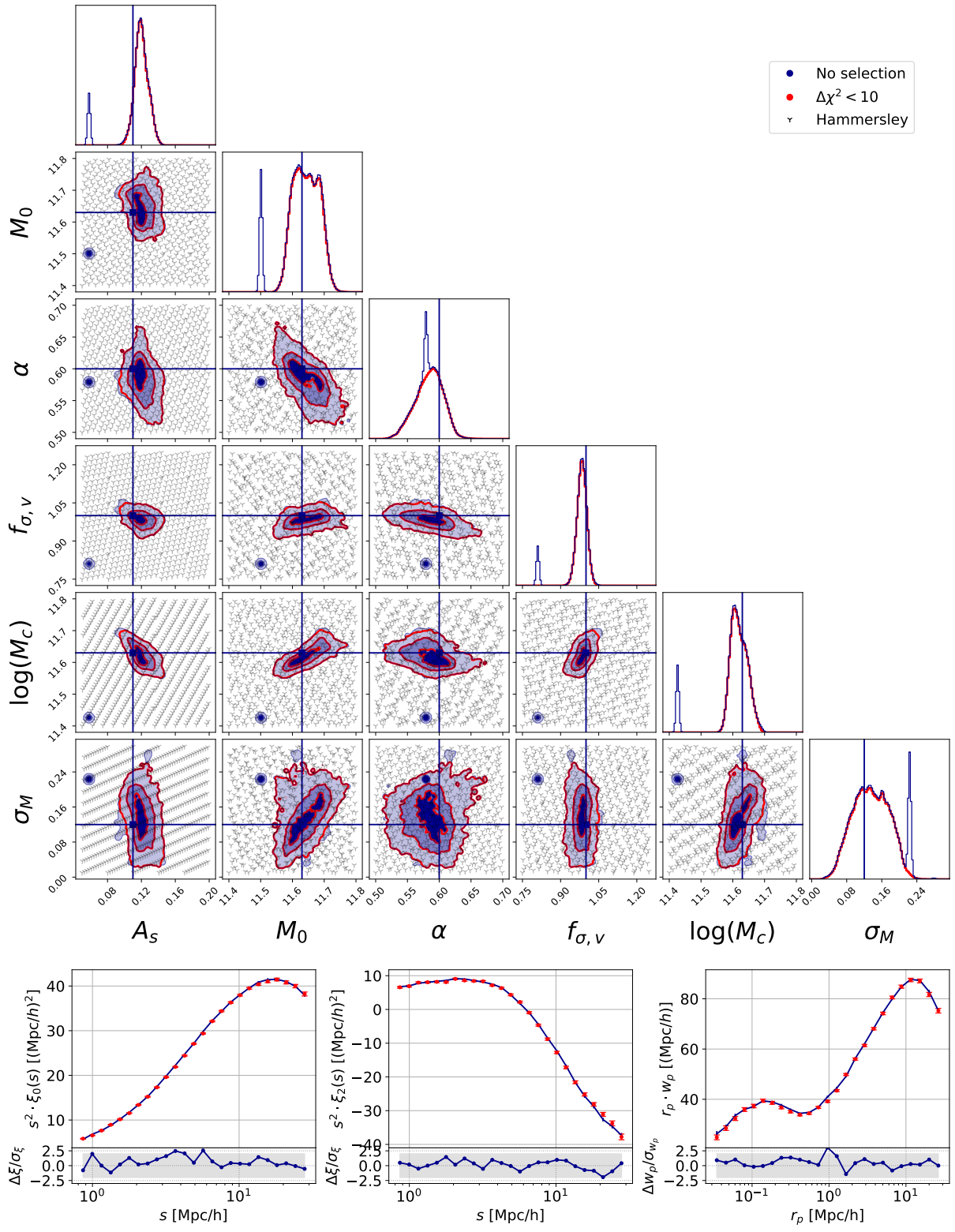

Figure 4.8: *Top: Contours and marginalised 1D-posteriors at iteration 800 for one fit from the reproducibility tests. Solid lines are the input parameter values. Grey points are the Hammersley training sample. The contours are obtained from MCMC chains after the burn-in phase, with no further selection in blue and excluding points with a large predicted χ^2 error in red. Bottom: Clustering measurements predicted by the HOD model given by the marginalised values from the red 1D-posteriors in the top plot. The shaded band encompasses $\pm 2\sigma$ residuals, with errors on pseudo-data (red) and model (blue) added in quadrature.*

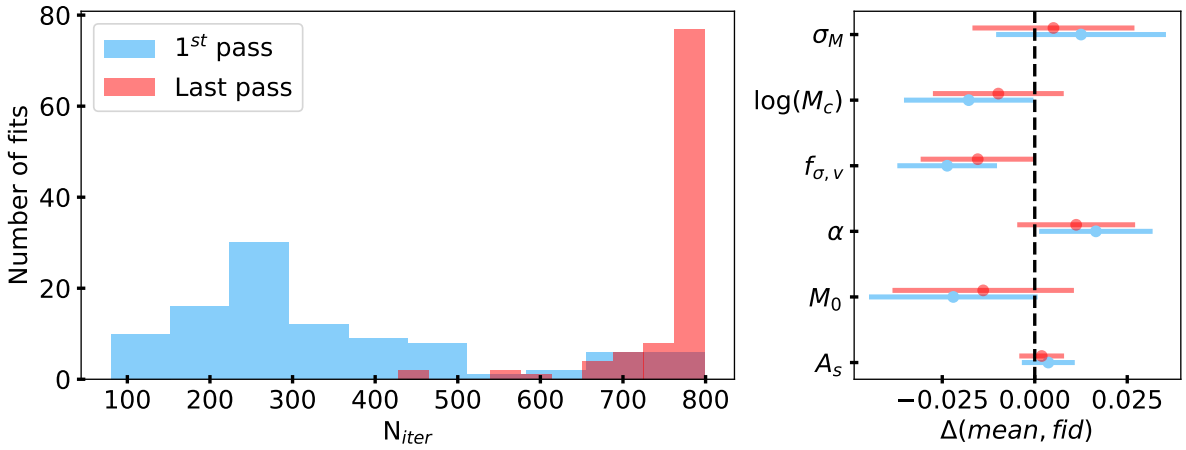

Figure 4.9: *Left: Distribution of first (in light-blue) and last (in red) iteration fulfilling the Kullback-Leibler stability criterion for a set of 100 fits with cosmic variance included. Right: Difference between the fiducial values and the mean results of the 100 fits, for each of the 6 HOD parameters. Results from fits stopped at the first (resp. last) iteration fulfilling the Kullback-Leibler criterion are reported in light-blue (resp. red). The error bars correspond to the ± one standard deviation of the 100 marginalised fit results. All fits had initial training of the Gaussian Processes based on Hammersley sampling of the HOD parameter space with 600 points.*

four fits which did not pass the KL criterion do not stand as clear outliers in any of the parameters, showing that the lack of stability does not necessarily imply a large offset in the measured parameters, nor larger error bars. Figure 4.9 also shows the distribution of the last iteration meeting the KL criterion (left-hand plot, red histogram), which is strongly peaked towards N_{max}. The right-hand plot in this figure compares the bias observed on each parameter when the fits are stopped at the first or last iteration meeting the KL criterion, instead of allowing all fits to go up to N_{max}. Stopping the fits at the first iteration reaching the KL stability criterion clearly leads to larger biases than stopping at the last one or going up to N_{max} (biases in these two cases are very similar), which justifies our choice of the latter option as a stopping criterion.

With this criterion, all HOD parameters are reconstructed with a mean bias either well within, or for f_{σ_v} at the level of, one standard deviation of the parameter distribution. More precisely, we find a mean bias of 0.29σ for A_s, 0.52σ for M_0, 0.69σ for α, 0.97σ for f_{σ_v}, 0.52σ for $\log(M_c)$ and 0.26σ for σ_M. In the above, σ is the standard deviation of the marginalised fit result distribution. This is also the expected statistical error for one fit, with the errors accounted for in the covariance matrix used in the fits, namely stochasticity, cosmic variance and galaxy sample size for a density close to that expected for DESI data but for a volume three times larger than that of the early DESI ELG data. We thus expect the HOD parameter values to be derived from these data with our procedure to have an accuracy much better than 1 σ of the data statistical uncertainty for most parameters, the worst case being the f_{σ_v} parameter for which the accuracy is expected to be about 0.6 σ.

The above reasoning assumes statistical errors from the fits to be normally distributed. As a cross-check, we compared the mean of the parameter errors from the fits to the standard deviation of the marginalised fit result distribution used in the above bias estimates. We found the mean error to be higher than the standard deviation by 60% for σ_M and 20-30% for all other

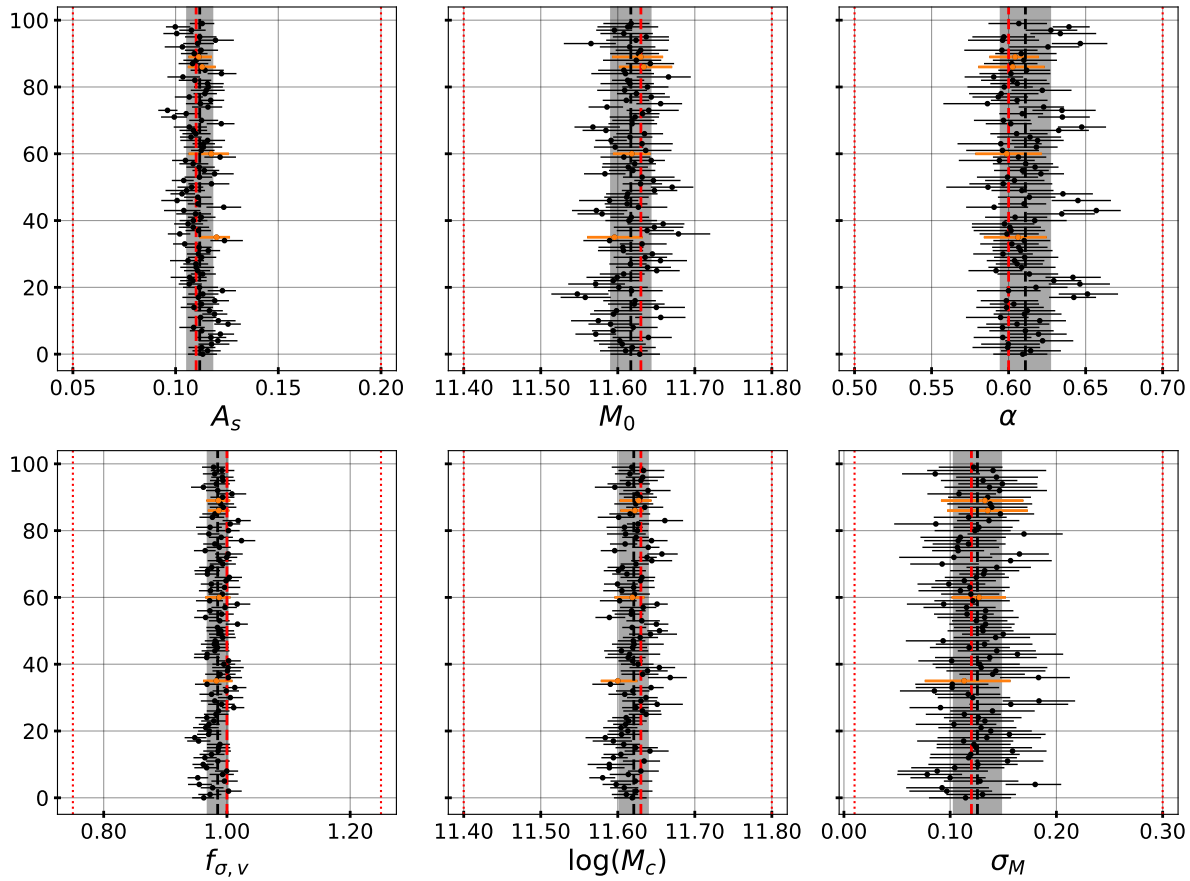

Figure 4.10: *Accuracy test from 100 fits with cosmic variance included. Dots are marginalised HOD parameter values with errors defined by the* 16% *and* 84% *percentiles of the fit posteriors. Orange dots stand for the four fits which did not converge. The grey band spans the* $\pm 1\sigma$ *interval around the average of the marginalised values given by the vertical black line. The red dashed line is the input HOD parameter values. Four series of fits to the same 25 pseudo-data mocks were run with a model drawn from a different sub-cube of the same large box for the first three series and from a sub-cube of a different large box in the fourth one. All fits were run up to 800 iterations after initial training of the Gaussian Processes based on Hammersley sampling of the HOD parameter space with 600 points. Red dotted lines indicate the fit priors.*

parameters. These departures are likely to be due to non Gaussian posteriors, as observed in most parameters (see Figure 4.6), and make our expected bias estimates conservative.

However, despite the slight biases observed in the HOD parameters, the procedure provides a very good modelling of the clustering statistics, as already shown in Figure 4.6 on one example. Altogether, in the four series of fits to the 25 pseudo-data mocks used in this section, the mean value of the computed reduced χ^2 for the best fit model (defined by the marginalised HOD parameter values) is $\sim$0.8 with a dispersion of $\sim$0.15.

4.4.3 More on stability

Figure 4.11 shows the superimposition of contours and marginalised 1D-posteriors from 50 iterations of the fit shown in Figures 4.5 and 4.6. We took every one iteration out of four between

iteration 600 and the stopping iteration, 800. This plot is complementary to Figure 4.5. It illustrates that, in the last quarter of iterations, the evolution of the GP surrogate model does not alter the marginalised median value of the posteriors but only changes slightly their $[16-84]$ percentile range.

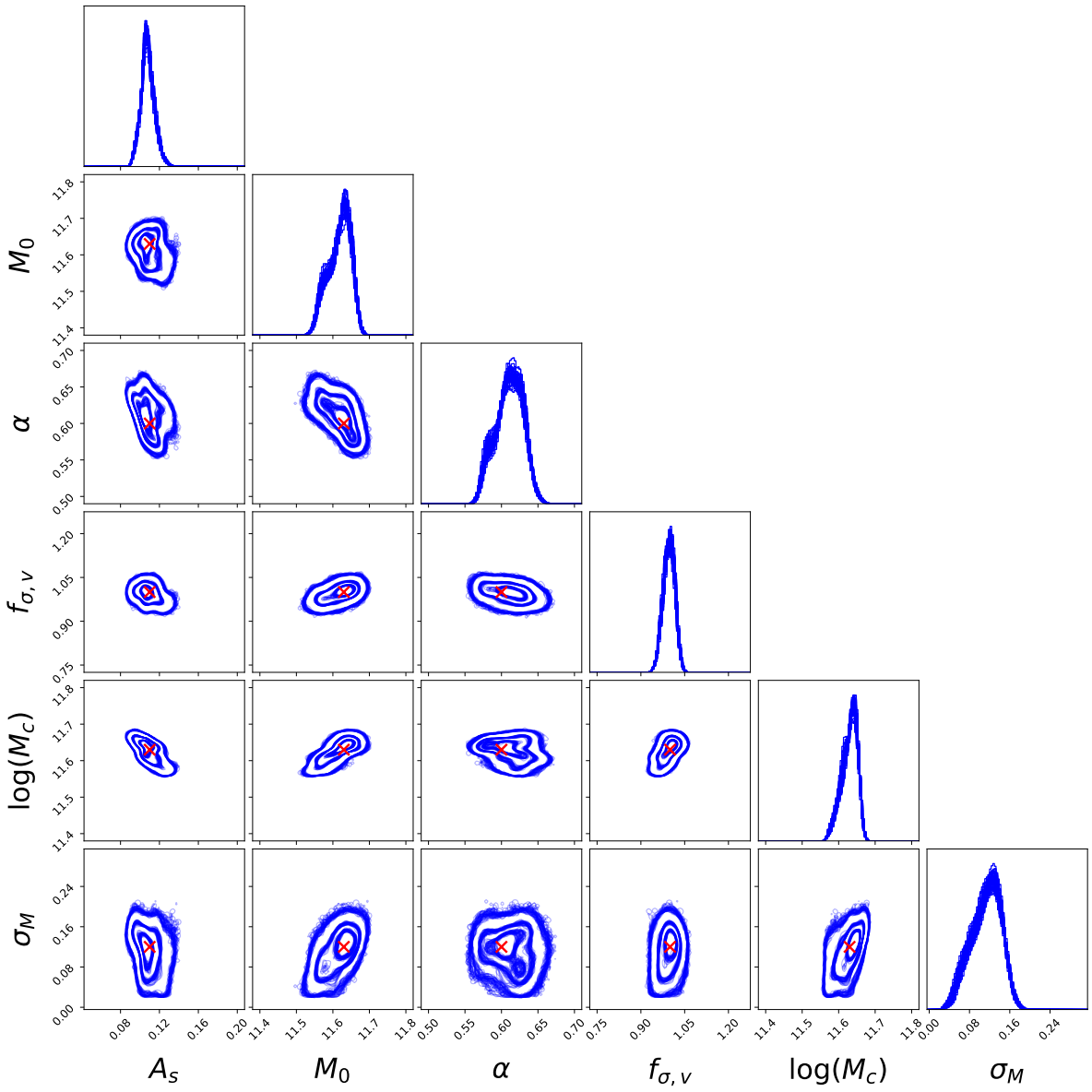

Figure 4.11: *Contours and marginalised 1D-posteriors superimposed taking one iteration out of four between iterations 600 and 800 (so 50 iterations in total), for the fit shown in Figure 4.5. The red cross is the parameter input values.*

4.4.4 Dependence on initial conditions and kernel

In this section, we test how the results of our procedure are affected by different initial conditions. We present results from different numbers of points in the initial training sample, from different GP kernels and different initial sampling algorithms. The priors being unchanged,

testing different numbers of points in the training phase amounts to testing different densities when sampling the HOD parameter space.

4.4.5 Initial training sample

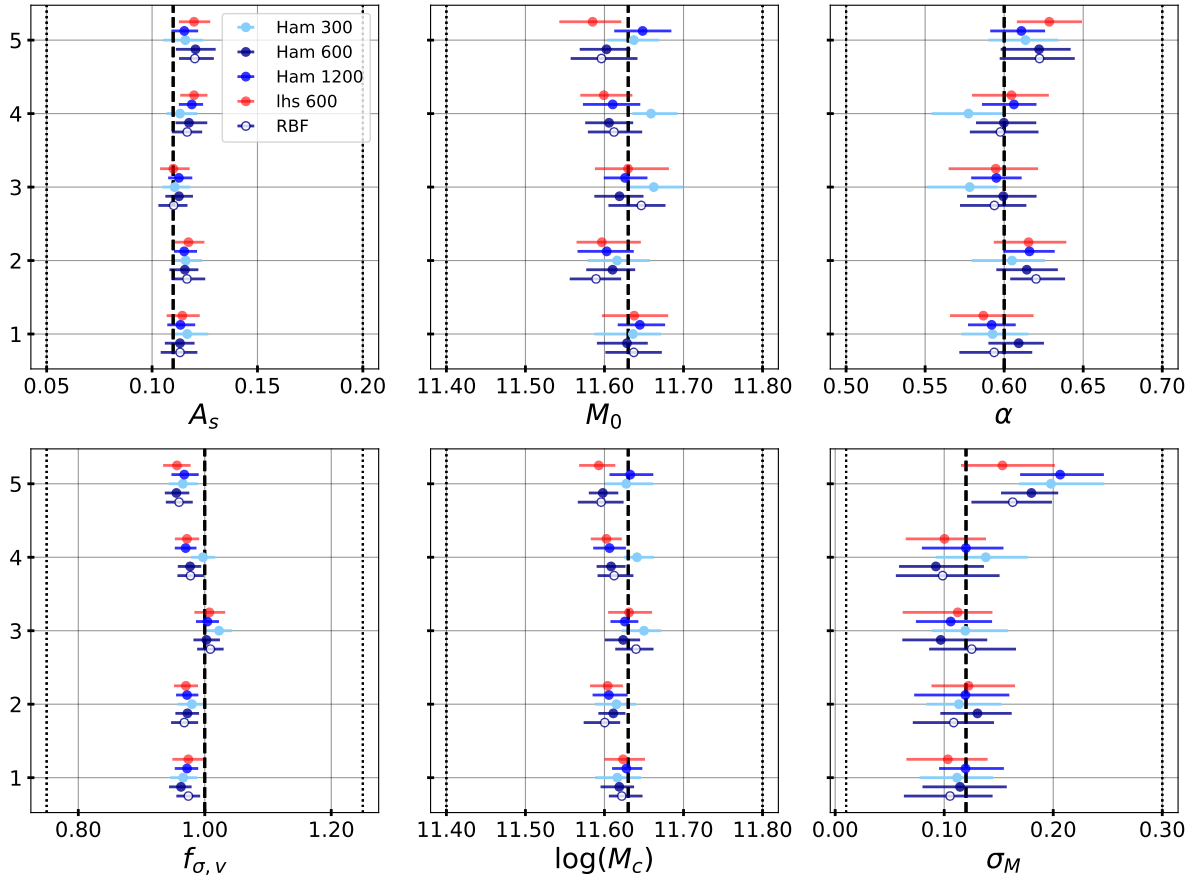

Figure 4.12: *Tests from fits with different training samples and different kernels for the Gaussian Processes. Five series of fits to the same 5 pseudo-data mocks with the same box for the model were run with different sizes of the Hammersley training sample (from 300 to 1200), or replacing Hammerlsey sampling by LHS for the baseline sample size (600) or using a Radial Basis Function kernel instead of the baseline Matérn kernel of index 5/2 for the baseline sample size (600). All fits were run up to 800 iterations. Dots are marginalised HOD parameter values with errors defined by the* 16% *and* 84% *percentiles of the fit posteriors. The black dashed line is the input HOD parameter values and the dotted ones indicate the fit priors.*

Our baseline option for the initial training sample is 600 points distributed according to Hammersley sampling, with A_s as the parameter with equidistant points. Using the first 5 pseudo-data mocks submitted to fits with cosmic variance (see section 4.4.2), we run fits with Hammersley training samples of 300 and 1200 points, and with LHS training samples of 600 points. All other fits conditions remain unchanged, notably running all fits up to 800 iterations, with the same box for the model.

Results are reported in figure 4.12. The dispersion of the results between different conditions of fits for a given pseudo-data mock is generally lower than the dispersion between mocks for the

same fitting conditions, which is dominated by cosmic variance. We note however that sampling with only 300 points appears to gives results less consistent with fits in other conditions for M_0, α and $\log(M_c)$. LHS sampling gives slightly larger errors than Hammersley sampling with the same number of points and there is no obvious gain in accuracy with 1200 points compared to 600 points in Hammerlsey sampling.

This means that there is no need to increase the density of the initial sampling infinitely, at some point what matters most is to increase the density in the region of interest, close to the likelihood maximum, which is the aim of the iterative procedure that follows initial sampling.

4.4.6 Choice of GP kernel

In a second test, the 5 pseudo-data mocks were submitted to fits with Hammersley sampling of 600 points and A_s as the parameter with equidistant points but with an RBF kernel instead of the baseline one, a Matérn kernel of index 5/2. Again, all other fits conditions remained unchanged. Results are reported in figure 4.12. Changing the kernel does not change the results significantly. We note that the RBF kernel leads to slightly larger uncertainties on the parameters, the average increase ranging from 4% for A_s to 23% for $\log(M_c)$.

4.4.7 Choice of parameter with equidistant points

In a third test, the same 5 pseudo-data mocks were submitted to fits with Hammersley sampling of 600 points, varying the parameter with equidistant points, all other fitting conditions remaining unchanged. Results are reported in figure 4.13. Except for pseudo-data mock 5, the dispersion of the results between the different choices for a given pseudo-data mock is small, but we note that choosing f_{σ_v} as the parameter with equidistant points can give results differing by 1σ from those with other choices, notably for M_0.

Finally, as already observed in section 4.4.2, there is a slight systematic offset in the fit values of f_{σ_v} in both figures 4.12 and 4.13, but it remains at the same level as in section 4.4.2, that is below the statistical uncertainty expected from the early DESI ELG sample.

4.5 Practical implementation

The HOD fitting procedure described in this chapter relies on an HOD pipeline and a fitting pipeline that I developed during this thesis work. Implementation details for both steps are given in the following.

4.5.1 HOD pipeline

The HOD pipeline[1] built for this work produces mock catalogues and clustering measurements from an N-body simulation box, given parameters of an HOD model. Besides the GHOD model defined in Section 4.3.1, the code supports other HOD models for central galaxies described in Section 3.4.1, all supplemented by a power law model for satellite galaxies. By default, satellites are drawn within randomly pre-computed r/r_s points from an NFW profile (Navarro et al.,

[1] https://github.com/antoine-rocher/GP_HODpy

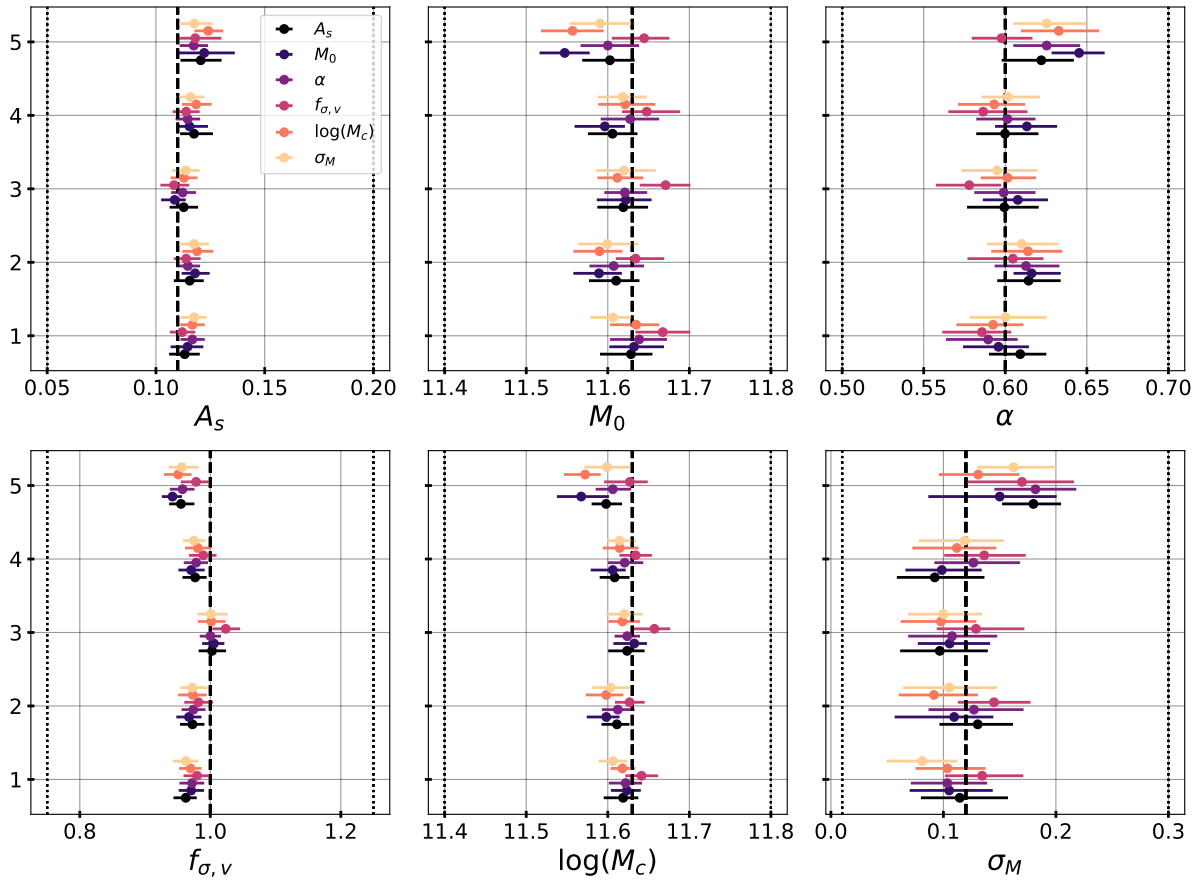

Figure 4.13: *Tests from fits with different training samples. Four series of fits to the same 5 pseudo-data mocks with the same box for the model were run with 600 points from Hammersley sampling, varying the parameter with equidistant sampling. Dots are marginalised HOD parameter values with errors defined by the* 16% *and* 84% *percentiles of the fit posteriors. Different colors indicate different choices for the parameter with equidistant sampling, A_s being the default. The black dashed line is the input HOD parameter values and the dotted ones indicate the fit priors.*

1996). Satellites populated on simulation DM particles are also supported as an alternative. For each mock computation, central and satellite galaxies are generated independently and concatenated in a Python dictionary. To optimize the mock generation we use `numba`[2], an open source Just In Time (JIT) compiler, that translates a subset of Python and `numpy` code into fast machine code, multi-threaded with automatic parallelization of JIT. We are also developing an MPI implementation of the mock generation that was not used for this study.

Since HOD fitting results depend on the observed density of objects, the density in mocks can be set to a given value, as used in this work. The code generates the exact chosen density by pre-computing the total number of galaxies, n_{gal}, using Equation (4.10) for the current set of HOD parameters. The amplitudes for centrals and satellites, A_c and A_s, are then rescaled by $n_{gal,exp}/n_{gal}$, where $n_{gal,exp}$ is the galaxy number expected for the chosen density. As the resulting mock clustering only depends on the ratio A_s/A_c, we can re-scale both amplitudes by the same factor to change the density while keeping the same clustering.

[2] `https://numba.pydata.org/`

Once a mock catalogue is computed, the HOD pipeline provides an easy way to compute clustering measurements for the projected 2 point correlation function, w_p and for the 2 point correlation function monopole and quadrupole, with user-defined separation ranges and binnings. We use the DESI wrapper PYCORR[3] around the `Corrfunc` package (Sinha & Garrison, 2020) to compute these measurements.

The fitting procedure requires $N = 20$ mocks to be created at each point of the HOD parameter space to compute the χ^2 value and its uncertainty. To speed up the fitting procedure, this step runs the N mocks in parallel using the `joblib` package[4]. Then, the N correlation functions are computed one after the other. Reproducibility of the results when using multi-threading can be quite complicated. As each thread will generate a different seed (even if a seed is fixed at initialisation), it becomes difficult to have reproducible results. We choose to adopt a solution easy to implement. We fix a single seed at initialisation and use it to determine a list of random integer numbers that will be used to initialise one seed per thread. This is easy to implement but the reproducibility of the results will depend on the number of threads.

4.5.2 Fitting pipeline

The Hammersley sampling of the HOD parameter space is performed with the PySMO sampling method[5] of the `idaes-pse` package. The Gaussian process part relies on the Gaussian Process Regressor[6] function from the `scikit-learn` package. The GP hyperparameters are the length scales and the kernel variance. All are initially set to one. The kernel variance σ_k^2 is allowed to vary between 10^{-5} and 10^5. The same range of variation is imposed for all length scales by the GP package and we set the it to be between 10^{-3} and 10. Note that the length scale values depend on the parameter values and prior ranges. Given our parameters, the chosen prior on length scales is large enough to describe reasonable variations, knowing that all our parameter values and ranges of variation are of order 10^{-2} to 1.

The MCMC component is ensured by the `emcee` package and runs 12 chains of 10,000 points each in parallel, the first 800 points being discarded in each chain.

4.5.3 Performance

The performance of the inference procedure are as follows. Our computer system uses two AMD EPYC 7513 32-core processors, which are multi-threaded by 2 (128 threads/node in total), clocked at 2.6 GHz and equipped with 256 GB DDR4 RAM. The tests described in this section were run with 24 threads on a single node. The CPU time consumption per point of the HOD parameter space breaks down as follows: ~ 25 sec to create 20 realisations of the HOD model from a cubic simulation box of 1 Gpc/h length, ~ 12 sec to compute the correlations and the χ^2 value based on these 20 realisations, ~ 20 sec to run the MCMC chains on the GP prediction. The CPU time consumption to derive the prediction of the GP depends on the number of points in the training sample. Altogether, for a 6-parameter fit based on 600 training points and 800

[3] https://github.com/cosmodesi/pycorr

[4] https://joblib.readthedocs.io/en/latest/generated/joblib.Parallel.html

[5] https://idaes-pse.readthedocs.io/en/1.5.1/surrogate/pysmo/pysmo_sampling_properties.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html

added points, the total CPU time consumption per iteration increases from ~ 50 seconds with the initial training to ~ 3 minutes at the final iteration.

4.6 Summary and prospects

In this chapter, we introduced a method to fit HOD parameters using Gaussian Processes (GP) to provide a model of the multidimensional likelihood function, in the framework of a stochastic HOD modelling technique based on mock galaxy catalogues built from N-body simulations.

Our two-step procedure starts with initial training of the GP with 600 points distributed in the HOD parameter space according to Hammersley sampling. The likelihood model provided by the GP from this initial training is further improved by an iterative procedure adding one point to the training sample at each iteration, the next point to be added being randomly selected in Monte Carlo Markov chains (MCMC) run on the likelihood posterior predicted by the GP at the current iteration. This ensures that the sampling is made denser close the maximum of the likelihood function so as to provide a good determination of both the likelihood maximum and the error contours, despite the stochastic nature of our HOD modelling. The iterative procedure is pushed until a total of 800 iterations is achieved.

The reproducibility and accuracy of the method were studied on simulated mocks built from the AbacusSummit suite of high-accuracy N-body simulations on cubic boxes of 1 Gpc/h length. These mocks are representative of the expected density of the DESI ELG sample, but cover a volume three times larger than that covered by the early DESI ELG data. The procedure was repeated on sets of simulated mocks corresponding to different realisations of the same 6-parameter HOD model suitable for ELGs. Results on the 6 HOD parameters, defined by the marginalised values from the posterior distributions extracted from the MCMC chains run at the final iteration, were found to be reproducible within ranges smaller than those expected when cosmic variance is also included. In the presence of cosmic variance, we reach accuracies on the HOD parameters which are below the statistical uncertainty expected for early DESI ELG data, reaching at most 60% of the statistical uncertainty in the worst case (one parameter out of six, the maximum bias for the other parameters being 40%).

We also explore the stability of the method when varying different ingredients. We find that the results do not depend on the sampling algorithm applied to define the training sample nor on the GP kernel. This is also true for the choice of the parameter with equidistant points in the initial training sample with our baseline Hammersley sampling. More dependence is found with respect to the choice of the number of points in the training sample and in the number of iterations after initial training. Different numbers of training points as well as numbers of iterations lower than 800 were tested. We find that there is no need to increase the density of the initial sampling infinitely. What matters most is to increase the density in the region of interest close to the likelihood maximum, once that region is roughly defined. In our framework this is achieved with our baseline of 600 points in initial training and 800 further iterations.

Finally, the fit progress towards stability during the iterative loop was monitored with the help of the Kullback-Leibler (KL) divergence between the MCMC chains. Requiring the KL divergence to be below 0.1 in a set of 20 consecutive iterations as a chain stability criterion, we observe that 96% of the fits pass this criterion well before iteration 800 and the few fits which fail are not outliers in any of the HOD parameters. On the other hand, if fits were stopped as soon

as the KL criterion was met, we would obtain larger biases in the HOD parameters, showing that this does not ensure unbiased results. Hence our choice to push fits up to a total of 800 iterations, for which our tests on simulation show that the expected bias in the HOD parameter values is reasonably below the statistical uncertainty we expect from data. More generally, this illustrates the difficulty to define a robust convergence criterion when inference is performed on a surrogate model of a likelihood posterior while the model is still under evolution and subject to noise in the likelihood estimates.

In the next chapter I apply this procedure to the ELG sample from the DESI 1% survey.

Bibliography

Angulo, R. E., Zennaro, M., Contreras, S., et al. 2021, Monthly Notices of the Royal Astronomical Society, 507, 5869–5881, doi: 10.1093/mnras/stab2018

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, MNRAS, 499, 5486, doi: 10.1093/mnras/staa2951

El Gammal, J., Schöneberg, N., Torrado, J., & Fidler, C. 2022, arXiv e-prints, doi: 10.48550/arXiv.2211.02045

Garnett, R., Osborne, M. A., & Roberts, S. J. 2008, Gaussian Processes for Global Optimization (Proccedings of Third International Conference on Learning and Intelligent Optimization (LION3))

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399, doi: 10.1051/0004-6361:2006617010.48550/arXiv.astro-ph/0608064

Jones, D. R., Schonlau, M., & Welch, W. J. 1998, Journal of Global optimization, 13, 455, doi: 10.1023/A:1008306431147

Kullback, S., & Leibler, R. 1951, Ann. Math. Statist., 22, 79, doi: 10.1214/aoms/1177729694

Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., et al. 2021, MNRAS, 508, 4017, doi: 10.1093/mnras/stab2484

McKay, M. D., Beckman, R. J., & Conover, W. J. 1979, Technometrics, 21, 239–245, doi: 10.1080/00401706.1979.10489755

Mockus, J., Tiesis, V., & Zilinskas, A. 1978, Towards Global Optimization, Vol. 2, The application of Bayesian methods for seeking the extremum (North-Holland Publishing Company), 117–129

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563, doi: 10.1086/177173

Neveux, R., Burtin, E., Ruhlmann-Kleider, V., et al. 2022, MNRAS, 516, 1910, doi: 10.1093/mnras/stac211410.48550/arXiv.2201.04679

Nishimichi, T., Takada, M., Takahashi, R., et al. 2019, ApJ, 884, 29, doi: 10.3847/1538-4357/ab3719

Pellejero-Ibañez, M., Angulo, R. E., Aricó, G., et al. 2020, MNRAS, 499, 5257, doi: 10.1093/mnras/staa307510.48550/arXiv.1912.08806

Rasmussen, C. E., & Williams, C. K. I. 2005, Gaussian Processes for Machine Learning (The MIT Press), doi: 10.7551/mitpress/3206.001.0001

Rocher, A., Ruhlmann-Kleider, V., Burtin, E., & de Mattia, A. 2023, J. Cosmology Astropart. Phys., 2023, 033, doi: 10.1088/1475-7516/2023/05/033

Sinha, M., & Garrison, L. H. 2020, MNRAS, 491, 3022, doi: 10.1093/mnras/stz3157

Sáez-Casares, I., Rasera, Y., & Li, B. 2023. http://arxiv.org/abs/2303.08899

Wong, T.-T., Luk, W.-S., & Heng, P.-A. 1997, Journal of Graphics Tools, 2, 9, doi: 10.1080/10867651.1997.10487471

Yuan, S., Garrison, L. H., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022, MNRAS, 510, 3301, doi: 10.1093/mnras/stab3355

Zheng, Z., & Guo, H. 2016, MNRAS, 458, 4015, doi: 10.1093/mnras/stw523

5

Results from the DESI One-Percent survey

Contents

This chapter presents the small-scale clustering analysis of the ELG sample of the DESI One-Percent survey that I did during my thesis. It was one of the science papers issued together with the Early Data Release of DESI on June 13, 2023 and has been submitted to publication in JCAP (Rocher et al., 2023).

5.1 Introduction

The ELG galaxy-halo connection has been previously studied using different approaches (e.g Avila et al., 2020, Gao et al., 2022, Gonzalez-Perez et al., 2018, Lin et al., 2023, Okumura et al., 2021). From these studies, ELGs are expected to reside in dark matter (DM) halos of mass $\sim 10^{12} M_{\odot}$, and the occupation of DM halos decreases when the halo mass increases. In the literature, a sizeable fraction of ELGs are considered to be satellites. Depending on the galaxy-halo connection model, the satellite fraction varies from $\sim 10\%$ to $\sim 30\%$. The purpose of this chapter is to study the HOD of the DESI ELG sample from the One-Percent survey. Based on previous work (Alam et al., 2021, Avila et al., 2020, Gonzalez-Perez et al., 2018), we study 4 different distributions for central galaxy occupation and allow for different modelling of galaxy satellite velocities. The impact of secondary parameters, such as assembly bias (Gao & White, 2007), based on the halo concentration, local halo density and density anisotropies is also investigated. We also test for departures from a pure NFW profile for satellite positioning. Finally, we study variation of the HOD parameters considering 3 different cosmologies. We use the HOD fitting pipeline based on Gaussian processes described in the previous chapter to derive the best-fitting parameters to DESI ELG data and the corresponding posterior contours.

5.2 ELG data sample

The ELG data sample studied in this chapter was collected during the One-Percent survey of DESI that was conducted at the end of the Survey Validation (SV) campaign in April and May of 2021 (DESI collaboration et al., 2023a) before the start of the main survey operations (see Section 2.6). Before SV, DESI had proven its ability to simultaneously measure spectra at 5000 specific sky locations, with fibres placed accurately using robotic positioners populating the DESI focal plane (Silber et al., 2023). During SV, the DESI data and operation teams proved their ability to optimise operations (E. Schlafly et al., 2023) and to efficiently process the spectra through the DESI spectroscopic pipeline (Guy et al., 2023). To obtain a high completeness, the footprint of the One-Percent survey was defined as a set of 20 non-overlapping regions of the sky, called rosettes in the following, which were observed at least 11 times each. Starting from an initial target list (Myers et al., 2023), the DESI fibre assignment algorithm (DESI Collaboration et al., 2022) places each fibre onto a reachable target within a 6 mm patrol radius around the nominal fibre position, so that only a subset of the targets can be observed in every visit. This leads to incompleteness, which decreases rapidly with the number of visits.

The One-Percent survey covered 140 deg^2 with final target selection algorithms and depths similar to those of the main survey. The ELG target selection (Raichoor et al., 2023) focuses on the redshift range $0.6 < z < 1.6$ and is designed to select galaxies with strong spectral emission lines. The [O II] doublet emission line allows precise redshifts to be measured by DESI. Higher priority in the spectroscopic measurements is given to objects expected in the interval $1.1 < z < 1.6$ where ELGs are the main tracer of DESI. Between 0.2 and 1.5 degrees from the centre of each rosette, spectra were

successfully obtained for 94.5% of ELG targets, while targets outside these regions were observed with fewer visits and thus lower completeness in fibre assignment. This sample is very appropriate to study ELGs inside halos as it provides precise measurements of the galaxy clustering down to very small scales.

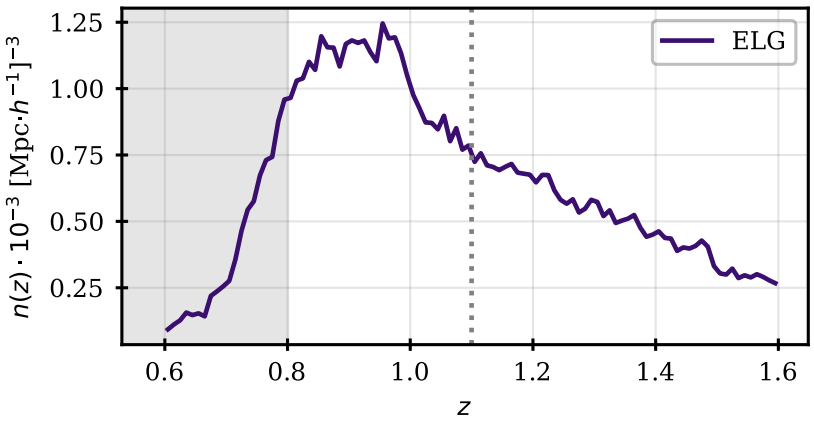

Figure 5.1: *Number density of the DESI One-Percent survey ELG data sample, as a function of redshift (corrected for completeness). The shaded region is not used in this work. The dotted line indicates the mean completeness-weighted redshift of the sample,* $\bar{z} = 1.13$.

In the following, we make use of the ELG sample collected during the One-Percent survey and spectroscopically confirmed in the redshift range from 0.8 to 1.6, over which the radial density distribution varies slowly, as shown in Figure 5.1. The region $z < 0.8$ is not considered in our final sample as it exhibits dependence of the redshift density with respect to imaging depth. This sample contains 244k spectra and has a mean density of $7 \times 10^{-4}(h/\mathrm{Mpc})^3$. Section 4 of DESI collaboration et al. (2023b) describes the construction of all EDR large scale structure catalogues, including the random catalogues. We highlight a few details on the random construction here. Random catalogues are first produced by the DESI targeting team (Myers et al., 2023) at a fixed density. These randoms are input to the DESI fibre assignment software, which processes them through each observed tile matching the state used during observations. Only the randoms that were identified as observable by this process are kept. Additional vetoes are applied for bright stars and other foregrounds (see DESI collaboration et al. (2023b) for the precise details). Therefore, the density and radial distribution of each rosette are the same (within only Poisson fluctuations). Finally, redshifts are assigned randomly in the random catalogue using the redshifts from the galaxy sample to keep the same $n(z)$ distribution.

5.2.1 Clustering statistics

The clustering of the selected sample was studied in configuration space with the 2-point statistics defined in the previous chapter, which we recall hereafter. We first define the galaxy two-point correlation function in two dimensions, $\xi(r_p, \pi)$, where π and r_p are the galaxy pair separation components along and perpendicular to the line-of-sight, respectively. We then introduce the projected correlation function, $w_p(r_p)$, obtained by integrating $\xi(r_p, \pi)$ over the line-of-sight, as well as the monopole and quadrupole of the two point correlation function $\xi(s, \mu)$, where s is the galaxy pair separation and μ the cosine of the angle between the line-of-sight and galaxy separation vector:

$$\begin{aligned} w_p(r_p) &= \int_{\pi_{min}}^{\pi_{max}} \xi(r_p, \pi) d\pi \\ \xi_l(s) &= \frac{2l+1}{2} \int_{-1}^{1} \xi(s, \mu) \mathcal{L}_l(\mu) d\mu \end{aligned} \tag{5.1}$$

where $l = [0, 2]$ and $\mathcal{L}_l(\mu)$ is the Legendre polynomial of order l.

We rely on PYCORR[1], the DESI implementation of the CORRFUNC package (Sinha & Garrison, 2020), to compute $\xi(r_p, \pi)$ and $\xi(s, \mu)$. For mocks, which are obtained from cubic boxes, they are computed with the natural estimator which compares galaxy pair counts to the expected pair count for a uniform distribution in the box volume. For data, the Landy-Szalay estimator is used (Landy & Szalay, 1993). For mocks, the z axis is chosen as line-of-sight for the application of redshift space distortions.

For $\xi(r_p, \pi)$, we use 17 logarithmic bins in r_p between 0.04 and 32 Mpc/h and 80 linear bins in π between -40 and 40 Mpc/h. The same binning and range are used for $w_p(r_p)$ so that $\pi_{max} = -\pi_{min} = 40\,\mathrm{Mpc}/h$ in Equation (5.1). For the multipoles, we use 27 logarithmic bins in s between 0.17 and 32 Mpc/h and 200 linear bins in μ between -1 and 1. Finally, in the galaxy pair count computation, whether in data or simulation, the fiducial cosmology used to convert galaxy redshift into distances is the Planck 2018 baseline Λ_{CDM} best-fit result (Planck Collaboration et al., 2020) with $h = 0.6736, A_s = 2.0830 \times 10^{-9}, n_s = 0.9649, \omega_{cdm} = 0.12, \omega_b = 0.02237$ and $\sigma_8 = 0.8079$.

5.2.2 Clustering measurements

The clustering of the One-Percent survey ELG sample is first illustrated in Figure 5.2 which shows the 2D correlation function in successive bins in π, as a function of r_p. This figure highlights several key points about the ELG clustering measurement from the One-Percent survey. A strong signal at small scales is visible at separations larger than $r_p = 0.03\mathrm{Mpc}/h$, the threshold below which target blending makes clustering measurements unreliable (see Section 2.9). The strong up-turn in the small-scale clustering appears for transverse separations below $r_p \sim 0.2\mathrm{Mpc}/h$ and is mostly due to separations along the line-of-sight below $\pi = 3\mathrm{Mpc}/h$. In this region, the incompleteness of the survey may bias the clustering measurements due to fibre collisions if the number of visits is limited. To illustrate this, measurements corresponding to the complete survey footprint (solid lines) are compared with those excluding regions of lower completeness, outside the interval between 0.2 and 1.5 degrees from the field centre of each rosette (dashed lines). The strong up-turn in the clustering signal appears also in the latter measurements, showing that incompleteness due to fibre collisions is not responsible for the strong ELG clustering at small scales that we observe.

Incompleteness can however bias the clustering measurements, especially at small scales. To limit that effect, we restrict the ELG sample to those targets observed in regions of high completeness, that is between 0.2 and 1.5 degrees from the field centre of each rosette. This reduces the sample size by 12% leaving 215k galaxies. Residual density inhomogeneities in that sample due to residual fibre assignment inefficiencies are corrected with a weighting procedure. This is illustrated in Figure 5.3 which shows how the clustering changes when fibre-assignment corrections are applied. These corrections are twofold. Incompleteness weights for individual galaxies and for galaxy pairs are computed as inverse probabilities of being targeted in a set of multiple realisations of the actual fibre assignment algorithm, as described in Bianchi & Percival (2017). These weights are complemented by angular up-weighting to treat the case of galaxy pairs with zero selection probability in the previous computation, as described in Percival & Bianchi (2017). Mohammad et al. (2020) showed that this weighting scheme provides an unbiased clustering down to $\sim 0.1\mathrm{Mpc}/h$. As anticipated from the removal of the regions of lower completeness in the rosettes, the fibre-assignment weights have a small impact on the measured clustering, visible essentially at scales lower than $\sim 1\mathrm{Mpc}/h$, as a result of fibre collisions.

The clustering measurements can be biased by other systematic effects, such as density inhomogeneities due to imaging conditions or redshift failure rate variations with spectroscopic observing conditions. We checked that the correcting weights associated with these effects have a negligible impact on the small-scale clustering measurements. Besides the completeness weights, we also apply Feldman-Kaiser-Peacock (FKP) weights (Feldman et al., 1994) that minimise variance in the clustering measurements (evaluated with $k_0 \sim 0.15$ and $P_0 = 4000$). We also checked with simulations that the small footprint of

[1] https://github.com/cosmodesi/pycorr

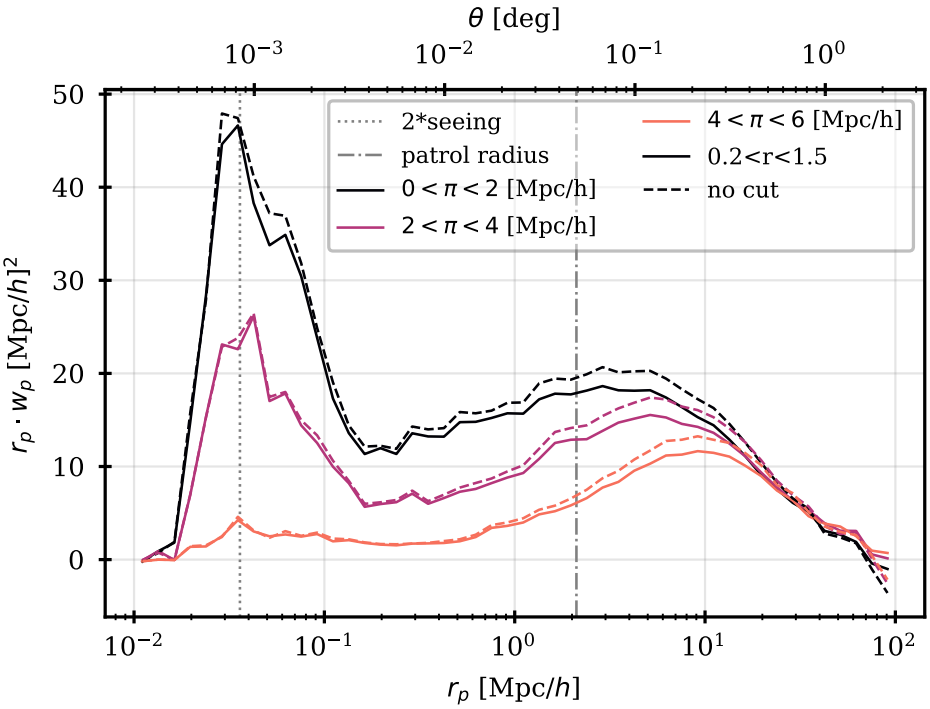

Figure 5.2: *DESI clustering measurements for the One-Percent survey ELG data sample restricted to the redshift range* $0.8 < z < 1.6$. *The 2D correlation function in successive bins of* $2\mathrm{Mpc}/h$ *in the galaxy-pair separation along the line-of-sight is shown as a function of the separation perpendicular to the line-of-sight,* r_p. *No correction weight has been applied. Measurements using the whole survey footprint (solid lines) are compared with measurements excluding the inner and outer regions of the rosettes where the survey was less incomplete (dashed lines). Also indicated are the separation corresponding to the fibre patrol radius (dot-dashed grey line) and the limit corresponding to twice the mean survey seeing (dotted grey line). Below this limit, target blending cannot be resolved, leading to a loss of power. This plot demonstrates that the strong increase in power at small scales (below* $0.2\mathrm{Mpc}/h$*) is not due to the (slight) incompleteness of the One-Percent survey.*

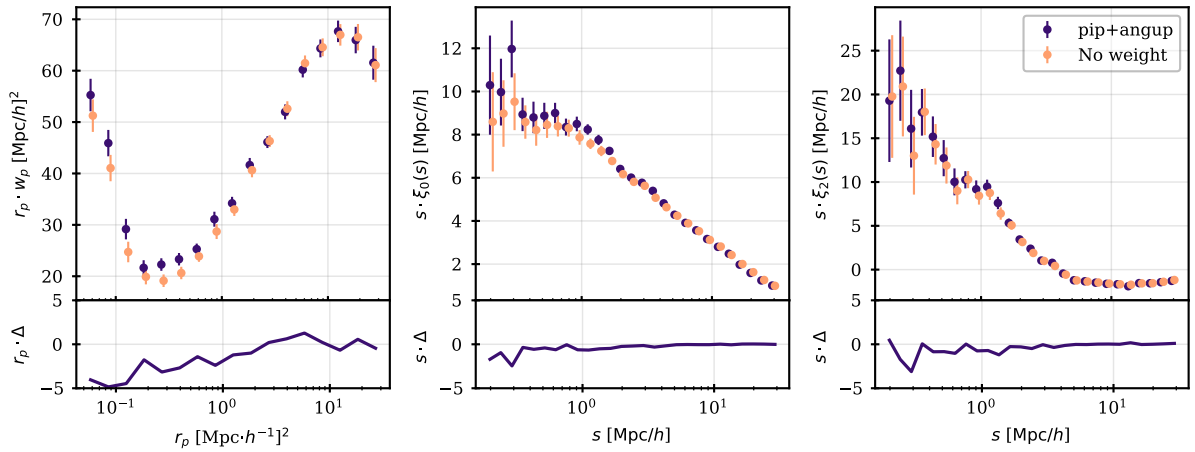

Figure 5.3: *Top: DESI clustering measurements for the One-Percent survey ELG data sample restricted to the redshift range* $0.8 < z < 1.6$ *and to regions of high completeness. Data are shown without (orange) and with (purple) tiling incompleteness weights. Errors are jackknife statistical uncertainties. Bottom: difference between clustering measurements without and with fibre assignment weights applied.*

the One-Percent Survey has a negligible integral constraint effect (de Mattia & Ruhlmann-Kleider, 2019) on these measurements in the range of separations used in this study.

Several observational studies of ELG clustering at redshifts $z \sim 1$ have already been published, using data from various surveys, such as COSMOS (Tinker et al., 2013), VIPERS (e.g Favole et al., 2016, Gao et al., 2022), eBOSS (Alam et al., 2021, Avila et al., 2020, Guo et al., 2019, Lin et al., 2023) or the HSC SSP survey (Okumura et al., 2021). But the clustering measurements provided by the DESI One-Percent survey are the first redshift space measurements that go down to transverse separation scales as low as 0.03Mpc/h, offering a direct and robust measurement of the one-halo term contribution to the clustering.

In the following, we first use the clustering measurements in the redshift range $0.8 < z < 1.6$ to test different prescriptions for the ELG HOD modelling and then fit the most promising one to the measurements, splitting the sample in two redshift intervals in Section 5.8 to test for a possible HOD parameter evolution with redshift.

5.3 Standard ELG HOD models

The standard HOD formalism describes the relation between galaxies and their dark matter halos as the probability that a halo with mass M hosts N such galaxies. Central and satellite galaxies are considered separately, with $\langle N_{cent}(M)\rangle$ and $\langle N_{sat}(M)\rangle$ their respective mean numbers per halo of a given halo mass.

5.3.1 Models for central galaxies

Based on previous studies of ELG clustering (Alam et al., 2021, Avila et al., 2020), we retain four possible HOD prescriptions for central galaxies, one with a Gaussian shape and three different functions producing an asymmetric shape (see also Section 3.4.1):

- a Gaussian HOD model (GHOD):

$$\langle N_{cent}(M)\rangle = \frac{A_c}{\sqrt{2\pi}\sigma_{\rm M}} \cdot e^{-\frac{(\log_{10} M - \log_{10} M_c)^2}{2\sigma_{\rm M}^2}} \equiv \left\langle N_{cent}^{GHOD}(M)\right\rangle \tag{5.2}$$

- a LogNormal HOD model (LNHOD): defining $x = \log_{10} M - (\log_{10} M_c - 1)$, the prescription for central galaxies is:

$$\langle N_{cent}(M)\rangle = \frac{A_c}{\sqrt{2\pi}\sigma_{\rm M} \cdot x} \cdot e^{-\frac{(\ln x)^2}{2\sigma_{\rm M}^2}} \quad \text{for } x > 0, \text{ and 0 otherwise} \tag{5.3}$$

- a Star Forming HOD model (SFHOD):

$$\langle N_{cent}(M)\rangle = \begin{cases} \left\langle N_{cent}^{GHOD}(M)\right\rangle & M \leq M_c \\ \frac{A_c}{\sqrt{2\pi}\sigma_{\rm M}} \cdot \left(\frac{M}{M_c}\right)^{\gamma} & M > M_c \end{cases} \tag{5.4}$$

- a modified High Mass Quenched model (mHMQ):

$$\langle N_{cent}(M)\rangle = \left\langle N_{cent}^{GHOD}(M)\right\rangle \cdot \left[1 + \text{erf}\left(\frac{\gamma(\log_{10} M - \log_{10} M_c)}{\sqrt{2}\sigma_{\rm M}}\right)\right] \tag{5.5}$$

  Note that this model is derived from the High Mass Quenched model of Alam et al. (2021) setting the quenching factor to infinity to only retain the asymmetric shape of the central distribution.

In the above formulas, A_c sets the size of the central galaxy sample, M_c is the characteristic mass for a halo to host a central galaxy, $\sigma_{\rm M}$ is the width of the distribution and γ, if present, controls its asymmetry. A Bernoulli distribution with mean equal to $\langle N_{\rm cent}(M)\rangle$ is used to generate either 0 or 1 central galaxy per halo.

5.3.2 Baseline model for satellite galaxies

For satellite galaxies, we adopt the same HOD as in Section 3.4.1 (Alam et al., 2021, Avila et al., 2020):

$$\langle N_{sat}(M)\rangle = A_s\left(\frac{M-M_0}{M_1}\right)^{\alpha} \tag{5.6}$$

where A_s sets the size of the satellite galaxy sample, M_0 is the cut-off halo mass from which satellites can be present and α controls the increase in satellite richness with increasing halo mass. M_1 is introduced for normalisation purpose and corresponds to the halo mass at which 1 satellite is expected if $A_s = 1$ and M_0 is negligible w.r.t. M_1. The above form (without the normalisation factor A_s) was first introduced in Kravtsov et al. (2004) based on N-body simulations and in Zheng et al. (2005) based on semi-analytical models and hydrodynamical simulations of galaxy formation, for it was found to provide a very good description of the occupation distribution of satellites predicted in these frameworks. Note that when $\alpha \sim 1$, this form gives a mean number of satellites which simply traces the halo mass. The normalisation factor A_s was introduced in later works as a way to model the incompleteness of the satellite sample. In this analysis, we use both A_c and A_s to impose a density constraint to our HOD models, as explained in the next section.

Throughout the chapter, unless stated otherwise, the actual number of satellite galaxies as a function of halo mass is drawn from a Poisson distribution with mean equal to $\langle N_{sat}(M)\rangle$. By default, several satellites can thus be present in the same halo, and satellites can be present even if there is no central galaxy in the halo. We note that in such a case, classifying them as satellites may appear inappropriate, but is no more than a convenience to refer to the parametrisation used. Beyond the above functional forms for the mean numbers of central and satellite galaxies as a function of halo mass, a prescription must be chosen to define how satellite positions and velocities are distributed. This is described in Section 5.5.

From the above equations, derived parameters can be calculated analytically, such as the expected total number density of the galaxy sample:

$$\bar{n}_{\rm gal} = \int \frac{{\rm d}n(M)}{{\rm d}M}\left[\langle N_{cent}(M)\rangle + \langle N_{sat}(M)\rangle\right]{\rm d}M \tag{5.7}$$

the fraction of satellites:

$$f_{sat} = \frac{1}{\bar{n}_{\rm gal}}\int \frac{{\rm d}n(M)}{{\rm d}M}\langle N_{sat}(M)\rangle\,{\rm d}M \tag{5.8}$$

or the average halo mass of the sample:

$$\langle M_h\rangle = \frac{1}{\bar{n}_{\rm gal}}\int \frac{{\rm d}n(M)}{{\rm d}M}\left[\langle N_{cent}(M)\rangle + \langle N_{sat}(M)\rangle\right]M{\rm d}M \tag{5.9}$$

where $\frac{{\rm d}n(M)}{{\rm d}M}$ is the halo mass function, taken from the N-body simulation. We also define an effective M_1' mass parameter that is equivalent to the M_1 mass scale in the original parametrisation for satellite occupation without the A_s parameter:

$$M_1' \equiv \frac{M_1}{A_s^{1/\alpha}} \tag{5.10}$$

M_1' is the halo mass scale to have one satellite on average if M_0 is negligible w.r.t. M_1'.

5.3.3 HOD free parameters and density constraint

The HOD parameters are $A_c, M_c, \sigma_{\rm M}$ (and possibly γ) for central galaxies and A_s, M_0, α, M_1 for satellite galaxies. M_1 being degenerate with A_s and α cannot be constrained in the fits. Unless otherwise stated, it is fixed to a value of $10^{13}M_\odot/h$ in the fits described in this chapter. The normalisation parameters A_c and A_s are used to impose a density constraint in the fitting procedure to match the density in DESI data, as explained below. All other parameters are left free to vary.

The galaxy sample number density in Equation (5.7) is governed by both A_c and A_s and the fraction of satellites in Equation (5.8) is controlled by their ratio. All other conditions being equal, the same clustering is obtained whatever A_c and A_s values, provided their ratio is fixed. The density constraint is introduced in the following way. At each point in the HOD parameter space, we set A_c to an initial value, while A_s is sampled from a flat prior range. We compute the total number density in Equation (5.7) for these initial values of A_c and A_s and rescale them by the same factor (to preserve the clustering) in order to normalize the galaxy density to $10^{-3}(h/\mathrm{Mpc})^3$, close to that of the DESI ELG sample. In our tables, we report A_c initial values, best-fit values of A_s which are unrescaled and we provide the corresponding rescaling factor used to set the density of the mocks to that of data. This factor is applied for the derived parameters and for mock creation.

5.4 Simulation

As in Chapter 4, mock catalogues generated from simulations according to the above HOD models are based on the AbacusSummit suite of high-accuracy cosmological N-body simulations (Maksimova et al., 2021) designed for the clustering analyses of DESI. We use the cleaned halo catalogues obtained with the CompaSO algorithm (Hadzhiyska et al., 2022a) applied to these simulations. The suite is defined primarily in the base Planck 2018 ΛCDM best-fit cosmology (Planck Collaboration et al., 2020) but contains also several variants, and proposes different resolutions and cubic box sizes.

usage	cosmology	box size	resolution	realisations
baseline modelling	Planck 2018 Λ_{CDM}	1.185 Gpc/h	4096^3	1
correlation matrix	Planck 2018 Λ_{CDM}	0.5 Gpc/h	1728^3	1800
cosmic variance	Planck 2018 Λ_{CDM}	2 Gpc/h	6912^3	25
high N_{eff}	$N_{\mathrm{eff}} = 3.7$	1.185 Gpc/h	4096^3	1
high N_{eff} cosmic variance	$N_{\mathrm{eff}} = 3.7$	2 Gpc/h	6912^3	6
low σ_8	Planck 2018 with $\sigma_8 = 0.75$	1.185 Gpc/h	4096^3	1
low σ_8 cosmic variance	Planck 2018 with $\sigma_8 = 0.75$	2 Gpc/h	6912^3	6

Table 5.1: *Cosmology, box size and mass resolution of the AbacusSummit simulations used in this work. The mass resolution is given as the number of particles in the box. The first column indicates the use of each set of simulations: baseline HOD modelling, correlation matrix for data, cosmic variance for the model covariance matrix. The last four sets are used to explore different cosmologies but with identical simulation initial conditions as in the baseline model.*

Table 5.1 presents the subset of simulations used in this analysis. They all have the same resolution, that is 6912^3 particles in a box of 2 Gpc/h length, which corresponds to a particle mass of about $2{\times}10^9 M_\odot/h$. This ensures that halos are well resolved down to $10^{11} M_\odot/h$ giving $\sim$50 particles/halo (Maksimova et al., 2021). Besides, the halos corresponding to best fitting results obtained in this work have a mass larger than $3 \times 10^{11} M_\odot/h$ which corresponds to 150 particles/halo. Note that throughout the chapter, we define the halo mass as the number of particles in the halo multiplied by the particle mass.

5.5 Fitting Methodology

The HOD fitting pipeline used in this work is that described in Chapter 4. It proceeds in two steps, HOD mock generation and HOD parameter fitting, based on Gaussian processes. The main features of the pipeline are summarised hereafter, with more details than in Chapter 4 about the prescription

used for satellite positioning, and the construction of the covariance matrix used in the fits is described afterwards.

5.5.1 Pipeline based on Gaussian processes

The fitting pipeline uses Gaussian processes (GP) to obtain a surrogate model of the likelihood surface describing the comparison of clustering measurements between data and HOD mocks. At each point of the HOD parameter space of a given model, DM halos from the baseline AbacusSummit 1.18 Gpc/h cubic box (see Table 5.1) are populated with galaxies according to the HOD parameters, in order to generate mock catalogues. Mocks are created with a fixed galaxy density of $10^{-3}(h/\mathrm{Mpc})^3$, close to that of the actual ELG sample. In this process, the HOD prescriptions for the mean numbers of central and satellite galaxies described in Section 5.3 are complemented by the following assumptions. Central galaxies are positioned at the centre of their halos. Satellite positions obey a Navarro-Frenk-White profile (Navarro et al., 1996) using r_{25}, the radius of a sphere that contains 25% of the halo particles, as a proxy for r_s, the scale radius of the profile. Since the mass enclosed in a sphere of radius r is divergent for the NFW profile, we further apply a cut-off at the halo virial radius $r = r_{\mathrm{vir}}$, taking r_{98}, the radius of a sphere containing 98% of the halo particles, as a proxy for r_{vir}. The above proxies were chosen because we observed that they provide a predicted clustering which is very close to that obtained with satellite positioning using DM particles, as illustrated in Appendix A. Satellite velocities are normally distributed around their mean halo velocity, with a dispersion equal to that of the halo dark matter particle velocities, rescaled by an extra free parameter denoted f_{σ_v}, following Alam et al. (2021).

At each point of the parameter space, we generate 20 mock catalogues and compare their clustering to that of data to produce one χ^2 value per mock. The 20 χ^2 values are then averaged and both the mean χ^2 value and the standard deviation of the mean are fed into the GP. The covariance matrix entering the χ^2 definition contains a data component and a model component that accounts for the stochastic noise of the mock creation and the cosmic variance to be expected for 1.18 Gpc/h cubic boxes. These two covariance matrices are discussed further in Section 5.5.2. In the χ^2 computation, each of the covariance matrix components is corrected for the Hartlap effect (Hartlap et al., 2007). The inverse covariance matrix is biased because of the number of mocks used to estimate it is finite. The amount of bias depends on the ratio between the number of measurements and the number of mocks, the more mocks the less bias for a given set of measurements.

Initial training of the GP is obtained from the χ^2 values and errors computed on a given set of points. Based on the conclusions of Chapter 4, the training sample is obtained from Hammersley sampling of the HOD parameter space using flat priors. After initial training, the GP model of the likelihood surface is further improved by an iterative procedure adding one point to the training sample at each iteration. The added point is randomly chosen in Monte Carlo Markov chains (MCMC) sampling the GP prediction. This allows us to obtain both an accurate minimisation of the χ^2 and reliable error contours in the HOD parameter space.

In the following, we use an initial training sample of 800 points from Hammersley sampling, followed by 800 iterations and check the fit convergence during the iterative step by means of the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) between the MCMC chains. As shown in Chapter 4, instabilities in the GP likelihood surface estimate can be generated in the course of the iterative procedure, due to learning phases triggered in small regions of the parameter space by the addition of the extra point. These instabilities do not strongly impact the iterative evolution of the marginalized parameter values but affect the KL divergence. To check the fit convergence, we thus compute the KL divergence from cleaned MCMC chains, where points with uncertainties above 10 in the χ^2 value predicted by the GP have been removed. We consider a fit to be converged when the KL divergence is below 0.1 in a set of 20 consecutive iterations and we define the final iteration as the last iteration of the last such set of iterations. When no such set of consecutive iterations is found, the final iteration is the last iteration with a KL divergence below 0.1. The fit results are defined by the marginalized HOD parameter values

at that final iteration, with statistical uncertainties given by the $[0.16 - 0.84]$ quantiles of the parameter posteriors at that same iteration.

As 2-point statistics, the GP pipeline uses the projected correlation function, $w_p(r_p)$, as well as the monopole $\xi_0(s)$ and quadrupole $\xi_2(s)$ of the two point correlation function, as introduced in Section 5.2.1.

5.5.2 Covariance matrix for data and model

A data covariance matrix appropriate for the ELG clustering measurements used by the GP pipeline was derived applying the delete-one Jackknife method to the One-Percent survey footprint divided into 128 independent regions, the maximum number of large enough regions given the small extent of the footprint. The jackknife regions were defined using a K-means sampler that cuts the footprint into regions of similar size in RA/DEC, as implemented in the DESI package `pycorr`. To recover an unbiased estimate of the covariance matrix, correction terms were applied as described in Mohammad & Percival (2022). As the off-diagonal terms of that matrix are affected by noise, a smooth correlation matrix was derived from simulations to replace that from data. For that purpose, we resorted to the 1800 small boxes from the

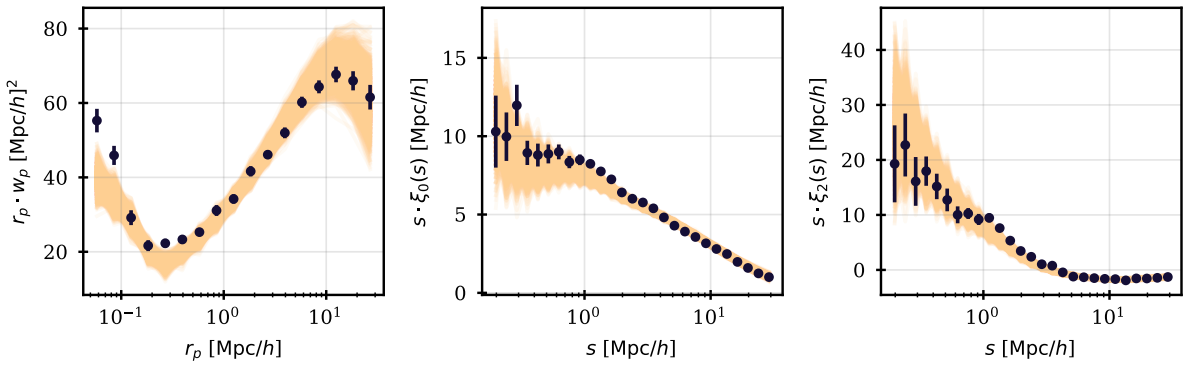

Figure 5.4: *DESI ELG clustering measurements from the One-Percent survey data sample. From left to right, we show the projected correlation function, the monopole and quadrupole of the correlation function. Data (dots with error bars) are compared to expectations (solid lines) from 1800 realisations of a HOD model obtained from a preliminary fit to these data using a pure Jackknife covariance matrix. Uncertainties are Jackknife errors.*

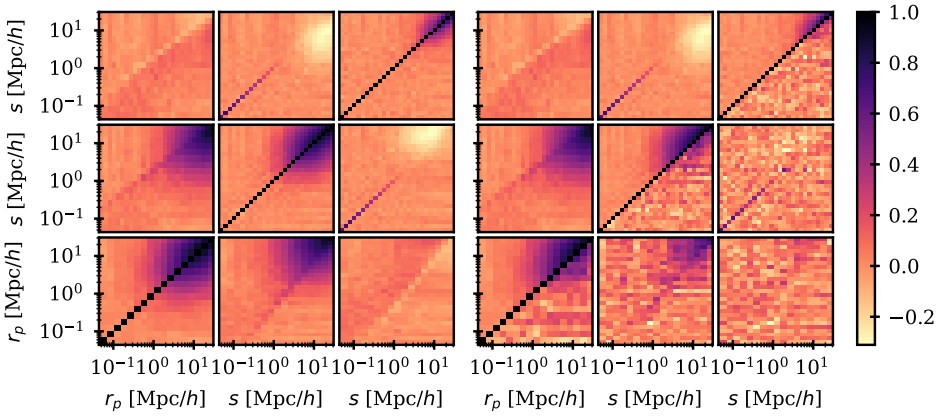

Figure 5.5: *Left: correlation matrix derived from 1800 mocks built from the HOD model in Figure 5.4. Right: correlations from the mock-based matrix (above the diagonal) compared with those from the pure Jackknife covariance matrix (below the diagonal).*

AbacusSummit simulations in Table 5.1 that allow cosmic variance to be included with good statistical precision.

The result of a preliminary HOD fit to the data using the Jackknife covariance matrix was first used to populate the small box halos, with a density identical to that in the data sample. Figure 5.4 compares the clustering from the data with the calculated clustering for the mocks used to determine the correlation matrix. The binning of the three statistics is that defined in Section 5.2.1. The off-diagonal terms of the mock-based correlation matrix are much smoother than the ones calculated from the data as shown in Figure 5.5. In the following, we define the data covariance matrix of the HOD fits, C_{data} from the mock-based correlation matrix, using the Jackknife diagonal errors to appropriately normalise variances and covariances.

The GP pipeline also considers a covariance matrix for the model, C_{model}, as described in Chapter 4. To build it, correlations are assumed to have small variations over the HOD parameter space and we first compute a fixed correlation matrix from 1000 realisations of the HOD model under test, at a given reference point in the parameter space. When scanning the HOD parameter space, the model covariance matrix at each point is then obtained by normalising all terms of the previous correlation matrix by the quadratic sum of two sets of diagonal errors. The first set contains the variances of the clustering measurements over the 20 realisations drawn to compute the χ^2 at the current point to take into account the stochasticity of the model, and thus accounts for stochastic noise. The second set contains the variances of the clustering measurements obtained from 48 realisations of the HOD model at the reference point, each drawn from a different sub-cube of 1 Gpc/h length cut out of 25 realisations of the same simulation box (see Table 5.1, third line) and corrected for volume effects to take into account the cosmic variance in our errors.

In the GP pipeline, the χ^2 computed at each point of the HOD parameter space is thus defined as follows:

$$\chi^2 = (\xi_{data} - \xi_{model})^\top \left[C_{data}/(1 - D_{data}) + C_{model}/(1 - D_{model})\right]^{-1} (\xi_{data} - \xi_{model}) \tag{5.11}$$

where ξ is a vector of clustering measurements, C the corresponding covariance matrix and D the Hartlap correction factor (Hartlap et al., 2007) based on the number of mocks used to derive the corresponding correlation matrix. This is averaged over 20 HOD realisations.

5.6 Standard HOD results

Best fitting clustering from the GP pipeline are presented in Figure 5.6 and the corresponding best-fit values of the HOD parameters are summarised in Table 5.2. We test the four prescriptions for central galaxies of Section 5.3, keeping the standard prescription for satellites (see Equation (5.6) and Section 5.5). Except for M_1 which is kept fixed in the fits, we used flat priors for all other HOD parameters. The exact prior ranges depend on the HOD models tested but we took care to choose them wide enough to get reasonably enclosed contours. Examples of prior ranges are shown in Appendix B.

The four best fitting models provide similar expectations for the ELG clustering, which agree reasonably well with data. Features difficult to model correctly are the slope of the projected correlation function between 0.2 and 10 Mpc/h and the bump at $s \sim 1-2$ Mpc/h in the monopole and quadrupole. This partially explains the high χ^2 values which average at $\sim$157 for 65 degrees of freedom, depending on the model. Since all models behave similarly, it implies that there are ingredients missing in the standard HODs for ELGs. This will be studied in the following sections.

Also shown in Figure 5.6 is the expected clustering computed from halos only, regardless of the galaxies they contain (dashed line). This highlights the fact that pairs of galaxies inside the same halo contribute, as expected, only at low scales in the three statistics. This contribution constitutes the so-called one-halo term of the galaxy-halo connection and is essential to reproduce the strong clustering measured at small scales in our data, notably the strong up-turn of the projected correlation function at $r_p < 0.3$ Mpc/h. Note however that between 0.3 and $\sim$1 Mpc/h in r_p, the measured clustering is above

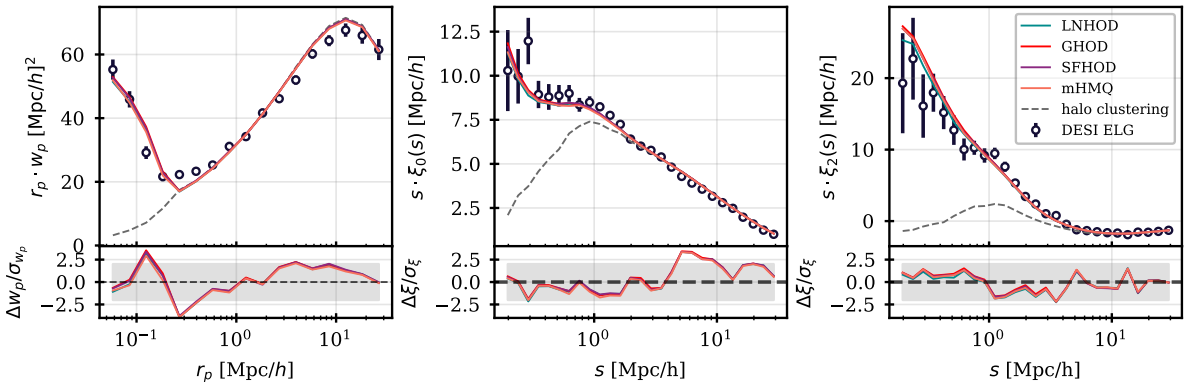

Figure 5.6: *Top: DESI ELG clustering measurements from the One-Percent survey data sample, compared to best fitting standard HOD models obtained with the GP pipeline. The models in solid line correspond to different prescriptions for the central galaxies, keeping the standard power-law prescription for satellites. The model in dashed line is the pure halo clustering, showing that pairs of galaxies from the one-halo term have a strong impact on the clustering at the lowest scales. Errors are Jackknife uncertainties only. Bottom: Fit residuals normalised by the diagonal errors of the full covariance matrix, that comprise Jackknife uncertainties for the data as well as stochastic noise and cosmic variance for the model, but no Hartlap factor corrections.*

parameter & χ^2	LNHOD	GHOD	SFHOD	mHMQ
A_c (resc.)	1 (0.08)	1 (0.08)	1 (0.08)	1 (0.08)
$\log_{10} M_0$	$11.78^{+0.04}_{-0.05}$	$11.72^{+0.03}_{-0.04}$	$11.73^{+0.03}_{-0.03}$	$11.70^{+0.03}_{-0.03}$
A_s	$0.09^{+0.01}_{-0.01}$	$0.08^{+0.04}_{-0.01}$	$0.09^{+0.04}_{-0.02}$	$0.10^{+0.04}_{-0.03}$
$\log_{10} M_c$	$11.87^{+0.01}_{-0.01}$	$11.89^{+0.02}_{-0.02}$	$11.87^{+0.03}_{-0.03}$	$11.72^{+0.06}_{-0.04}$
α	$-0.28^{+0.03}_{-0.03}$	$-0.31^{+0.08}_{-0.05}$	$-0.28^{+0.06}_{-0.04}$	$-0.26^{+0.08}_{-0.08}$
f_{σ_v}	$1.29^{+0.07}_{-0.06}$	$1.23^{+0.06}_{-0.06}$	$1.27^{+0.07}_{-0.07}$	$1.27^{+0.07}_{-0.06}$
σ_M	$0.08^{+0.02}_{-0.01}$	$0.11^{+0.02}_{-0.02}$	$0.07^{+0.04}_{-0.02}$	$0.22^{+0.08}_{-0.11}$
γ	-	-	$-4.42^{+0.99}_{-0.76}$	$7.06^{+1.33}_{-1.97}$
$\log_{10} M_1'$	5.37	6.12	5.57	4.77
f_{sat}	$0.10^{+0.02}_{-0.02}$	$0.12^{+0.03}_{-0.02}$	$0.11^{+0.02}_{-0.02}$	$0.12^{+0.02}_{-0.02}$
f_{1h}	$0.041^{+0.007}_{-0.005}$	$0.040^{+0.005}_{-0.006}$	$0.039^{+0.005}_{-0.006}$	$0.039^{+0.009}_{-0.008}$
$\log_{10} \langle M_h \rangle$	$11.87^{+0.01}_{-0.01}$	$11.87^{+0.01}_{-0.01}$	$11.88^{+0.01}_{-0.01}$	$11.87^{+0.02}_{-0.01}$
χ^2 (ndf)	156.0 ± 1.0 (65)	157.6 ± 1.3 (65)	155.5 ± 1.2 (64)	158.2 ± 1.0 (64)

Table 5.2: *Results of standard HOD fits to the DESI ELG clustering measurements from the One-Percent survey. The first line provides the initial fixed value of A_c and the rescaling factor applied to impose the density constraint in the fits. The following six or seven parameters are the free HOD parameters, the next four are derived parameters. $\log_{10} M_1'$ is given for best-fit values of α and A_s (the latter after rescaling). f_{sat} is the fraction of galaxies which are satellites and f_{1h} is the fraction of galaxies which are not alone in their halos. All masses are in units of $(M_\odot/h)$.*

the predicted halo clustering, meaning that the one-halo contribution arising from the NFW profile is not sufficient to describe the data in this region.

These results are further illustrated in Table 5.2, which provides the best-fit values of the model parameters, and in Figure 5.7, which shows the four best fitting HOD models and the distributions of the number of galaxies per halo mass bin for halos populated according to these HOD models. The four models exhibit similar features. The HOD for centrals peak at a mass slightly below $10^{12} M_{\odot}/h$ and span a short interval of halo masses as shown by the low values of σ_M. The minimal mass to populate halos with satellites is slightly below M_c and the satellite HOD has a negative power-law index. Both features reflect the need to have close pairs of galaxies in low mass halos in order to reproduce the ELG clustering at small scales, and translate into a one-halo component of the distribution of the number of galaxies per populated halo mass bin that peaks at low halo mass, as shown in the right-hand plot in Figure 5.7. The mean halo mass of the galaxy sample and the satellite fraction are analytically calculated using Equation (5.8) and Equation (5.9) from the [16-84] quantiles of the best-fit Markov chains. The fraction of galaxies which are not alone in their halos, f_{1h}, is found to be about 4% in the four models tested. This fraction is computed numerically from 50 mocks generated from random HOD parameters drawn from the 1σ errors of the best fitting HOD models. Note that the value of f_{1h} depends on the number density of the mocks, which is constrained to be that of our data sample $\sim 10^{-3} (h/\mathrm{Mpc})^3$.

Previous small-scale clustering studies of ELG samples at redshifts ~ 1 were performed in different frameworks, either HOD (e.g Avila et al., 2020, Okumura et al., 2021, Tinker et al., 2013), Abundance Matching (e.g Favole et al., 2016, Gao et al., 2022, Lin et al., 2023) or conditional stellar mass function method (Guo et al., 2019). They find consistent results about the mean mass of halos hosting such galaxies, $\log_{10} \langle M_h \rangle \sim 12$. They reported satellite fractions ranging from 13 to 22% for standard HOD prescriptions but extended ones can increase significantly these numbers (Avila et al., 2020) showing that the satellite fraction does not provide a robust way to make precise comparisons between different analyses. For these two parameters, our findings are similar, namely $\log_{10} \langle M_h \rangle \sim 11.9$ and $f_{sat} \sim 12\%$.

All previous HOD studies also reported a satellite HOD that increases at high halo mass (Avila et al., 2020, Gao et al., 2022, Lin et al., 2023, Okumura et al., 2021), or possibly becomes uniform (Guo et al., 2019), while we find a significant decrease (see Figure 5.7). This decrease is also responsible for the

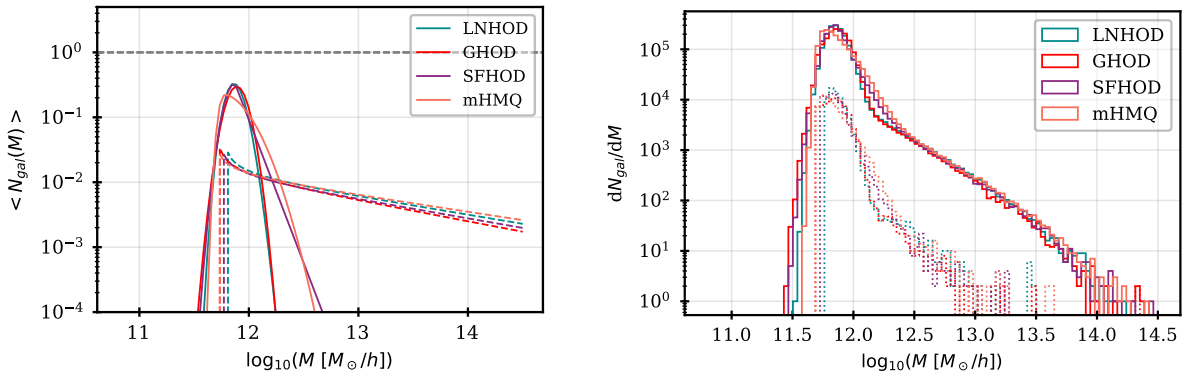

Figure 5.7: *Left: Best fitting HOD models to the DESI One-Percent ELG sample obtained with standard prescriptions for central (solid lines) and satellite (dashed lines) galaxies. We recall that satellites can populate halos even if no central galaxy is present. Four models for central galaxies were used and give similar results. Most noticeably, the satellite average number decreases with increasing halo mass. Right: Number of galaxies per halo mass bin for halos populated according to the four HOD models on the left. The simulation box volume is* $1.66\,(Gpc/h)^3$. *The full distributions are in solid lines. The dashed lines show the contribution of halos hosting more than one galaxy, that is the one-halo component of the full distributions. The four prescriptions for central galaxies lead to similar results, both for the full distribution or for its one-halo component.*

meaningless values of the effective $\log_{10} M_1'$ parameter reported in Table 5.2, as the mass scale for having one satellite on average cannot be found at high halo mass. As pairs of galaxies from the one-halo term dominates the clustering at small-scales (see Figure 5.6), we attribute this decrease to the strong signal observed by DESI in a range of scales which were not previously probed and that we can model only with pairs of galaxies preferentially in low mass halos.

However, physically motivated models of ELGs, either based on semi-analytical modelling (e.g Contreras et al., 2019, Favole et al., 2020, Gonzalez-Perez et al., 2018, 2020) or hydrodynamical simulations (e.g Hadzhiyska et al., 2021) do predict an increasing satellite HOD at high halo mass for ELGs at redshifts ~ 1. We thus interpret our negative index result as a sign of an inadequate HOD model to describe DESI ELGs. In the next section we modify the model to include central-satellite conformity, that is the fact that satellite occupation may be conditioned by the presence of central galaxies of the same type, an hypothesis corroborated by hydrodynamical simulations (Hadzhiyska et al., 2022b). Note that indications of conformity between central and satellite galaxies related to their types have already been reported in the literature (Weinmann et al., 2006).

5.7 Results in extended HOD models

In this section, we modify the standard prescription for satellite occupation. We first test conformity bias, as suggested by the results previously described and by studies from hydrodynamical simulations. We then test other possible changes in an attempt to better fit the clustering measurements in the problematic regions spotted in the above section.Throughout this section, the baseline model for central galaxies is the mHMQ prescription. We also describe a cross-check of our results using the AbacusHOD pipeline as an alternative fitting tool.

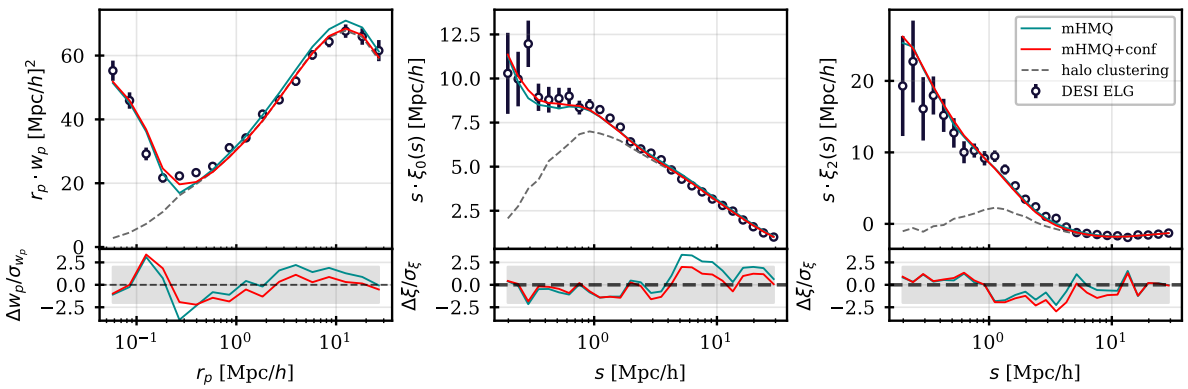

Figure 5.8: *Top: DESI ELG clustering measurements from the One-Percent survey data sample, compared to best fitting mHMQ models obtained with the GP pipeline, without (green line) and with (red line) strict conformity bias. The dashed line is the pure halo clustering. The agreement between data and expectations is slightly improved by requiring strict conformity, that is by conditioning satellite occupation to the presence of a central galaxy. Errors are Jackknife uncertainties only. Bottom: Fit residuals normalised by the diagonal errors of the full covariance matrix, that comprise Jackknife uncertainties for the data as well as stochastic noise and cosmic variance for the model, but no Hartlap factor corrections.*

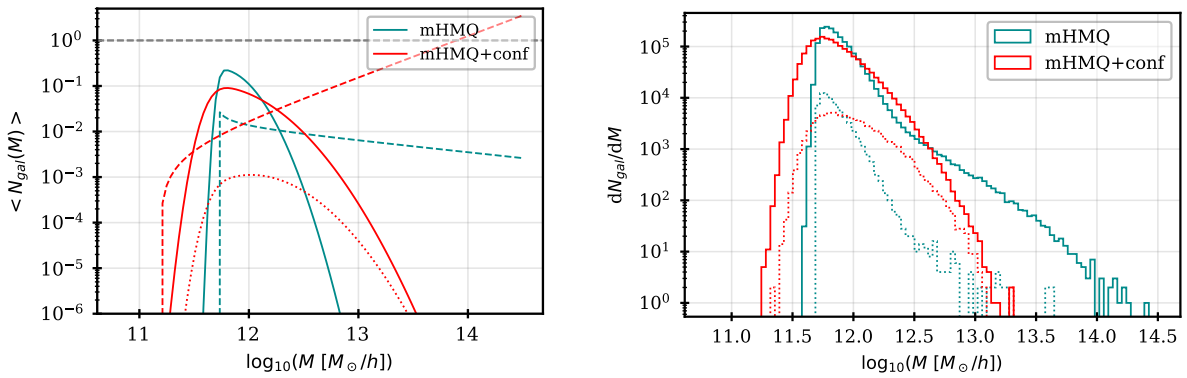

Figure 5.9: *Left: Best fitting HOD models to the DESI One-Percent ELG sample obtained without (green line) and with (red line) strict conformity bias between central (solid lines) and satellite (dashed lines) galaxies. The mHMQ prescription for centrals is used. In the case of conformity, the satellite HOD (red dashed line) corresponds to the mean number of satellites in halos already populated by a central. To better reflect the mean number of satellites with strict conformity, the product of the mean numbers of centrals and satellites is shown as the red dotted line. Right: Number of galaxies per halo mass bin for halos populated according to the two HOD models on the left. The simulation box volume is* $1.66\,(Gpc/h)^3$*. The full distributions are in solid lines. The dashed lines show the contribution of halos hosting more than one galaxy, that is the one-halo component of the full distributions. Requiring strict conformity drastically changes the HOD models and the distributions of the number of galaxies per populated halo mass bin: satellites are forced to populate only halos with central galaxies and thus are spread over a wider range of halo masses.*

5.7.1 Strict conformity bias

Best fitting clustering with strict central-satellite conformity from the GP pipeline are presented in Figure 5.8 and compared to previous results without conformity. In this extended model, satellites can populate a halo only if a central galaxy is already present. Best-fit values of the model parameters are reported in Table 5.3. Strict conformity only slightly improves the agreement with data and the best-fit χ^2 value. On the other hand, the shape of the HOD model and that of the distribution of galaxies per populated halo mass bin are significantly modified, as shown in Figure 5.9. With strict conformity, pairs of satellites in halos with no central galaxy are forbidden. To obtain the strong one-halo term needed to reproduce the small scale clustering, when conformity is required, pairs of galaxies are distributed over a wider range of halo masses at both low and high halo mass, as can be seen in the right-hand panel of Figure 5.9 (see distributions in dashed lines). This translates into a satellite HOD that increases linearly ($\alpha = 0.91^{+0.14}_{-0.11}$) with halo mass, as expected in physically motivated HOD models. Note also that the mass scale for having one satellite on average is now obtained at large halo mass, as can be seen directly on the left panel in Figure 5.9 and from the value of the effective $\log_{10} M_1'$ parameter in Table 5.2. We recall that even though the HOD of satellites increases with halo mass, strict conformity can only populate halos that already have a central. The product of the mean numbers of centrals and satellites, which better represents the expected number of satellites with strict conformity, is shown as the red dotted line in the left panel of Figure 5.9. In the following figures of the same kind, we only show the standard satellite HOD curves (dashed-lines), which hold only for halos populated by a central for models with strict conformity.

As a consequence, with strict conformity, the HOD parameters are all changed, except for the velocity dispersion parameter, f_{σ_v} which we discuss further in section 5.7.2. The fraction of satellites f_{sat} is five times smaller than without conformity, as a result of trading satellites alone in their halos for central-satellite pairs. On the other hand the one-halo term fraction f_{1h} and the mean halo mass remain very close

to their values without conformity. This shows that these are model-independent characteristics that can be constrained by clustering measurements. As such they provide suitable quantities to compare results from analyses done in different frameworks. Note that, by definition, in the case of strict conformity, $f_{1h} = 2f_{sat}$ for halos hosting one central and one satellite, which is typical of our ELG sample, cases with more than 1 satellite being rare.

parameter & χ^2	mHMQ	mHMQ+conformity
A_c (resc.)	1 (0.08)	0.1 (0.63)
$\log_{10} M_0$	$11.70^{+0.03}_{-0.03}$	$11.19^{+0.12}_{-0.10}$
A_s	$0.10^{+0.04}_{-0.03}$	$0.31^{+0.15}_{-0.08}$
$\log_{10} M_c$	$11.72^{+0.06}_{-0.04}$	$11.64^{+0.04}_{-0.04}$
α	$-0.26^{+0.08}_{-0.08}$	$0.91^{+0.14}_{-0.11}$
f_{σ_v}	$1.27^{+0.07}_{-0.06}$	$1.34^{+0.08}_{-0.08}$
σ_M	$0.22^{+0.04}_{-0.02}$	$0.39^{+0.08}_{-0.10}$
γ	$7.06^{+1.33}_{-1.97}$	$4.50^{+1.49}_{-1.29}$
$\log_{10} M_1'$	4.77	13.78
f_{sat}	$0.12^{+0.02}_{-0.02}$	$0.024^{+0.030}_{-0.017}$
f_{1h}	$0.039^{+0.009}_{-0.008}$	$0.048^{+0.010}_{-0.012}$
$\log_{10} \langle M_h \rangle$	$11.87^{+0.02}_{-0.01}$	$11.86^{+0.02}_{-0.02}$
χ^2 (ndf=64)	156.0 ± 1.0	152.5 ± 1.1

Table 5.3: *Results of mHMQ fits without and with strict conformity bias between central and satellite galaxies. The first line provides the initial fixed value of A_c and the rescaling factor applied to impose the density constraint in the fits. The following seven parameters are the free HOD parameters, the next four are derived parameters.* $\log_{10} M_1'$ *is given for best-fit values of α and A_s (the latter after rescaling). f_{sat} is the fraction of galaxies which are satellite galaxies. f_{1h} is the fraction of galaxies which are not alone in their halos. All masses are in units of $(M_\odot/h)$.*

Finally, Figure 5.10 presents the best fitting HOD models and distributions of the number of galaxies per populated halo mass bin for the four prescriptions we can use for central galaxies. The four models show an increase of the satellite HOD with increasing halo mass. The LNHOD model converges toward a triangular shaped HOD for centrals showing a sharp cut-off in mass for halos to host centrals, $\log_{10} M_h > \log_{10} M_c - 1$ which originates from the large best-fit value of σ_M. This deviates substantially from physically inspired ELG models although the shape of the best fitting clustering statistics is almost indiscernable from the three other HOD models. This is reflected in the best-fit χ^2 values that are similar in the four models, ~ 152 for mHMQ and LNHOD, ~ 156 for GHOD and ~ 161 for SFHOD. With strict conformity, the minimal halo mass to host satellites, $\log_{10} M_0$ is around ~ 11.2 for the four models, compared to 11.7 without conformity. This decrease is to be expected since with conformity the value of M_0 is driven by the minimum mass of halos hosting a central galaxy and reflects the need for having galaxy pairs in low-mass halos. We emphasize that all conformity models strongly favour putting satellites as soon as the halo is populated by central galaxies. The values of the characteristic halo mass for centrals, $\log_{10} M_c$, which were similar in the four models without conformity, are around 11.8 except for the LNHOD model which gives 12.6 due to the skewness of the HOD shape.

5.7.2 Velocity bias

In the GP pipeline, satellite velocities are normally distributed around their halo velocity, computed as the mean halo dark matter particle velocities. The satellite velocity dispersion is that of the particle

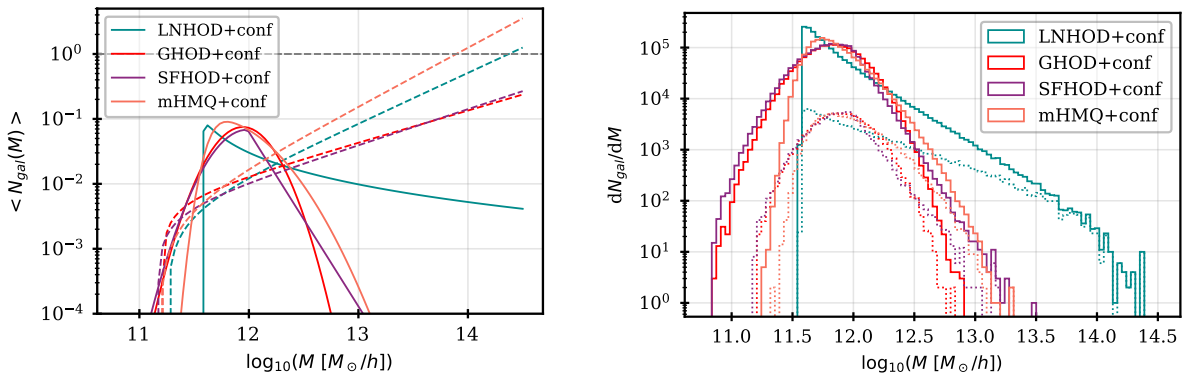

Figure 5.10: *Left: Best fitting HOD models to the DESI One-Percent ELG sample obtained with strict conformity bias between central and satellite galaxies. Four prescriptions for central galaxies are used. They reproduce clustering data equally well but give different HOD shapes. Right: Number of galaxies per halo mass bin for halos populated according to the four HOD models on the left. The simulation box volume is* $1.66\,(Gpc/h)^3$*. The full distributions are in solid lines. The dashed lines show the one-halo component of the full distributions. The four models show an increase of the satellite HOD with increasing halo mass. The LNHOD model converges towards a triangular shaped HOD for centrals showing a sharp cut-off in mass for halos to host centrals,* $\log_{10} M_h > \log_{10} M_c - 1$ *which originates from the large best-fit value of* σ_M*.*

velocities rescaled by the f_{σ_v} parameter which is left free to vary in the fits, namely:

$$\vec{v}_{\rm sat} \sim \mathcal{N}(\vec{v}_h, f_{\sigma_v} \cdot \sigma_{v_h}) \tag{5.12}$$

This parameter represents a simple way to make ELG satellites hotter or cooler than dark matter particles, an hypothesis which was tested in studies of the eBOSS ELG sample (Alam et al., 2021, Avila et al., 2020). The GP pipeline results previously presented show that, without or with conformity bias, and whatever the HOD prescription for central galaxies, the best-fit value for f_{σ_v} is significantly higher than 1, which is in line with what was reported in Avila et al. (2020). The best-fit values range from 1.2 to 1.5 depending on the model, with an error around ± 0.1.

We check the impact of satellite velocities on this result using two other prescriptions. Instead of drawing satellite velocities according to Equation (5.12), we set them to the halo velocity and add a circular velocity drawn from a NFW profile as defined in Navarro et al. (1996):

$$\vec{v}_{\rm sat}(r) = \vec{v}_h + \sqrt{\frac{GM_h}{r_{vir}}}\sqrt{\frac{g(c_h \cdot r)}{r \cdot g(c_h)}}\vec{u}_{circ} \quad \text{with} \quad g(x) = \ln(1+x) - \frac{x}{(1+x)} \tag{5.13}$$

$\vec{u}_{circ}$ is a unitary vector perpendicular to the vector joining the halo centre to the satellite position and whose orientation in this plane is randomly chosen. In the above equation, r is the satellite radial position (in unit of r_{vir}), r_{vir} is the virial halo radius and c_h its concentration. As mentioned in Section 5.5.1, we take r_{98} as a proxy for r_{vir} and r_{25} as a proxy for r_s, so that $c_h \equiv r_{vir}/r_s = r_{98}/r_{25}$. As a second choice, we first draw a satellite velocity $\vec{u}_{sat}$ according to Equation (5.12) and add to it a common infall velocity $\vec{v}_{\rm infall}$ defined along the line between the satellite position to the halo centre:

$$\vec{u}_{\rm sat} \sim \mathcal{N}(\vec{v}_h, \sigma_{v_h}) \quad \text{then} \quad \vec{v}_{\rm sat} = \vec{u}_{\rm sat} + \vec{v}_{\rm infall} \quad \text{with} \quad \vec{v}_{\rm infall} = v_{\rm infall} \cdot \frac{\vec{r}_h - \vec{r}_{\rm sat}}{|\vec{r}_h - \vec{r}_{\rm sat}|} \tag{5.14}$$

This model is a good approximation of the prediction presented in Orsi & Angulo (2018) based on semi-analytical models of star-forming galaxies. The latter predict that among star-forming galaxies those which were accreted the latest could have a net infall velocity towards the halo centres.

In order to illustrate the impact of satellite velocities on the clustering statistics, Figure 5.11 compares the DESI data clustering to the best fitting mHMQ model with strict conformity bias found in the previous section (purple curve) and to predictions from that model where we modify the satellite velocity prescription, keeping the other HOD parameters fixed, without refitting the data. The satellite velocity is modified according to Equation (5.13) and Equation (5.14), using a value of 170 km/s for v_{infall} in the latter case. Also shown is the predicted clustering with f_{σ_v} set to 1 to remove any velocity bias (green curve). The four models predict the same projected clustering, as expected since velocities have no effects on this statistic. Taking a NFW profile for velocities (red curve) does not provide a good model of the 2-point correlation functions multipoles. Not rescaling the velocity dispersion (green curve) provides a good model of the monopole only, while up-scaling the dispersion (purple curve) allowing to model both multipoles correctly. Last, there is practically no difference in the predicted clustering between an up-scaling of the particle velocity dispersion with a factor of 1.34 and a net infall velocity of $\sim$ 170 km/s added to velocities normally distributed around the halo velocity with a dispersion equal to that of the particle velocities (orange curve). These two models, although different, have quite similar impact on the clustering and cannot be disentangled with the statistics we are using. Note that random errors in the ELG redshift determination (J. Yu et al., 2023) are equivalent to a 60 km/s velocity dispersion along the line of sight and only accounts for 0.03 on the observed shift in f_{σ_v} w.r.t. 1. We conclude that a velocity dispersion larger than that of DM particles is needed to reproduce the clustering of the DESI One-Percent survey ELG sample. Interestingly, a velocity bias was also reported by SDSS for main galaxies at low redshifts ($z < 0.2$) and LRGs at intermediate ($z \sim 0.5$) redshifts (Guo et al., 2015a,b) but the effect goes in the opposite direction, with satellites moving more slowly than particles by a factor that depends on the galaxy luminosity, the bias being stronger for more luminous galaxies.

In the following, we continue with the baseline prescription for satellite velocities, expressed as a rescaling of dark matter particle velocities by a factor f_{σ_v}.

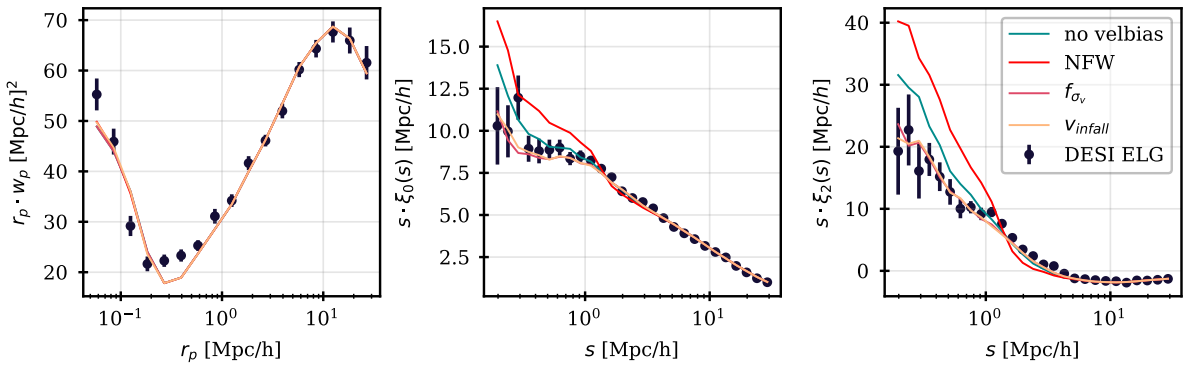

Figure 5.11: *DESI ELG clustering measurements from the One-Percent survey data sample compared to HOD models differing only by their satellite velocity prescriptions. We show the best fitting mHMQ model with conformity bias found in Section 5.7.1 (purple) which corresponds to rescaling the dispersion of dark matter particles by a factor f_{σ_v}=1.34 to describe the satellite velocities. Other models correspond to the following changes: f_{σ_v}=1 (green), drawing satellite velocities from a NFW profile (red) and assuming a common infall velocity (yellow) of v_{infall} = 170 km/s. Errors are Jackknife uncertainties. The four models give exactly the same projected clustering but produce differences in the 2-point correlation function multipoles at small-scales. We note that the same clustering can be obtained by rescaling the particle velocity dispersion or assuming a common infall velocity.*

5.7.3 Comparison to AbacusHOD pipeline

At this point, we cross-check our results with the AbacusHOD pipeline (Yuan et al., 2022), which is particle-based and highly efficient. Designed specifically for multi-tracer analyses and HOD-cosmology combined analyses, it takes advantage of the large volume and precision of the AbacusSummit simulations by optimising computational efficiency.

The baseline HOD prescription for ELG central galaxies in AbacusHOD is the HMQ model of Alam et al. (2021). In the present work, we restrict to the simpler mHMQ model of section 5.3 (with A_c renamed to p_{max}). For the satellite galaxies, we adopt the baseline power law model of Equation (5.6) except we reparametrise $M_0 = \kappa M_c$. Central galaxies are assigned the position and velocity vector of the centre of mass of the largest sub-halo while satellite galaxies are assigned to DM particles of the halo with equal probabilities. Each halo can only host at most one central galaxy and each particle can also host at most one satellite.

The AbacusHOD implementation of ELG central-satellite conformity introduces one extension parameter to the standard satellite HOD to modulate the strength of the conformity effect. Specifically, we modulate the M_1 parameter, which controls the overall amplitude of satellite occupation, by whether the halo hosts a central ELG or not:

$$\langle N_{sat}(M)\rangle = \begin{cases} \left(\frac{M-\kappa M_c}{M_{1,\mathrm{EE}}}\right)^{\alpha} & \text{if ELG central} \\ \left(\frac{M-\kappa M_c}{M_1}\right)^{\alpha} & \text{if not.} \end{cases} \tag{5.15}$$

where $M_{1,EE}$ is the new parameter that modulates the ELG-ELG conformity strength. If there is no conformity, then $M_{1,EE} = M_1$, and if there is maximal conformity, i.e. ELG satellites only occupy halos with ELG centrals, then $M_{1,EE} \ll M_1$. In principle, another conformity term between ELG satellites and LRG centrals is also possible but it was not included in the present work.

Velocity bias prescriptions are different between the two pipelines. For the GP pipeline, bias on velocities are changed only for satellites, through the scaling parameter f_{σ_v}, as described in Equation (5.12). The AbacusHOD pipeline allows both for central and satellite velocity biases, through parameters α_c and α_s, respectively. Those impact velocities as $v_{\mathrm{cent}} = v_h + \alpha_c \delta_v(\sigma_{vh})$ for centrals, where $\delta_v(\sigma_{vh})$ is the Gaussian scatter of the velocity dispersion of the halo, and $v_{\mathrm{sat}} = v_{\mathrm{particles}} + \alpha_s(v_{\mathrm{particles}} - v_h)$, as described in equations 8 & 9 in Yuan et al. (2022).

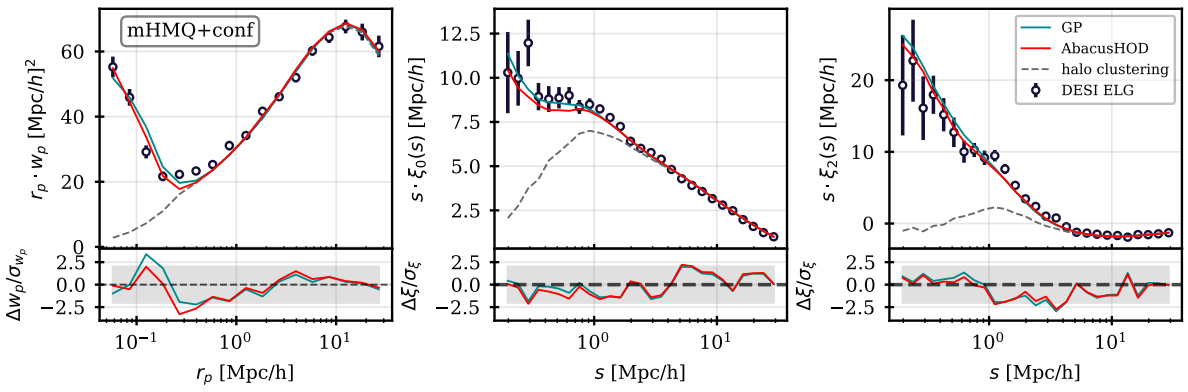

Figure 5.12: *Top: DESI ELG clustering measurements from the One-Percent survey data sample compared to best fitting mHMQ models with parametrised central-satellite conformity in the AbacusHOD pipeline (red) and with strict conformity bias in the GP pipeline (green). Bottom: Fit residuals normalised by the diagonal errors of the full covariance matrix, that comprise Jackknife uncertainties for the data as well as stochastic noise and cosmic variance for the model, but no Hartlap factor corrections.*

Although the baseline statistics of the AbacusHOD pipeline is the galaxy two-point correlation fonction in two dimensions, for this cross-check it is run using the same 2-point statistics (see Section 5.2.1) and the same data covariance matrix as the GP pipeline matrix (see Section 5.5.2). The model covariance matrix of the GP pipeline is ignored in the fits but is used to compute best-fit χ^2 values provided below. We compare the mHMQ best fitting results from the two pipelines in Figure 5.12 for the predicted clustering and in Table 5.4 for the HOD and derived parameters. Best-fit parameters from the AbacusHOD pipeline are derived using global optimisation chains using Gaussian priors so no error bars are provided. The two pipelines produce quite similar best fitting clustering predictions and goodness of fit results, despite the completely different nature of the pipelines and their different prescriptions for some parameters of the mHMQ model.

Most parameters treated in the same way in both pipelines have similar best fitting values, except for the γ parameter that controls the asymmetry of the central HOD. This difference is reflected in the shape of the distribution of the number of galaxies per populated halo mass bin, whose asymmetry is more pronounced for the GP pipeline result, as can be seen in Figure 5.13. On the other hand, the γ parameter is hardly constrained in the fits (see error bars in Table 5.4 and γ posteriors in Appendix B), which means that our clustering statistics are not very sensitive to the asymmetric character of the HOD distributions, so that distributions of the number of galaxies per populated halo mass bin as different as those in Figure 5.13 can produce very similar clustering signals (see Figure 5.12).

Although the velocity bias prescriptions are different, both pipelines end up with the same conclusion, namely that the satellite velocity dispersion is higher than that of halo particles. As for central velocities, the AbacusHOD pipeline result shows that allowing for a velocity dispersion of centrals is not really mandatory. As for central-satellite conformity, the AbacusHOD parametrised bias indicates clearly a preference for conformity since $M_{1,EE}$ is lower than M_1 by more than 5 units, making the strict conformity of the GP pipeline implementation a good approximation. Remarkably, both pipelines agree well on the derived parameters, the satellite fraction, one-halo term fraction and the mean halo mass value of the sample. Finally, we note that the χ^2 of the AbacusHOD result is slightly better than that of the GP pipeline but does not significantly improve the goodness of fit.

This means that the reason for the poor goodness of fit of our results so far is not to be found in the fitting methodology but rather in the HOD model itself. In the following, we test other extensions of the model in the GP pipeline to check whether an improvement can be found.

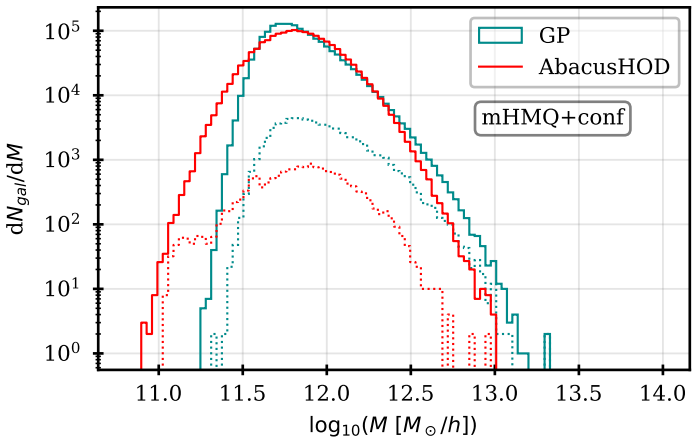

Figure 5.13: *Number of galaxies per halo mass bin for halos populated according to the best fitting mHMQ models to the DESI One-Percent ELG sample, from AbacusHOD with parametrised conformity bias (red) and from the GP pipeline with strict conformity bias (green). The simulation box volume is* $1.66\,(Gpc/h)^3$. *The full distributions are in solid lines and the dashed lines show the one-halo component of the full distributions.*

parameter & χ^2	AbacusHOD pipeline	GP pipeline
$p_{max} = A_c$ (resc.)	0.08	0.10 (0.63)
$\log_{10} M_0$	11.03 ($\kappa = 0.19$)	$11.19^{+0.12}_{-0.10}$
A_s	1 (fixed)	$0.31^{+0.15}_{-0.08}$
$\log_{10} M_c$	11.75	$11.64^{+0.04}_{-0.04}$
α	0.72	$0.91^{+0.14}_{-0.11}$
α_c, α_s or f_{σ_v}	0.19, 1.49	$1.34^{+0.08}_{-0.08}$
$\log_{10} M_1$	19.83	13 (fixed)
σ_M	0.31	$0.39^{+0.08}_{-0.10}$
γ	1.39	$4.50^{+1.49}_{-1.29}$
$\log_{10} M_{1,EE}$	14.25	-
f_{sat}	0.020	$0.024^{+0.030}_{-0.017}$
f_{1h}	0.040	$0.048^{+0.010}_{-0.012}$
$\log_{10} \langle M_h \rangle$	11.89	$11.86^{+0.02}_{-0.02}$
χ^2 (ndf)	143.53 (62)	152.5 ± 1.1 (64)

Table 5.4: *Results of mHMQ fits with parametrised central-satellite conformity from the AbacusHOD pipeline (left) and with strict conformity bias from the GP pipeline (right). The upper ten rows list HOD parameters, the next three give derived parameters. f_{sat} is the fraction of galaxies which are satellite galaxies. f_{1h} is the fraction of galaxies which are not alone in their halos. All masses are in units of $(M_\odot/h)$.*

5.7.4 Assembly bias

HOD modelling is primarily a function of halo mass only but semi-analytical models and hydrodynamical simulations predict dependencies in other properties that are referred to as secondary biases in the literature. In this section, we explore assembly bias which introduces a dependence related to the halo assembly history. We test dependencies either in halo concentration, local halo density or local halo density anisotropies, using the parametrisation suggested in Hadzhiyska et al. (2022c):

$$\langle N'_{cent}(M) \rangle = [1 + a_{cen} f_a (1 - \langle N_{cent}(M) \rangle)] \langle N_{cent}(M) \rangle \tag{5.16}$$

$$\langle N'_{sat}(M) \rangle = [1 + a_{sat} f_a] \langle N_{sat}(M) \rangle \tag{5.17}$$

where $\langle N_{cent}(M) \rangle$ and $\langle N_{sat}(M) \rangle$ are given in Section 5.3. In the above equations, f_a is introduced to materialize the property of each halo in a normalized way. In a given halo mass bin, halos are first ranked by decreasing values of the halo property and each halo is attributed a different value of f_a, assuming that the latter decreases linearly between 0.5 and -0.5 when going from the top ranked halo to the last one.

The halo properties we consider are the halo concentration, $c_h = r_{98}/r_{25}$ and the halo environment that we first characterize by the local halo density. To compute the latter, we project all halos in the simulation box onto a grid of 5 Mpc/h mesh using a count-in-cell resampling algorithm and calculate the density in each grid cell. Each halo is then attributed the local density of the grid cell it belongs to. As a third halo property, we consider local halo density anisotropies deduced from the so-called adaptive halo shear, computed from the smoothed local density field as described in Hadzhiyska et al. (2022c), using a smoothing scale of 1.5 Mpc/h.

Figure 5.14 presents the clustering predicted by the best fitting mHMQ models with strict conformity bias obtained without and with the three assembly bias prescriptions (see Table 5.5 for HOD and derived parameters). The mHMQ goodness of fit does not improve significantly when adding assembly bias

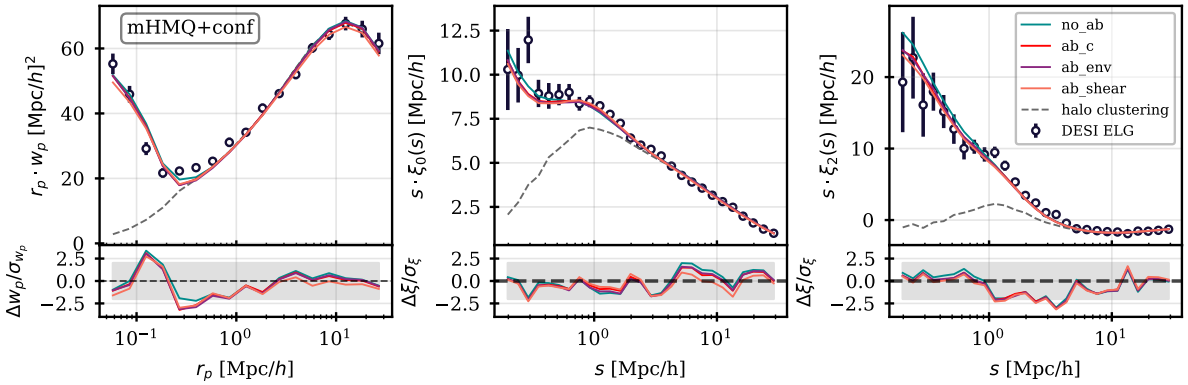

Figure 5.14: *Top: DESI ELG clustering measurements from the One-Percent survey data sample compared to different best fitting mHMQ models with strict conformity bias: baseline result found in Section 5.7.1 (green), model with assembly bias for both centrals and satellites as a function of concentration (red), local halo density (purple) and local halo density anisotropies (orange). Errors are Jacknnife uncertainties only. Bottom: Fit residuals normalised by the diagonal errors of the full covariance matrix, that comprise Jackknife uncertainties for the data as well as stochastic noise and cosmic variance for the model, but no Hartlap factor corrections.*

parameter & χ^2	no assembly	c_h assembly	ρ assembly	'shear' assembly
A_c (resc.)	0.1 (0.63)	0.1 (0.60)	0.1 (0.64)	0.1 (0.69)
$\log_{10} M_0$	$11.19^{+0.12}_{-0.10}$	$11.17^{+0.13}_{-0.10}$	$11.19^{+0.12}_{-0.11}$	$11.19^{+0.13}_{-0.11}$
A_s	$0.31^{+0.15}_{-0.08}$	$0.35^{+0.10}_{-0.07}$	$0.28^{+0.07}_{-0.05}$	$0.27^{+0.07}_{-0.05}$
$\log_{10} M_c$	$11.64^{+0.04}_{-0.04}$	$11.63^{+0.04}_{-0.03}$	$11.61^{+0.04}_{-0.03}$	$11.66^{+0.03}_{-0.03}$
α	$0.91^{+0.14}_{-0.11}$	$0.93^{+0.10}_{-0.07}$	$0.86^{+0.06}_{-0.06}$	$0.92^{+0.08}_{-0.10}$
f_{σ_v}	$1.34^{+0.08}_{-0.08}$	$1.35^{+0.08}_{-0.09}$	$1.34^{+0.10}_{-0.09}$	$1.31^{+0.08}_{-0.08}$
σ_M	$0.39^{+0.08}_{-0.10}$	$0.39^{+0.08}_{-0.09}$	$0.44^{+0.13}_{-0.11}$	$0.41^{+0.10}_{-0.08}$
γ	$4.50^{+1.49}_{-1.29}$	$4.54^{+1.20}_{-0.87}$	$5.76^{+1.13}_{-1.19}$	$6.05^{+1.04}_{-1.13}$
a_{cen}	-	$0.75^{+0.12}_{-0.25}$	$-0.02^{+0.22}_{-0.24}$	$0.10^{+0.05}_{-0.05}$
a_{sat}	-	$-0.32^{+0.59}_{-0.42}$	$0.02^{+0.63}_{-0.65}$	$0.00^{+0.61}_{-0.57}$
$\log_{10} M_1'$	13.78	13.72	13.87	13.79
f_{sat}	$0.024^{+0.030}_{-0.017}$	$0.022^{+0.024}_{-0.015}$	$0.021^{+0.024}_{-0.015}$	$0.021^{+0.022}_{-0.017}$
f_{1h}	$0.048^{+0.010}_{-0.012}$	$0.044^{+0.009}_{-0.013}$	$0.042^{+0.013}_{-0.008}$	$0.042^{+0.015}_{-0.01}$
$\log_{10} \langle M_h \rangle$	$11.86^{+0.02}_{-0.02}$	$11.84^{+0.02}_{-0.02}$	$11.83^{+0.02}_{-0.02}$	$11.82^{+0.02}_{-0.02}$
χ^2 (ndf)	152.5 ± 1.1 (64)	144.8 ± 1.0 (62)	150.4 ± 1.4 (62)	147.98 ± 1.14 (62)

Table 5.5: *Results of mHMQ fits with strict conformity bias between central and satellite galaxies without (left) and with assembly bias as a function of halo concentration (c_h), local density (ρ) and local density anisotropies ('shear'). The first line provides the initial fixed value of A_c and the rescaling factor applied to impose the density constraint in the fits. The following seven or nine parameters are the free HOD parameters, the next four are derived parameters. $\log_{10} M_1'$ is given for best-fit values of α and A_s (the latter after rescaling). f_{sat} is the fraction of galaxies which are satellite galaxies. f_{1h} is the fraction of galaxies which are not alone in their halos. All masses are in units of $(M_\odot/h)$.*

based on halo concentration, local density or local density anisotropies. HOD parameters and derived parameters are within 1σ of their values in the model without assembly bias. As a result all models exhibit similar clustering (almost indistinguishable). We note that the model with assembly bias using

halo concentration slightly improves the χ^2 value with a preference for highly concentrated halos, a_{cen} being close to 1. This constitutes a mild preference for assembly bias, but as the effect on clustering statistics is small, this preference cannot be established unambiguously. The best fit for assembly bias using halo local density points towards no dependence with the density as a_{cen} is found to be compatible with 0. In the case of local density anisotropies, best-fit results indicate a preference for halos with a slightly positive shear, a_{cen} being positive, but this preference is weaker than that for halo concentration. Lastly, the a_{sat} parameter is consistent with 0 and poorly constrained in the three models as a consequence of the fact that the satellite fraction with strict conformity bias is small ($\sim 2\%$).

5.7.5 Satellite positioning with a modified NFW profile

None of the extensions of the HOD model studied in the previous sections succeeds in producing extra pairs of galaxies at scales $r_p = [0.1, 1]\mathrm{Mpc}/h$ as required by data.

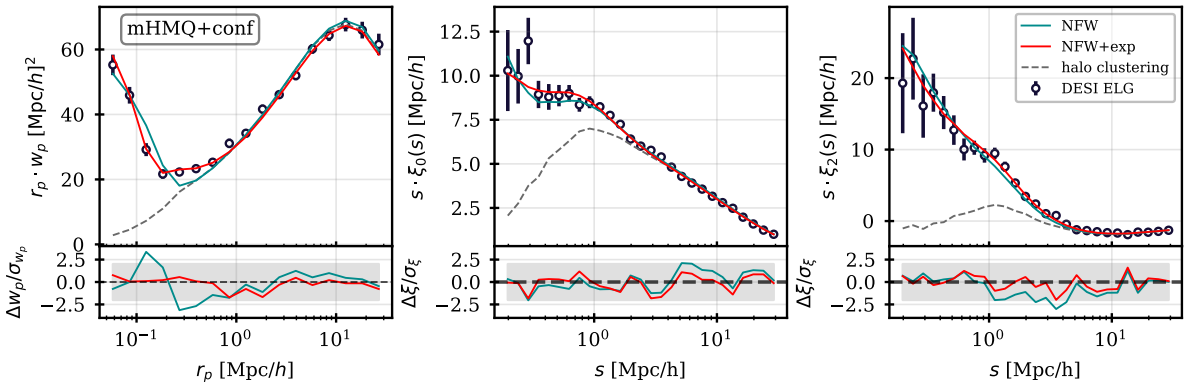

Figure 5.15: *Top: DESI ELG clustering measurements from the One-Percent survey data sample compared to best fitting mHMQ models with strict conformity bias: baseline result found in Section 5.7.1 (green) and model with satellite positioning according to a modified NFW profile (red). Errors are Jackknife uncertainties only. Bottom: Fit residuals normalised by the diagonal errors of the full covariance matrix, that comprise Jackknife uncertainties for the data as well as stochastic noise and cosmic variance for the model, but no Hartlap factor corrections.*

Nevertheless, it is possible to overcome this by changing the radial profile of satellites. Orsi & Angulo (2018) suggest that, whatever the halo mass, ELGs populate preferentially the outskirts of their host halos, galaxies accreted more recently being found further away from the halo centre. This is explained by the fact that satellite galaxies can present high star formation rates only for a short period once the galaxy gas has been depleted by tidal and ram pressure stripping. As a consequence, star-forming satellite ELGs are expected to be preferentially located in the outskirts of their halo where recently accreted subhalos free of the above processes can be found. On the observational side, results showing that the quenched fraction of the specific star formation rate distribution of galaxies is radially dependent within a halo were already reported for SDSS galaxies (Blanton & Berlind, 2007, Wetzel et al., 2012).

Inspired by the above publications, we test a modified NFW profile to position ELG satellites. The number of satellites for a given halo is first drawn according to the standard prescription in Equation (5.6). A fraction of them, f_{exp} have radial positions drawn from an exponential law:

$$\frac{dN(r)}{dr} = e^{-r/(\tau \cdot r_s)} \tag{5.18}$$

where r is the distance between the satellite and the halo centre, and τ governs the length scale of the exponential and acts on the extension of the profile. Radial positions of the remaining satellites obey a NFW profile with the same proxy for r_{vir} as in Section 5.5.1 but squeezing the proxy for r_s by a

parameter & χ^2	NFW profile	modified profile
A_c (resc.)	0.1 (0.63)	0.1 (0.51)
$\log_{10} M_0$	$11.19^{+0.12}_{-0.10}$	$11.20^{+0.11}_{-0.09}$
A_s	$0.31^{+0.15}_{-0.08}$	$0.41^{+0.10}_{-0.15}$
$\log_{10} M_c$	$11.64^{+0.04}_{-0.04}$	$11.64^{+0.04}_{-0.04}$
α	$0.91^{+0.14}_{-0.11}$	$0.81^{+0.08}_{-0.14}$
f_{σ_v}	$1.34^{+0.08}_{-0.08}$	$1.63^{+0.11}_{-0.10}$
σ_M	$0.39^{+0.08}_{-0.10}$	$0.30^{+0.09}_{-0.07}$
γ	$4.50^{+1.49}_{-1.29}$	$5.47^{+1.37}_{-1.58}$
f_{exp}	-	$0.58^{+0.06}_{-0.05}$
τ	-	$6.14^{+1.11}_{-1.20}$
λ_{NFW}	-	$0.67^{+0.06}_{-0.06}$
$\log_{10} M_1'$	13.78	13.84
f_{sat}	$0.024^{+0.030}_{-0.017}$	$0.034^{+0.010}_{-0.012}$
f_{1h}	$0.048^{+0.010}_{-0.012}$	$0.069^{+0.020}_{-0.024}$
$\log_{10} \langle M_h \rangle$	$11.86^{+0.02}_{-0.02}$	$11.86^{+0.03}_{-0.03}$
χ^2 (ndf)	152.5 ± 1.1 (64)	87.91 ± 1.84 (61)

Table 5.6: *Results of mHMQ fits with strict conformity bias using a standard NFW profile for satellite positioning (left) and our modified profile (right). The first line provides the initial fixed value of A_c and the rescaling factor applied to impose the density constraint in the fits. The following seven or ten parameters are the free HOD parameters, the next four are derived parameters. $\log_{10} M_1'$ is given for best-fit values of α and A_s (the latter after rescaling). f_{sat} is the fraction of galaxies which are satellite galaxies. f_{1h} is the fraction of galaxies which are not alone in their halos. All masses are in units of $(M_\odot/h)$.*

factor λ_{NFW}, namely $r_s \to r_s/\lambda_{NFW}$. This is almost equivalent to extending the profile cut-off with respect to r_{vir} into $r_{cutoff} = \lambda_{NFW} \cdot r_{vir}$ and allows for modifications of the profile extension. The three parameters f_{exp}, τ and λ_{NFW} are left free to vary in the fits. Note that galaxies positioned beyond the halo virial radius are improperly called satellites but we keep that denomination here to reflect the HOD parametrisation component they come from.

The best fitting mHMQ results with strict conformity and the above prescription are compared with the baseline results using a pure NFW profile in Figure 5.15 for the clustering predictions and in Table 5.6 for the HOD and derived parameters. The modified positioning of satellites translates into a significant improvement of the agreement between data and predictions, with a χ^2 value dropping from ∼152 to ∼88 (p-value of 1.4%). The improvement is most notable in the region of the up-turn of the projected correlation function (see residuals in Figure 5.15) showing that additional pairs of galaxies have been generated at these scales with the extended profile, with no degradation of the agreement elsewhere.

An example of satellite density profile corresponding to the best fitting parameters is represented in Figure 5.16 as a function of the radial position of the satellites with respect to the halo centre projected perpendicular to the line of sight. The profile clearly shows that the exponential component acts at projected scales between 0.03 and 1 Mpc/h, the region of the up-turn in w_p. Note that the scales covered by our clustering measurements are more sensitive to the region close to the halo virial radius (hence to the cut-off applied to the NFW profile) than to the shape of the profile deep in the halo core.

Table 5.6 shows that the HOD parameters as well as the derived parameters are similar between the two models, except for a 20% increase of the value of f_{σ_v}, meaning that the extended profile of satellites leads to a higher satellite velocity dispersion. This provides a coherent picture as recently-accreted subhalos in the outskirts of halos are expected to have higher velocities than the virial velocity

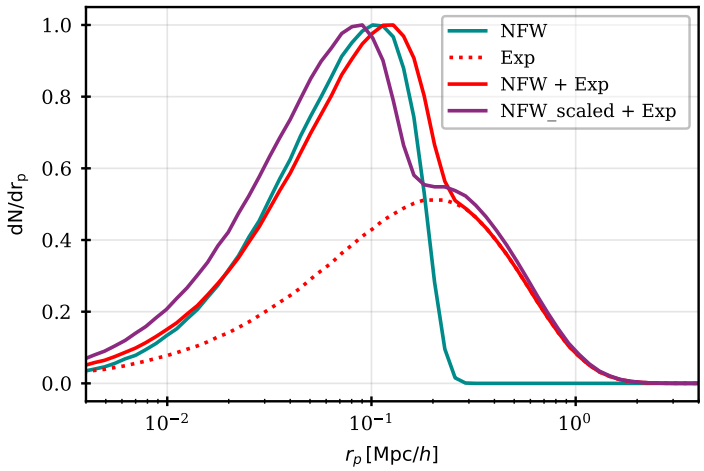

Figure 5.16: *Normalized satellite density profile for best-fit parameters in the mHMQ model with strict conformity and our modified NFW profile prescription for satellites, as a function of the projected galaxy-halo centre distance perpendicular to the line of sight. Once this profile is embedded into a HOD model, this distance is also the projected separation of central-satellite pairs. In this example, we consider a halo of concentration 5 and* $r_s = 0.06 Mpc/h$ *(corresponding to halo masses around* $10^{12} M_\odot/h$*, close to the mean halo mass value of our sample). Curves (all normalized at a maximal value of 1) are for the NFW profile (blue), the added exponential law (dotted), the combination of the two with no scaling of the NFW cut-off (green) and the complete modified model (red).*

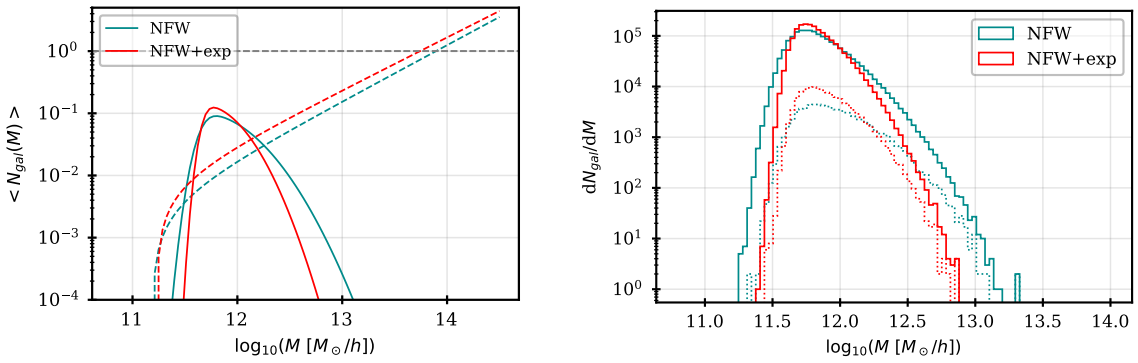

Figure 5.17: *Left: Best fitting HOD models to the DESI One-Percent ELG sample with strict conformity bias, obtained with a standard NFW profile for satellites (green) and with our modified NFW profile (red). Solid (resp. dashed) lines represent central (resp. satellite) galaxies. The mHMQ prescription is used for centrals. Right: Number of galaxies per halo mass bin for halos populated according to the mHMQ models on the left. The simulation box volume is* $1.66\,(Gpc/h)^3$*. The full distributions are in solid lines. The dashed lines show the one-halo component of the full distributions. The two satellite profiles produce similar results, with a larger scatter in populated halo masses for the modified profile.*

of the halo. The comparison between the two models is further illustrated in Figure 5.17, which presents the HOD and the distribution of the number of galaxies per populated halo mass bin of the two models. The only difference is a larger scatter in populated halo masses for with the modified NFW profile.

The profile parameters, f_{exp}, τ and λ_{NFW}, are all well constrained by data and their best fitting values are in favour of a departure from a standard NFW profile. We find that the exponential profile contains around 60% of the satellites and a fraction of these (approximately 12% of the total number of satellites, as measured in the mocks at best fitting HOD parameters) are placed beyond our proxy for the halo virial radius (see Figure 5.16). The above modified profile is empirical and can most probably be replaced by a more physics driven modelling. Nevertheless, our main finding is that the ELG clustering measured by the DESI One-Percent survey clearly favours a fraction of ELGs residing in the outskirts of halos, as suggested by Blanton & Berlind (2007), Wetzel et al. (2012) and Orsi & Angulo (2018).

5.8 Testing for redshift evolution

The ELG clustering measurements are produced in two separate redshift bins, from 0.8 to 1.1 and 1.1 to 1.6, with completeness-weighted redshifts of 0.95 and 1.32, respectively. The clustering measurements for the two redshift bins including completeness and FKP weights are shown in Figure 5.18. Measurements in the two redshift bins agree for most separations but exhibit significant differences in the monopole up to 10 Mpc/h and in the projected correlation function around the up-turn scale of 0.3 Mpc/h. It is thus interesting to fit the two bins in redshift separately to see how the agreement between HOD modelling and data evolves. The HOD model in each bin is calculated from the N-body simulation snapshot closest to the mean completeness-weighted redshift of the bin (i.e. snapshots at z=0.95 and z=1.325, respectively).

Best fitting results in the two redshift bins from the mHMQ model with strict conformity and our modified NFW profile for satellite positioning are presented in Figure 5.18 and summarised in Table 5.7. With respect to results obtained in the full redshift bin (see right column in Table 5.7, p-value of 1.4%), the goodness of fit is similar in the low redshift bin (p-value of 0.6%) and much better in the high redshift bin (p-value of 35%). Variations of the HOD parameters and the derived parameters with redshift appear to be moderate, parameter values in the two redshift bins being all within 1σ. The same is true for the distribution of the number of galaxies per populated halo mass bin as shown in Figure 5.19. In the

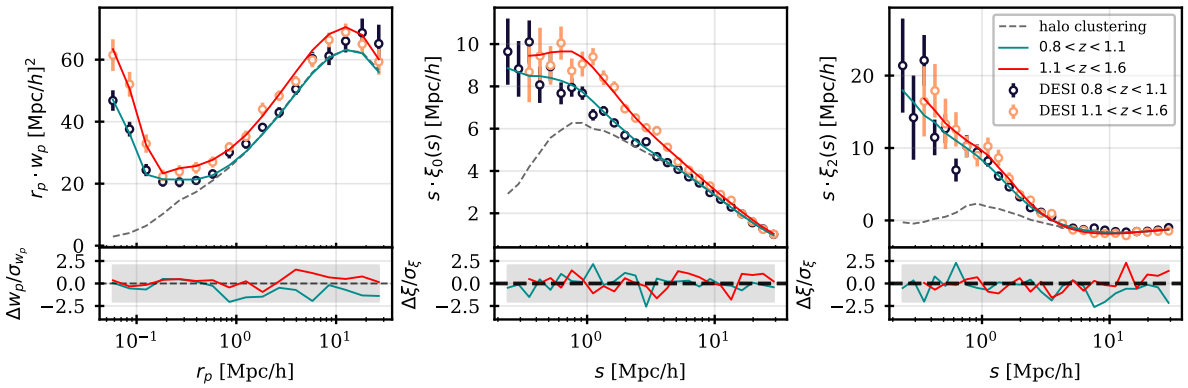

Figure 5.18: *Top: DESI ELG clustering measurements from the One-Percent survey data sample in two different redshift bins from 0.8 to 1.1 (black) and 1.1 to 1.6 (orange), compared to best fitting mHMQ models with strict conformity bias and our modified NFW profile for satellite positioning, for the redshift 0.8 to 1.1 (green) and 1.1 to 1.6 (red). The dashed line is the pure halo clustering for the low redshift bin. Errors are Jackknife uncertainties. Bottom: Fit residuals normalised by the diagonal errors of the full covariance matrix (calculated for each redhsift bin), that comprise Jackknife uncertainties for the data as well as stochastic noise and cosmic variance for the model, but no Hartlap factor corrections.*

parameter & χ^2	$0.8 < z < 1.1 \quad \bar{z} = 0.95$	$1.1 < z < 1.6 \quad \bar{z} = 1.325$	$0.8 < z < 1.6 \quad \bar{z} = 1.1$
A_c (resc.)	0.1 (0.43)	0.1 (0.51)	0.1 (0.51)
$\log_{10} M_0$	$11.10^{+0.05}_{-0.04}$	$11.23^{+0.16}_{-0.14}$	$11.20^{+0.11}_{-0.09}$
A_s	$0.38^{+0.04}_{-0.04}$	$0.47^{+0.13}_{-0.13}$	$0.41^{+0.10}_{-0.15}$
$\log_{10} M_c$	$11.62^{+0.02}_{-0.04}$	$11.67^{+0.04}_{-0.04}$	$11.64^{+0.04}_{-0.04}$
α	$0.74^{+0.07}_{-0.05}$	$0.85^{+0.08}_{-0.10}$	$0.81^{+0.08}_{-0.14}$
f_{σ_v}	$1.71^{+0.11}_{-0.14}$	$1.71^{+0.20}_{-0.15}$	$1.63^{+0.11}_{-0.10}$
σ_M	$0.21^{+0.10}_{-0.05}$	$0.29^{+0.11}_{-0.08}$	$0.30^{+0.09}_{-0.07}$
γ	$6.49^{+0.69}_{-1.39}$	$5.10^{+1.51}_{-1.20}$	$5.47^{+1.37}_{-1.58}$
f_{exp}	$0.70^{+0.10}_{-0.09}$	$0.55^{+0.10}_{-0.09}$	$0.58^{+0.06}_{-0.05}$
τ	$5.69^{+1.72}_{-2.00}$	$7.22^{+1.77}_{-3.14}$	$6.14^{+1.11}_{-1.20}$
λ_{NFW}	$0.60^{+0.09}_{-0.09}$	$0.67^{+0.07}_{-0.07}$	$0.67^{+0.06}_{-0.06}$
$\log_{10} M_1'$	14.05	13.73	13.84
f_{sat}	$0.026^{+0.005}_{-0.005}$	$0.035^{+0.010}_{-0.011}$	$0.034^{+0.010}_{-0.012}$
f_{1h}	$0.053^{+0.009}_{-0.009}$	$0.069^{+0.019}_{-0.021}$	$0.069^{+0.020}_{-0.024}$
$\log_{10} \langle M_h \rangle$	$11.78^{+0.03}_{-0.04}$	$11.86^{+0.05}_{-0.05}$	$11.86^{+0.03}_{-0.03}$
χ^2 (ndf)	89.78 ± 0.66 (59)	58.35 ± 0.41 (55)	87.91 ± 0.84 (61)

Table 5.7: *Results of mHMQ fits with strict conformity bias and our modified NFW profile for satellite positioning, presented separately in two redshift bins and compared to the results with the whole redshift bin (right). The first line provides the initial fixed value of A_c and the rescaling factor applied to impose the density constraint in the fits. The following ten parameters are the free HOD parameters, the next four are derived parameters. $\log_{10} M_1'$ is given for best-fit values of α and A_s (the latter after rescaling). f_{sat} is the fraction of galaxies which are satellite galaxies. f_{1h} is the fraction of galaxies which are not alone in their halos. The number of degrees of freedom is different in the three bins and indicated in brackets in the χ^2 row. All masses are in units of $(M_\odot/h)$.*

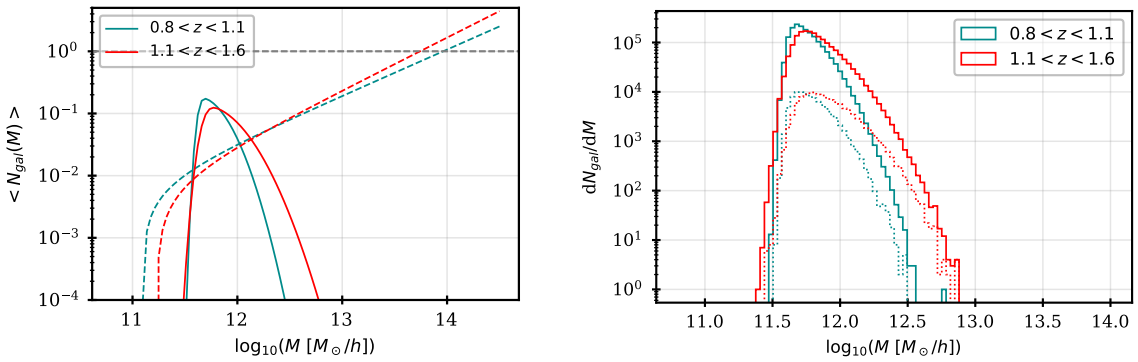

Figure 5.19: *Left: Best fitting HOD models to the DESI One-Percent ELG sample split in two separate redshift bins, 0.8-1.1 (green) and 1.1 to 1.6 (red), with strict conformity bias and our modified NFW profile for satellite positioning. Solid (resp. dashed) lines represent central (resp. satellite) galaxies. The mHMQ prescription is used for centrals. Right: Number of galaxies per halo mass bin for halos populated according to the mHMQ model on the left. The simulation box volume is $1.66\,(Gpc/h)^3$. The full distributions are in solid lines. The dashed lines show the one-halo component of the full distributions. The two redshift bins exhibit similar distributions, the higher redshift bin (1.1 to 1.6) showing a larger scatter towards higher populated halo masses.*

high redshift bin, there is a small increase of the scatter in the latter towards higher populated halo masses, which is reflected in the higher halo mass scale of the ELG sample, $11.86^{+0.05}_{-0.05}$ vs $11.78^{+0.03}_{-0.04}$ but the difference is at the level of 1σ. For completeness, we show in Appendix B the contour plots of the mHMQ fits with strict conformity and our modified NFW profile for satellite positioning obtained at final iteration in the two redshift bins.

To conclude, changes of the ELG sample with redshift in terms of the mean halo mass or in the one-halo term fraction are at the level of 1σ and thus cannot be considered as significant.In the companion paper H. Gao et al. (2023), the ELG sample of the DESI One-Percent survey was split in narrower redshift bins but did not show a significant variation with redshift of the characteristic halo mass hosting ELGs either. In a second companion analysis J. Yu et al. (2023), the luminosity of that sample (from [O II] emission) was also found to evolve very mildly with redshift (see their Figure 9).We discuss further the results from the companion analyses in Section 5.10. Using a sample of [O II] emitters at $z > 1$ in the Subaru HSC survey, Okumura et al. (2021) also found a constant mass across redshifts bins, in agreement with our findings.

5.9 Testing for cosmology dependence

In this section, we study how the previous results evolve when changing the reference cosmology both in the simulation box (used for the modelling) and in the fiducial cosmology (used to convert redshifts to distances). We test one cosmology with a high $N_{\rm eff}$ value and one with a low σ_8 value (see Table 5.8 for the complete list of cosmological parameter values). In this section, we continue with the mHMQ model with strict conformity bias and the extended NFW profile for satellite positioning but perform fits in the full redshift bin.

Cosmologies	$\Omega_{cdm}h^2$	$\Omega_b h^2$	σ_8	n_s	h	w_0	w_a
baseline	0.1200	0.02237	0.811355	0.9649	0.6736	-1	0
high $N_{\rm eff}$(c003)	0.1291	0.02260	0.855190	0.9876	0.7160	-1	0
low σ_8 (c004)	0.1200	0.02237	0.753159	0.9649	0.6736	-1	0

Table 5.8: *Parameter values of the three cosmologies used in Section 5.9. Indicated are the present-day densitites of cold dark matter and baryons, the normalisation today of the linear power spectrum in spheres of radius* 8Mpc/h*, the spectral index of the primordial matter power spectrum, the reduced value of the Hubble constant and the dark energy equation of state parameters.*

Best fitting results are presented in Figure 5.20 and summarised in Table 5.9. Despite the change of cosmology, the data clustering can be modeled with similar goodness of fit as in the baseline cosmology, showing that the tested changes have a negligible impact on clustering. Changing the cosmology does not lead to significant changes for most HOD and derived parameters. The largest changes are for $\log_{10} M_c$ and f_{σ_v}, with shifts between 1 and 2σ. For the derived parameters, both the satellite and one-halo fractions have consistent values. As a consequence of the variation of $\log_{10} M_c$, the mean halo mass, $\log_{10} \langle M_h \rangle$, varies by at most 2.7σ (0.08 dex) with the cosmological changes tested. Figure 5.21 shows the distribution of the number of galaxies per populated halo mass bin at best fit for the three cosmologies. The spread of the distribution is different in the three cases, the largest spread being observed for the low σ_8 cosmology. The baseline and low σ_8 cosmologies differ only by their values of the σ_8 parameter which has a direct impact on structure formation. A higher σ_8 value is expected to generate fewer small-mass halos and more large-mass halos at the same redshift and thus may explain the reduced spread at smaller halo masses for the baseline cosmology. At large mass, the spread evolves in the opposite direction to that expected from σ_8 values and may be governed more by the clustering to be modelled. The same argument holds for the N_{eff} cosmology, although in that case, several other parameters have different values than

in the baseline cosmology (lower Ω_m and higher n_s, h) which have a different effect on structure formation that can partly compensate for the effect of σ_8.

parameter & χ^2	high $N_{\rm eff}$	low σ_8	Planck 2018
A_c (resc.)	0.1 (0.63)	0.1 (0.49)	0.1 (0.51)
$\log_{10} M_0$	$11.20^{+0.17}_{-0.13}$	$11.22^{+0.12}_{-0.12}$	$11.20^{+0.11}_{-0.09}$
A_s	$0.42^{+0.14}_{-0.11}$	$0.46^{+0.11}_{-0.12}$	$0.41^{+0.10}_{-0.15}$
$\log_{10} M_c$	$11.67^{+0.03}_{-0.02}$	$11.51^{+0.02}_{-0.02}$	$11.64^{+0.04}_{-0.04}$
α	$0.97^{+0.12}_{-0.14}$	$0.80^{+0.09}_{-0.11}$	$0.81^{+0.08}_{-0.14}$
f_{σ_v}	$1.40^{+0.09}_{-0.10}$	$1.65^{+0.13}_{-0.17}$	$1.63^{+0.11}_{-0.10}$
σ_M	$0.42^{+0.10}_{-0.08}$	$0.59^{+0.09}_{-0.08}$	$0.30^{+0.09}_{-0.07}$
γ	$4.36^{+0.9}_{-0.88}$	$5.02^{+1.18}_{-1.38}$	$5.47^{+1.37}_{-1.58}$
f_{exp}	$0.57^{+0.07}_{-0.07}$	$0.55^{+0.05}_{-0.05}$	$0.58^{+0.06}_{-0.05}$
τ	$6.01^{+1.07}_{-1.04}$	$8.05^{+1.18}_{-1.62}$	$6.14^{+1.11}_{-1.20}$
λ_{NFW}	$0.64^{+0.06}_{-0.06}$	$0.63^{+0.07}_{-0.06}$	$0.67^{+0.06}_{-0.06}$
$\log_{10} M_1'$	13.60	13.81	13.84
f_{sat}	$0.034^{+0.009}_{-0.009}$	$0.034^{+0.008}_{-0.010}$	$0.034^{+0.010}_{-0.012}$
f_{1h}	$0.067^{+0.018}_{-0.017}$	$0.067^{+0.019}_{-0.016}$	$0.069^{+0.020}_{-0.024}$
$\log_{10} \langle M_h \rangle$	$11.94^{+0.03}_{-0.03}$	$11.84^{+0.02}_{-0.02}$	$11.86^{+0.03}_{-0.03}$
χ^2 (ndf=61)	78.23 ± 0.90	93.80 ± 0.83	87.91 ± 0.84

Table 5.9: *Results of mHMQ fits with strict conformity bias between central and satellite galaxies in our baseline cosmology (right), in the high $N_{\rm eff}$ cosmology (left) and in the low σ_8 cosmology (middle). The first line provides the initial fixed value of A_c and the rescaling factor applied to impose the density constraint in the fits. The following ten parameters are the free HOD parameters, the next four are derived parameters.* $\log_{10} M_1'$ *is given for best-fit values of α and A_s (the latter after rescaling). f_{sat} is the fraction of galaxies which are satellite galaxies. f_{1h} is the fraction of galaxies which are not alone in their halos. All masses are in units of $(M_\odot/h)$.*

5.10 Comparing to companion DESI analyses

Two companion analyses studied the clustering of the One-Percent DESI ELG sample in the same redshift range as in the present analysis, but with different methodologies, SHAM in J. Yu et al. (2023) and a novel abundance matching method based on the stellar-halo mass relation (SHMR-AM) in H. Gao et al. (2023). Despite differences in methodology, N-body simulation, reference cosmology, clustering statistics and separation ranges included in the analysis, their findings on the mean halo mass scale of the DESI ELG sample, 11.90 ± 0.06 in the SHAM analysis and ~ 12.07 in the SHMR-AM one, agree with ours, $11.86^{+0.02}_{-0.01}$.

The satellite fraction we find without central-satellite conformity - that is allowing for satellite ELG galaxies with no central ELG galaxy in their halo - is $12\% \pm 2\%$. This result becomes $3.4\% \pm 1.0\%$ with central-satellite conformity. Note that both companion SHAM analyses include satellite galaxies (living in subhalos) with no ELG central galaxy in the main halo, which is comparable to no central-satellite conformity. The SHMR-AM analysis uses measurements of w_p above 0.1 Mpc/h in r_p and multipole measurements on scales above 0.3 Mpc/h and measures a satellite fraction $\sim 15\%$, which is consistent with our result. The SHAM analysis uses multipole measurements on scales above 5 Mpc/h and thus can only derive a predicted fraction of satellites. Their result is $3.4\% \pm 2.0\%$, which does not agree with the above results, most probably as a result of too high a threshold on scales included in their fits. This

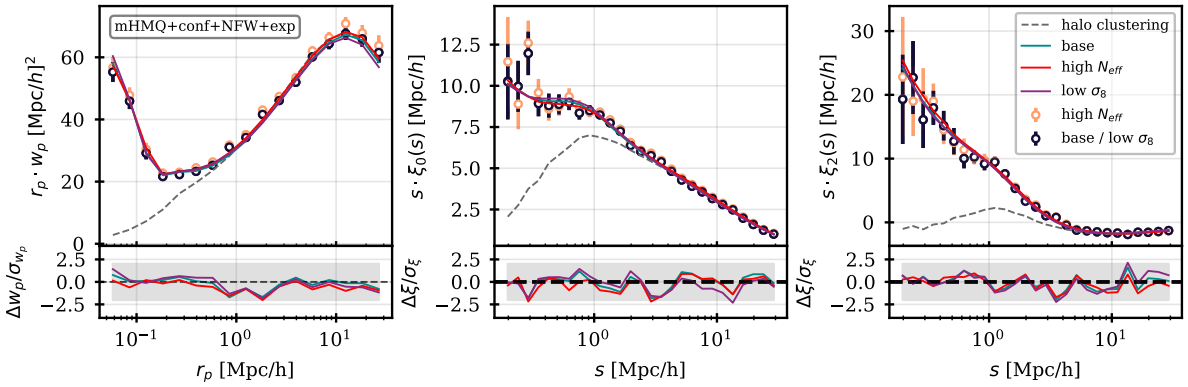

Figure 5.20: *Top: DESI ELG clustering measurements from the One-Percent survey data sample in different cosmologies, high N_{eff} (orange dots) and low σ_8 (dark blue dots). The distance-redshift relation in the low σ_8 cosmology is the same as in the baseline cosmology. Data are compared to best fitting HOD models obtained in the baseline (green), low σ_8 (purple) and high N_{eff} cosmologies. The HOD model is the mHMQ model with strict conformity bias and our modified NFW profile for satellite positioning. The dashed line is the pure halo clustering. Errors are jackknife uncertainties. Top: high $N_{\rm eff}$ cosmology Bottom: Fit residuals normalised by the diagonal errors of the full covariance matrix (calculated for each cosmology), that comprise Jackknife uncertainties for the data as well as stochastic noise and cosmic variance for the model, but no Hartlap factor corrections.*

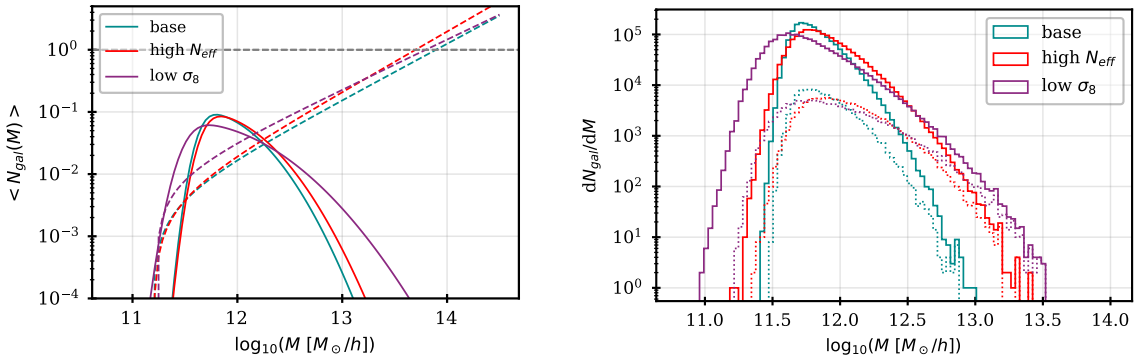

Figure 5.21: *Left: Best fitting HOD models to the DESI One-Percent ELG sample with strict conformity bias and our modified NFW profile for satellite positioning, obtained with different cosmologies: baseline (green), high N_{eff} (red) and low σ_8 (purple). Solid (resp. dashed) lines represent central (resp. satellite) galaxies. The mHMQ prescription is used for centrals. Right: Number of galaxies per halo mass bin for halos populated according to the best fitting HOD models on the left. The simulation box volume is 1.66 $(Gpc/h)^3$. The full distributions are in solid lines. The dashed lines show the contribution the one-halo component of the full distributions.*

high threshold also makes it impossible to achieve a good modelling of the w_p up-turn at small-scales (see Figure C4 in J. Yu et al. (2023)). Despite their using small-scale measurements in their fits, the SHMR-AM analysis also struggles to correctly reproduce the w_p clustering at the smallest scales (see Figure 11 in H. Gao et al. (2023)). Work is underway to include central-satellite conformity in the SHMR-AM analysis, which should improve the results.

Using our best fitting mHMQ model with strict conformity bias and our modified NFW profile, we also compute the predicted linear bias factor of the DESI One-Percent ELG galaxy sample. To do so, we produce 100 mocks with HOD parameters randomly selected in the MCMC chains at the fit final iteration, convert them to real-space and compare the 2PCF from these mocks to the predicted real space 2PCF from linear theory (at the same cosmology), which are related by the squared value of the linear bias factor of the galaxy sample:

$$\xi^r_{mocks}(s) = b^2 \xi^r_{linear}(s) \tag{5.19}$$

Using this equation for s between 40 and 80 Mpc/h, we fit the value of b for each mock and average them over all mocks. In order to propagate the uncertainties from the measured clustering and the fitting methodology (which are reflected in the pool of HOD parameter values used to produce the mocks), the dispersion over the mocks is taken as the error on the reported value of b. Our results are presented in Figure 5.22 as a function of the redshift of the simulation snapshot used for the modelling. We find the following values: $b_{0.95} = 1.20^{+0.04}_{-0.04}$ for the low redshift bin, $b_{1.1} = 1.33^{+0.03}_{-0.03}$ for the complete redshift bin and $b_{1.325} = 1.45^{+0.03}_{-0.03}$ for the high redshift bin.We also indicate the evolution with redshift of the inverse of the linear growth factor, with arbitrary normalisation. The bias deduced from our HOD study has an evolution consistent at the 1σ level with that of the growth factor.

Figure 5.22 also presents the results derived in two companion analyses, both SHAM analyses, the first one already mentioned J. Yu et al. (2023) based on the UNIT simulation, and the second one F. Prada et al. (2023) using the UCHUU simulation and the ELG data sample restricted to the redshift range between 0.8 and 1.34. Note that in the latter case, the reported errors are errors on the mean bias measured from a set of best-fit SHAM lightcones and thus do not include clustering measurements errors from data. Despite the differences between the analyses already outlined at the beginning of this section, the predictions with error bars are in reasonable agreement. The set of results with incomplete error bars provides a qualitative cross-check.

5.11 Conclusions

The sample of $\sim 270k$ ELGs collected by the DESI One-Percent survey in the redshift range between 0.8 and 1.6 (averager redshift of 1.13) is used to study the ELG small-scale clustering in the HOD framework. Thanks to the high completeness of the sample, the clustering measurements can be pushed down to scales never probed before in redshift space, 0.04 Mpc/h in r_p for the projected correlation function w_p and 0.17 Mpc/h in separation s for the two even multipoles of the 2PCF. A strong one-halo signal is observed at the smallest scales, below 0.2 Mpc/h in r_p and below 1 Mpc/h in s. To correctly model the strong one-halo term signal requires putting close pairs of galaxies in small-mass halos.

For central galaxies, we consider different prescriptions, a pure Gaussian distribution and three asymmetric ones, the strongest skewness being achieved with a log normal distribution. For satellites, we use a standard power law and do not require the presence of a central galaxy to put a satellite in the halo. Satellite positioning follow a NFW profile with a cut-off set at the halo virial radius, and we allow for velocity dispersion biased w.r.t that of the halo dark matter particles. Several extensions of these models are also explored.

In our baseline settings, whatever the different prescriptions for the central HOD, we achieve a good modelling of the measured clustering down to the smallest scales but obtain satellite HODs that decrease at large halo mass, contrary to expectations from semi-analytical ELG models. We recover satellite occupation distributions that agree with expectations if we introduce central-satellite conformity, that is if we require that satellite occupation is conditioned by the presence of central galaxies of the same type.

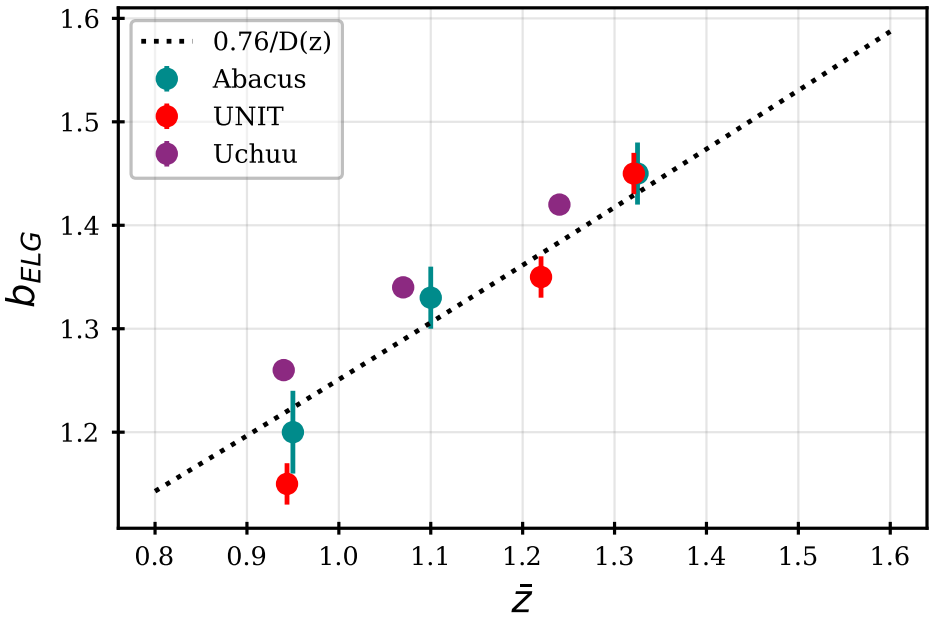

Figure 5.22: *Linear bias factor of the DESI One-Percent survey ELG sample as a function of redshift, as found in this work (green dots with errors) and in two companion DESI analyses which explored the galaxy-halo connection with a different methodology (red diamonds from J. Yu et al. (2023) and blue dots from F. Prada et al. (2023)). Errors for the green dots and red diamonds (resp. blue points) include (resp. do not include) statistical errors from the measured clustering. The dashed line is the predicted evolution of the inverse of the linear growth factor $D(z)$ (in the baseline cosmology of the present analysis) arbitrarily normalized.*

With or without conformity, whatever the prescriptions for central HOD, satellite velocity dispersion and secondary biases, when the standard NFW profile is used for satellites, our modelling of the measured clustering, although good, exhibit residuals with a reproducible pattern between 0.1 and 1 Mpc/h, showing that extra pairs of galaxies are lacking in our predictions for this region. A much better modelling is obtained with a modified NFW profile, allowing for ELG positioning outside of the halo virial radius, following a decreasing exponential law. With this prescription, we find that the measured ELG clustering clearly indicates that around 0.5% of ELGs reside in the outskirts of halos. The significant improvement in the goodness of fit with the modified satellite profile leaves the other parameters of the HOD modelling unchanged.

Moreover, with or without conformity, and whatever the model for central galaxies, we find that the satellite velocity dispersion must be enhanced w.r.t. that of dark matter particles to correctly reproduce the measured clustering. We show that this model cannot be disentangled from a coherent satellite infall velocity inside halos. The velocity bias reaches ~ 1.6 when our modified NFW profile for satellite positioning is used, and ~ 1.3 otherwise. Note that an increased velocity dispersion is coherent with the picture of ELGs residing in the outskirts of halos as recently-accreted sub-halos in these regions are expected to have higher velocities than the virial velocity of the halo.

The above findings are the main results of our work. With our best fitting HOD modelling, that is with central-satellite conformity, an extended NFW profile for satellite positioning and satellite velocity bias, the average halo mass of the ELG sample is $\log_{10}\langle M_h\rangle \sim 11.9$, the linear bias factor at a redshift of 1.1 is ~ 1.3 and the fraction of galaxies which are not alone in their halos (the so-called one-halo component) is $\sim 7\%$. The fraction of satellites is $\sim 3\%$ but is highly dependent on the details of the HOD modelling, and would be $\sim 12\%$ without central-satellite conformity.

We also investigate secondary biases and do not observe significant differences in our results when allowing for assembly bias as a function of halo concentration, local density or local density anisotropies. Although we report a slight improvement in the χ^2 value for assembly bias as a function of halo concentration, this effect has a small impact on clustering statistics (almost indistinguishable).

Spliting the ELG sample in two redshift bins, from 0.8 to 1.1 and 1.1 to 1.6 moderately changes the HOD and derived parameters. We do see a slight change across redshift in terms of halo mass populated with ELGs (0.08 dex), which we do not consider as significant.

The above results are obtained using simulation boxes from the AbacusSummit suite generated at the baseline Planck 2018 cosmology but we investigate two other cosmologies, with higher $N_{\rm eff}$ and lower σ_8 values respectively. These moderate change in the simulation cosmology have no significant impact on the one-halo term fraction and most HOD parameters, except for $\log_{10} M_c$ and f_{σ_v}, and thus for the predicted average halo mass of the sample which varies at most by 0.08 dex, which again cannot be considered as significant. This effect may be related to the different σ_8 values in the three cosmologies tested. However, despite the change of cosmology, the data clustering can be modeled with similar goodness of fit.

Finally, in the DESI framework, this study will be used to generate a large suite of accurate DESI-like mocks, varying the HOD models. These mocks will be useful to study the impact of observational systematics, test the corresponding mitigation algorithms and to study the impact of the complexity of galaxy formation and evolution on cosmological inference.

A Proxies for r_s and r_{vir} in the NFW profile

We further discuss our proxy choice for r_s and r_{vir} in the NFW profile used in our analysis. Figure 23 shows the predicted projected 2-point correlation function w_p on scales $r_p < 0.4$ Mpc/h for the same HOD model, changing the proxy for r_{vir} and r_s. For r_{vir}, we test two different choices, either r_{98}, the radius of a sphere enclosing 98% of the halo particles and r_{so}, the radius of a sphere containing the total halo mass M_{vir}, computed as the sum of the halo particle masses and expressed as an overdensity Δ:

$$r_{so} \equiv \left(\frac{3}{4\pi} \frac{M_{vir}}{\Delta \rho_c(z)} \right)^{1/3} \tag{20}$$

where ρ_c is the critical density. The overdensity is provided for each AbacusSummit snapshot, e.g. for the snapshot corresponding to the effective redshift $z = 1.1$ of the ELG sample, $\Delta = 223$. For the r_s proxy, we use the radius r_x of a sphere encompassing different percentages of the halo particles, with $x = 50, 33, 25$ and 10%. We compare the above predictions to that from a particle based mock (where the satellite assignment is based on particles inside the halo) for the same HOD model. The shaded grey region represents the $\pm 1\sigma$ measurement error for the actual DESI ELG sample in the redshift range between 0.8 and 1.6. From this comparison, the proxy that best reproduces the particle based mock corresponds to $r_{vir} = r_{98}$ and $r_s = r_{25}$.

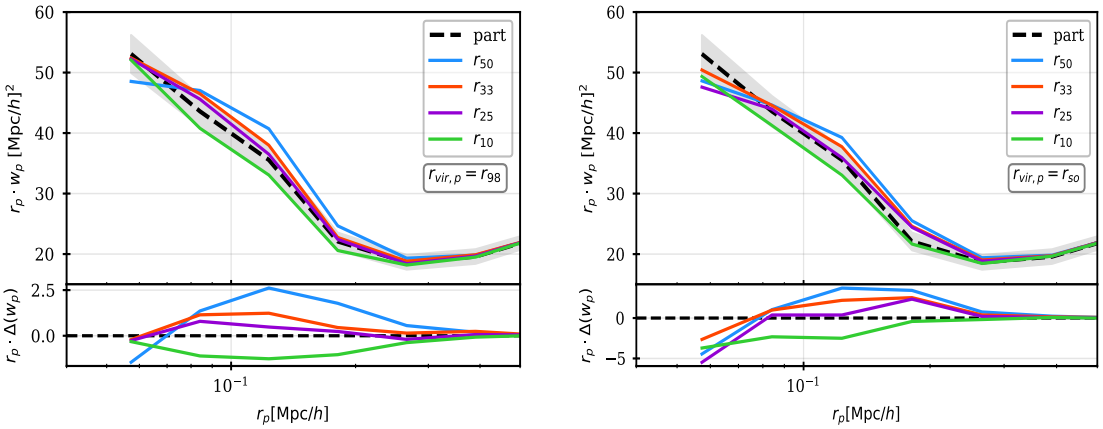

Figure 23: *Top: Predicted w_p clustering on scales $r_p < 0.4$ Mpc/h for the same HOD model, using as a proxy for r_{vir} either r_{98} (left) or r_{so} (right). Predictions for different proxies for r_s, corresponding to the radius of a sphere that contains $50, 33, 25$ and 10% of the halo particles (in blue, red, purple and green, respectively) are compared to the clustering of one mock where the satellite assignment is based on DM particles (dashed black line). Bottom: w_p difference between mocks with different r_s proxies and the particle based mock, multiplied by r_p. The shaded grey area corresponds to the $\pm 1\sigma$ error of DESI data as shown in Figure 5.3.*

B Contour plots of the mHMQ fits

Figures 24 and 25 show the contours obtained at final iteration in the Gaussian Process (GP) pipeline for mHMQ with and without strict conformity bias fits to the DESI One-Percent Survey ELG sample. Most contours are well enclosed our prior ranges. The notable exceptions are γ and $\log_{10} M_0$ for the conformity case. For $\log_{10} M_0$, the prior range is limited by the minimum halo mass available in the simulations, 10.86 and the fact that $\log_{10} M_0$ is not constrained if its value is below the minimum mass of halos that

can be populated with central galaxies. γ is degenerated with σ_M and has a weak impact on the shape of the HOD compared to σ_M.

The parameters we constrain the most are α and its degeneracy with A_s, $\log_{10} M_c$ and its degeneracy with σ_M, $f_{\sigma,v}$ and $\log_{10} M_0$ (only for the case without conformity for the latter two parameters).

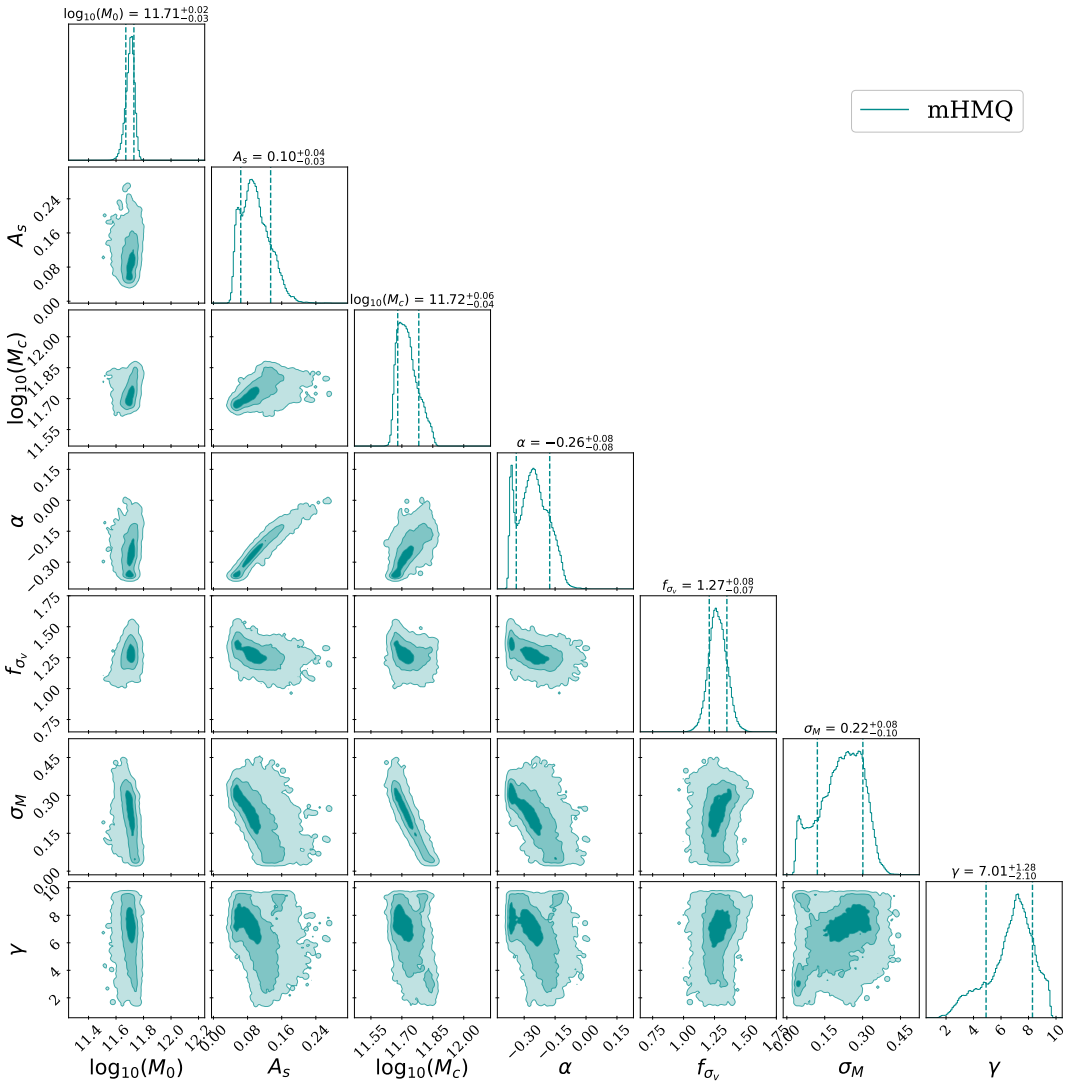

Figure 24: *Contours (at 1,2 and 3σ level) and marginalised 1D posteriors at final iteration obtained in the GP pipeline for the mHMQ fit to the One-Percent DESI survey ELG data for the whole redshift bin* $0.8 < z < 1.6$*, without conformity bias between central and satellite galaxies.*

Figure 26 shows the contours obtained at final iteration in the Gaussian Process (GP) pipeline for mHMQ fits to the DESI One-Percent Survey ELG sample in the two redshift bins considered in this paper, $0.8 < z < 1.1$ and $1.1 < z < 1.6$. Strict conformity is applied as well as our modified NFW profile for satellite positioning. The HOD parameters are well constrained in the lower redshift bin, while the constraints are less stringent in the higher bin, where we constrain only $\log_{10} M_c$ and its degeneracy with σ_M, α and its degeneracy with A_s, f_{σ_v} and two of the satellite profile parameters, f_{exp} and λ_{NFW}.

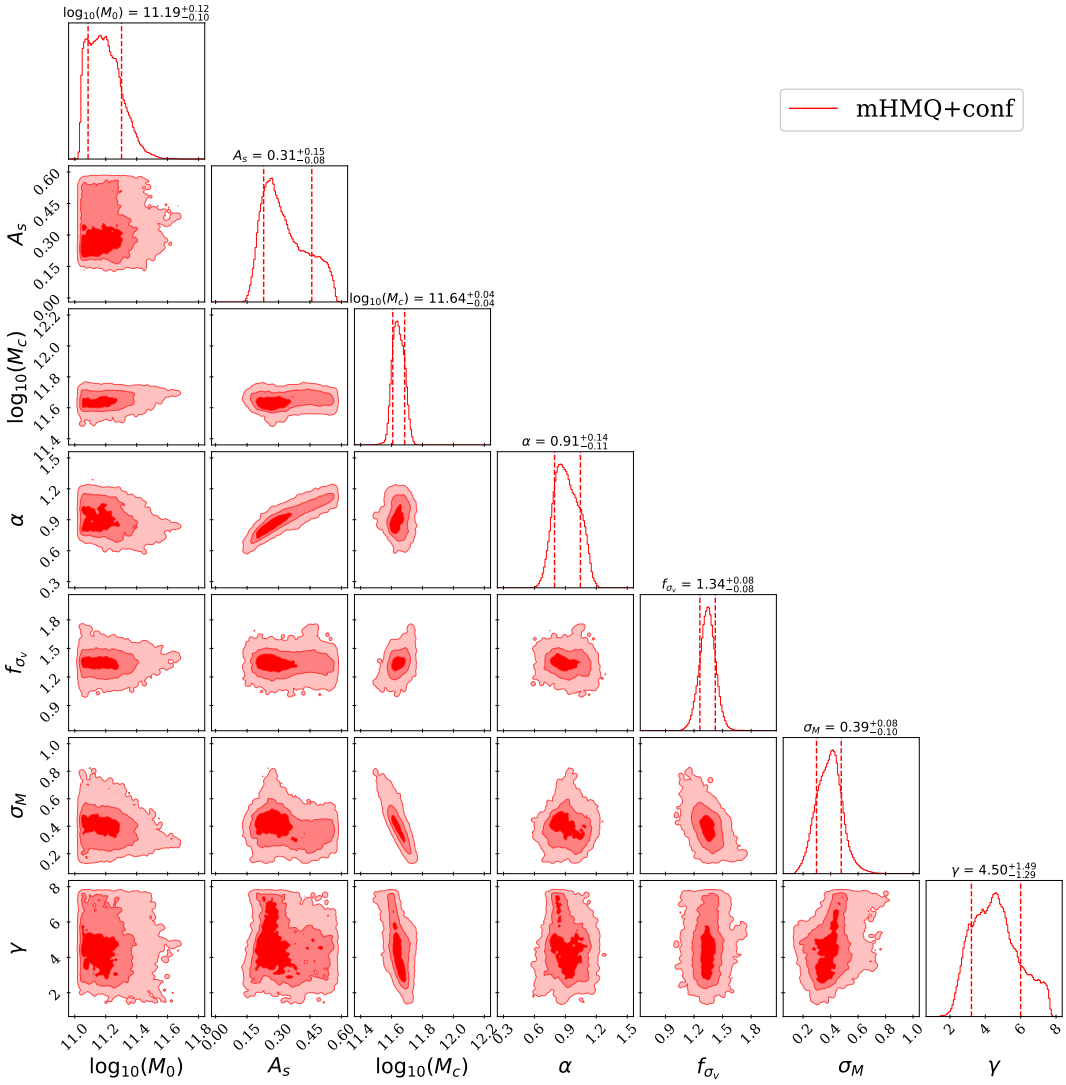

Figure 25: *Same as Figure 24 for the mHMQ model with strict conformity.*

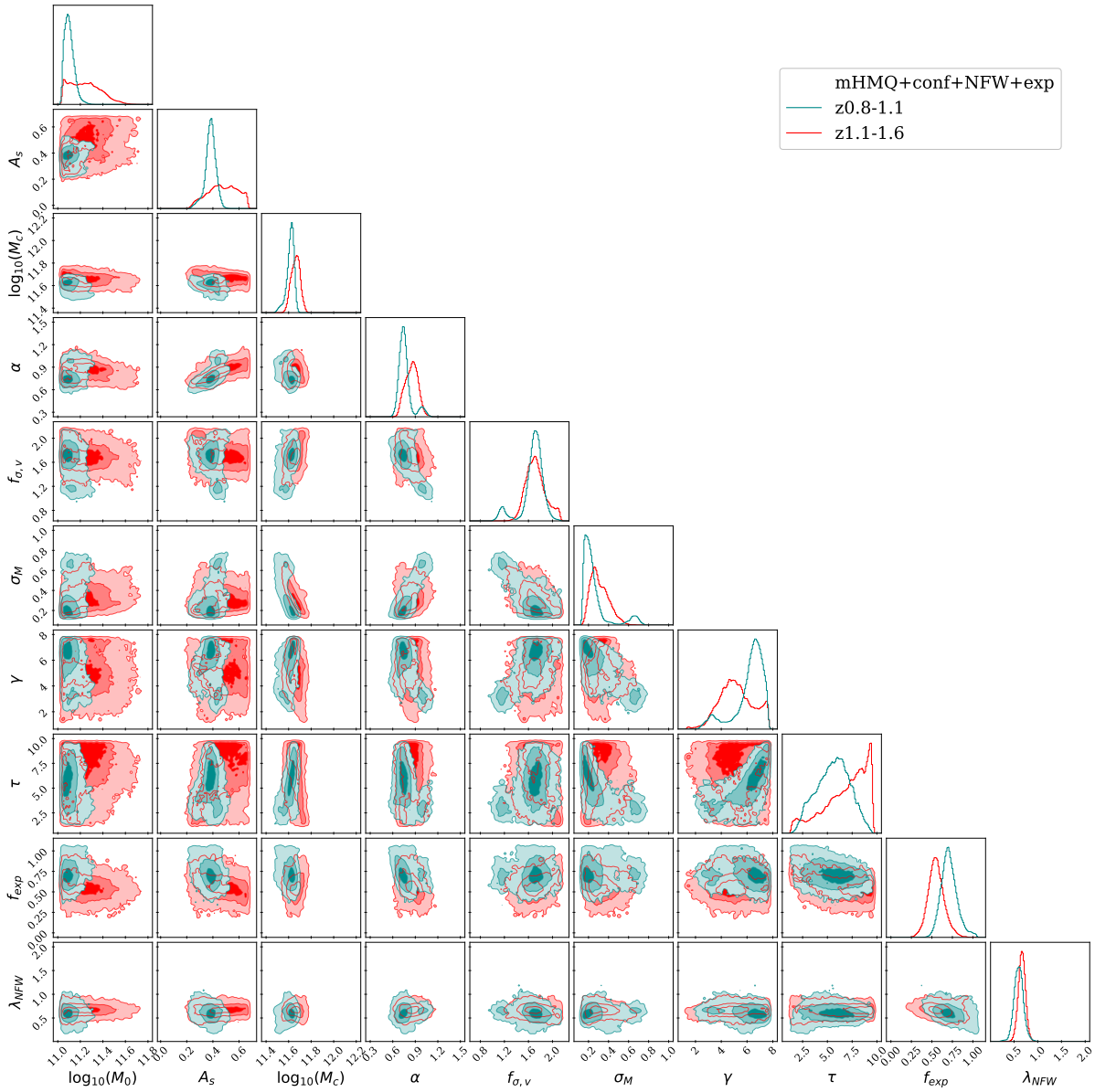

Figure 26: *Contours (at 1,2 and 3σ level) and marginalised 1D posteriors at final iteration obtained in the GP pipeline for the mHMQ fits to the One-Percent DESI survey ELG data with redshifts between* $0.8 < z < 1.1$ *in green and* $1.1 < z < 1.6$ *in red. The mHMQ model in this plot has strict conformity bias between central and satellite galaxies and a modified NFW profile for satellite positioning. This figure shows the small evolution of the HOD parameters with redshift.*

Bibliography

Alam, S., de Mattia, A., Tamone, A., et al. 2021, MNRAS, 504, 4667, doi: 10.1093/mnras/stab1150

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, MNRAS, 499, 5486, doi: 10.1093/mnras/staa2951

Bianchi, D., & Percival, W. J. 2017, MNRAS, 472, 1106, doi: 10.1093/mnras/stx2053

Blanton, M. R., & Berlind, A. A. 2007, ApJ, 664, 791, doi: 10.1086/512478

Contreras, S., Zehavi, I., Padilla, N., et al. 2019, MNRAS, 484, 1133, doi: 10.1093/mnras/stz018

de Mattia, A., & Ruhlmann-Kleider, V. 2019, J. Cosmology Astropart. Phys., 2019, 036, doi: 10.1088/1475-7516/2019/08/036

DESI Collaboration, Abareshi, B., Aguilar, J., et al. 2022, AJ, 164, 207, doi: 10.3847/1538-3881/ac882b

DESI collaboration et al. 2023a, submitted to AJ. https://arxiv.org/abs/2306.06307

—. 2023b, submitted to AJ. https://arxiv.org/abs/2306.06308

E. Schlafly et al. 2023, to be submitted. https://arxiv.org/abs/2306.06309

F. Prada et al. 2023, submitted to MNRAS. https://arxiv.org/abs/2306.06315

Favole, G., Comparat, J., Prada, F., et al. 2016, MNRAS, 461, 3421, doi: 10.1093/mnras/stw1483

Favole, G., Gonzalez-Perez, V., Stoppacher, D., et al. 2020, MNRAS, 497, 5432, doi: 10.1093/mnras/staa2292

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23, doi: 10.1086/174036

Gao, H., Jing, Y. P., Zheng, Y., & Xu, K. 2022, ApJ, 928, 10, doi: 10.3847/1538-4357/ac501b

Gao, L., & White, S. D. M. 2007, , 377, L5–L9, doi: 10.1111/j.1745-3933.2007.00292.x

Gonzalez-Perez, V., Comparat, J., Norberg, P., et al. 2018, MNRAS, 474, 4024, doi: 10.1093/mnras/stx2807

Gonzalez-Perez, V., Cui, W., Contreras, S., et al. 2020, MNRAS, 498, 1852, doi: 10.1093/mnras/staa2504

Guo, H., Zheng, Z., Zehavi, I., et al. 2015a, MNRAS, 446, 578, doi: 10.1093/mnras/stu2120

—. 2015b, MNRAS, 453, 4368, doi: 10.1093/mnras/stv1966

Guo, H., Yang, X., Raichoor, A., et al. 2019, ApJ, 871, 147, doi: 10.3847/1538-4357/aaf9ad

Guy, J., Bailey, S., Kremin, A., et al. 2023, AJ, 165, 144, doi: 10.3847/1538-3881/acb212

H. Gao et al. 2023, submitted to ApJ. https://arxiv.org/abs/2306.06317

Hadzhiyska, B., Eisenstein, D., Bose, S., Garrison, L. H., & Maksimova, N. 2022a, MNRAS, 509, 501, doi: 10.1093/mnras/stab2980

Hadzhiyska, B., Tacchella, S., Bose, S., & Eisenstein, D. J. 2021, MNRAS, 502, 3599, doi: 10.1093/mnras/stab243

Hadzhiyska, B., Hernquist, L., Eisenstein, D., et al. 2022b, submitted to MNRAS, doi: 10.48550/arXiv.2210.10068

Hadzhiyska, B., Eisenstein, D., Hernquist, L., et al. 2022c, submitted to MNRAS, doi: 10.48550/arXiv.2210.10072

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399, doi: 10.1051/0004-6361:20066170

J. Yu et al. 2023, submitted to MNRAS. https://arxiv.org/abs/2306.06313

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, ApJ, 609, 35, doi: 10.1086/420959

Kullback, S., & Leibler, R. 1951, Annals Math. Statist., 22, 79

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64, doi: 10.1086/172900

Lin, S., Tinker, J. L., Blanton, M. R., et al. 2023, arXiv e-prints, doi: 10.48550/arXiv.2302.09199

Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., et al. 2021, MNRAS, 508, 4017, doi: 10.1093/mnras/stab2484

Mohammad, F. G., & Percival, W. J. 2022, MNRAS, 514, 1289, doi: 10.1093/mnras/stac1458

Mohammad, F. G., Percival, W. J., Seo, H.-J., et al. 2020, MNRAS, 498, 128, doi: 10.1093/mnras/staa2344

Myers, A. D., Moustakas, J., Bailey, S., et al. 2023, AJ, 165, 50, doi: 10.3847/1538-3881/aca5f9

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563, doi: 10.1086/177173

Okumura, T., Hayashi, M., Chiu, I. N., et al. 2021, PASJ, 73, 1186, doi: 10.1093/pasj/psab068

Orsi, Á. A., & Angulo, R. E. 2018, MNRAS, 475, 2530, doi: 10.1093/mnras/stx3349

Percival, W. J., & Bianchi, D. 2017, MNRAS, 472, L40, doi: 10.1093/mnrasl/slx135

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6, doi: 10.1051/0004-6361/201833910

Raichoor, A., Moustakas, J., Newman, J. A., et al. 2023, AJ, 165, 126, doi: 10.3847/1538-3881/acb213

Rocher, A., Ruhlmann-Kleider, V., Burtin, E., et al. 2023, arXiv e-prints, arXiv:2306.06319, doi: 10.48550/arXiv.2306.06319

Silber, J. H., Fagrelius, P., Fanning, K., et al. 2023, AJ, 165, 9, doi: 10.3847/1538-3881/ac9ab1

Sinha, M., & Garrison, L. H. 2020, MNRAS, 491, 3022, doi: 10.1093/mnras/stz3157

Tinker, J. L., Leauthaud, A., Bundy, K., et al. 2013, ApJ, 778, 93, doi: 10.1088/0004-637X/778/2/93

Weinmann, S. M., van den Bosch, F. C., Yang, X., & Mo, H. J. 2006, MNRAS, 366, 2, doi: 10.1111/j.1365-2966.2005.09865.x

Wetzel, A. R., Tinker, J. L., & Conroy, C. 2012, MNRAS, 424, 232, doi: 10.1111/j.1365-2966.2012.21188.x

Yuan, S., Garrison, L. H., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022, MNRAS, 510, 3301, doi: 10.1093/mnras/stab3355

Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, ApJ, 633, 791, doi: 10.1086/466510

6

Conclusions & Prospects

Over the past decade, the large scale structures of the Universe have become one of the most promising cosmological probes of dark energy and gravity models, through the precise determination of the baryon acoustic oscillation scale and the measure of the non-linear growth rate of structure from redshift space distortions (Alam et al., 2021).

Chapter 1 presents the standard cosmological model ΛCDM and the different surveys that can probe the nature of the cosmic acceleration and constrain cosmological parameters. Among the different approach, large structure of the Universe is a key cosmological probe, clustering measurements from spectroscopic galaxy surveys allow cosmological parameters to be constrained through measurements of the baryon acoustic oscillation (BAO) scale and from the full shape analyses of 2-point statistics which translate into constraints on both BAO scales and the linear growth rate of structure $f\sigma_8$ through redshift space distortions. The growth rate can be used as a direct test of the underlying theory of gravitation, general relativity (GR) in the standard cosmological model. At small scales, clustering measurements are invaluable to study the galaxy-halo connection and to provide precise measurements of $f\sigma_8$. They also allow realistic mock catalogues to be produced that are used to prepare large scale analyses and to assess their systematic errors related to the complexity of galaxy formation and evolution.

The core of my PhD work was to provide accurate model of the small-scale clustering of the ELG sample collected by the Dark Energy Spectroscopic Instrument (DESI) survey. Chapter 2 gives an overview of the DESI, which will over 5 years, observe $\sim$ 40M galaxies and quasars over $0.1 < z < 3.5$, including $\sim$ 17M ELGs in $0.6 < z < 1.6$, to strongly constrain dark energy models. In only two months of early observations, DESI observed 267k ELGs which is the largest ELG spectroscopic sample to date. Thanks to its high fibre arrangement completeness, this early data sample –The DESI One-Percent survey– allows precise clustering measurements down to very small scales, 0.03 Mpc/h. During my PhD, I actively took part of the DESI collaboration. In particular, I have made a major contribution to the generation of mocks (simulated galaxy catalogues) for the DESI collaboration, studying the galaxy-halo connection of the ELGs sample. I worked closely with the data, participating in the creation of the ELGs data clustering catalogues, and carry out numerous tests to check systematic effects.

The Chapter 3 of this thesis gives an overview of theoretical formalism of structure formation in the linear and non-linear regime, describes the different simulation techniques to simulate the non-linear evolution of the dark matter field (and baryons). In a second part, I introduce the galaxy-halo connection and the various modelling techniques, either based on semi-analytical models or on empirical relations used to populate dark matter halos from N-body simulations. In this PhD work I used the Halo Occupation Distribution (HOD) formalism, to study the galaxy-halo connection of the ELG sample from the DESI One-Percent survey.

Then, in Chapter 4 I developed a novel and promising HOD Gaussian process based fitting pipeline to reproduce the small scale clustering in an accurate and efficient way and use it to study the DESI ELG sample. I first developed an efficient multi-threaded code to generate galaxy mocks using HOD models suitable for different matter tracers (LRGs, ELGs, QSOs). These models differ in the functions describing the probability to populate halos with central or satellite galaxies, but by default all probabilities are only functions of the halo mass. This pipeline was further complemented by an iterative fitting procedure based on Gaussian processes to create a surrogate model of the expected likelihood multidimensional function. The whole procedure was tested on simulations, showing that HOD parameters for clustering 2-point statistics of a DESI ELG-like sample are recovered with better precision than expected from fitting DESI data, while evaluating ≈ 100 times less points in the HOD parameters space than standard techniques based on Monte Carlo Markov Chains.

In Chapter 5, we apply this method to the DESI ELG sample from the One-Percent survey. DESI ELGs shows an unexpected strong clustering signal at small scales $r_p < 0.3$ Mpc/h. I investigated potential sources of systematics (foreground effects, blending) which could have impacted the clustering at these scales, but found the signal robust to such contaminations. I demonstrated that physically motivated ELG HOD models for central and satellite galaxies cannot reproduce this behaviour. To reproduce the small scale clustering of ELGs, I then demonstrated that we must introduce close pairs of galaxies in low mass halos ($< 10^{12} M_\odot$), which was unexpected based on previous ELGs studies Avila et al.

(2020), Gonzalez-Perez et al. (2018). To propose a physical model to the clustering excess apparent at small scales, I investigated the effect of a potential conformity bias. Conformity adds prior information to the satellite probability function depending on whether the DM halo already hosts a central galaxy. This property slightly improves the modelling of the small-scale clustering data while keeping the satellite halo mass dependence in agreement with physical expectations. Our results are in agreement with what very recent hydrodynamical studies find on conformity between ELGs central and satellite galaxies Hadzhiyska et al. (2021). We also report a satellite velocity dispersions about $\sim 50\%$ higher than that of dark matter particles. Other extensions to standard HOD models (eg. secondary biases, changing fiducial cosmology) bring no significant change to our results, except when we allow satellite ELGs to lie outside of the halo virial radius. It is with this assumption that we obtain the best model of the measured clustering, corresponding to $\sim 0.5\%$ of the sample ELG sample (and $\sim 12\%$ of the satellites) residing in the halo outskirts. This work was submitted as part of the Early Data Release of the DESI collaboration.

As part of my work on HOD ELG fitting, I provided the DESI collaboration with a set of official mocks representative of the ELG clustering as measured in the One-Percent survey. These mocks are currently used within DESI to test standard BAO and RSD clustering analyses for DESI Year 1 data release and I participate in the discussion and the analysis. Besides the above work, I contributed to many efforts within the DESI collaboration that should allow me to be a continuing participant of the DESI collaboration. I participated in the study of possible systematic effects in the ELG spectroscopic redshift success rate measurement as a function of spectroscopic observing conditions which will also lead to a supporting paper for DESI Year 1 analysis. During the survey validation process (prior to run the main survey), I also contributed to the visual inspection of galaxy spectra to validate the pipeline of redshift determination Lan et al. (2022).

Future perspectives in cosmology

I hereafter present what could be my future activities in cosmology. The guidelines of my future work will focus on how to improve cosmological measurement using information from the smallest scales.

➢ Joint inference of HOD and cosmological parameters

Small scale clustering studies will bring numerous promising applications for cosmological analysis. **Various recent studies have investigated the combined inference of the galaxy-halo connection and cosmological parameters using emulator techniques (mostly based on Gaussian processes) DeRose et al. (2019), Lange et al. (2022), Yuan et al. (2022a).** Yuan et al. (2022a) showed an improvement in the measure of the growth rate by 30–40% compared to previous studies, by adding non-linear scales using an emulator and expect these constraints should tighten by at least 50% thanks to the statistical power of DESI Year 1.

In our work Rocher et al. (2023), we show that even at fixed cosmology, cosmic variance, stochasticity and degeneracies between HOD parameters complicate the measurement of HOD parameters. Adding secondary biases in the HOD model can introduce some degeneracy and bias the results when performing the combined inference of HOD and cosmological parameters Cuesta-Lazaro et al. (2022). One of the major challenges will be to correctly understand and model the degeneracies between HOD and cosmological parameters to avoid biases in the results. I intend to expand this work and perform combined inference of HOD and cosmological parameters. Based on my expertise of DESI data and HOD modelling, I plan to use the 85 available N-body simulations at different cosmologies to construct an emulator of the clustering statistics (2PCF) in order to perform combined inference of HOD and cosmological parameters for DESI Year 1 ELG data. I want to investigate several aspects of the ELG HOD modelling, including secondary biases, performing a scale-dependent study and testing the emulator robustness and accuracy.

This study is intend to be perform using Year 1 DESI data and expected to be done for the Y1 data release. It can be done during the time available for independent projects and will be also useful for other studies related with weak lensing analysis.

➢ Refine the galaxy-halo connection using hydrodynamical simulations

HOD models are an efficient way to describe the connection between tracers and dark-matter halos. However it neglects baryonic processes that arise from galaxy formation and evolution. In that sense, hydrodynamical simulations which resolve simultaneously dark matter and baryonic physics, are a natural path to improve our understanding of the galaxy-halo connection. Studies from the Illustris-TNG simulations show very promising results for the galaxy halo connection of ELGs Hadzhiyska et al. (2021), Yuan et al. (2022b). Recently, using the State-of-the-art hydrodynamical simulation, Millenium-TNG Hernández-Aguayo et al. (2022), the conformity between ELGs central and satellites was evidenced in Hadzhiyska et al. (2022), **which support our findings in DESI data.** The results of hydrosimulations provide us with new lines of investigation, i.e. super-Poisson distribution of satellites, better modelling of radial profile, radial velocity profile or velocity dispersion profile of ELG satellites. Implementing these dependencies in our model and applying it to fit the DESI data will greatly improve our understanding of the physical processes at play in the galaxy-halo connection.

➢ Extracting information beyond 2-point statistics

Non standard statistics can bring relevant additional information to constrain cosmological parameters and the galaxy-halo connection. One example is the use of higher order statistics in the correlation function. With the results from early ELG DESI data, i.e. the need for satellite pairs at small separations, the use of the 3PCF should improve contraints on the galaxy-halo connection, and I plan to take a significant role in these studies. Another very interesting result when combining HOD+cosmological parameters is that of Storey-Fisher et al. (2022), which finds that adding non standard statistics, i.e. density-dependent correlation function, improves the precision of the measurement by a factor of 2. Complementary to my involvement in the 3PCF, I intend to use and develop these techniques to improve the amount of information that we can get from data.

A different approach to improve cosmological constraints is to take advantage of cross-correlations between galaxy tracers. In DESI, first results from multi-tracer analyses for RSD or BAO studies show improvement of 20% on the growth rate estimation for low redshift galaxies and we can expect more from optimising the analysis. To go in that direction, I plan to extend my studies of the ELG halo connection by adding the cross-correlation with other DESI tracers and participate to the creation of high fidelity mocks to prepare the cosmological analyses with the DESI Year 3 sample expected at the end of 2024. My expertise will be highly valuable and I would like to study in more details the issues related to the conformity effects for the different tracers. This study will be also a critical step for performing joint modelling of galaxy clustering and galaxy-galaxy lensing as I describe in the following.

➢ Joint galaxy clustering and weak lensing analysis

In the near future, upcoming experiments like the Vera Rubin Observatory LSST or EUCLID space telescope will bring a massive amount of photometric data and will give constraints on the dark energy equation of state 10 times tighter than those from the latest results from weak lensing (WL) with the Dark Energy Survey (DES) Collaboration (2005), that is sub % precision. Cosmological information with large photometric surveys is extracted using the 3x2 point correlation function (cosmic shear x galaxy clustering x cross-correlation galaxy-galaxy lensing). Latest results from DES Collaboration et al. (2022) suggest a mild tension with Planck and spectroscopic results that has been highly discussed in the community. One of the potential investigations for this tension that I wish to undertake concerns the lensing-is-low effect Lange et al. (2021), Leauthaud et al. (2017), Yuan et al. (2021): accurate small

scale measurements of galaxy clustering provide incorrect galaxy-galaxy lensing predictions compared to observations. An explanation for this effect may reside in secondary biases known as assembly bias, i.e the fact that in addition to halo mass, the clustering of halos may also depend on environment, concentration or merger history Gao & White (2007), Jespersen et al. (2022), Lacerna & Padilla (2011).

Furthermore, WL analyses usually assume a linear relation between galaxy tracers and the total matter field (linear galaxy bias). Therefore, they cannot exploit small scales, which contain a lot of cosmological information, so better understanding and modelling of small scales (including baryonic effects) is needed to improve the results.

Building on my expertise on HOD modelling and spectroscopic surveys, I intend to explore how implementations of assembly bias in HOD models can improve the galaxy-galaxy lensing prediction. Having information from both clustering (spectroscopic) and future weak lensing data will shrink the constraints and break degeneracies between HOD and cosmological parameters Delgado et al. (2022). Trying to understand the discrepancy in clustering amplitude by simultaneously fitting galaxy-halo connection models to both the galaxy-galaxy lensing and galaxy clustering will be challenging. We will need to develop highly realistic simulations that reproduce both galaxy clustering and galaxy-galaxy lensing. These simulations will be also very useful to check systematic effects. The first science results of such analysis will be pioneering work for cross experiment analyses, and I would be very delighted to take part.

Last words

During my Ph.D., I developed an expertise in large scale structure cosmology, using spectroscopic surveys to constrain cosmological parameters from the halo scale to the largest scales. Within the DESI collaboration I communicated with many collaborators around the world (USA, France, Spain, Korea...). I acquired a broad understanding of the many different LSS probes and an expertise with the DESI instrument, which is highly valuable as tensions on dark energy measurements between LSS, cosmic microwave background and supernova measurements have arisen during the last decade. With the new generation of galaxy surveys for stage IV cosmology, and the increase of computational resources, we are entering the era of statistical precision cosmology. It will be a very exciting time for cosmological research !

Finally, I would particularly like to thank my thesis supervisors, Vanina and Etienne, for their patience and unfailing support, without which this thesis would not have been the same.

Bibliography

Alam, S., Aubert, M., Avila, S., et al. 2021, Physical Review D, 103, 083533, doi: 10.1103/PhysRevD.103.083533

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, Monthly Notices of the Royal Astronomical Society, 499, 5486–5507, doi: 10.1093/mnras/staa2951

Collaboration, D., Abbott, T. M. C., Aguena, M., et al. 2022, Physical Review D, 105, 023520, doi: 10.1103/PhysRevD.105.023520

Collaboration, T. D. E. S. 2005, arXiv e-prints, astro. https://ui.adsabs.harvard.edu/abs/2005astro.ph.10346T

Cuesta-Lazaro, C., Nishimichi, T., Kobayashi, Y., et al. 2022. http://arxiv.org/abs/2208.05218

Delgado, A. M., Wadekar, D., Hadzhiyska, B., et al. 2022, Monthly Notices of the Royal Astronomical Society, 515, 2733–2746, doi: 10.1093/mnras/stac1951

DeRose, J., Wechsler, R. H., Tinker, J. L., et al. 2019, The Astrophysical Journal, 875, 69, doi: 10.3847/1538-4357/ab1085

Gao, L., & White, S. D. M. 2007, Monthly Notices of the Royal Astronomical Society: Letters, 377, L5–L9, doi: 10.1111/j.1745-3933.2007.00292.x

Gonzalez-Perez, V., Comparat, J., Norberg, P., et al. 2018, Monthly Notices of the Royal Astronomical Society, 474, 4024–4038, doi: 10.1093/mnras/stx2807

Hadzhiyska, B., Tacchella, S., Bose, S., & Eisenstein, D. J. 2021, Monthly Notices of the Royal Astronomical Society, 502, 3599–3617, doi: 10.1093/mnras/stab243

Hadzhiyska, B., Hernquist, L., Eisenstein, D., et al. 2022. http://arxiv.org/abs/2210.10068

Hernández-Aguayo, C., Springel, V., Pakmor, R., et al. 2022, doi: 10.48550/arXiv.2210.10059

Jespersen, C. K., Cranmer, M., Melchior, P., et al. 2022, doi: 10.48550/arXiv.2210.13473

Lacerna, I., & Padilla, N. 2011, Monthly Notices of the Royal Astronomical Society, 412, 1283–1294, doi: 10.1111/j.1365-2966.2010.17988.x

Lan, T.-W., Tojeiro, R., Armengaud, E., et al. 2022, doi: 10.48550/arXiv.2208.08516

Lange, J. U., Hearin, A. P., Leauthaud, A., et al. 2022, Monthly Notices of the Royal Astronomical Society, 509, 1779–1804, doi: 10.1093/mnras/stab3111

Lange, J. U., Leauthaud, A., Singh, S., et al. 2021, Monthly Notices of the Royal Astronomical Society, 502, 2074–2086, doi: 10.1093/mnras/stab189

Leauthaud, A., Saito, S., Hilbert, S., et al. 2017, Monthly Notices of the Royal Astronomical Society, 467, 3024–3047, doi: 10.1093/mnras/stx258

Rocher, A., Ruhlmann-Kleider, V., Burtin, E., & de Mattia, A. 2023. http://arxiv.org/abs/2302.07056

Storey-Fisher, K., Tinker, J., Zhai, Z., et al. 2022. http://arxiv.org/abs/2210.03203

Yuan, S., Garrison, L. H., Eisenstein, D. J., & Wechsler, R. H. 2022a, Monthly Notices of the Royal Astronomical Society, 515, 871–896, doi: 10.1093/mnras/stac1830

Yuan, S., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022b, Monthly Notices of the Royal Astronomical Society, 512, 5793–5811, doi: 10.1093/mnras/stac830

Yuan, S., Hadzhiyska, B., Bose, S., Eisenstein, D. J., & Guo, H. 2021, Monthly Notices of the Royal Astronomical Society, 502, 3582–3598, doi: 10.1093/mnras/stab235

7

Résumé en français

– Antoine Rocher, *2023*

Contents

7.1 Introduction à la cosmologie des grandes structures

La cosmologie est une science extraordinaire. C'est la phrase que Vanina, ma directrice, m'a dite lorsque je lui ai demandé quelle phrase elle souhaitait voir figurer dans mon manuscrit. Je suis d'accord avec elle et je vais essayer d'expliquer pourquoi. La cosmologie (du grec : *kosmos*, Univers et logos, *théorie*) est une science fondamentale qui vise à répondre à des questions simples dont les réponses sont difficiles.

L'Univers est en constante évolution, comme la connaissance. Il n'est pas immuable, mais évolue avec le temps, les observations et le progrès. Depuis un siècle, avec la théorie de la relativité générale d'Einstein, la découverte de l'expansion de l'Univers par Edwin Hubble et Georges Lemaître, notre compréhension de l'Univers a changé. De l'hypothèse d'Einstein d'un Univers statique à l'observation d'une expansion accélérée, nos connaissances ont continué à se développer jusqu'à aujourd'hui et continueront à le faire grâce à de nouveaux moyens d'observation et à de nouvelles découvertes scientifiques.

L'Univers est homogène et isotrope à grande échelle. C'est le premier principe cosmologique. Il signifie que son apparence générale ne dépend pas de la position de l'observateur ni de la direction d'observation. Cela peut être difficile à admettre, car nous voyons des milliards d'étoiles dans notre galaxie et des milliards de galaxies à l'extérieur. Nous pouvons même observer des super-structures, telles que des amas de galaxies ou des vides cosmiques géants qui peuvent atteindre quelques dizaines de mégaparsec (Mpc) et former, au total, une toile de nœuds et de filaments, que l'on appelle la *toile cosmique*. Le parsec (pc) ou même le méga-pc (Mpc) est l'unité standard de distance en cosmologie. Le parsec est défini par la distance des objets astronomiques (c'est-à-dire les étoiles), qui ont un déplacement angulaire sur le ciel de 1 seconde d'arc (") lorsque la Terre se déplace sur la moitié d'une orbite du Soleil (également connu sous le nom de parallaxe). Un parsec représente environ 3,26 années-lumière, soit environ 31 billions (10^{12}) de kilomètres. Pour donner un ordre d'idée, si 1 km correspondait à la taille d'un atome, 1 parsec serait la distance Terre-Lune ! Mais ces objets géants sont petits par rapport à la taille de l'Univers observable ($\sim$ 14,300 Mpc pour la distance comobile entre la Terre et le bord de l'Univers observable). À cette échelle l'Univers est le même partout (homogène) dans toutes les directions (isotrope). La principale preuve observable de ce principe est le fond diffus cosmologique, un rayonnement homogène et isotrope représentant la première lumière de l'Univers.

Aujourd'hui, le modèle cosmologique standard, ΛCDM, décrit le contenu et la dynamique de l'Univers. Ce modèle est basé sur seulement six paramètres, la gravité est régie par la relativité générale (RG) et différentes contributions constituent le contenu énergétique de l'Univers aujourd'hui, comme le montre Figure 7.1:

- **La matière baryonique** : Elle représente la matière ordinaire, celle que nous pouvons voir, c'est-à-dire les planètes, les étoiles, les galaxies... et ne représente que $\sim$ 5% du contenu énergétique de l'Univers aujourd'hui. L'autre partie $\sim$ 95% est la face cachée de l'Univers, celle que nous ne pouvons que deviner par son effet sur la matière baryonique.

- **Matière noire froide** : c'est la composante majeure de la masse de l'Univers – $\sim$ **85**% de la masse – et $\sim$ 25% du contenu énergétique de l'Univers aujourd'hui. Détectée uniquement par son impact à travers les effets gravitationnels, sa nature est encore inconnue. Il pourrait s'agir de particules au-delà du modèle standard de la physique des particules ou d'objets astrophysiques qui doivent être formés avant la nucléosynthèse primordiale (par exemple les trous noirs primordiaux).

- **L'énergie noire** : C'est la composante principale du contenu énergétique de l'Univers aujourd'hui, $\sim$ 70%. C'est une forme d'énergie qui est responsable de l'accélération tardive de l'expansion de l'Univers, dont l'origine précise est inconnue.

- **La radiation** : Elle englobe toutes les espèces relativistes (c.-à-d. les photons, les neutrinos) provenant de l'Univers primitif chaud et dense. Aujourd'hui, sa contribution est négligeable.

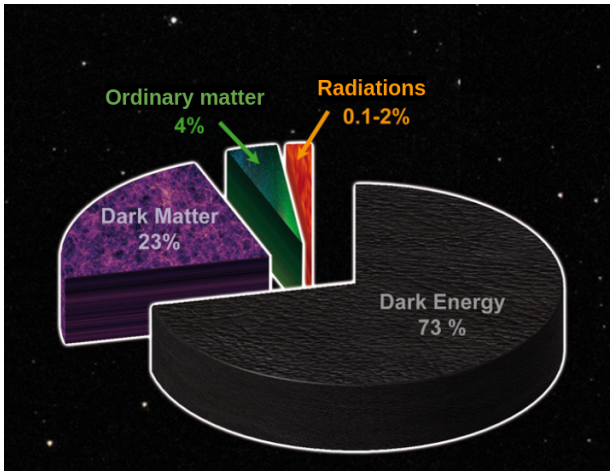

Figure 7.1: *Contenu énergétique de l'Univers aujourd'hui. Il est principalement dominé par une forme d'énergie inconnue appelée énergie noire* $\sim 70\%$. *Les autres* $\sim 30\%$ *sont des composants de la matière :* $\sim 25\%$ *de matière noire froide (CDM) et* $\sim 5\%$ *de matière baryonique (ou ordinaire). Une partie négligeable* ($< 10^{-4}$) *du budget énergétique provient du rayonnement, c'est-à-dire des photons et des neutrinos relativistes, mais au début de l'Univers, c'était la partie dominante. Cette figure est adaptée du site web: https://www.spacecentre.co.uk/news/space-now-blog/what-s-in-the-dark/*

Le modèle ΛCDM peut décrire l'Univers depuis les premiers instants, lorsque les baryons et la matière noire ont été condensés dans un plasma très chaud, jusqu'à la formation des galaxies et des structures à grande échelle que nous observons aujourd'hui. Il repose sur trois contraintes observationnelles fortes, appelées les trois piliers cosmologiques :

- **l'expansion de l'Univers** : une récession des galaxies à une vitesse proportionnelle à leur distance par rapport à nous,
- **la nucleosynthèse primordiale** : ceci explique l'abondance des éléments chimiques dans l'Univers,
- **le fond diffus cosmologique** : c'est la première lumière de l'Univers, émise $\sim 380,000$ ans après le Big Bang.

Pour décrire l'évolution et les propriétés de l'Univers, le modèle ΛCDM dépend de seulement six paramètres libres :

- A_s : l'amplitude du spectre de puissance primordial,
- n_s : l'indice spectral du spectre de puissance primordial,
- θ_* : l'échelle angulaire sur le ciel correspondant à l'horizon sonore en mouvement lors de la recombinaison,
- $\Omega_b h^2$: la densité de baryons dans l'Univers aujourd'hui,
- $\Omega_{\rm cdm} h^2$: la densité de matière noire dans l'Univers aujourd'hui,
- τ : la profondeur optique lors de la réionisation.

La métrique de **Friedmann-Lemaître–Robertson–Walker** (FLRW) permet de décrire mathématiquement un Univers homogène, isotrope et en expansion. Elle est définie par:

$$ds^2 = dt^2 - a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right] \tag{7.1}$$

Cette métrique définit la géométrie de l'Univers dans un espace à trois dimensions spatiales (en coordonnéed sphériques $[r, \theta, \phi]$) et une dimension temporelle t, le temps cosmique. La partie radiale de la métrique peut être affectée par la courbure de l'espace-temps k. L'Univers peut être ouvert, plat ou fermé (respectivement $k < 0$, $k = 0$, $k > 0$).

Dans un Univers en expansion, nous considérons les positions des objets comme fixes dans un espace en expansion et nous définissons un facteur d'échelle $a(t)$ qui décrit l'expansion de l'espace lui-même à un moment donné. En pratique, les objets conservent les mêmes coordonnées (appelées *coordonnées comobiles*) à tout moment, de sorte que leur *distance comobile* restera la même, tandis que leur *distance propre* (ou distance physique), c'est-à-dire la distance qui serait mesurée à un moment donné à l'aide d'une règle rigide, augmentera en raison de l'expansion de l'Univers. La distance d'un objet est caractérisée grâce au *redshift* (ou décalage vers le rouge) z[1], qui correspond au rapport entre la longueur d'onde observée λ_{obs} et la longueur d'onde d'émission λ_e d'un objet:

$$1 + z = \frac{\lambda_{obs}}{\lambda_e} = \frac{1}{a(t)} \tag{7.2}$$

La mesure du redshift est utilisée pour déterminer la distance des galaxies et donc permet de cartographier les galaxies dans l'Univers. Étudier la distribution spatiale des galaxies permet de connaître la composition de l'Univers et de contraindre les modèles cosmologique comme on peut le voir sur la Figure 7.2. En cosmologie, on décrit la distribution des galaxies (ou de la matière) en définissant un champ de contraste de densité $\delta(\mathbf{x})$ défini par :

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \langle\rho\rangle}{\langle\rho\rangle} \tag{7.3}$$

où $\langle\rho\rangle$ correspond à la densité moyenne de l'Univers.

Comme les baryons ne représentent que $\sim 25\%$ de la masse totale de l'Univers (l'autre partie étant de la matière noire invisible), les galaxies suivent la distribution de la matière dans l'Univers et donc la distribution de matière noire. Elles résident principalement au centre de *halos de matière noire*, qui sont des régions surdenses de la toile cosmique. Les galaxies sont donc des *traceurs* de la distribution de matière dans l'Univers, ce qui signifie que là où il y a une galaxie, il doit y avoir aussi de la matière noire. Cependant, le fait qu'il n'y ait pas de galaxies ne signifie pas qu'il n'y a pas de matière noire. Ainsi, le champ de galaxies est *biaisé* par rapport au champ de matière totale. La prescription standard pour modéliser cet effet est la suivante :

$$\delta_g = b_g \delta_m \tag{7.4}$$

ou b_g est le biais linéaire des galaxies et δ_g, δ_m sont les contrastes de densité des galaxies et de la matière. La distribution spatiale des galaxies peut être décrite en statistique par la fonction de corrélation à deux points (2PCF) $\xi(r)$:

$$\xi(r) = \langle\delta(\mathbf{x})\delta(\mathbf{x} + \mathbf{r})\rangle \tag{7.5}$$

$\xi(r)$ mesure l'excès de probabilité, par rapport à une probabilité aléatoire, que deux galaxies soient séparées par une distance r. La distribution spatiale des galaxies dépend de la composition de l'Univers. Son étude permet donc de connaître la composition de l'Univers et de contraindre les modèles cosmologique comme on peut le voir sur la Figure 7.2.

Au lieu de calculer la 2PCF en fonction de la séparation entre 2 galaxies r, nous pouvons décomposer la distance r en deux composantes, s et μ, où s est la distance observée entre les paires de galaxies et

[1]Les longueur d'ondes émisse par un object lointain sont étendues et donc décalées vers le rouge à cause de l'expansion de l'Univers

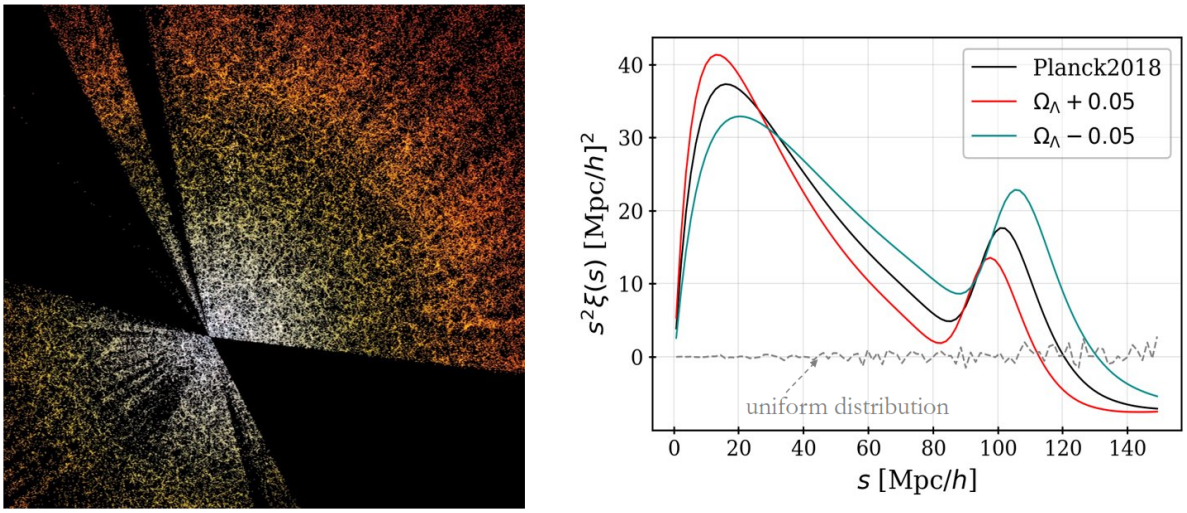

Figure 7.2: *Gauche: Distribution spatiale des galaxies vue par DESI, l'observateur est au centre et chaque point représente une galaxie. Droite: Exemple de fonction de corrélation à 2 points; la courbe noire correspond au modèle cosmologique ajusté sur les données du fond diffus cosmologique mesurées par Planck (Planck Collaboration et al., 2020). Les courbes rouge et noire montrent comment la 2PCF (donc la distribution spatiale des galaxies) changent en ajoutant (en rouge) ou retirant (vert) de l'énergie noire dans la composition de l'Univers.*

$\mu = \cos(\theta)$ où θ est l'angle de séparation entre les 2 galaxies sur le ciel. Nous pouvons ensuite développer la fonction de corrélation en polynômes de Legendre $\mathcal{L}_\ell(\mu)$ pour obtenir l'estimations des multipoles de la 2PCF :

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1}\xi(s,\mu)\mathcal{L}_\ell(\mu)d\mu \tag{7.6}$$

Les moments multipolaires fournissent un mécanisme de compression de l'anisotropie dans la fonction de corrélation. Le monopôle $\xi_{\ell=0}$ est la composante isotrope de la 2PCF, tandis que le quadrupôle $\xi_{\ell=2}$ (et les ordres pairs supérieurs) contient des informations sur les anisotropies de la fonction de corrélation. Selon le principe cosmologique, la distribution des galaxies devrait être presque isotrope. Cependant, les vitesses particulières des galaxies induisent des anisotropies dans la distribution observée des galaxies, connues sous le nom d'effet de distorsion dans l'espace des redshifts (RSD) qui conduit à des multipoles pairs non nuls. En théorie linéaire (donc à grande échelle) Kaiser (1987) a montré que les anisotropies dans la 2PCF sont proportionnelles à la composante isotrope de la 2PCF (dans l'espace réel) par un facteur $\propto \mu^4$ (voir Equation (2.16)), ce qui signifie qu'il n'y a pas de contribution plus élevée que μ^4. Ainsi, le signal cosmologique est porté par le monopôle, le quadruopôle et l'hexadécapôle ($\ell = 0, 2, 4$).

Dans cette thèse j'étudie la distribution spatiale des galaxies, et plus particulièrement les galaxies à raies d'émissions (ELGs). Les ELGs sont majoritairement des galaxies qui forment des étoiles car les raies d'émission fortes dans les spectres de galaxies sont corrélées avec le taux de formation stellaire des galaxies (Moustakas et al., 2006). Je me concentre sur l'étude des corrélations spatiales à petite échelle (de l'ordre de $\sim$ 1Mpc) pour voir comment ces galaxies sont connectées à la distribution de matière totale de l'Univers et aux halos de matière noire. Les études de la connexion galaxie-halo sont importante pour comprendre les processus physiques de la connexion entre les galaxies et la matières sous-jacente, pour générer des catalogues de galaxies simulées qui permettent de tester les analyses cosmologiques ainsi que pour étudier l'impact des modèle de connexion galaxie-halo sur les paramètres cosmologiques.

Pour éviter l'impact des vitesses particulières des galaxies sur les petites échelles, nous pouvons utiliser la fonction de corrélation projetée $w_p(r_p)$. Au lieu de décomposer la distance r entre les galaxies en (s, μ), nous pouvons décomposer ses composantes le long de, et perpendiculairement à, la ligne de visée, π et r_p. La fonction de corrélation projetée est obtenue en intégrant $\xi(r_p, \pi)$ le long de la ligne de visée :

$$w_p(r_p) = \int_{\pi_{min}}^{\pi_{max}} \xi(r_p, \pi) d\pi \tag{7.7}$$

La fonction de corrélation projetée est largement utilisée dans les études de la connexion galaxie-halo car elle a l'avantage d'être presque insensible à la vitesse particulière des galaxies aux petites échelles (Bosch et al., 2013).

7.2 L'échantillon des ELGs du relevé 1% DESI

7.2.1 DESI

L'instrument spectroscopique de l'énergie noire (en anglais: le Dark Energy Spectroscopic Instrument, DESI) est un instrument spectroscopique robotisé, alimenté par des fibres optiques, qui fonctionne sur le télescope Mayall de 4 mètres (monture équatoriale) au Kitt Peak National Observatory (KPNO) sur la montagne Iolkam Du'ag (Kitt peak) en Arizona (États-Unis). DESI est conçu pour mesurer simultanément les spectres de 5000 objets sur un champ de ~ 3 degrés et mène actuellement une étude de cinq ans sur 14 000 deg^2 (environ un tiers du ciel), pour obtenir les spectres d'environ 40 millions de galaxies et de quasars dans une gamme de redshift $0 < z < 3.5$. DESI vise à créer une carte tridimensionnelle de la distribution de la matière couvrant un volume sans précédent, en ciblant différents types de galaxies.

A faible redshift, $z < 0,5$, DESI réalise un relevé de galaxies brillante, créant un échantillon limité en magnitude de $\sim$ 13M galaxies pour étudier la structure cosmique à l'époque dominée par l'énergie noire avec un échantillonnage à haute densité. A un décalage vers le rouge plus élevé, DESI ciblera au total $\sim$ 8M de *galaxies rouges lumineuses* (LRGs) entre $0.4 < z < 1.1$, $\sim$17M *galaxies à raies d'émission* (ELGs) entre $0.6 < z < 1.6$, et $\sim$3M *quasars* ou *objets quasi stellaires* (QSOs) entre $0.8 < z < 3.5$, produisant des contraintes strictes sur la distribution des galaxies à grande échelle pour essayer de déchiffrer la nature de l'énergie noire.

7.2.2 Le relevé 1% de DESI

Le relevé 1% de DESI a couvert $\sim$ 140 deg^2 avec des algorithmes de sélection de cibles et des profondeurs d'image similaires à celles du relevé principal. Elle a été menée pendant deux mois (avril et mai 2021) avant le début du relevé principal en juin 2021. Comme son nom l'indique, le relevé 1% vise à reproduire 1% du relevé principal. La sélection des cibles ELG (Raichoor et al., 2023) se concentre sur l'intervalle de redshift $0.6 < z < 1.6$ et est conçue pour sélectionner des galaxies avec des lignes d'émission spectrale fortes. La ligne d'émission du doublet [O II] permet de mesurer les redshifts précisément avec DESI. La géométrie de ce relevé est constituée de vingt régions non superposées, de la taille d'un plan focal, appelées *rosettes*, représentées en rouge sur Figure 7.3. Chaque région fait l'objet d'au moins $\sim$ 11 visites afin d'obtenir un degré élevé de complétude de l'attribution des fibres (qui est bien supérieur à celui du relevé principal). En effet, chaque fibre ne peut accéder que à une cible dans le ciel, donc si 2 cibles sont côte à côte, il faut repasser une deuxième fois sur la même région pour pourvoir observer les deux. Pour chaque visite, les centres sont légèrement déplacés afin d'augmenter la complétude dans les régions du plan focal qui ont peu de fibres (au centre et au bord). Au final, le relevé 1% a collecté $\sim$ 270k ELGs

à des redshifts $0.8 < z < 1.6$, ce qui en fait l'échantillon spectroscopique d'ELG le plus grand observé aujourd'hui.

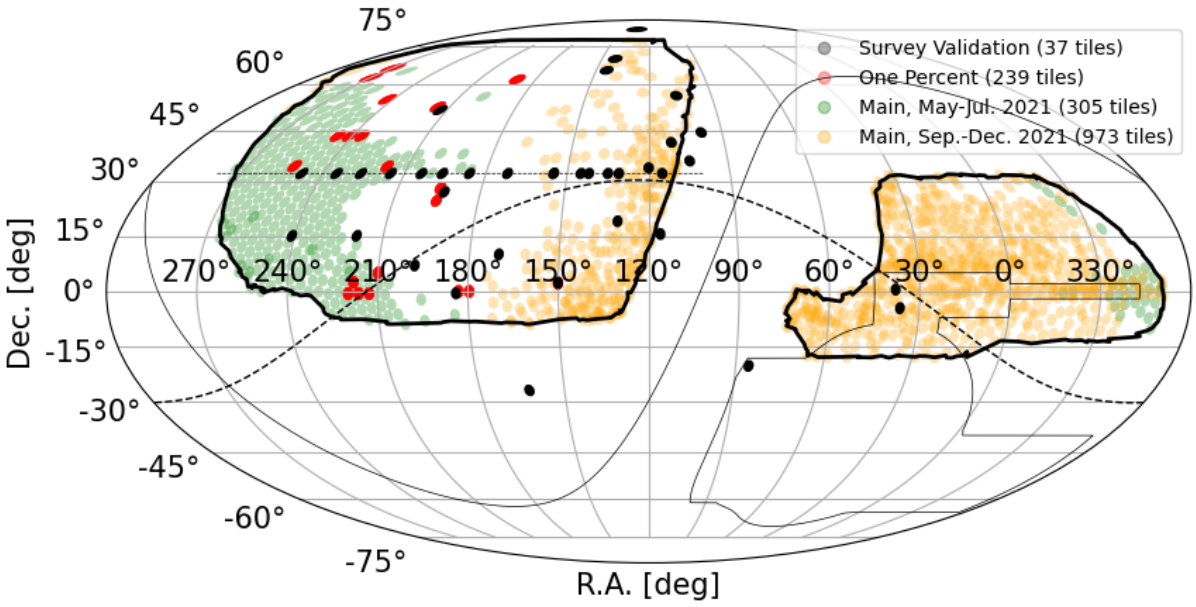

Figure 7.3: *En haut : Distribution dans le ciel des tuiles observées par DESI pendant le relevé 1% en rouge et pendnat la phase de validation du relevé en noir. La couvertue du relevé principal au tout début de la campagne est représentée en vert (2021 mai-juillet) et en orange (2021 septembre-décembre). Cette figure est tirée de (Raichoor et al., 2023).*

Pour les besoins de cette thèse, et l'étude de la connexion galaxie-halo, nous avons besoin de mesurer les corrélations spatiales des galaxies à petite échelle. Grâce à la grande complétude du relevé 1% de DESI, nous pouvons mesurer les corrélations spatiales des galaxies jusqu'à de très petites échelles, 0.04 Mpc$/h$ en distance perpendiculaire r_p pour la fonction de corrélation projetée w_p et 0.17 Mpc$/h$ en séparation de paires de galaxies s pour le monopôle et le quadrupôle de la 2PCF. La Figure 7.4 montre le clustering projeté (intégré entre -40 et 40 Mpc$/h$), le monopôle et le quadrupôle de la 2PCF pour l'échantillon ELG du relevé 1% de DESI, restreint à des redshifts $0.8 < z < 1.6$. Cet échantillon est donc très approprié pour étudier les ELGs à l'intérieur des halos de matière noire car il fournit des mesures précises des corrélations spatiales des galaxies jusqu'à de très petites échelles.

7.3 Connexion galaxie-halo des ELGs

L'hypothèse de base de notre vision actuelle de la formation des galaxies est que les galaxies se forment dans des halos de matière noire. Par conséquent, la croissance, les propriétés internes et la distribution spatiale des halos de matière noire peuvent être liées à celles des galaxies. La *connexion galaxie-halo* est un concept fondamental en cosmologie qui explore la relation entre les galaxies et les halos de matière noire dans lesquels elles résident. Les codes de simulation à N-corps constituent la base des modèles de formation des galaxies, et diverses techniques sont ensuite appliquées pour relier les galaxies et les halos de matière noire dans les simulations.

Une représentation schématique de la connexion galaxie-halo est montrée dans la Figure 7.5 et montre différents modèles pour connecter les galaxies au halos de matière noire. Ces techniques sont ensuite utilisées pour contraindre la connexion galaxie-halo à partir des données, fournissant des informations inestimables sur la physique de la formation des galaxies. Ces contraintes sont également essentielles pour garantir la robustesse des résultats cosmologiques des études de galaxies, car elles nous permettent

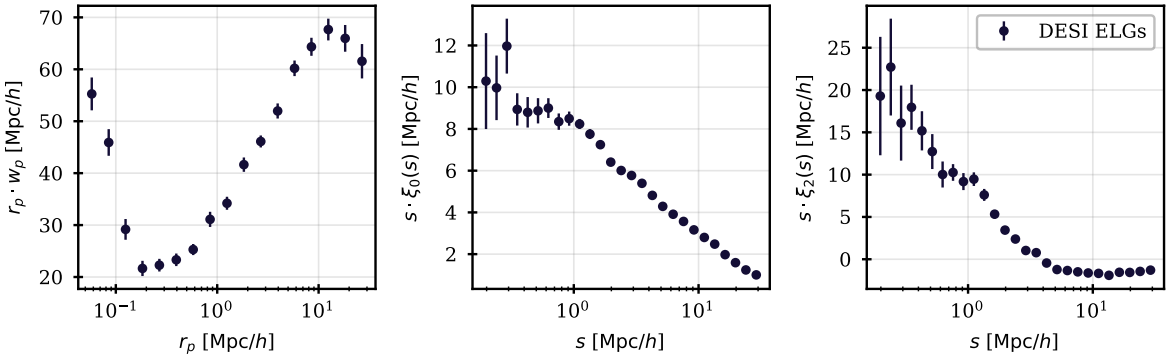

Figure 7.4: *Mesure de la 2PCF des ELGs à partir du relevé 1% de DESI dans l'intervalle de redshift* $0,8 < z < 1,6$. *A Gauche : Fonction de corrélation projetée* $w_p(r_p)$ *en fonction de* r_p *(intégrée entre* $\pi_{min} = -40$ *et* $\pi_{max} = 40$ *Mpc/h). Au milieu : monopôle de la 2PCF* $\xi_0(s) \times s$. *A droite : Quadrupôle de la 2PCF* $\xi_2(s) \times s$. *Sur chaque panneau, les barres d'erreur sont calculées à l'aide de la méthode du Jackknife en divisant le relevé en 128 régions indépendantes de même volume.*

de produire des mocks fiables (c'est-à-dire des catalogues de galaxies simulées) utilisés pour tester les analyses de clustering et pour dériver leur budget d'incertitudes systématiques (Alam et al., 2021).

Dans notre analyse, nous utilisons les simulations à N-corps AbacusSummit[1] (Maksimova et al., 2021) qui ont été construites pour répondre aux exigences scientifiques de DESI. Il s'agit d'une vaste suite de simulations cosmologiques à N-corps de haute précision produites avec le code Abacus N-corps, sur le supercalculateur Summit de l'Oak Ridge Leadership Computing Facility. Cette suite est composée de 150 boîtes de simulation, couvrant 97 modèles cosmologiques. Nous utilisons le modèle de distribution d'occupation des halos (en anglais: Halo Occupation distribution, HOD) pour peupler les halos de matières noires dans les simulations par des galaxies.

7.3.1 Modèle de distribution d'occupation des halos

La distribution d'occupation des halos (HOD) est un formalisme empirique qui décrit la relation entre une classe de galaxies et les halos de matière noire, comme la probabilité qu'un halo de masse M_h contienne N galaxies. Les modèles HOD ont des contributions de deux populations de galaxies, à savoir les centrales et les satellites, avec $\langle N_{cent}(M_h)\rangle$ et $\langle N_{sat}(M_h)\rangle$ leurs nombres moyens respectifs hébergés par halo d'une masse donnée. Une fois que le nombre moyen de galaxies par halo est calculé, une fonction de distribution de probabilité est utilisée pour affecter les galaxies centrales et les galaxies satellites à un halo. Les choix standard sont une distribution de Bernoulli pour les galaxies centrales et une distribution de Poisson pour les galaxies satellites. Les galaxies centrales sont typiquement placées au centre du halo avec une vitesse donnée par la vitesse particulière du halo, tandis que les satellites sont placés en supposant un profil de halo (principalement NFW (Navarro et al., 1997)) ou assignés à une particule aléatoire du halo.

L'occupation des halos pour les ELG centrales est proche d'une distribution gaussienne asymétrique (Cowley et al., 2016, Geach et al., 2012, Gonzalez-Perez et al., 2018a, Hadzhiyska et al., 2021a, Osato & Okumura, 2022, Yuan et al., 2022) et celle des satellite est bien représentée par une loi de puissance (Zheng et al., 2007):

$$\langle N_{sat}(M_h)\rangle = \begin{cases} \langle N_{cent}(M_h)\rangle \left(\frac{M_h - M_0}{M_1}\right)^{\alpha} & \text{if } M_h > M_0 \\ 0 \quad \text{sinon} \end{cases} \tag{7.8}$$

[1] https://abacussummit.readthedocs.io/en/latest/index.html

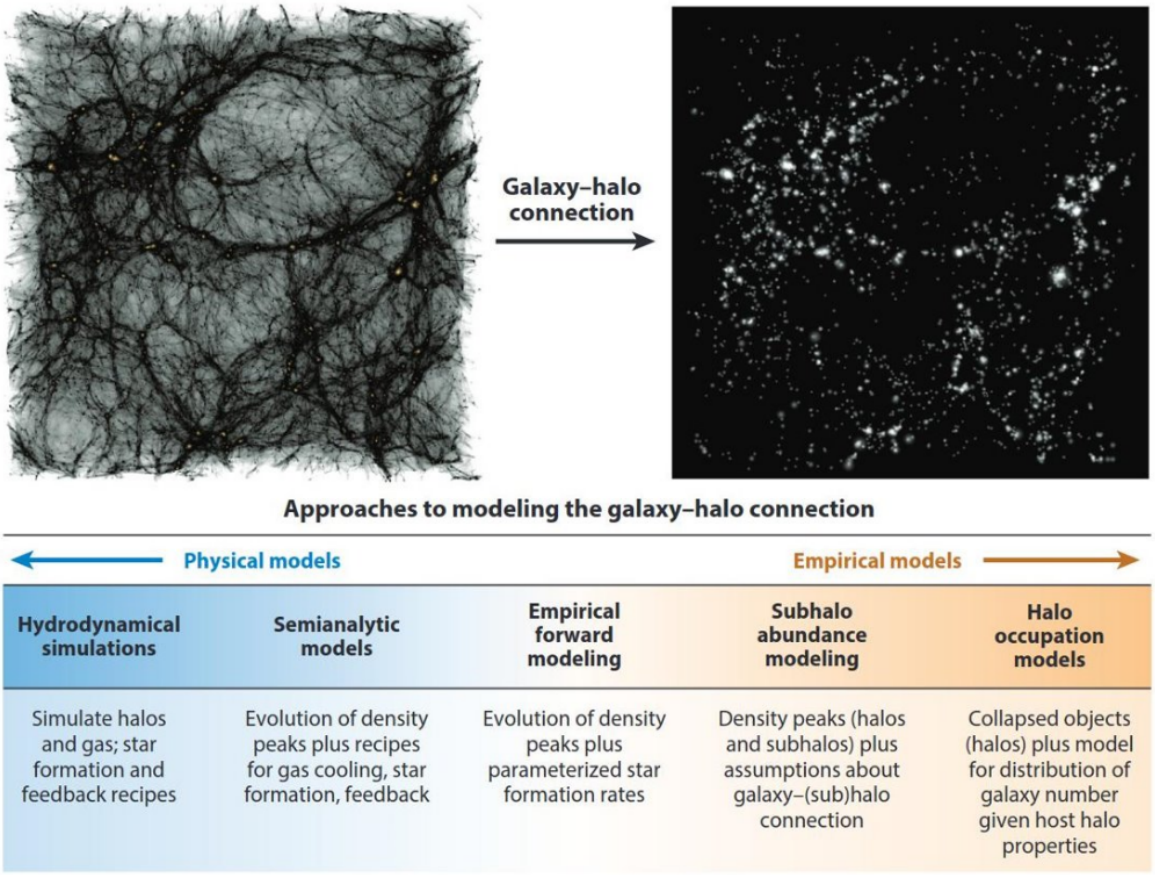

Figure 7.5: *Differentes approches pour modéliser la connexion entre galaxies et halos de matière noire, des plus physiques (à gauche) aux modèles les plus empiriques (à droite). Cette figure est tirée de Wechsler & Tinker (2018).*

Différents modèles HOD ont été développés pour reproduire la forme gaussienne de la distribution centrale des ELG. Dans une expérience précédente, eBOSS, Avila et al. (2020) et Alam et al. (2020) ont utilisé différents modèles HOD pour les ELGs centrales afin de reproduire la forme HOD obtenue dans les modèles semi-analytique (SAM) (Gonzalez-Perez et al., 2018a) :

➢ HOD Gaussien (GHOD):

$$\langle N_{cent}(M)\rangle = \frac{A_c}{\sqrt{2\pi}\sigma_{\rm M}} \cdot e^{-\frac{(\log_{10} M - \log_{10} M_c)^2}{2\sigma_{\rm M}^2}} \equiv \langle N_{cent}^{GHOD}(M)\rangle \tag{7.9}$$

Dans ce modèle, $\langle N_{cent}\rangle$ est simplement une fonction gaussienne avec une moyenne de M_c, une largeur de σ_M et A_c définit l'amplitude de la distribution. Ce modèle est comparé aux prédictions SAM dans le panneau gauche de Figure 7.6 (étiqueté comme HOD-2).

➢ Star-Forming HOD (SFHOD):

$$\langle N_{cent}(M)\rangle = \begin{cases} \langle N_{cent}^{GHOD}(M)\rangle & M \leq M_c \\ \frac{A_c}{\sqrt{2\pi}\sigma_{\rm M}} \cdot \left(\frac{M}{M_c}\right)^{\gamma} & M > M_c \end{cases} \tag{7.10}$$

Ce modèle est une combinaison d'une distribution gaussienne pour les halos de faible masse $< M_c$ et d'une loi de puissance décroissante pour les halos de masse élevée $> M_c$. Le résultat est une forme asymétrique

(voir le panneau gauche de la Figure 7.6, étiqueté HOD-3) où l'asymétrie est contrôlée par le paramètre γ. Cette fonction décrit bien les prédictions SAM, mais présente l'inconvénient d'être discontinue à $M = M_c$.

➢ **High mass quenched modifié (mHMQ):**

$$\langle N_{cent}(M)\rangle = \langle N_{cent}^{GHOD}(M)\rangle \cdot \left[1 + \mathrm{erf}\left(\frac{\gamma(\log_{10} M - \log_{10} M_c)}{\sqrt{2}\sigma_{\mathrm{M}}}\right)\right] \tag{7.11}$$

Ce modèle est dérivé du modèle "High Mass Quenched" de Alam et al. (2021), qui a une forme de gaussienne asymétrique.

➢ **HOD LogNormal (LNHOD):** En plus des 3 modèles décrit ci-dessus, j'utilise un modèle basé sur une distribution lognormal (LNHOD):

$$\langle N_{cent}(M)\rangle = \frac{A_c}{\sqrt{2\pi}\sigma_{\mathrm{M}} \cdot x} \cdot e^{-\frac{(\ln x)^2}{2\sigma_{\mathrm{M}}^2}} \quad \text{pour } x > 0, \text{ et 0 sinon} \tag{7.12}$$

où $x = \log_{10} M - (\log_{10} M_c - 1)$.

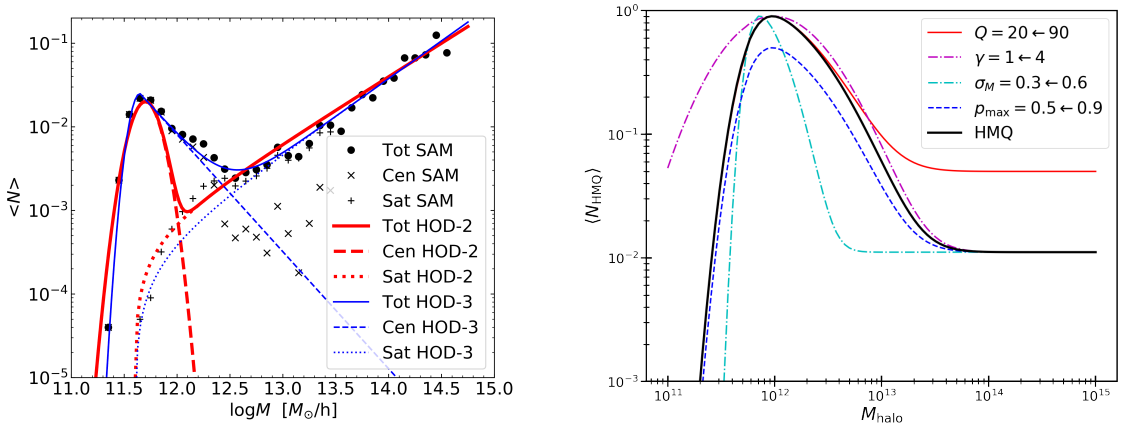

Figure 7.6: *À gauche: nombre moyen de galaxies ELG en fonction de la masse du halo. Les points sont les résultats de modèle semi-analytique (Gonzalez-Perez et al., 2018a) avec les contributions centrales et satellites indiquées séparément. Les lignes rouges correspondent au modèle HOD gaussien (étiqueté HOD-2) de Equation (7.9) et les lignes bleues au modèle HOD de formation d'étoiles (étiqueté HOD-3) de Equation (7.10). Pour les deux modèles HOD, la contribution des centrales (resp. des satellites) est représentée en tiret (resp. en pointillés). A droite: HOD du modèle "high mass quench" de (Alam et al., 2020). L'effet de la variation des paramètres individuels est illustré par des lignes colorées, tandis que la ligne noire continue représente le modèle de départ. Les lignes rouges pleines, magenta pointillées, cyan pointillées et bleues pointillées montrent l'impact des paramètres Q, γ, σ_M et p_{max} respectivement, lorsqu'ils sont modifiés par rapport aux valeurs d'entrées indiquées dans la légende.*

Dans cette thèse j'utilise ces 4 modèles HOD pour peupler les simulations à N-corps AbacusSummit et reproduire les corrélations spatiales des ELGs aux petites échelles. Pour reproduire la distributions des ELGs des données de DESI, j'ai développé une techniques d'ajustement qui utilise des processus gaussiens.

7.4 Methode d'ajustement des modelès HOD avec des processus gaussiens

Dans la littérature, plusieurs approches ont été proposées pour effectuer des ajustements de modèles HOD plus efficacement, en particulier pour limiter la stochasticité de la procédure. Une technique populaire est la méthode tabulée qui pré-calcule les corrélations des halos et des particules de la simulation et la convolue ensuite avec la distribution d'occupation des halos (Zheng & Guo, 2016). D'autres techniques utilisent des codes optimisés et parallélisés pour calculer les modèles HOD et les corrélations, comme par exemple AbacusHOD (Yuan et al., 2022)).

Dans cette thèse, j'ai développé une nouvelle méthode pour ajuster les paramètres HOD sur des mesures de corrélations à petite échelle en utilisant des processus gaussiens (GP) (Rasmussen & Williams, 2005). Cette méthode permet de faire des ajustements précis des paramètres du modèle HOD tout en minimisant le temps de calcul. Elle est inspiré d'un algorithme appelée "Efficient Global Optimization" (Jones et al., 1998). C'est une procédure en deux étapes qui utilise les processus gaussiens (GP) pour prédire le résultat d'une fonction, dans notre cas le χ^2 de notre ajustement défini par:

$$\chi^2 = (\xi_{data} - \xi_{model})^\top \left[\mathbf{C}_{data}/(1 - D_{data}) + \mathbf{C}_{model}/(1 - D_{model})\right]^{-1} (\xi_{data} - \xi_{model}) \tag{7.13}$$

où ξ représente le vecteur de données utilisé pour l'ajustement ($\xi = w_p, \xi_0, \xi_2$), $\mathbf{C}_{model}$ et $\mathbf{C}_{data}$ sont les matrices de covariance du modèle et des données, respectivement et D est le facteur de correction de Hartlap (Hartlap et al., 2007) appliqué à l'inverse de la matrice de covariance.

Dans un premier temps, nous échantillonnons l'espace des paramètres HOD suivant une séquence de Hammersley et calculons les χ^2 correspondants, pour fournir un échantillon d'entraînement initial au GP. Le modèle prédit par le GP est échantillonné par des chaînes de Markov (MCMC) pour obtenir une prédiction des distribution postérieurs des paramètres HOD. Ces distributions sont ensuite améliorées par itérations successives. A chaque itération, un point aléatoire est tiré depuis les chaînes MCMC, son χ^2 est calculé et ce nouveau point est rajouté à l'échantillon d'entraînement pour améliorer la prédiction du GP jusqu'à ce que les valeurs de chi^2 prédites deviennent suffisamment stables pour que les valeurs marginales des paramètres HOD et les contours postérieurs puissent être dérivés de manière fiable.

Comme les modèles HOD sont aléatoire, pour chaque jeu de paramètres HOD le χ^2 est mesuré 20 fois. On calcule ensuite la moyenne de ces valeurs de χ^2 et leur dispersion divisée par $\sqrt{20}$ pour estimer le χ^2 moyen et son incertitude. La plage dynamique des variations de χ^2 étant importante dans l'espace des paramètres HOD, cela peut rendre difficile la modélisation de la vraisemblance a posteriori. Nous utilisons donc le logarithme naturel des valeurs moyennes de χ^2 et les erreurs correspondantes comme données d'entrée du GP. Nous fournissons au GP $\boldsymbol{x} = \{\log_{10}(M_c), \alpha, A_s, \log_{10}(M_0), \log_{10}(M_1), \sigma_{textscm}, f_{\sigma_v}\}$, $y = \ln(\chi^2)$, $\epsilon = \delta(\ln(\chi^2))$.

7.4.1 Test de la méthode

Pour tester la précision de la méthode dans des conditions réalistes, nous avons utilisé 100 jeu de données simulées où nous connaissons les paramètres HOD d'entrée qui proviennent de 100 boîtes de simulation à N-corps différentes, à la même cosmologie. Les sources d'erreur de ces tests sont la variance cosmique due aux différente boîtes utilisées et la stochasticité des tirages HODs. Nous prenons un échantillon d'entraînement de 600 points dans une séquence de Hammersley et nous laissons tourner l'ajustement jusqu'à 800 itérations. Les résultats des 100 ajustements sont présentés sur la Figure 7.7.

Tous les paramètres HOD sont reconstruits avec un biais moyen qui se situe à l'intérieur ou, pour f_{sigma_v}, au niveau d'un écart-type de la distribution des paramètres. Plus précisément, nous trouvons un biais moyen de 0,29 pour A_s, 0,52 pour M_0, 0,69 pour α, 0,97 pour f_{σ_v}, 0,52 pour $\log(M_c)$ et 0,26 pour σ_M.

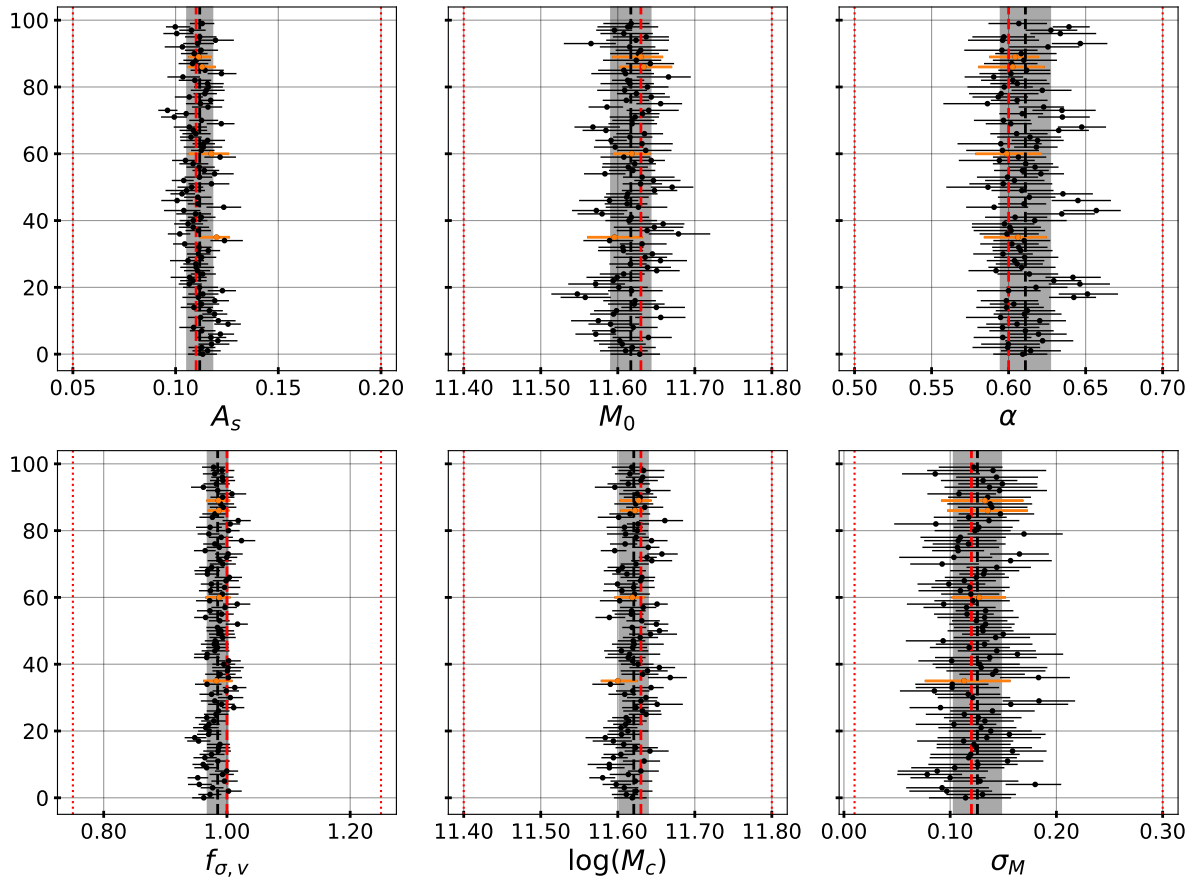

Figure 7.7: *Test de précision à partir de 100 ajustements. Les points représentent les valeurs marginales des paramètres HOD, les erreurs étant définies par les quantiles* 16% *et* 84% *des postérieurs de l'ajustement. La bande grise couvre l'intervalle* $\pm 1\sigma$ *autour de la moyenne des valeurs marginalisées, donnée par la ligne verticale noire. La ligne rouge en pointillés représente les valeurs des paramètres HOD d'entrée. Tous les ajustements ont été exécutés jusqu'à 800 itérations après l'entraînement initial des processus gaussiens basé sur l'échantillonnage de Hammersley de l'espace des paramètres HOD avec 600 points. Les lignes pointillées rouges indiquent les "priors" d'ajustement.*

Dans ce qui précède, σ est l'écart-type de la distribution marginalisée des résultats de l'ajustement. Il s'agit également de l'erreur statistique attendue pour un ajustement, les erreurs étant prises en compte dans la matrice de covariance utilisée dans les ajustements, à savoir la stochasticité, la variance cosmique et la densité de galaxies proche de celle attendue pour les données DESI ($\sim 10^{-3}$ gal/Mpc) mais pour un volume trois fois plus grand que celui du relevé 1% des ELGs de DESI. Nous nous attendons donc à ce que les valeurs des paramètres HOD dérivées de ces données avec notre procédure aient une précision bien meilleure que $\sim 1\sigma$ de l'incertitude statistique des données pour la plupart des paramètres, le pire cas étant le paramètre f_{σ_v} pour lequel la précision devrait être d'environ $\sim 0.6\sigma$.

Nous avons aussi testé la méthode en changant le nombre de point (300, 600 et 800) dans l'échantillonnage initial et la techniques d'échantillonnage (hypercube latin). De ces tests, nous avons conclu qu'il faut une densité de points initiale dans l'espace des paramètres assez haute, 300 n'étant pas suffisant pour retrouver des résultats non biasés, mais qu'il n'est pas nécessaire d'ajouter trop de points au départ, avec 1200 points initiaux les performances de l'ajustement sont presque similaires qu'avec 600 points.

Cette méthode étant fiable sur nos simulations, nous l'utilisons ensuite sur les données du relevé 1% de DESI.

7.5 Résultats sur le relevé 1% des ELG de DESI

Pour cette étude nous utilisons les 4 modèles HOD pour les centrales ainsi que la loi de puissance pour le HOD satellite décrit dans la Section 7.3.1. Nous ajustons ces modèles avec la technique utilisant des processus gaussiens décrite et testée précédement.

7.5.1 Résultats pour des HODs standards

Les résultats de l'ajustements sur les données de DESI sont présentés dans la Figure 7.8. A l'exception du paramètre M_1 qui est maintenu fixe dans les ajustements, nous avons utilisé des priors plats pour tous les autres paramètres HOD et pour tous les ajustements présentés ci-dessous.

Les quatre modèles les mieux ajustés fournissent des résultats similaires, qui s'accordent raisonnablement bien avec les données. La Figure 5.6 montre les corrélations calculées à partir des halos uniquement, indépendamment des galaxies qu'ils contiennent (ligne pointillée grise). Ceci met en évidence le fait que les paires de galaxies à l'intérieur d'un même halo ne contribuent, comme attendu, qu'aux petites échelles dans les trois statistiques. Cette contribution constitue ce que l'on appelle le terme 1-halo de la connexion galaxie-halo et est essentielle pour reproduire le fort signal mesuré aux petites échelles dans nos données, notamment celui dans la fonction de corrélation projetée à $r_p < 0.3$ Mpc/h. Pour reproduire ce signal, tous les modèles favorisent des paires de galaxies dans des halos de faible masse $< 10^{12}\ M_\odot/h$, ce qui est inattendu comparé aux études précédente de la connection galaxie-halo des ELGs.

On note aussi qu'entre 0.3 et ∼1 Mpc/h en r_p, le clustering mesuré est supérieur au clustering prédit à partir des halos, ce qui signifie que la contribution d'un seul halo provenant du profil NFW n'est pas suffisante pour décrire les données dans cette région. Les autres caractéristiques difficiles à modéliser correctement sont la pente de la fonction de corrélation projetée entre 0.2 et 10 Mpc/h et la bosse à $s \sim 1-2$ Mpc/h dans le monopôle et le quadrupôle. Cela explique en partie les valeurs élevées de χ^2 qui sont en moyenne de ∼157 pour 65 degrés de liberté. Comme tous les modèles se comportent de manière similaire, cela implique qu'il manque des ingrédients dans les HODs standard pour les ELGs.

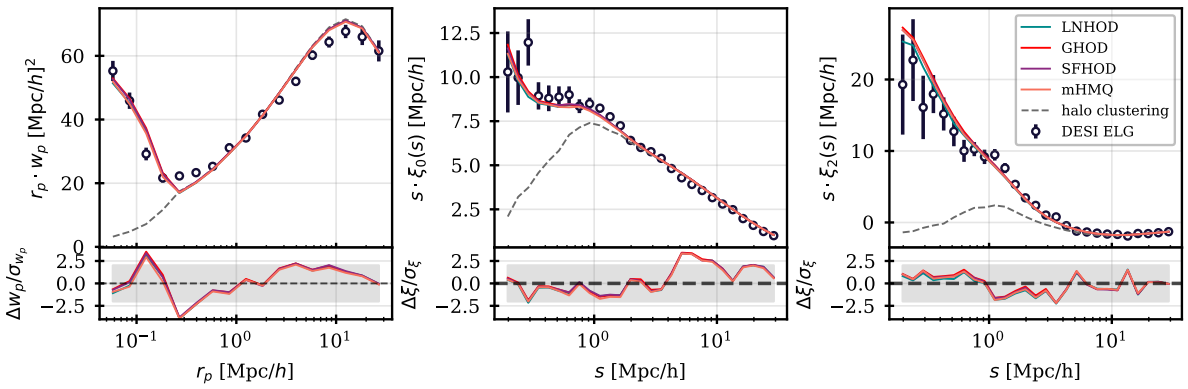

Figure 7.8: *En haut: Mesure des corrélations du relevé 1% des ELGs de DESI, comparées aux modèles HOD standard les mieux ajustés obtenus avec la méthode d'ajustement par GP. Les modèles en trait plein correspondent à différentes prescriptions pour les galaxies centrales, en conservant la prescription standard de la loi de puissance pour les satellites. Le modèle en ligne pointillée correspond aux corrélations des halos uniquement, montrant que les paires de galaxies issues d'un halo unique ont un fort impact sur la distribution des galaxies aux très petites échelles. Les erreurs sont uniquement des incertitudes de la méthode Jackknife. En bas: Résidus d'ajustement normalisés par les erreurs diagonales de la matrice de covariance complète, qui comprennent les incertitudes du Jackknife pour les données ainsi que le bruit stochastique et la variance cosmique pour le modèle, mais pas les corrections du facteur de Hartlap.*

Un autre résultat intéressant de ces ajustements est que, quelle que soit la prescription HOD pour les galaxies centrales, la valeur la mieux ajustée pour f_{σ_v} est significativement supérieure à 1. Ceci qui est en accord avec ce qui a été rapporté dans Avila et al. (2020). Les valeurs les mieux ajustées se situent entre 1,2 et 1,5 selon le modèle, avec une erreur d'environ $\pm 0,1$, et montre donc que les ELGs ont des dispersions de vitesse supérieurs à celle des particules de leurs halos.

Dans les sections suivantes nous utilisons uniquement le HOD mHMQ pour les galaxies centrales et nous changeons les prescriptions pour peupler les galaxies satellites.

7.5.2 Ajout d'un modèle de conformité

Dans cette section nous utilisons le principe de conformité. La conformité change la valeur moyenne du nombre de galaxies satellite en utilisant la connaissance a priori du fait que le halo contient une centrale ou non. Ici, nous utilisons un modèle très simple de stricte conformité : les satellites ne peuvent peupler un halo que si une galaxie centrale est déjà présente. Cela permet de forcer les galaxies satellites à être en paire avec des centrales et donc de générer naturellement du signal 1-halo. Les résultats de l'ajustements du modèle mHMQ avec stricte conformité sur les données de DESI sont présentés dans la figure 7.9. Une conformité stricte n'améliore que légèrement l'accord avec les données, comparé au modèle sans conformité. Par contre, lorsque la conformité est requise, les paires de galaxies intra-halo sont réparties sur une gamme plus large de masse de halo, à la fois à faible et à forte masse de halo.

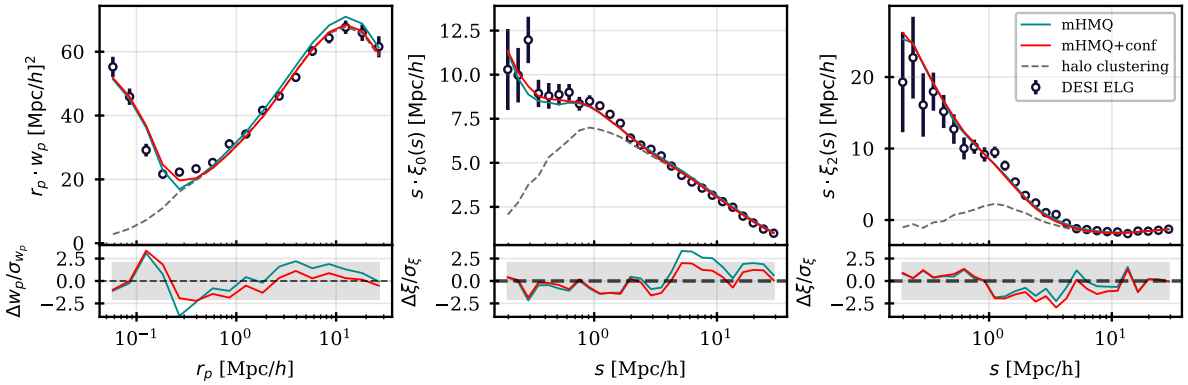

Figure 7.9: *En haut: Mesure des corrélations du relevé 1% des ELGs de DESI, comparée aux modèles mHMQ les mieux ajustés sans (ligne verte) et avec (ligne rouge) conformité stricte. La ligne pointillée correspond aux corrélations des halos uniquement. L'accord entre les données et les modèles est légèrement amélioré en ajoutant la conformité stricte, c'est-à-dire en conditionnant l'occupation des satellites à la présence d'une galaxie centrale. Les erreurs sont uniquement des incertitudes de Jackknife. En bas: Résidus d'ajustement normalisés par les erreurs diagonales de la matrice de covariance complète, qui comprennent les incertitudes du Jackknife pour les données ainsi que le bruit stochastique et la variance cosmique pour le modèle, mais aucune correction du facteur de Hartlap.*

7.5.3 Ajout du bias d'assemblage des halos

Bien que les modèles HODs ne soient basés que sur la masse des halos, ils permettent une bonne modélisation des données jusqu'à de très petites échelles. Cependant, les modèles semi-analytiques et les simulations hydrodynamiques montrent que des caractéristiques autres que la masse du halo ont un impact sur la formation des galaxies. Chaque halo et chaque galaxie sont uniques et ont leur propre histoire. Au cours de leur évolution, les galaxies et les halos connaissent une grande variété d'histoires dans leur

parcours d'assemblage qui peuvent influencer des propriétés autres que la masse du halo. Les modèles semi-analytiques et des simulations hydrodynamiques prédisent des corrélations entre la distribution spatiale des galaxies dans des halos de même masse et des propriétés secondaires du halo. Ce phénomène est connu sous le nom de *biais d'assemblage* (Croton et al., 2007, Gao & White, 2007, Wechsler et al., 2002, 2006). Des études récentes (par exemple Hadzhiyska et al. (2022), Hearin et al. (2016), Yuan et al. (2018)) ont développé des modèles HOD étendus qui prennent en compte les propriétés secondaires des halos telles qu'elles sont observées dans la simulation afin de modifier le nombre moyen de halos à une masse donnée en fonction de ces propriétés. Dans la littérature, plusieurs propriétés secondaires ont été étudiées, telles que la concentration du halo, l'environnement de densité, le cisaillement (les anisotropies dans la distribution de matière locale), le paramètre de spin du halo, le taux d'accrétion maximal... Mao et al. (2018) présente un résumé des corrélations entre plusieurs proxies de l'histoire de l'assemblage et les biais secondaires du halo.

Pour notre étude, nous testons 3 propriétés secondairess des halos : la concentration, de la densité locale et les anisotropies locales de densité, en utilisant la paramétrisation suggérée dans Hadzhiyska et al. (2022) :

$$\langle N'_{cent}(M)\rangle = [1 + a_{cen} f_a (1 - \langle N_{cent}(M)\rangle)] \langle N_{cent}(M)\rangle \tag{7.14}$$

$$\langle N'_{sat}(M)\rangle = [1 + a_{sat} f_a] \langle N_{sat}(M)\rangle \tag{7.15}$$

où $\langle N_{cent}(M)\rangle$ et $\langle N_{sat}(M)\rangle$ sont donnés dans la Section 7.3.1. Dans les équations ci-dessus, f_a est introduit pour matérialiser la propriété secondaire normalisée pour chaque halo. Dans une tranche de masse de halo donnée, les halos sont d'abord classés par valeur décroissante de la propriété secondaire considérée et chaque halo se voit attribuer une valeur différente de f_a, cette dernière diminuant linéairement entre 0,5 et $-0,5$ en allant du halo le mieux classé au dernier.

La figure 7.10 présente les corrélations prédites par les modèles mHMQ les mieux ajustés avec un biais de conformité strict, obtenus sans et avec les trois prescriptions de biais d'assemblage. La qualité de l'ajustement des modèles mHMQ ne s'améliore pas de manière significative lorsqu'on ajoute l'un des trois biais d'assemblage. Les paramètres HOD et les paramètres dérivés se situent à 1σ près des valeurs dans le modèle sans biais d'assemblage. Par conséquent, tous les modèles présentent des résultats de clustering similaires (presque impossibles à distinguer). Nous notons que le modèle avec biais d'assemblage utilisant la concentration des halos améliore légèrement la valeur du χ^2 avec une préférence pour les halos très concentrés, $a_{cen} \sim 0.75$. Cela constitue une légère préférence pour le biais d'assemblage, mais comme l'effet sur les statistiques de clustering est faible, cette préférence ne peut pas être établie de manière robuste. Le meilleur ajustement pour le biais d'assemblage en utilisant la densité locale du halo montre que les ELGs préfèrent un environnement neutre puisque a_{cen} est compatible avec 0. Dans le cas des anisotropies de densité locale, les résultats du meilleur ajustement indiquent une préférence pour les halos avec un cisaillement légèrement positif, $a_{cen} \sim 0.1$, mais cette préférence est plus faible que celle pour la concentration du halo. Enfin, le paramètre a_{sat} est compatible avec 0 et peu contraint dans les trois modèles en raison du fait que la fraction de satellites avec un biais de conformité strict est faible ($\sim 2\%$).

7.5.4 Changement de profil des halos

Aucune des extensions au modèle HOD étudiées dans les sections précédentes ne parvient à produire des paires supplémentaires de galaxies à des échelles $r_p = [0.1, 1]\mathrm{Mpc}/h$ comme l'exigent les données.

Néanmoins, il est possible d'y remédier en modifiant le profil radial de position des satellites dans les halos. Orsi & Angulo (2018) suggèrent que, quelle que soit la masse du halo, les ELGs peuplent préférentiellement la périphérie de leurs halos hôtes, les galaxies accrétées plus récemment se trouvant plus éloignées du centre du halo. Cela s'explique par le fait que les galaxies satellites ne peuvent présenter des taux élevés de formation d'étoiles que pendant une courte période, achevée une fois que le gaz de la galaxie a été épuisé par les effets de marée et de forte pression ("ram pressure"). En conséquence, les

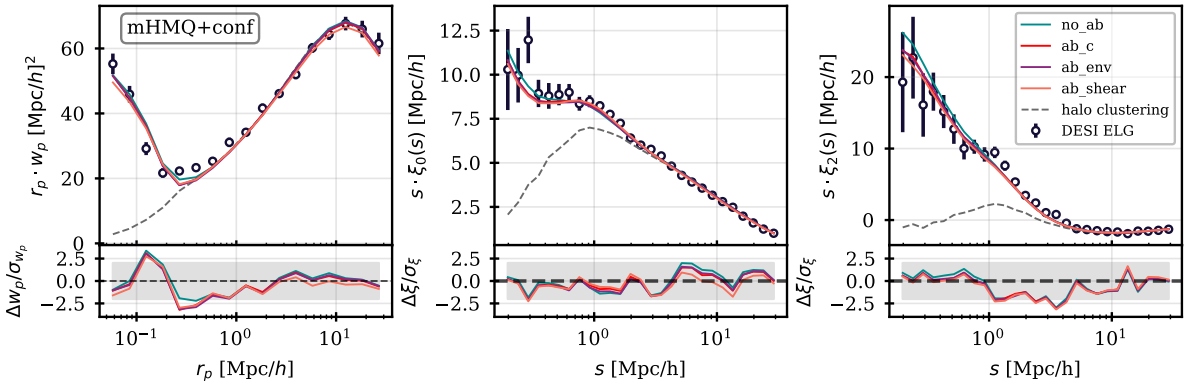

Figure 7.10: *En haut: Mesure des corrélations du relevé 1% des ELGs de DESI, comparées à différents modèles mHMQ les mieux ajustés avec un biais de conformité strict (ligne verte) et avec biais d'assemblage pour les centrales et les satellites en fonction de la concentration (rouge), de la densité locale du halo (violet) et des anisotropies de la densité locale du halo (orange). Les erreurs sont uniquement des incertitudes de Jacknnife. En bas: Résidus d'ajustement normalisés par les erreurs diagonales de la matrice de covariance complète, qui comprennent les incertitudes Jackknife pour les données ainsi que le bruit stochastique et la variance cosmique pour le modèle, mais aucune correction du facteur de Hartlap.*

galaxies satellites qui forment des étoiles devraient être préférentiellement situées à la périphérie de leur halo. Du point de vue observationnel, des résultats montrent que la distribution de la fraction éteinte du taux spécifique de formation d'étoiles des galaxies varie en fonction de la distance radiale à l'intérieur d'un halo (Blanton & Berlind, 2007, Wetzel et al., 2012).

Inspirés par les publications ci-dessus, nous testons un profil de halo NFW modifié pour positionner les satellites ELG. Le nombre de satellites pour un halo donné est d'abord déterminé selon la prescription standard dans Equation (5.6). Une fraction d'entre eux, f_{exp} ont des positions radiales tirées d'une loi exponentielle :

$$\frac{dN(r)}{dr} = e^{-r/(\tau \cdot r_s)} \tag{7.16}$$

où r est la distance entre le satellite et le centre du halo, et τ contrôle l'échelle de longueur de l'exponentielle et agit sur l'extension du profil. Les positions radiales des satellites restants obéissent à un profil NFW mais en modifiant l'approximation pour r_s d'un facteur λ_{NFW}, à savoir $r_s \to r_s/\lambda_{NFW}$. Ceci est presque équivalent à l'extension de la coupure du profil par rapport à r_{vir} en $r_{halo} = \lambda_{NFW} \cdot r_{vir}$ et permet de modifier l'extension du profil. Ce profil modifié est montré sur la Figure 7.11. Les trois paramètres f_{exp}, τ et λ_{NFW} sont laissés libres de varier dans les ajustements. Notons que les galaxies positionnées au-delà du rayon viriel du halo sont improprement appelées satellites, mais nous conservons cette dénomination ici pour refléter la composante de paramétrisation HOD dont elles sont issues.

Les modèles mHMQ les mieux ajustés avec une conformité stricte en modifiant la prescription du profil des satellites comme décrit ci-dessus sont comparés dans la Figure 7.12 aux résultats de base utilisant un profil NFW. Le positionnement modifié des satellites se traduit par une amélioration significative de l'accord entre les données et les modèles, avec une valeur de χ^2 passant de ∼152 à ∼88. L'amélioration est la plus notable dans la région de la courbe ascendante de la fonction de corrélation projetée (voir les résidus dans la Figure 7.12) montrant que des paires supplémentaires de galaxies ont été générées à ces échelles avec le profil étendu, sans dégradation par ailleurs.

Nous constatons que le profil exponentiel contient environ 60% des satellites et qu'une fraction de ceux-ci (environ 12% du nombre total de satellites, tel que mesuré dans les simulations au meilleur ajustement des paramètres HOD) sont placés au-delà du rayon du halo donné par la simulation.Le profil modifié ci-dessus est empirique et peut très probablement être remplacé par un modèle plus physique. Néanmoins, notre principale conclusion est que le regroupement des ELGs mesuré par le relevé 1% de DESI favorise

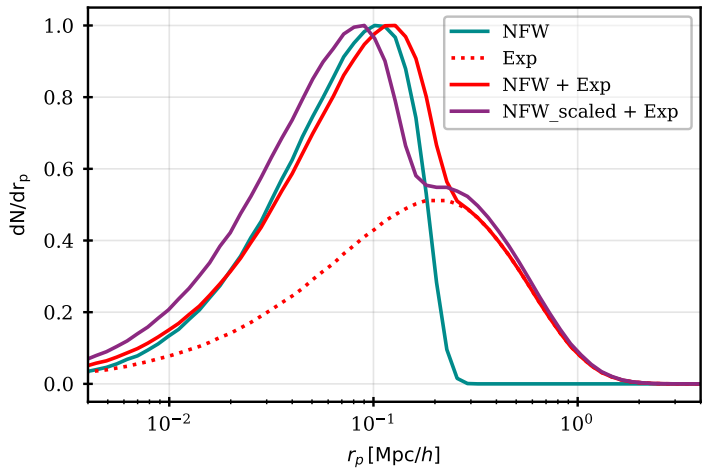

Figure 7.11: *Profil de densité de satellites normalisé pour les paramètres les mieux ajustés dans le modèle mHMQ avec une conformité stricte et notre prescription modifiée de profil NFW pour les satellites, en fonction de la distance projetée entre la galaxie et le centre du halo, perpendiculairement à la ligne de visée. Une fois ce profil intégré dans un modèle HOD, cette distance est également la séparation projetée des paires centrale-satellite. Dans cet exemple, nous considérons un halo de concentration 5 et* $r_s = 0.06 Mpc/h$ *(correspondant à des masses de halo autour de* $10^{12} M_\odot/h$*, proche de la valeur moyenne de la masse des halos de notre échantillon). Les courbes (toutes normalisées à une valeur maximale de 1) correspondent au profil NFW (vert), à la loi exponentielle ajoutée (en pointillés), à la combinaison des deux sans mise à l'échelle de la coupure NFW (rouge) et au modèle modifié complet (violet).*

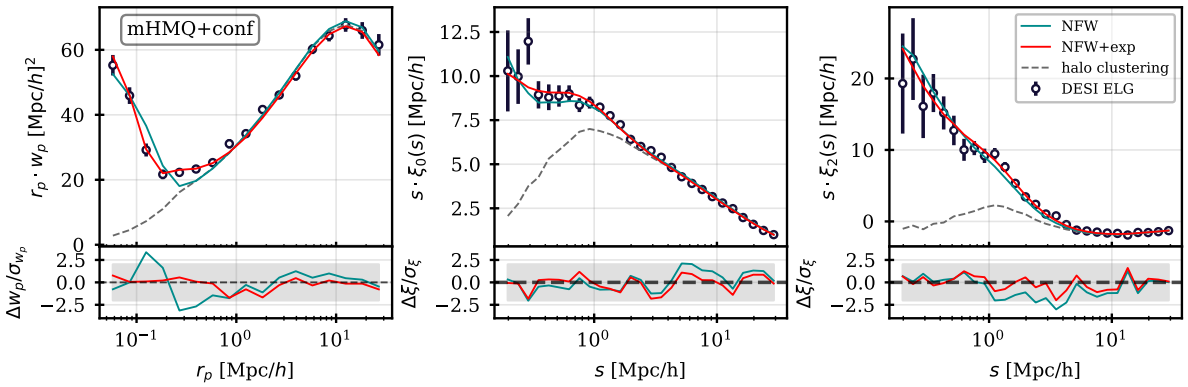

Figure 7.12: *En haut: Mesure des corrélations du relevé 1% des ELGs de DESI, comparées à différents modèles mHMQ les mieux ajustés avec un biais de conformité strict sans (vert) et avec (rouge) le profil de halo modifié pour le positionnement des satellites. Les erreurs sont uniquement des incertitudes de Jacknnife. En bas: Résidus d'ajustement normalisés par les erreurs diagonales de la matrice de covariance complète, qui comprennent les incertitudes du Jackknife pour les données ainsi que le bruit stochastique et la variance cosmique pour le modèle, mais aucune correction du facteur de Hartlap.*

clairement une fraction des ELGs résidant à la périphérie des halos, comme le suggèrent Blanton & Berlind (2007), Wetzel et al. (2012) et Orsi & Angulo (2018).

En complément, nous avons étudié les modèles HODs en coupant l'échantillon des ELGs en deux tranches de redshifts, $0.8 < z < 1.1$ et $1.1 < z < 1.6$, et utilisé 2 autres modèles de cosmologie pour les simulations à N-corps. La division de l'échantillon ELG en deux tranches de redshift modifie modérément les paramètres HOD. Nous observons un léger changement en fontion du redshift en termes de masse moyennes des halos peuplés par les ELGs (0.08 dex), que nous ne considérons pas comme significatif. Des changements modérés dans la cosmologie des simulations utilisées n'ont pas d'impact significatif sur les résultats, les corrélations des données sont modélisées avec une qualité d'ajustement similaire.

7.6 Conclusions

Au cours des dernières décennies, les structures à grande échelle de l'Univers sont devenues l'une des sondes cosmologiques les plus prometteuses pour contraindre les modèles d'énergie noire et de gravité (Alam et al., 2021).

Le cœur de mon travail de doctorat consiste à fournir un modèle précis du clustering à petite échelle de l'échantillon des galaxie à raies d'émissions (ELG) collecté par l'expérience DESI (Dark Energy Spectroscopic Instrument) pendant le relève 1% de l'instrument. Cette expérience vise à observer pendant 5 ans $\sim$ 40M galaxies et quasars à des redshifts entre $0.1 < z < 3.5$, y compris $\sim$ 17M d'ELGs entre $0.6 < z < 1.6$, pour contraindre fortement les modèles d'énergie noire. En seulement deux mois d'observations, DESI a observé 267k ELGs, ce qui constitue le plus grand échantillon spectroscofipique d'ELGs à ce jour. Grâce à la complétude de ce relevé, ce premier échantillon de données ELGs permet d'avoir des mesures précises des corrélations entre les distance des galaxies jusqu'à de très petites échelles, $0,03\mathrm{Mpc}/h$. Pendant mon doctorat, j'ai participé activement à la collaboration DESI. En particulier, j'ai apporté une contribution majeure à la génération de mocks (catalogues de galaxies simulées) pour la collaboration DESI, en étudiant la connexion galaxie-halo de l'échantillon ELG. J'ai travaillé directement sur les premières données de DESI, en participant à la création des catalogues de galaxies, et j'ai effectué de nombreux tests pour vérifier les potentiels effets d'erreurs systématiques.

Dans un premier temps, j'ai développé une nouvelle méthode d'ajustement prometteur basé sur les processus gaussiens pour ajuster les modèles d'occupation des halos (HOD) (cadre grandement utilisé pour étudier la connection galaxie-halo) de manière précise et efficace. J'ai d'abord développé un code parallélisé, efficace pour générer des catalogues de galaxies simulées à l'aide de modèles HOD. J'ai ensuite développé une procédure d'ajustement itérative basée sur des processus gaussiens afin de créer un modèle de substitution de la fonction multidimensionnelle de vraisemblance attendue. L'ensemble de la procédure a été testé sur des simulations, montrant que les paramètres HOD pour les corrélations des statistiques à 2 points d'un échantillon similaire à celui des ELGs de DESI sont retrouvés sans biais significatif par rapport à l'incertitude statistique attendue pour l'ajustement sur les données du relevé 1%, tout en évaluant environ 100 fois moins de points dans l'espace des paramètres HOD que les techniques standard basées sur les chaînes de Markov (MCMC).

Dans un deuxième temps, j'ai appliqué cette méthode à l'échantillon des ELGs du relévé 1% de DESI. Cet échantillon montre un signal de clustering fort et inattendu aux petites échelles $r_p < 0.3$ Mpc/h. Pour reproduire ce signal à petite échelle des ELGs, j'ai montré que nous devons introduire des paires de galaxies intra-halo dans des halos de faible masse ($< 10^{12} M_\odot$). Ce résultat était inattendu au vu des études précédentes sur les ELGs (Avila et al., 2020, Gonzalez-Perez et al., 2018b). Pour proposer un modèle plus physique, j'ai étudié l'effet de la conformité entre les galaxies proches. La conformité ajoute une information préalable à la fonction de probabilité des galaxies satellites selon que le halo héberge déjà une galaxie centrale ou non. Cette propriété améliore légèrement la modélisation des données à petite échelle tout en maintenant la dépendance de la masse du halo satellite en accord avec les attentes physiques. Nos résultats sont en accord avec ce que des études hydrodynamiques très récentes ont trouvé sur la conformité entre les galaxies centrales et satellites des ELGs (Hadzhiyska et al., 2021b). Nous

montrons également que la dispersion des vitesses des satellites est environ $\sim 30\%$ plus élevée que celle des particules de matière noire dans leurs halos. D'autres extensions aux modèles HOD standards, par exemple en prenant en compte les biais d'assemblage ou changement de cosmologie des simulations à N-corps n'apportent pas de changement significatif à nos résultats. La seule extension qui montre une amélioration significative des résultats est quand nous permettons aux satellites ELG de se situer en dehors du rayon du halo. En mettant $\sim 0.5\%$ des galaxies de l'échantillon ELG (et $\sim 12\%$ des galaxies satellites) à la périphérie des halos, nous obtenons le meilleur accord entre le modèle et les données. Ce travail a été publié dans JCAP lors de la publication des données du relevé 1% de la collaboration DESI.

Dans le cadre de mon travail sur l'ajustement des modèles HOD pour les ELGs, j'ai fourni à la collaboration DESI un ensemble de simulations officielles représentatives de la distribution des ELGs telle que mesurée dans le relevé 1%. Ces modèles sont actuellement utilisés au sein de DESI pour tester les analyses cosmologiques (mesure de l'oscillation acoustique des baryons (BAO) et du taux de croissance des structures (RSD)) pour la publication des données de la première année de DESI, et je participe activement aux discussions et à l'analyse en cours. Outre les travaux susmentionnés, j'ai contribué à de nombreux efforts au sein de la collaboration DESI. J'ai participé à l'étude des effets systématiques possibles dans la mesure du taux de réussite du redshift spectroscopique des ELGs en fonction des conditions d'observation spectroscopique, ce qui conduira également à un article pour l'analyse des données de la première année d'observation de DESI. Au cours du processus de validation de l'expérience (avant de commencer le relevé principal), j'ai également contribué à l'inspection visuelle des spectres des ELGs pour valider le code de détermination des redshifts Lan et al. (2022).

Au cours de mon doctorat, j'ai développé une expertise en cosmologie des structures à grande échelle (LSS), en utilisant des études spectroscopiques pour contraindre les paramètres cosmologiques depuis l'échelle du halo jusqu'aux plus grandes échelles (avec les effets de distorsions dans l'espace des redshifts et les échelles d'oscillation acoustique des baryons). Au sein de la collaboration DESI, j'ai interagi avec de nombreux collaborateurs dans le monde entier (États-Unis, France, Espagne, Corée...). J'ai acquis une large compréhension des différentes sondes LSS et une expertise de l'instrument DESI. Avec la nouvelle génération de relevés de galaxies pour la cosmologie de stade IV et l'augmentation des ressources informatiques, nous entrons dans l'ère de la cosmologie de précision. Ce sera une période très excitante pour la recherche cosmologique !

Enfin, je tiens à remercier tout particulièrement mes directeurs de thèse, Vanina et Etienne, pour leur patience et leur soutien sans faille, sans lesquels cette thèse n'aurait pas été la même.

Bibliography

Alam, S., Peacock, J. A., Kraljic, K., Ross, A. J., & Comparat, J. 2020, Monthly Notices of the Royal Astronomical Society, 497, 581, doi: 10.1093/mnras/staa1956

Alam, S., de Mattia, A., Tamone, A., et al. 2021, Monthly Notices of the Royal Astronomical Society, 504, 4667, doi: 10.1093/mnras/stab1150

Alam, S., de Mattia, A., Tamone, A., et al. 2021, MNRAS, 504, 4667, doi: 10.1093/mnras/stab1150

Alam, S., Aubert, M., Avila, S., et al. 2021, Physical Review D, 103, 083533, doi: 10.1103/PhysRevD.103.083533

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, Monthly Notices of the Royal Astronomical Society, 499, 5486, doi: 10.1093/mnras/staa2951

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, MNRAS, 499, 5486, doi: 10.1093/mnras/staa2951

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, Monthly Notices of the Royal Astronomical Society, 499, 5486–5507, doi: 10.1093/mnras/staa2951

Blanton, M. R., & Berlind, A. A. 2007, ApJ, 664, 791, doi: 10.1086/512478

Bosch, F. v. d., More, S., Cacciato, M., Mo, H., & Yang, X. 2013, Monthly Notices of the Royal Astronomical Society, 430, 725, doi: 10.1093/mnras/sts006

Cowley, W. I., Lacey, C. G., Baugh, C. M., & Cole, S. 2016, Monthly Notices of the Royal Astronomical Society, 461, 1621, doi: 10.1093/mnras/stw1069

Croton, D. J., Gao, L., & White, S. D. M. 2007, Monthly Notices of the Royal Astronomical Society, 374, 1303, doi: 10.1111/j.1365-2966.2006.11230.x

Gao, L., & White, S. D. M. 2007, Monthly Notices of the Royal Astronomical Society: Letters, 377, L5, doi: 10.1111/j.1745-3933.2007.00292.x

Geach, J. E., Sobral, D., Hickox, R. C., et al. 2012, Monthly Notices of the Royal Astronomical Society, 426, 679, doi: 10.1111/j.1365-2966.2012.21725.x

Gonzalez-Perez, V., Comparat, J., Norberg, P., et al. 2018a, Monthly Notices of the Royal Astronomical Society, 474, 4024, doi: 10.1093/mnras/stx2807

—. 2018b, Monthly Notices of the Royal Astronomical Society, 474, 4024–4038, doi: 10.1093/mnras/stx2807

Hadzhiyska, B., Tacchella, S., Bose, S., & Eisenstein, D. J. 2021a, Monthly Notices of the Royal Astronomical Society, 502, 3599, doi: 10.1093/mnras/stab243

—. 2021b, Monthly Notices of the Royal Astronomical Society, 502, 3599–3617, doi: 10.1093/mnras/stab243

Hadzhiyska, B., Hernquist, L., Eisenstein, D., et al. 2022, The MillenniumTNG Project: Refining the one-halo model of red and blue galaxies at different redshifts, arXiv. http://arxiv.org/abs/2210.10068

Hadzhiyska, B., Eisenstein, D., Hernquist, L., et al. 2022, submitted to MNRAS, doi: 10.48550/arXiv.2210.10072

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399, doi: 10.1051/0004-6361:2006617010.48550/arXiv.astro-ph/0608064

Hearin, A. P., Zentner, A. R., Bosch, F. C. v. d., Campbell, D., & Tollerud, E. 2016, Monthly Notices of the Royal Astronomical Society, 460, 2552, doi: 10.1093/mnras/stw840

Jones, D. R., Schonlau, M., & Welch, W. J. 1998, Journal of Global optimization, 13, 455, doi: 10.1023/A:1008306431147

Kaiser, N. 1987, Monthly Notices of the Royal Astronomical Society, 227, 1, doi: 10.1093/mnras/227.1.1

Lan, T.-W., Tojeiro, R., Armengaud, E., et al. 2022, doi: 10.48550/arXiv.2208.08516

Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., et al. 2021, Monthly Notices of the Royal Astronomical Society, 508, 4017, doi: 10.1093/mnras/stab2484

Mao, Y.-Y., Zentner, A. R., & Wechsler, R. H. 2018, Monthly Notices of the Royal Astronomical Society, 474, 5143, doi: 10.1093/mnras/stx3111

Moustakas, J., Kennicutt, Jr., R. C., & Tremonti, C. A. 2006, The Astrophysical Journal, 642, 775, doi: 10.1086/500964

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, The Astrophysical Journal, 490, 493, doi: 10.1086/304888

Orsi, Á. A., & Angulo, R. E. 2018, MNRAS, 475, 2530, doi: 10.1093/mnras/stx3349

Osato, K., & Okumura, T. 2022, Monthly Notices of the Royal Astronomical Society, 519, 1771, doi: 10.1093/mnras/stac3582

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6, doi: 10.1051/0004-6361/201833910

Raichoor, A., Moustakas, J., Newman, J. A., et al. 2023, AJ, 165, 126, doi: 10.3847/1538-3881/acb213

Raichoor, A., Moustakas, J., Newman, J. A., et al. 2023, The Astronomical Journal, 165, 126, doi: 10.3847/1538-3881/acb213

Rasmussen, C. E., & Williams, C. K. I. 2005, Gaussian Processes for Machine Learning (The MIT Press), doi: 10.7551/mitpress/3206.001.0001

Wechsler, R. H., Bullock, J. S., Primack, J. R., Kravtsov, A. V., & Dekel, A. 2002, The Astrophysical Journal, 568, 52, doi: 10.1086/338765

Wechsler, R. H., & Tinker, J. L. 2018, Annual Review of Astronomy and Astrophysics, 56, 435, doi: 10.1146/annurev-astro-081817-051756

Wechsler, R. H., Zentner, A. R., Bullock, J. S., Kravtsov, A. V., & Allgood, B. 2006, The Astrophysical Journal, 652, 71, doi: 10.1086/507120

Wetzel, A. R., Tinker, J. L., & Conroy, C. 2012, MNRAS, 424, 232, doi: 10.1111/j.1365-2966.2012.21188.x

Yuan, S., Eisenstein, D. J., & Garrison, L. H. 2018, Monthly Notices of the Royal Astronomical Society, 478, 2019, doi: 10.1093/mnras/sty1089

Yuan, S., Garrison, L. H., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022, MNRAS, 510, 3301, doi: 10.1093/mnras/stab3355

Yuan, S., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022, Monthly Notices of the Royal Astronomical Society, 512, 5793, doi: 10.1093/mnras/stac830

Zheng, Z., Coil, A. L., & Zehavi, I. 2007, The Astrophysical Journal, 667, 760, doi: 10.1086/521074

Zheng, Z., & Guo, H. 2016, MNRAS, 458, 4015, doi: 10.1093/mnras/stw523