UNIVERSIDAD DE GUANAJUATO
CAMPUS LEÓN

DIVISIÓN DE CIENCIAS E INGENIERÍAS

Synthetic Lyman-α quasar spectra datasets as a tool for cosmological studies.

Hiram Kalid Herrera Alcantar

Proyecto de tesis como requisito
para obtener el título profesional de
Doctor en Física

Asesora:
Dra. Alma X. González-Morales
División de Ciencias e Ingenierías
Universidad de Guanajuato

JULIO, 2024

Dedicada a toda mi familia,

especialmente a los pilares de mi vida,
mis padres Victoria y Javier.

Con amor para mi cometa,
mi amada esposa, María bonita.

Declaratory

I, Hiram Kalid Herrera Alcantar, hereby declare that the work presented in this thesis is my own and is based on the research conducted during my PhD between 2020 and 2024.

This thesis utilizes several books and article references on cosmology, baryon acoustic oscillations (BAO), and the Lyman-α forest as a guide for writing. Some figures were taken directly from these references, while others were created by the DESI Collaboration and specific members and included in this thesis with permission. In all cases, due credit is given to the authors, and relevant references are properly cited.

Part of the content presented in Chapter 4 and the first three sections of Chapter 5 is based on an article product of my PhD (Herrera-Alcantar et al. 2024). I tried my best to summarize, re-phrase and re-write the content of this article to avoid self-plagiarism. Part of the content of Section 5.4 is based in work presented in an article where I made major contributions (Cuceu et al. 2024). Some figures presented in this thesis were already presented in the aforementioned articles, this is mentioned in the caption of the relevant figures.

I also acknowledge the use of Artificial Intelligence (AI) assistants such as ChatGPT from OpenAI and TeXGPT from Overleaf for providing information, for text suggestions and enhancements, and for proofreading.

Hiram Kalid Herrera Alcantar

Date: May/31/2024

Agradecimientos

En primer lugar, me gustaría agradecer a mi asesora, Alma González. Alma, te agradezco infinitamente por todo el tiempo, apoyo y guía que me has brindado a lo largo de mi posgrado, ha sido un verdadero placer y honor trabajar contigo. Te agradezco mucho por presentarme a la colaboración DESI y al grupo de trabajo de Lyman-α que hicieron de mi posgrado una experiencia muy fructífera.

Agradezco a Ana Laura Benavides, Mariana Vargas, Alberto Diez, Gustavo Niz, Luis Ureña y Carlos Vaquera, por leer esta tesis, y por sus comentarios y sugerencias.

A mis compañeros y amigos del posgrado en Física, por las discusiones, ocurrencias y reuniones que ciertamente hicieron que mi posgrado fuera una experiencia divertida. Una mención especial a Andrea Muñoz, por convertirse en una gran amiga para mí, estar en los momentos que necesitaba ser escuchado, y por aclarar mis dudas respecto a metales y astrofísica.

También, me gustaría agradecer al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) y a la Universidad de Guanajuato, en especial a la Dirección de Apoyo a la Investigación y al Posgrado (DAIP), por el apoyo que se me dio para realizar mi doctorado.

A la colaboración DESI y al grupo de trabajo de Lyman-α por permitirme realizar mi investigación con los datos del instrumento y su software, además de las grandes enseñanzas y aprendizajes que he tenido. Gracias a Andreu Font Ribera, Ignasi Pérez i Ràfols y Julián Bautista. Merci à Eric Armengaud, Julien Guy et Corentin Ravoux. Danke an Michael Walther. Thanks to Jim Rich, Paul Martini, David Kirkby, Naim Karaçaylı and all of my DESI colleagues. Special mention to Andrei Cuceu, for all his help regarding running mocks (specially the hundreds of realizations we did), for guiding me with `Vega`, and for the fruitful discussions we had.

Por último, me gustaría agradecer infinitamente a toda mi familia. Especialmente a mis padres, Javier Herrera y Victoria Alcantar, los pilares de mi vida, les agradezco de todo corazón siempre estar a mi lado y por alentarme a seguir en el camino aún en los momentos difíciles. A mi esposa, María Mendoza, es muy bello compartir contigo detalles sobre mi trabajo y, ciertamente, será un placer compartir mi vida contigo.

Contents

Introduction

Over the past decades, the ΛCDM model has been the most successful theory for describing our Universe. This model describes an expanding Universe, composed of Dark Energy, Cold Dark Matter (CDM), baryonic matter, photons, and neutrinos. Furthermore, this model explains various important observations such as the accelerated expansion of the Universe, the existence of the Cosmic Microwave Background (CMB), the large-scale structure of the Universe (LSS), among others.

Like any theory in physics, the ΛCDM model must be tested. In this regard, baryon acoustic oscillations (BAO) stand out as an important cosmological probe. BAO arises during the early stages of the Universe, when photons and baryons were tightly coupled due to Compton scattering. Overdense regions in the Universe attracted baryons due to gravitational interaction, while thermal pressure, resulting from photon interactions, pushed the baryonic matter outwards. This interaction generated oscillations that propagated as wavefronts through the photon-baryon plasma, leaving imprints on the matter distribution at a characteristic scale, that serves as a standard ruler for measuring distances in the Universe.

Galaxy spectroscopic surveys provide a great opportunity to measure the BAO imprint on the distribution of galaxies. The most recent effort in this kind of survey is the Dark Energy Spectroscopic Instrument (DESI). It is expected to measure the spectra of approximately 40 million galaxies, including about 700 thousand Lyman-α quasars. Lyman-α quasars are a type of galaxy that contains a set of absorption features in their spectra caused by neutral hydrogen, known as the Lyman-α forest. The distribution of these absorption lines is tightly linked to the distribution of matter.

The main goal of this thesis is twofold: first, to describe the Lyman-α forest, highlighting its importance as a matter distribution tracer and its use in studying the expansion history of the Universe. And second, but most importantly, to provide a description of the methodology used to generate realistic Lyman-α synthetic datasets, detail the contributions I made to this methodology during my PhD, and showcase their role as a tool for testing and improvement of cosmological study methods involving the Lyman-α forest.

The structure of this thesis is as follows. Chapter 1 gives a brief description of the fundamental concepts of cosmology including general relativity, the ΛCDM model, and the BAO scale. Chapter 2 introduces the Lyman-α forest, its physical origin, its role as a matter distribution tracer, the state-of-the-art pipeline used to estimate the Lyman-α correlation function that provides insight on the BAO scale, and the physical model of these correlations. Chapter 3 presents the DESI survey as an outstanding experiment on the cosmic expansion study, reviewing the components of the instrument, the target types observed, the strategies followed to acquire data, and giving an overview of the survey progress from the beginning of the survey up to the date of the writing of this manuscript.

Chapter 4 presents a review of the state-of-the-art methodology used to generate synthetic Lyman-α forest quasar datasets. A more detailed description is provided in a paper product from my PhD. This method includes the flexibility to mirror a given survey in terms of object density, footprint, redshift distribution, magnitude distribution, and signal-to-noise ratio. In this regard, chapter 5 presents an example of this methodology applied to generate datasets emulating the characteristics of the DESI survey during its early (EDR) and first-year (DR1) data releases, along with the expected characteristics of the completed survey (DESI-Y5). A key product of this thesis is the production of 150 mock datasets of DESI-DR1 that were used to validate the analysis pipeline of the DESI 2024 Lyman-α BAO studies.

Chapter 6 showcases various possible uses of synthetic Lyman-α datasets including the characterization of the impact of contaminants on the Lyman-α correlation function, testing the performance of some algorithms used in cosmological studies. For example, redshift estimators, object classifiers, and algorithms for detecting astrophysical contaminant features in spectra. The mock datasets methodology presented in this thesis can also be used to perform forecasts of the constraining power of any given survey. This is exemplified by presenting a forecast methodology and its validation with observed data.

1 Fundamentals of Cosmology

The Cambridge dictionary defines cosmology as *the study of the nature and origin of the Universe* [1]. This field of physics has evolved rapidly over the years, benefiting from both theoretical predictions and the increased precision and statistics of the measurement of several observable quantities in the vast cosmic inventory, thus becoming one of the most precise sciences in physics. The main goal of this chapter is to introduce basic concepts of cosmology that will be useful for the content discussed throughout this thesis. This chapter is based on lectures on general relativity [2] and on books providing an introduction to cosmology [3] and a modern cosmology description [4], and various courses related to these topics that I took during my PhD.

1.1 The Einstein Field Equations

General relativity, hereafter referred to simply as GR, has been the most successful gravity theory over the past decades. Its classical tests include the precession of Mercury's perihelion [5] and the bending of light by the Sun, first measured in 1919 [6] during a solar eclipse. More recent tests have studied and confirmed the existence of phenomena such as strong gravitational lensing, initially observed in 1979 [7]; gravitational waves, first detected in 2015 [8]; and black holes, first directly imaged in 2019 [9], to name a few.

GR is based on the Equivalence Principle, which in its weak form (WEP) states that inertial mass and gravitational mass are equivalent. This equivalence lays the foundation for describing gravity as a consequence of the curvature of a four-dimensional manifold, known as spacetime, which results from the presence of mass (or energy).

To describe the curvature of spacetime, we first need to define a metric tensor $g_{\mu\nu}$, that depends on the chosen coordinate system x^μ and defines distances between

nearby points along the spacetime geometry.[1] An infinitesimal distance segment squared ($\mathrm{d}s^2$) between two nearby points in spacetime is invariant independently of which metric and coordinate system is used, this is defined by

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu\,, \tag{1.1}$$

where $\mathrm{d}x^\mu$ and $\mathrm{d}x^\nu$ are infinitesimal differences in the coordinate system.[2]

In regimes where gravitational fields are weak, such as in special relativity, flat spacetime is assumed and the metric is represented by the Minkowski metric defined as $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$. Note that we will follow the convention of a positive signed metric $(-, +, +, +)$ throughout this section.

In curved spacetime, the trajectory of an object moving freely through the x^μ coordinate system is described by the geodesic equation,

$$\frac{d^2x^\lambda}{d\tau^2} + \Gamma^\lambda{}_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0, \tag{1.2}$$

where $\mathrm{d}\tau = c\,\mathrm{d}s$ is an infinitesimal interval of proper time, and c is the speed of light in vacuum. The term $\Gamma^\lambda{}_{\mu\nu}$ are the Christoffel symbols, which can be obtained from the definition of the metric by

$$\Gamma^\lambda{}_{\mu\nu} = \frac{1}{2}g^{\lambda\rho}\left(\partial_\mu g_{\rho\nu} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu}\right), \tag{1.3}$$

where $\partial_\mu \equiv \frac{\partial}{\partial x^\mu}$.

Describing the curvature of the spacetime due to mass-energy, requires defining the Riemann curvature tensor, Ricci tensor, and the Ricci scalar, all of which are defined at the moment the metric is chosen.

The Riemann curvature tensor ($R^\rho{}_{\sigma\mu\nu}$) is defined in terms of the Christoffel symbols and their derivatives,

$$R^\rho{}_{\sigma\mu\nu} = \partial_\mu\Gamma^\rho{}_{\nu\sigma} - \partial_\nu\Gamma^\rho{}_{\mu\sigma} + \Gamma^\rho{}_{\mu\lambda}\Gamma^\lambda{}_{\nu\sigma} - \Gamma^\rho{}_{\nu\lambda}\Gamma^\lambda{}_{\mu\sigma}. \tag{1.4}$$

[1] Throughout this section we will use Greek letters (e.g μ and ν), running from 0 to 3, to describe the 4-dimensional spacetime, where there the 0-component corresponds to a time-like coordinate (e.g ct), and the 1,2 and 3-components correspond to space (e.g x, y, z), also denoted by Latin letters, that run from 1 to 3.

[2] Note that we have adopted Einstein's summation convention where the $\sum$ sign is omitted and the sum is performed over repeated indices.

Next, we define the Ricci tensor ($R_{\mu\nu}$) from the Riemann tensor by,

$$R_{\mu\nu} = R^{\rho}{}_{\mu\rho\nu}. \tag{1.5}$$

At last, the Ricci scalar (R), also known as the curvature scalar, is defined from the Ricci tensor by

$$R = g^{\mu\nu} R_{\mu\nu}. \tag{1.6}$$

The deformation of spacetime by the presence of mass-energy is described by the Einstein Field Equations (EFE)

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}, \tag{1.7}$$

where G is Newton's gravitational constant. The left-hand side describes the curvature of spacetime by the Einstein tensor $G_{\mu\nu}$, defined in terms of the Ricci tensor and Ricci scalar by,

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R. \tag{1.8}$$

The term $T_{\mu\nu}$ on the right-hand side of the EFE is known as the energy-momentum tensor, which describes the mass energy content causing the space-time deformation.

1.2 The ΛCDM Model

Throughout this section, we will elaborate on the standard model of cosmology, also known as the ΛCDM model, which describes the composition and evolution of our Universe. This model is based on several key assumptions. First, the laws of physics are the same in every point of the Universe. Second, Albert Einstein's General Relativity [10] accurately describes gravity at all scales. Third, the Universe is well described by the cosmological principle, meaning that at large scales it is statistically homogeneous and isotropic. Fourth, the Universe began from a hot, dense state known as the Big Bang and has been expanding since. Fifth, the Universe is composed of five main components: Dark Energy, Dark Matter, Baryons, Photons, and Neutrinos. More assumptions and details of the ΛCDM can be found in Section 3.1 of [11].

1.2.1 The homogeneous, and isotropic expanding Universe

As introduced earlier, a fundamental assumption of the standard cosmological model is the statistical homogeneity and isotropy of the Universe at scales beyond roughly 100 Mpc. Furthermore, Edwin Hubble's 1929 measurements confirmed the Universe's expansion [12]. The metric tensor that describes such a Universe is known as the Friedmann–Lemaître-Robertson–Walker metric (FLRW) which in spherical coordinates (r, θ, ϕ) takes the form

$$ds^2 = -c^2\,\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2)\right]. \tag{1.9}$$

In this case (r, θ, ϕ) are a comoving coordinate system, this means that their value remains constant and the evolution of the spatial coordinates due to expansion of the Universe is accounted for only by the scale factor $a(t)$, whose value is normalized to $a_0 = 1$ at present day. The curvature constant k defines the Universe's geometry: $k > 0$ for a closed Universe, $k < 0$ for an open Universe, and $k = 0$ for a flat (Euclidean) Universe.

For a homogeneous, isotropic, and smooth Universe, it is useful to describe its matter-energy content as a perfect fluid. For this kind of fluid the energy-momentum takes the form

$$T^{\mu\nu} = \left(\rho + \frac{P}{c^2}\right)u^\mu u^\nu + P g^{\mu\nu}, \tag{1.10}$$

where ρ is the energy density, P the pressure and u^μ the 4-velocity of the fluid. In general, density and pressure may be constituted by various kind of materials, then $\rho = \sum_i \rho_i$ and $P = \sum_i P_i$.

In GR it is expected that the covariant derivative of the energy-momentum tensor to vanish, this is

$$\nabla_\nu T^{\mu\nu} = \partial_\nu T^{\mu\nu} + \Gamma^\mu{}_{\lambda\nu} T^{\lambda\nu} + \Gamma^\nu{}_{\lambda\nu} T^{\mu\lambda} = 0, \tag{1.11}$$

which gives a local conservation law for energy and momentum. For the FLRW metric, Equation (1.11) results in the Conservation Equation

$$\dot{\rho} + 3\frac{\dot{a}}{a}\left(\rho + \frac{P}{c^2}\right) = 0, \tag{1.12}$$

note that we are using the dot notation for time derivatives (e.g $\dot{\rho} \equiv \frac{\partial \rho}{\partial t}$).

Now, we will focus on the left-hand side of the EFE (Equation (1.7)). From the

description we have done so far, this equation is missing from an ingredient: the cosmological constant Λ, which gives part of the ΛCDM model name.

In 1917, Einstein introduced the cosmological constant Λ into the field equations to allow the possibility of a static Universe [13]. The equations then take the form:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}. \tag{1.13}$$

However, after the evidence of the expansion of the Universe, Einstein referred to this term as his "biggest blunder" and removed it from the equations. About 70 years later, in 1998, two teams independently measured the distances of various Type Ia supernovae. One team was led by Saul Perlmutter [14], and the other by Brian Schimidt and Adam Riess [15] and both concluded that not only the Universe is expanding, but it is also doing so in an accelerating manner. Thus the cosmological constant was re-introduced in the theory to account for this acceleration. During Section 1.2.2 we will explore what this cosmological constant is associated with.

After solving the Christoffel symbols from the FLRW metric, computing the Ricci tensor, and scalar, we will notice that the time-time component of the EFE takes the form:

$$H^2(t) \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3}, \tag{1.14}$$

while if we take the trace of the EFE (i.e taking $g^{\mu\nu}G_{\mu\nu}$), we obtain

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{\Lambda c^2}{3}. \tag{1.15}$$

These two equations are known as the Friedmann Equations that describe the dynamics of the expansion of the Universe. Note that in Equation (1.14), we have defined $H(t) \equiv \dot{a}/a$, known as the Hubble parameter.

Note that, so far, we have three equations describing the evolution of the Universe and its content: The Friedmann Equations (1.14) and (1.15), and the Conservation Equation (1.12). However, the third one can be derived from the first two and therefore is not linearly independent. In other words: we have two linear independent equations and three free dynamic variables: a, ρ, and P. This means that we need another equation for solving this equation system.

Then, we introduce the equation of state, that relates density and pressure. In the ΛCDM model, a barotropic equation of state is usually assumed. This means that the pressure P is assumed to be a function of the density ρ only. Moreover, it is also common to assume a linear relationship between them. Therefore, for the i-th

component of the total content of the Universe, the equation of state takes the form

$$P_i = w_i \rho_i c^2, \tag{1.16}$$

which at the same time leads to Equation (1.12) have the solution

$$\rho_i = \rho_{i,0} a^{-3(1+w_i)}, \tag{1.17}$$

where $\rho_{i,0}$ is the density of the i-th component of the total content of the Universe at present day.

Now, we will introduce the concept of the critical density, ρ_c, defined as the density that the Universe should have in order to be spatially flat

$$\rho_c \equiv \frac{3H_0^2}{8\pi G} = 1.88 h^2 \times 10^{-26} \text{ kg/m}^3, \tag{1.18}$$

where $H_0 = 100h$ km/sMpc^{-1} is the expansion rate of the Universe at present day, also known as the Hubble constant, the parameter h is the dimensionless Hubble parameter.

With these definitions and Equation (1.14) we obtain a relation of the density parameters Ω at present day,

$$\sum_i \Omega_{i,0} + \Omega_{\Lambda,0} + \Omega_{k,0} = 1, \tag{1.19}$$

where $\Omega_{i,0} \equiv \rho_{i,0}/\rho_c$, with ρ_i the density of the i-th component of the total content in the Universe, $\Omega_{\Lambda,0} \equiv \Lambda c^2/3H_0^2$, and $\Omega_{k,0} \equiv -k^2c^2/H_0^2$. In practice, the definitions of the density parameters and Equation (1.19) remain valid over time by simply taking $H(t)$ instead of H_0 in the defintion of critical density (Equation (1.18)).

Equation (1.19) in conjunction with Equation (1.17) leads to another version of the Friedmann Equation:

$$H^2 = H_0^2 \left(\sum_i \Omega_{i,0} a^{-3(1+w_i)} + \Omega_{k,0} a^{-2} + \Omega_\Lambda \right). \tag{1.20}$$

Note that throughout this whole chapter we will be using the notation $\Omega_{i,0}$ for clarity to refer to the present day density parameters values of each species i. In the following chapters we will drop the suffix 0, and obviate that the density parameters Ω_i refer to the present-day values.

1.2.2 Cosmic inventory of the Universe

Now we will describe the components of the total content of the Universe, usually represented by the density parameters Ω_i in Equation (1.14), and their respective equation of state parameters w_i. Additionally, we will describe one possible energy content on the Universe associated with the cosmological constant. We will not discuss the curvature density Ω_k in this section as it is a geometric property rather than a physical component. Moreover, the ΛCDM model posits a flat spatial curvature, supported by various experimentst [e.g 16, 17]. The most recent results from Planck Collaboration provide a $\Omega_k = 0.001 \pm 0.002$ value [18], consistent with a spatially flat Universe.

Photons

The equation of state for photons (γ) takes the value $w_\gamma = 1/3$, and thus, according to Equation (1.17) their density parameter evolves as

$$\Omega_\gamma = \Omega_{\gamma,0} a^{-4}. \tag{1.21}$$

Currently, photons constitute a minor component of the Universe, with the vast majority of them coming from the Cosmic Microwave Background, to be briefly reviewed in Section 1.2.3. The radiation coming from the CMB follows a blackbody distribution, and therefore it is straightforward to know its energy density by determining its temperature at present day T_0 from the relation

$$\rho_\gamma = \frac{\pi^2 k_B^4}{15\hbar^3 c^3} T_0^4, \tag{1.22}$$

where k_B is the Boltzmann constant and $\hbar$ is the reduced Planck constant. Note that a consequence of this equation, and the fact that the photon density evolves as a^{-4} is that the temperature of photons scales as $T_\gamma \propto a^{-1}$, this can also be derived from the laws of thermodynamics.

The Far InfraRed Absolute Spectrophotometer (FIRAS) experiment provided the most precise measurement of the temperature of the CMB at present day with a value of $T_0 = 2.7260 \pm 0.0013$ K [19]. This value in conjunction with Equation (1.22) divided by the critical density (Equation (1.18)), results in a photon density parameter

$$\Omega_{\gamma,0} h^2 = 2.47 \times 10^{-5}. \tag{1.23}$$

Baryonic Matter

Baryonic matter (b) is constituted by all ordinary matter that we observe: all atoms, individual neutrons, protons, electrons, etc. Note that we refer everything to as baryons even though that we know from particle physics that electrons are leptons. However, the majority of the mass content of an atom comes from its nuclei, constituted mostly by protons, and neutrons, and therefore it is reasonable to assume that most of the ordinary matter mass content of the Universe comes from baryons.

Baryons in the Universe are distributed among the galaxies, stars, and planets and diffuse gas distributed among the interstellar, circumgalactic and intergalactic medium. It is safe to assume that Baryons are mostly non-relativistic, and therefore, can be described by a fluid without thermodynamic pressure, i.e taking $w_b = 0$ in Equation (1.16), and therefore their density parameter evolves as

$$\Omega_b = \Omega_{b,0} a^{-3}. \tag{1.24}$$

Determining the total content of baryons in the Universe at present day is more difficult than photons, given that they do not follow a particular distribution function. And directly measuring it includes various considerations such as the thermal state of the Universe that complicate the measurement. The Planck Collaboration [18] presented a value

$$\Omega_{b,0} h^2 = 0.0224 \pm 0.0001. \tag{1.25}$$

Neutrinos

Neutrinos (ν) are neutrally electrical charged leptons that interact with matter only through weak interaction and gravity. There are three flavors of neutrinos associated with each lepton: the electron neutrino ν_e, muon neutrino ν_μ, and the tau neutrino ν_τ.

During the early stages of the Universe, neutrinos behave in a relativistic manner, thus following an equation of state and density evolution similar to photons ($w_\nu = 1/3$, $\rho_\nu \propto a^{-4}$). As the Universe cools down, they become non-relativistic and then follow an equation of state and density evolution similar to baryonic matter ($w_\nu = 0$, $\rho_\nu \propto a^{-3}$).

Neutrinos were long thought to be massless. Following this assumption, taking into account their 3 flavors, and the fact that they follow a Fermi-Dirac distribution, their density parameter can be expressed in terms of the density parameter of photons

as

$$\Omega_{\nu,0}h^2 = 3 \times \frac{7}{8} \times \left(\frac{4}{11}\right)^{4/3} \Omega_{\gamma,0}h^2 = 1.68 \times 10^{-5}, \tag{1.26}$$

see section 2.4.4 of [4] for a detailed derivation of this result.

However, this assumption is not valid as an analysis on atmospheric neutrino data made in 1998 determined that they do indeed have a mass [20]. In such conditions the neutrino density is numerically

$$\Omega_{\nu,0}h^2 = \frac{\sum_\nu m_\nu}{94 \text{ eV}}. \tag{1.27}$$

The exact mass of neutrinos remains an open question in cosmology, with various experiments trying to set constraints on the total sum of their masses. The results of the Planck Collaboration yield $\sum_\nu m_\nu < 0.12$ eV.

Dark Matter

It might seem intuitive to assume that baryons and neutrinos constitute the majority, if not the entirety, of the mass content in the Universe. However, observational evidence suggests the existence of a missing constituent of matter in the Universe that does not interact with particles of the standard model, or if it does, it is only via the weak interaction. This can be inferred from its gravitational interaction with visible matter. We refer to this 'missing" material as Dark Matter.

The concept of Dark Matter was initially proposed in 1933 by Fritz Zwicky, who studied the radial velocities of galaxies within the Coma cluster [21]. His analysis led to the conclusion that the mass required to maintain the observed velocities of galaxies without them escaping the cluster had to be approximately 400 times greater than what was visually observed. Zwicky attributed this discrepancy to a "Dunkle Materie" (Dark Matter in German).

In 1970, Vera Rubin and Kent Ford conducted studies on galaxy rotation curves, starting with the Andromeda Galaxy [22]. Their observations revealed flat rotation curves, implying that stars within galaxies were moving at speeds inconsistent with the visible mass alone, which provided another hint of Dark Matter, although this term was not mentioned in their initial investigation.

More observational evidence includes gravitational lensing, galaxy clustering, the formation of structures in the Universe, the dynamics of galaxy or galaxy cluster collisions (for example, the Bullet cluster [23]), and the imprint that Dark Matter leaves on the CMB through its gravitational interaction with baryons coupled with

radiation during the early stages of the Universe, to mention some.

The nature of Dark Matter remains a mystery, with various models proposed, such as Weakly Interacting Massive Particles (WIMPs), axions, Scalar-Field Dark Matter (SFDM), primordial black holes, and sterile neutrinos, to mention a few. Additionally, these models are further classified based on the velocity (or free streaming length) of the Dark Matter particles as: Cold Dark Matter (CDM) for non-relativistic particles, Hot Dark Matter (HDM) for relativistic particles, and Warm Dark Matter (WDM) for an intermediate case. Refer to [24] for a detailed review on Dark Matter models and observational evidences of its existence.

The CDM in the ΛCDM model indicates that this model adopts Cold Dark Matter as the baseline scenario, as various experiments favor this kind of Dark Matter to explain the observed large-scale structure of the Universe. This is because if Dark Matter were completely hot, the Universe would look different to what we see at present, for example, galaxy clusters would form earlier than individual galaxies on the Universe evolution, while the opposite its observed to occur in reality. Mixed and WDM models remain a possibility, but are studied as extensions of the ΛCDM model.

The Planck collaboration most recent results [18] give a value

$$\Omega_{c,0}h^2 = 0.120 \pm 0.001, \tag{1.28}$$

which compared to Baryionic Matter density (Equation (1.25)) constitutes about 84% of the total matter content of the Universe. In any case, given its characteristics, the equation of state of Dark Matter adopts the form of a pressureless fluid $w_c = 0$, and its density evolves as a^{-3} similar to baryonic matter.

Dark Energy

As mentioned before, there exists observational evidence that our Universe is expanding in an accelerated manner, this is accounted for by the cosmological constant Λ that we introduced in Equation (1.13). If we add up density parameters of all the mass-energy components we have discussed so far and follow the relation Equation (1.20) we will notice that we have only accounted for approximately 31.3% of the total content of the Universe (4.9% baryons, 26.4% Dark Matter, and a minimal contribution of neutrinos and photons). The remaining 68.7% is comes from cosmological constant density parameter.

In the ΛCDM model, the cosmological constant has an equation of state where

$w_\Lambda = -1$, and its energy content is associated with a "Dark Energy" that permeates space and that causes the accelerating expansion. However, the nature of Dark Energy remains a mystery, with the possibility of Λ not being a constant (e.g the wCDM model), or inclusively the case of it not having constant equation of state for Λ (e.g the $w_0 w_a$CDM model).

The presence of Dark Energy and its contribution to the total content of the Universe, can be inferred from experiments such as measuring the distance to supernovae type Ia as mentioned before [e.g 14, 15], and through its contribution on expansion history of the Universe [e.g 16, 17]. The inferred value for the density of Dark Energy on the Universe given by the latest results of the Planck Collaboration [18] is

$$\Omega_{\Lambda,0} h^2 = 0.3107 \pm 0.0082. \tag{1.29}$$

1.2.3 The evolution of the Universe

The very early Universe and inflation.

The Big Bang theory is part of the fundamental assumptions of the ΛCDM model. This theory states that the Universe's origin comes from an infinitely hot and dense state and started expanding from there. The very early moments after the Big Bang, are open to speculation, and their description goes far beyond the current knowledge of physics. Refer to Section 11.4 of [25] and to [26] for a more detailed description of these epochs.

The ΛCDM model framework adopts the inflationary theory, introduced by [27], which proposes that at some point during its evolution, the Universe underwent an exponential expansion process known as cosmic inflation [27–29]. This theory addresses three main problems associated with the Hot Big Bang model: the horizon problem, the flatness problem, and the monopole problem.

- **The horizon problem.** According to the cosmological principle, the Universe appears statistically homogeneous and isotropic on large scales, yet regions separated by distances greater than their causal horizon should not be in thermal equilibrium. This poses a challenge to the explanation of the uniformity of the CMB radiation observed across the sky, to be discussed later on this same section.

- **The flatness problem.** The ΛCDM model adopts a description of a spatially flat Universe, which is further supported by observational evidence. However,

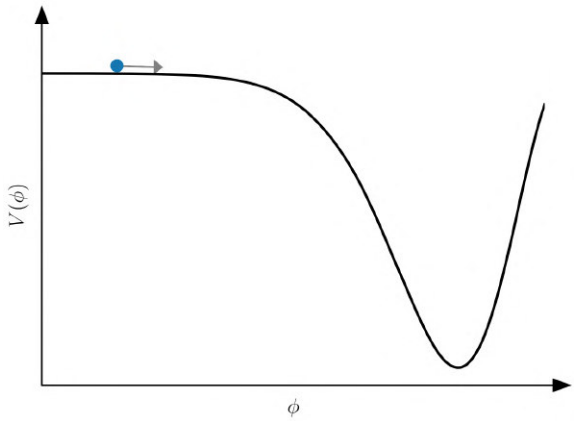

Figure 1.1: A potential $V(\phi)$ that could drive the inflationary epoch, with the scalar field rolling down through the potential. Figure taken from [4].

even the smallest deviations from flatness would have been amplified during cosmic expansion as a factor of a^{-2}, leading to curvature that should be observed nowadays.

- **The monopole problem.** It is thought that a brief moment after the Big Bang, gravity separated from the other three fundamental forces, leading to the Grand Unification epoch, where electromagnetic forces, and nuclear weak and strong forces acted as a single force. The conditions at this epoch provide a perfect scenario for the creation of magnetic monopoles as predicted by certain Grand Unification Theories (GUT, [30]). Yet, none have been observed, posing a challenge to these predictions.

The simplest physical model for cosmic inflation proposes the existence of an inflaton field $\phi(\mathbf{r}, t)$ associated with a potential energy $V(\phi)$ that dominates the energy density of the early Universe [4, 28, 29]. This potential energy drives the exponential expansion of space, similar to a cosmological constant. Figure 1.1 illustrates a potential that could drive this mechanism, with the scalar field rolling down the potential.

As the inflaton field slowly evolves through a condition known as slow-roll, the stored potential energy gradually converts into kinetic energy. This slow-roll phase is crucial for sustaining inflation, allowing it to last long enough to address the aforementioned problems. During this phase, quantum fluctuations in the inflaton field are stretched to macroscopic scales, leading to tiny density perturbations that function as the seeds for the formation of structure later on the evolution of the Universe.

An important prediction of inflation, is that the perturbations generated during this era produce a curvature perturbation $\mathcal{R}$, which defines a primordial power spectrum $P_{\mathcal{R}}$ given by

$$P_{\mathcal{R}} = \frac{2\pi^2}{k^3} A_s \left(\frac{k}{k_p}\right)^{n_s - 1}, \tag{1.30}$$

where n_s is the scalar spectral index, that describes how the perturbations vary with scale; A_s is the amplitude of the primodial power spectrum; and k_p is a pivot scale, usually set to $k_p = 0.05$ Mpc^{-1}.

As inflation ends, the inflaton field oscillates around the minimum of its potential until it settles at the critical point, decaying into radiation and fundamental particles that thermalize the Universe, in a process known as reheating. This transition leaves the Universe in its initial thermal state, and marks the beginning of an era where radiation dominates the energy content, where the Universe is filled mostly with radiation and relativistic particles, setting the stage for further evolution.

Nucleosynthesis and the Matter-Radiation equality.

During the first seconds of the radiation-dominated era following the Big Bang, quarks that had formed during inflation began to bind together, forming the first hadrons, including protons and neutrons. Around the same time, neutrinos decoupled from the rest of the radiation, leading to the formation of the theorized and unobserved Cosmic Neutrino Background ($C\nu B$). A few minutes later, as the Universe continued to cool, conditions became favorable for the nuclear fusion of protons and neutrons, resulting in the formation of the first atomic nuclei in a process known as nucleosynthesis.

Following nucleosynthesis, the Universe continued to cool, but the temperature remained too high for neutral atoms to form. At this stage, the Universe was an opaque plasma because photons were tightly coupled to electrons, with a very small mean free path. This state persisted for several hundred thousand years until the Universe had cooled sufficiently for the formation of the first stable atoms, predominantly hydrogen and helium. During this period, the densities of matter and radiation reached the same value at an epoch known as matter-radiation equality, marking the start of a matter-dominated era with a slower cosmic expansion rate.

Recombination: The Cosmic Microwave Background.

The temperature from the Universe kept cooling until it was such that photons no longer had the sufficient energy to prevent atoms from forming, this marks the start

of an epoch known as recombination. About 380 thousand years after the Big Bang, the energy of the Universe dropped to a level where most of the baryonic content of the Universe was contained into neutral atoms. At this point the mean free path of photons increased, allowing them to travel freely through space giving birth to the Cosmic Microwave Background (CMB).

The existence of the CMB was predicted during 1948 [31], and confirmed after its detection in 1965 [32]. The CMB serves as a outstanding cosmological probe. Various key aspects about our knowledge of the Universe come from studying it, including the density parameters quoted in Section 1.2.2. With various experiments dedicated for its study, including the Cosmic Background Explorer (COBE, [33]), the Wilkison Microwave Anisotropy Probe (WMAP, [34]), the Planck satellite [11], the Atacama Cosmology Telescope (ACT, [35]), and the South Pole Telescope (SPT, [36]), to mention some.

These studies are mainly based on the identification of anisotropies on the temperature of the CMB across the sky, shown in the top panel of Figure 1.2. Including the computation of its angular power spectrum shown at the bottom panel of Figure 1.2, whose shape heavily depends on the cosmological parameters that define the ΛCDM model, to be described in Section 1.2.4. Some other features are understood through the study of the cross-spectrum of the temperature anisotropies with the polarization of the CMB.

Dark Ages and Reionization

Following recombination, the Universe entered the period known as the Dark Ages. During this epoch, the Universe was primarily populated by neutral hydrogen and helium. There were no significant sources of light other than the photons released from the Cosmic Microwave Background (CMB), and therefore, the Universe remained dark. Determining the exact period during which the Dark Ages occurred is challenging, with one of the few observational probes being the 21-cm line emission from neutral hydrogen.

As the Dark Ages progressed, gravitational collapse intensified the small fluctuations in the density field that were seeded during inflation. These fluctuations eventually gathered sufficient Dark Matter and baryonic matter to form the first stars and, later, the first galaxies. The intense radiation emitted by these early celestial bodies began to ionize the surrounding medium. This phase, known as reionization, gradually led to a fully ionized state of the intergalactic medium, as observed today.

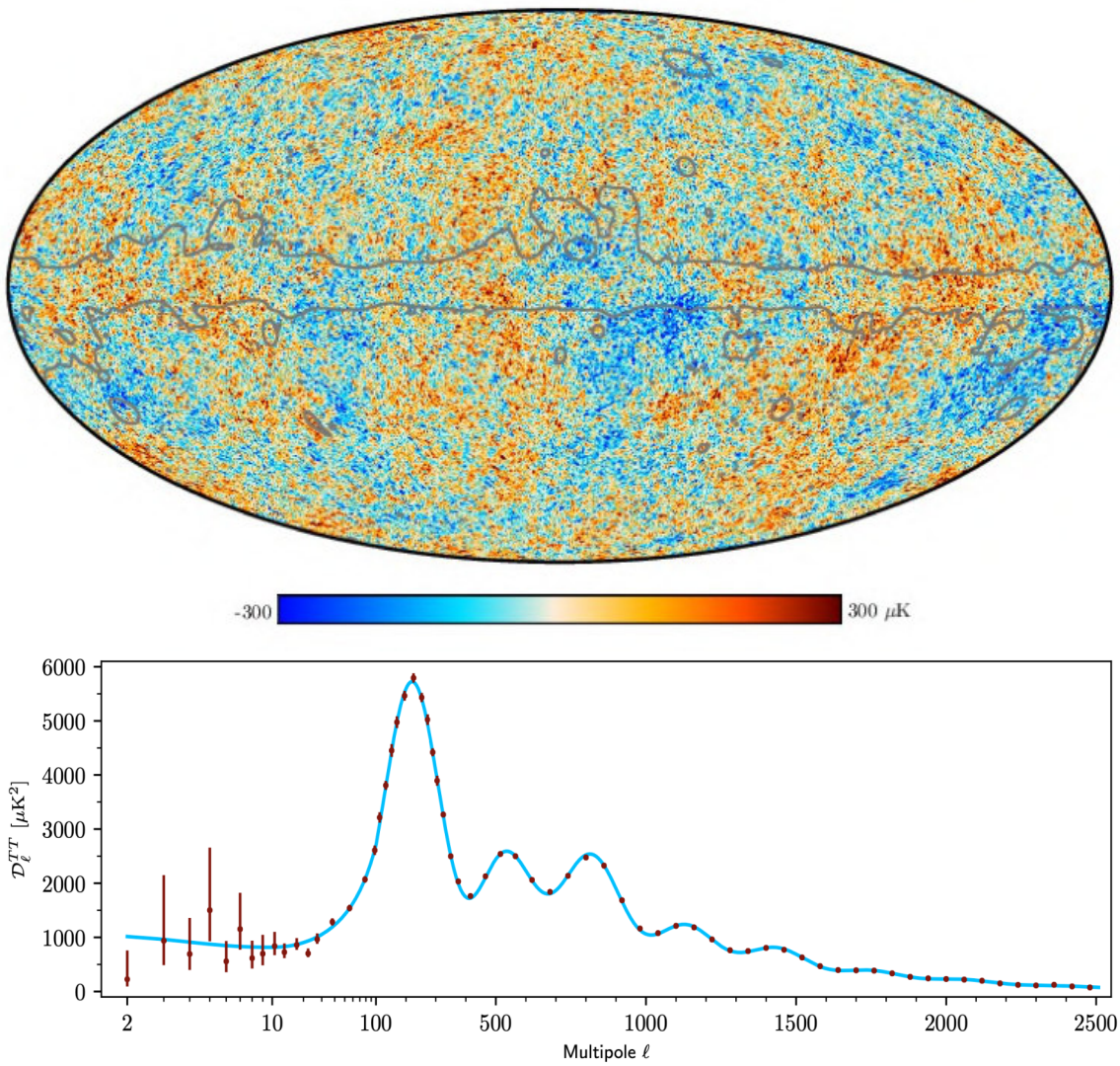

Figure 1.2: Top: Temperature map of the CMB as measured by the Planck collaboration in 2018. The gray lines display regions around the galactic plane that were masked and inpainted. Bottom: CMB temperature angular power spectrum. Figures taken from [11].

Present Days.

Today, we observe a mostly ionized Universe, where stars and galaxies have clustered forming the Large-Scale Structure (LSS) of the Universe. This is clearly appreciable in Figure 1.3, that showcases an slice in the galaxy map obtained by the Dark Energy Spectroscopic Instrument (DESI) during its first year of operations. This consists of the largest galaxy spectroscopic sample up to the date this thesis is being written.

At some point during the matter-dominated era, the matter density decreased sufficiently for Dark Energy to dominate the energy content of the Universe, starting the accelerated expansion process that we observe today. Current cosmological insights are heavily supported by observations of the CMB radiation, and more recently by the Baryon Acoustic Oscillations feature imprinted in galaxy clustering, as we will discuss at the end of this chapter. Yet, there remain a lot of mysteries about the Universe that must be unveiled.

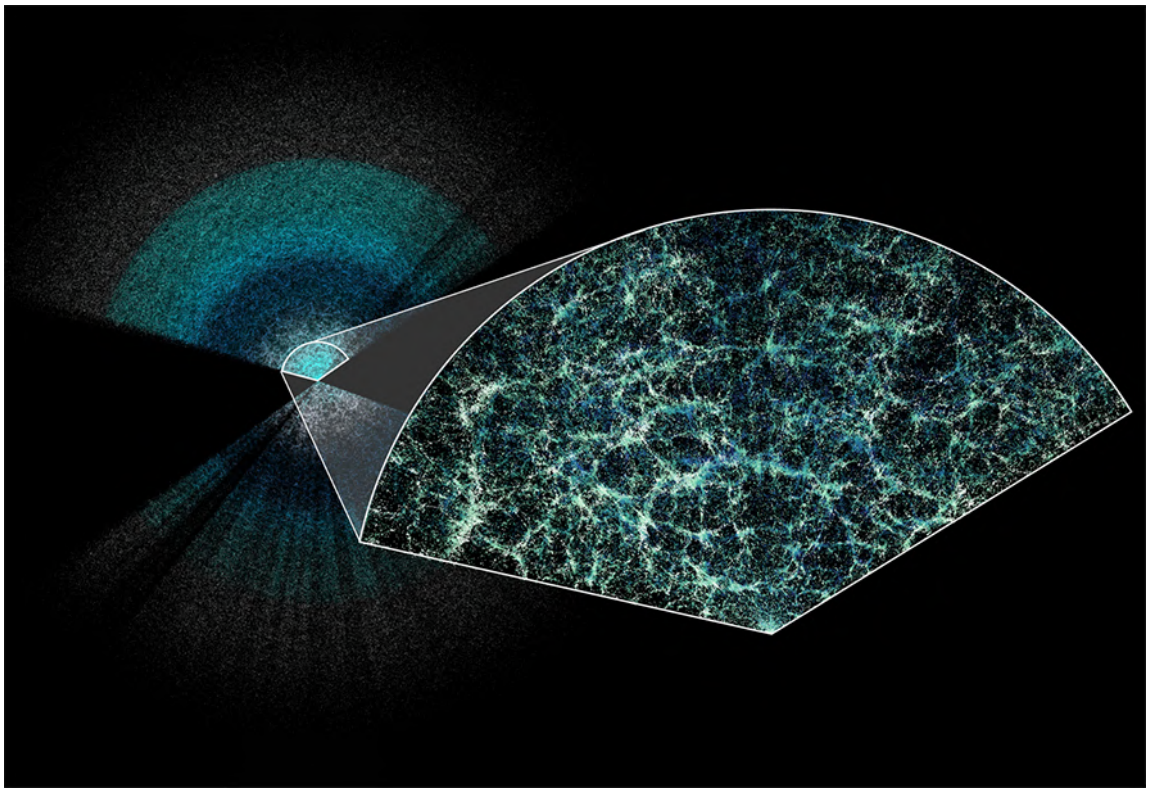

Figure 1.3: Slice spanning 190 degrees in right ascension and 14 degrees in declination of the 3D galaxy map as obtained by DESI during its first year of operations. The background image colors correspond to different kinds of tracers. Colors in zoom-in region correspond to the depth in the image to highlight the 3-dimensional Large-Scale Structure of the Universe. Credits to Claire Lamman and the DESI collaboration.

1.2.4 Cosmological parameters

The ΛCDM model is constructed around six free parameters, referred to as the cosmological parameters, that describe the Universe's composition, geometry, and dynamics. The parameters includes the Dark Matter $\Omega_{c,0}h^2$ and Baryon density parameters $\Omega_{b,0}h^2$, discussed in Section 1.2.2; the angular scale of acoustic fluctuations defined as $\theta_* = r_s(z_*)/D_M(z_*)$, where $r_s(z_*)$ is the BAO scale (to be defined during Section 1.4) at the recombination epoch redshift (z_*); the optical depth τ of the Universe at reionization epoch; and the spectral amplitude A_s, and spectral index n_s of the primordial power spectrum set during inflation (Equation (1.30)). All other physical quantities may be derived from these six main parameters.

The values measured by the Planck collaboration are presented in Table 1.1, usually defined as a reference for other cosmological studies.

Table 1.1: Cosmological Parameters in the ΛCDM Model as reported by the Planck Collaboration [18].

Parameter	Symbol	Value
Dark Matter physical density	$\Omega_{c,0}h^2$	0.1200 ± 0.0012
Baryon physical density	$\Omega_{b,0}h^2$	0.02237 ± 0.00015
Angular size of the sound horizon at z_*	$100\theta_*$	1.04110 ± 0.00031
Optical depth at reionization	τ	0.0544 ± 0.0073
Primordial power spectrum amplitude	$\ln(10^{10}A_s)$	3.044 ± 0.014
Scalar spectral index	n_s	0.9649 ± 0.0042

1.3 Distances in Cosmology

Throughout this thesis we will focus on the study of the expansion history of the Universe by measuring distances between objects scattered among the vast Universe space. Therefore in this section we will review on useful distance definitions and how they are measured with various techniques.

1.3.1 Redshift

In cosmology it is useful to determine the redshift (z) of objects in order to know their distance from us. The term redshift refers to the phenomenon in which the observed wavelength of light from an object is increased (or shifted to the red end of the spectrum). The opposite effect is known as blueshift. In terms of equations,

the redshift of an object is given by

$$1 + z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{rest}}}, \tag{1.31}$$

where λ_{obs} is the observed wavelength, and λ_{rest} is the wavelength at the emission point (rest-frame wavelength).

The three main causes of redshift in cosmology are:

- **Doppler redshift (z_{Doppler}):** This type of redshift occurs as a consequence of the movement of an object relative to the observer. If the object moves away from the observer, then it will exhibit a redshift, while if it moved towards the observer, then it would be blueshifted.

  This is caused by the Doppler effect for electromagnetic waves, and thereby receives the name of Doppler redshift. This redshift is related to the velocity v of the object with respect to the observed by

  $$1 + z_{\text{Doppler}} = \sqrt{\frac{1 + \frac{v}{c}}{1 - \frac{v}{c}}}. \tag{1.32}$$

  If the peculiar velocity of the object is non-relativistic, then this redshift reduces simply to $z_{\text{Doppler}} \approx v/c$. The most common source of this kind of redshift is given by the peculiar velocities of galaxies moving through space, in addition to their movement caused by the expansion of the Universe. In most cases, peculiar velocities of galaxies are in the order of a few hundred km/s, which makes this type redshift be considerable only for objects close the Earth (i.e low redshift $z \ll 1$).

- **Gravitational redshift (z_{grav}):** This kind of redshift occurs as a consequence of light losing energy due to its interaction with a gravitational potential distorting spacetime.

  This effect was predicted by Einstein in 1907 during a study of the effect of gravity on the propagation of light [37] and is related to the mass M of the object causing the spacetime distortion by

  $$1 + z_{\text{grav}} = \left(1 - \frac{2GM}{rc^2}\right)^{-1/2}, \tag{1.33}$$

  where r is the radial coordinate of the photon's emission point. While this type of redshift serves as an important classical test of GR, in cosmology it is neg-

ligible compared to the aforementioned Doppler redshift, and the cosmological redshift to be described next, except for some effects like the Sach-Wolfe effect in CMB studies.

- **Cosmological redshift (z_{cosmo}):** The major contribution of an object's redshift, at cosmological scales, comes from the expansion of the Universe itself.

  As the Universe expands, the space between objects increases, and consequently the wavelength of an emitted photon traveling through space also suffers from this effect, causing an observable redshift in the spectrum of the object. This redshift is related to the scale factor (a) by

$$1 + z_{\text{cosmo}} = a^{-1}, \tag{1.34}$$

We can use the cosmological redshift definition from Equation (1.34), and the density evolution of the contents of the Universe discussed in Section 1.2.2 to rewrite Equation (1.20) as

$$H(z) = H_0\left[\Omega_m(1+z)^3 + \Omega_r(1+z)^4 + \Omega_k(1+z)^2 + \Omega_\Lambda\right]^{1/2}, \tag{1.35}$$

note that we have written the total matter density as $\Omega_m = \Omega_b + \Omega_c$, and the total radiation density as $\Omega_r = \Omega_\gamma + \Omega_\nu$, considering neutrinos as relativistic particles.

1.3.2 Distance definitions

Now we have sufficient tools to define distances in the Universe. The first useful definition will be the Hubble Distance D_H, which in practical terms defines the size of the observable Universe at the time-scale defined by the redshift z as

$$D_H(z) = \frac{c}{H(z)}, \tag{1.36}$$

at present time this defines the Hubble length $D_{H_0} = c/H_0 \approx 3000$ Mpc$/h$, also referred to as the Hubble distance, however, for this thesis we will reserve that term for the value of $D_H(z)$.[3]

Now, we will define the comoving radial distance D_C, related to the distance between a distant object and an observer placed at Earth ($z = 0$), which also takes

[3]Throughout this thesis it will be useful to write distances in terms of Mpc$/h$ units, where h is the dimensionless Hubble parameter.

into account the expansion of the Universe,

$$D_C(z) = \int_0^z \frac{c\,\mathrm{d}z'}{H(z')}, \tag{1.37}$$

this distance remains constant in time, although it does not take into account the contributions by the peculiar motion of the object.

If this same object has an apparent angular size $\delta\theta$ in sky, the comoving distance between two opposite sides of the object would be $D_M\delta\theta$, where D_M is the comoving transverse distance $D_M(z)$ of the object, which as the radial comoving distance remains constant in time, and is defined by

$$D_M(z) = \begin{cases} \frac{c}{H_0\sqrt{|\Omega_k|}} \sin\left(\sqrt{|\Omega_k|}\frac{H_0}{c}D_C\right), & \text{for } \Omega_k < 0, \\ D_C, & \text{for } \Omega_k = 0, \\ \frac{c}{H_0\sqrt{\Omega_k}} \sinh\left(\sqrt{\Omega_k}\frac{H_0}{c}D_C\right), & \text{for } \Omega_k > 0, \end{cases} \tag{1.38}$$

depending on the curvature density Ω_k of the Universe.

It is also useful to define the angular diameter distance D_A, related to the comoving transverse distance by

$$D_A(z) = \frac{D_M(z)}{1+z}, \tag{1.39}$$

which at the same time is related to the physical size s of the object and its observed angular size θ by $s \approx D_A\theta$ for small angles.

Finally, consider a distant object with absolute luminosity L. In a completely flat and static Universe, the observed luminosity at Earth of this object would simply decay as the inverse of the distance to the object D_L, referred to as the luminosity distance, giving an observed flux

$$F = \frac{L}{4\pi D_L^2}. \tag{1.40}$$

In reality, one must take into account the expansion of the Universe in the total distance traveled by a photon emitted at the distant object. In such a case, the luminosity distance is related to the comoving transverse distance by

$$D_L(z) = (1+z)D_M(z), \tag{1.41}$$

this distance is useful to perform a model independent measurement of the distance

to bright objects relatively close to the Earth (low redshift), where the effects of the expansion of the Universe are sub-dominant compared to the peculiar motion of the object. Therefore, the distance can be inferred only from the observed flux F and a known luminosity L of a physical phenomenon.

1.3.3 The Cosmic Distance Ladder

Now that we have introduced the concept of distances in the Universe, we will focus on the methods used to measure them. As noted already, measuring distances in the Universe is a fundamental task for understanding its nature. Given the vast range of scales in the Universe, no single method is capable of precisely determining distances across all cosmic scales. Instead, one must employ a hierarchy of techniques, collectively known as the Cosmic Distance Ladder [38], where each step represents a method suitable for determining distances at a specific scale. Moreover, each step relies on the previous ones for calibration and standardization.

Astronomical Unit: The first steps on the ladder are based on direct measurements. The simplest among them is the average distance between Earth and the Sun, which defines the fundamental unit of distance in astronomy: the Astronomical Unit (AU), approximately equal to 1.496×10^{11} m. The method for determining the exact value of the astronomical unit has varied over time, including observations of transits of Venus and the application of Kepler's laws of planetary motion. The most recent techniques are based on radar measurements of the relative positions of Earth to other planets or nearby asteroids, providing a precise measurement of the astronomical unit.

Parallax: Observing the position of a nearby star relative to distant objects on one night, and then comparing it to its position six months later, reveals a slight shift known as parallax. This apparent displacement of an object due to the observer's change in position provides a direct method for measuring distances. By measuring the parallax angle shift and using trigonometric relations between this angle and the astronomical unit, the distance to the object can be determined. Distances in cosmology are often given in parsecs (pc), defined as the distance at which an object would have a parallax angle of 1 arcsecond, meaning 1 pc $\approx 3.086 \times 10^{16}$ m, based on the definition of an AU.

Standard Candles: Parallax offers a direct distance measurement for objects relatively nearby, but at larger scales, the parallax angle becomes nearly indistinguishable. For these distances, it is useful to define distances in terms of objects of known luminosity, known as standard candles. By convention, the difference between the apparent magnitude m of a source and its absolute magnitude M is related to the luminosity distance by $m - M = 5\log(D_L/10\text{ pc}) + K$, where K is a correction for the shift in the spectrum caused by the expansion of the Universe and $m - M$ is known as the distance modulus. The most common types of standard candles are Cepheid Variables and Type Ia supernovae.

Standard Sirens: A new step on the cosmic distance ladder was added with the detection of gravitational waves in 2016. This new step involves measuring gravitational waves emitted by the merger of compact binary systems, such as neutron stars or black holes. During the merger of these binary systems they emit gravitational waves with a frequency that increases as the objects get closer, this is known as the inspiral phase, until they merge, at which point the frequency suddenly decreases. The distance to these binary systems can then be inferred by comparing the amplitude of the observed signal to the frequency evolution related to the system's total mass (chirp mass) and output energy. These objects are known as standard sirens, in analogy to standard candles, given that in this case we are observing frequencies and amplitudes, similar to the sounds of an ambulance's siren.

Standard Ruler: The cosmic distance ladder also includes the use of the Baryon Acoustic Oscillations (BAO) scale feature imprinted in the distribution of matter in the Universe as an standard ruler to measure distances at the scale of galaxy clustering. Which has the main advantage that is does not require to be calibrated with the other steps of the distance ladder as the other methods do, and the only drawback being that the size of the BAO scale depends on the cosmological model adopted. This is due to the fact that its scale size depends on the photon-to-baryon ratio of the Universe, and its evolution in time only depends on the expansion of the Universe. We will explore further on BAO during Section 1.4.

1.4 Baryon Acoustic Oscillations (BAO)

During the early stages of the Universe's evolution, photons and baryons were tightly coupled due to Compton scattering. In overdense regions, gravity attracted baryons,

while thermal pressure, which results from photon interactions, pushed them outwards. This competition between gravitational attraction and thermal pressure generated an oscillatory behavior, known as Baryon Acoustic Oscillations (BAO), that created wavefronts that propagated through the primordial photon-baryon plasma at a velocity given by

$$c_s(z) = \frac{c}{\sqrt{3}}[1 + R(z)]^{-1/2}, \tag{1.42}$$

where c is the speed of light in vacuum, and $R(z) = 3\rho_b/4\rho_\gamma$ represents the baryon-to-photon density ratio at redshift z.

The size of the wavefronts is referred to as the sound horizon scale (r_s) are defined by,

$$r_s(z) = \int_\infty^z \frac{c_s(z)}{H(z)}\, dz. \tag{1.43}$$

These wavefronts traveled through the primordial plasma until the Universe expanded and cooled down sufficiently for photons and baryons to decouple. Shortly thereafter, the size of the wavefront was frozen, as photons no longer exerted thermal pressure on baryons, leaving an overdense shell of baryons. This period is known as the drag epoch, occurring at a redshift z_d, and the size of the sound horizon at this epoch, $r_d \equiv r_s(z_d)$, is referred to as the BAO scale.

Two important aspects are worth noting. First, after the drag epoch, the evolution of the BAO scale was influenced only by the expansion of the Universe, meaning that its size in comoving coordinates remains constant. Second, since Dark Matter interacts primarily through gravity, the initial overdense regions continued to grow as Dark Matter clustered in these regions from the beginning of the Universe's expansion. These Dark Matter enriched regions, along with the agglomeration of baryons at the BAO scale, became the seed for galaxy formation later in the Universe's evolution history.

1.4.1 Mathematical description of BAO

To describe this effect mathematically, first we start by considering small perturbations in the density field of the Universe given by

$$\rho(\mathbf{x}) = \bar{\rho}[1 + \delta(\mathbf{x})], \tag{1.44}$$

where $\bar{\rho}$ is the mean density content of a given material in the Universe in the homogeneous case, and $\delta(\mathbf{x}) \ll 1$ is a small perturbation on the density field that depends on the position (anisotropy). Next, we consider a small scalar perturbation

on the FLRW metric given by

$$g_{\mu\nu} = g_{\mu\nu}^{\text{FLRW}} + h_{\mu\nu}(\mathbf{x}, t), \tag{1.45}$$

where $g_{\mu\nu}^{\text{FLRW}}$ is the unperturbed FLRW metric described in Equation (1.9) $h_{\mu\nu}$ is a spacetime dependent scalar perturbation defined by two fields Φ and Ψ by

$$\begin{aligned} h_{00}(\mathbf{x}, t) &= -2\Psi(\mathbf{x}, t), \\ h_{0i}(\mathbf{x}, t) &= 0, \\ h_{ij}(\mathbf{x}, t) &= 2a^2(t)\delta_{ij}^K \Phi(\mathbf{x}, t), \end{aligned} \tag{1.46}$$

where δ_{ij}^K is a Kronecker delta. Here, Ψ refers to a gravitational fluctuation and Φ to a perturbation of the spatial curvature. Finally, we must consider the interaction of a non-ideal, non-isotropic, and non-homogeneous fluid. In this regard, we must compute the energy-momentum tensor to be able to solve the perturbed Einstein Field Equations. This requires considering temperature fluctuations in the fluid described by $T(\mathbf{x}, \hat{\mathbf{p}}, t) = T(t)[1 + \Theta(\mathbf{x}, \hat{\mathbf{p}}, t)]$, with $\Theta(\mathbf{x}, \hat{\mathbf{p}}, t)$ a small temperature perturbation. And solving the Boltzmann equation defined by

$$\frac{\mathrm{d}f(\mathbf{x}, \mathbf{p}, t)}{\mathrm{d}t} \equiv \frac{\partial f}{\partial t} + \dot{\mathbf{x}} \cdot \boldsymbol{\nabla}_x f + \dot{\mathbf{p}} \cdot \boldsymbol{\nabla}_p f = C[f], \tag{1.47}$$

for tightly coupled photons and baryons including collision terms due to Compton scattering. Here f is the particle distribution function of the fluid in phase-space that takes into account the temperature fluctuations Θ. $\boldsymbol{\nabla}_x$ and $\boldsymbol{\nabla}_p$ denote the gradient with respect to the coordinates $\mathbf{x}$ and the momentum $\mathbf{p}$. The term $C[f]$ is known as the collision term that describes how particles move from one phase-space element to another.

With these three considerations, the Einstein Field Equations yield to the equations of motion that govern the BAO phenom. According to section 9.3 of [4], the equation of motion for photons take the form,

$$\Theta_0'' + \frac{a'}{a}\frac{R}{1+R}\Theta_0' + k^2 c_s^2 \Theta_0 = -\frac{k^2}{3}\Psi - \frac{a'}{a}\frac{R}{1+R}\Phi' - \Phi'', \tag{1.48}$$

in Fourier space. Additionally, according to Equation 1 of [39], the equation of motion for baryons takes the form,

$$\delta_b'' + \frac{a'}{a}\frac{R}{1+R}\delta_b' + k^2 c_s^2 \delta_b = -k^2\Psi - 3\frac{a'}{a}\frac{R}{1+R}\Phi' - 3\Phi''. \tag{1.49}$$

In Equations (1.48) and (1.49) k is the wave number, δ_b denotes fluctuations in the baryon density, and $\Theta_0(\mathbf{x}, t) = \frac{1}{4\pi}\int \Theta(\mathbf{x}, \hat{\mathbf{p}}, t)\, \mathrm{d}\Omega$ is the monopole of temperature fluctuations $\Theta(\mathbf{x}, \hat{\mathbf{p}}, t)$, where $\mathrm{d}\Omega$ in this case denotes the solid angle differential. Note that for these particular equations we have adopted units where the speed of light c, the reduced Planck constant $\hbar$ and the Boltzmann constant k_B are set to unity for convenience. The tilde notation corresponds to the derivative with respect to the conformal time defined by $\eta = \int \mathrm{d}t\,/a$.

Notice that Equations (1.48) and (1.49) are similar to those of a damped driven harmonic oscillator. Here $k^2 c_s^2$ is the frequency of the oscillator, $\frac{a'}{a}\frac{R}{1+R}$ acts as the damping factor, and the right-hand side of the equations as the driving force. See Chapter 9 of [4] and [40] for a more detailed discussion on these equations, including their derivation.

1.4.2 BAO as a cosmological Probe

Since BAO provides an standard ruler to measure distances in the Universe, it can be used as a tool to measure the cosmic expansion of the Universe, and therefore it is a great tool to test cosmological models. In this section, we will exemplify the use of BAO as a cosmological probe by reviewing its effect on the CMB temperature anisotropies power spectrum, and galaxy clustering.

BAO in the CMB

The Cosmic Microwave Background (CMB) stands out as one of the most precise probes for the ΛCDM model, providing tight constraints on cosmological parameters. The CMB temperature anisotropies and polarization carry valuable information about the early Universe and its evolution.

As a consequence of the tight couple between baryons and radiations, BAO leaves an imprint on the CMB temperature anisotropy power spectrum, distinguished as the characteristic peaks and troughs visible in the bottom panel of Figure 1.2.

The content of Baryonic and Dark matter may be inferred by the position of the peaks and troughs, and the relative heights between them, as they are tightly related, see Figure 1.4. Therefore, the CMB temperature anisotropies power spectrum provides a high precision measurement that allows the determination of the matter content of the Universe, and thereby the BAO scale. In particular, the first peak seen approximately at $l \sim 200$ is corresponds to the scale of the sound horizon at the recombination computed from Equation (1.43) at redshift z_*. The latest results

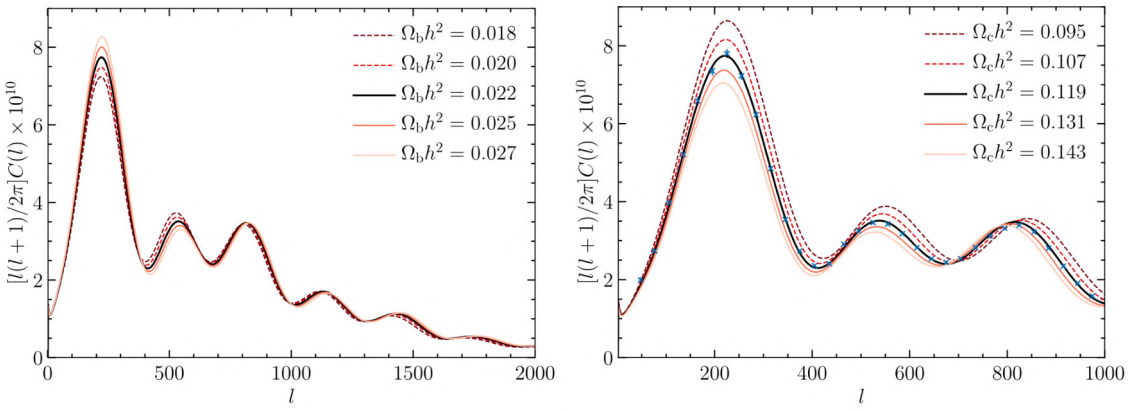

Figure 1.4: Effect of variations on the Baryonic $\Omega_b h^2$ (left panel) and Dark $\Omega_c h^2$ (right panel) in the CMB power spectrum. The blue dots in the right panel show the results obtained by the Planck Collaboration [18]. Figures taken from [4].

obtained by the CMB power spectrum study from Planck Collaboration [18] gave a value for the BAO scale of

$$r_d = 147.09 \pm 0.26 \text{ Mpc}. \tag{1.50}$$

BAO on the clustering of matter

As mentioned before, the BAO scale feature imprinted in the baryonic matter distribution along with the Dark Matter enriched regions located at the position of the perturbations on the density field became the seeds for forming the first stars and galaxies.

In this regard, one would expect that if we perform a statistical analysis of the comoving separation between galaxies or other tracers of matter on the large-scale structure of the Universe, such as the Lyman-α forest (to be described during Chapter 2), we should obtain the BAO scale. This is the foundation of modern spectroscopic surveys, whose goal is to build 3-dimensional maps of the Universe by measuring the angular positions and redshift of several galaxies in order to understand the properties of the Universe.

Since the mapping of galaxies is built in redshift-space, this introduces an effect where their peculiar velocities may change their position in real-space, this effect is known as the Redshift Space distortions (RSD). Furthermore, galaxies are a biased tracer of matter, which in conjunction with RSD means that their power spectrum P_g can be related to the linear matter power spectrum P_L by the Kaiser form [41]

$$P_g(\mathbf{k}) = (b + f\mu_k^2)^2 P_m(k) = b^2(1 + \beta\mu_k^2)^2 P_L(\mathbf{k}), \tag{1.51}$$

where $\mathbf{k} = (k_\parallel, k_\perp)$ is the wave-number vector of magnitude k, and $\mu_k = k_\parallel/k$, b is a bias parameter that linearly relate the overdensities of galaxies δ_g with the overdensities of matter δ_m by $\delta_g = b\delta_m$. In the expression at the middle of Equation (1.51), f is the growth rate of structure, related to the evolution of the density perturbation in matter distribution. In the rightmost expression, we have defined the RSD parameter β related to the bias and the growth rate by $\beta = f/b$.

Most BAO scale studies through matter tracer clustering rely on computing a correlation function of their positions, related to the linear power spectrum simply by a Fourier transform. The first ever detection of the BAO feature on the correlation function of galaxies was performed during 2004 by measuring the correlation function from 46.7k Luminous Red galaxies from the Sloan Digital Sky Survey [42], from that day until now the efforts on enhancing the methods to measure the correlation function, characterizing and reducing the systematic errors on the analysis, reaching percent level accuracy by augmenting the size of the galaxy samples, and including more types of galaxies to be used on the measurement of the expansion rate of the Universe at different redshift epochs, which provide an invaluable observable to constrain the nature of the Universe. We will briefly review the history of galaxy surveys during Chapter 3, while paying special attention to DESI, the latest effort on spectroscopic galaxy surveys. During its first year, DESI provided tight constraints on the cosmic expansion history of the Universe by measuring the BAO scale at 7 different redshifts from the correlation function of 6 types of tracers, as seen in Figure 1.5. The cosmological interpretation of these results are fully detailed in [17], which reports a measured BAO scale value of

$$r_d = 101.8 \pm 1.3 \text{ Mpc/h}, \tag{1.52}$$

and several other important results, including the nature of Dark Energy, to be further confirmed on the coming years.

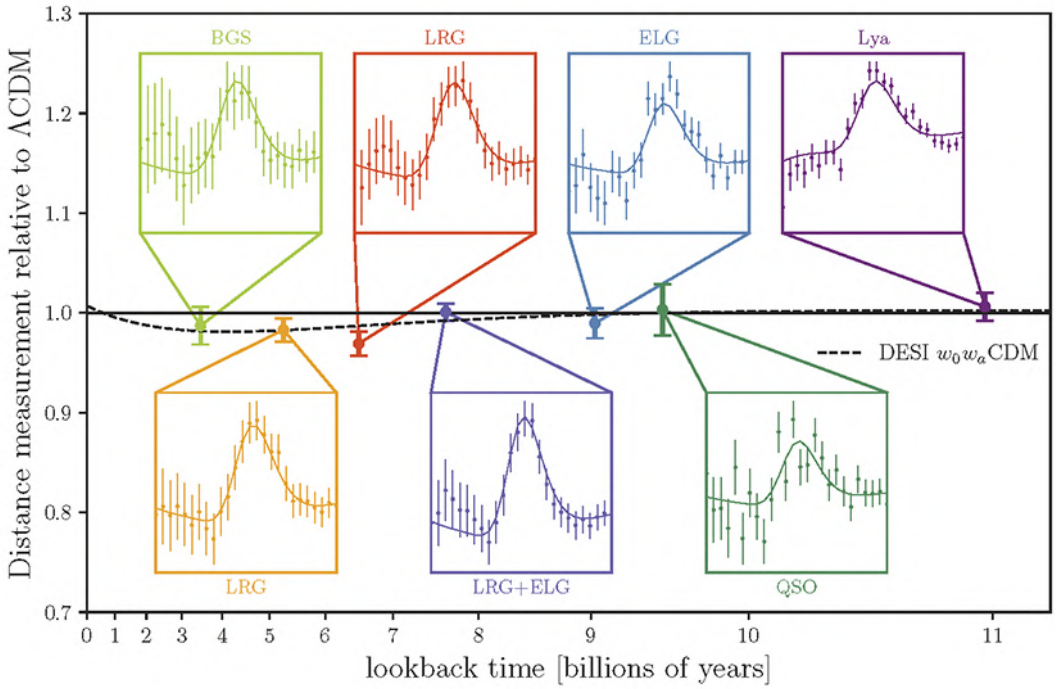

Figure 1.5: Expansion history of the Universe as measured by DESI during its first year of operations, relative to the predictions of the ΛCDM model. The figure displays the correlation functions measured on 6 different types of tracers (and combinations of them) at 7 different redshift ranges. Credits to Arnaud de Mattia and the DESI Collaboration.

2

The Lyman-α Forest

Among the galaxy repertoire in the Universe, quasars, also referred to as quasi-stellar objects (QSOs),[1] stand out as some of the most luminous and energetic objects in the distant Universe. Quasars belong to the group of Active Galactic Nuclei (AGN) galaxies, distinguished by hosting a super massive black hole (SMBH) at their core, surrounded by gas and dust in their accretion disk. Moreover, the characteristic brightness of AGNs results from the interaction of the material in the accretion disk with the SMBH, which is heated up by the gravitational and frictional forces. AGNs can outshine the entire host galaxy, emitting non-thermal radiation across the complete wavelength range from the radio to the X-ray regime. Including characteristic emission lines caused by transitions such as the Lyman series, C II, C IV, O VI, Mg II, Si IV, among others.

The high luminosity of quasars makes it possible to detect them from great distances, making them the perfect candidates to explore the high redshift stages of the Universe. Quasar spectra often present a series of absorption features blueward of the Lyman-α transition line ($\lambda_{\mathrm{Ly}\alpha} = 1215.67$ Å) caused by the interaction of the light of the quasar with neutral Hydrogen (HI) in the Inter-Galactic Medium (IGM) along the line of sight tracer between the quasar and the observer. The collection of these lines is called the Lymanα forest. The Lyman-α forest of sufficiently high redshift quasars ($z > 2.1$) can be observed by optical telescopes in the visible wavelength range of their spectra.

The Lyman-α forest has proven to be an important tool in our understanding of the Universe at redshifts $2 < z < 4$. The goal of this chapter is to introduce the physical aspects of the Lyman-α forest, its use as a matter distribution tracer, and the methods used to measure the BAO scale feature imprinted in it, including the Lyman-α 3-dimensional correlation function estimator and physical model.

[1] Throughout the extension of this thesis we will use the term quasars or QSO instinctively.

2.1 Physics of the Lyman-α forest

First, we will describe the underlying physical processes that give rise to the Lyman-α forest. Some of the contents in this section are based on a review of the Evolution of the IGM [43] and a Galaxy Formation and Evolution book [Chapter 16 of 44].

2.1.1 The Lyman series

The first key part to understanding the physical process behind the Lyman-α forest comes from the nature of the Hydrogen atom itself. From quantum mechanics we know that the energy levels of neutral Hydrogen are quantized and can be described by

$$E_n = -\frac{m_e e^4}{2(4\pi\epsilon_0\hbar)^2}\frac{1}{n^2} = -\frac{13.6 \text{ eV}}{n^2}, \tag{2.1}$$

where m_e is the mass of the electron, ϵ_0 is the electric permittivity of vacuum, $\hbar$ is the reduced Planck constant and e is the electron charge. The right-hand side of Equation (2.1) is obtained by expressing the fundamental constants $(m_e, e, \epsilon_0, \hbar)$ in terms of electron-volt units. The difference between an energy level n and a higher energy level m is then given by

$$\Delta E_{n,m} = 13.6 \text{ eV}\left(\frac{1}{n^2} - \frac{1}{m^2}\right). \tag{2.2}$$

Combining Equation (2.2) with the energy of a photon $E = 2\pi\hbar c/\lambda$, where c is the speed of light, results in

$$\frac{1}{\lambda} = R_{\text{H}}\left(\frac{1}{n^2} - \frac{1}{m^2}\right), \tag{2.3}$$

where $R_{\text{H}} = 13.6 \text{ eV}/(2\pi\hbar c)$ is the Rydberg constant for Hydrogen. Considering the transition between an excited state m and the ground state $n = 1$ of Hydrogen in Equation (2.3) gives the equation known as the Lyman series with wavelengths given by

$$\lambda_m = \frac{1}{R_{\text{H}}}\left(\frac{m^2}{m^2 - 1}\right) = 911.75 \text{ Å}\left(\frac{m^2}{m^2 - 1}\right), \tag{2.4}$$

where the transition the first excited state $m = 2$ and the ground state receives the name of the Lyman-α transition with wavelength $\lambda_\alpha = 1215.67$ Å, $m = 3$ receives the name of Lyman-β transition with wavelength $\lambda_\beta = 1025.72$ Å, and so on until the Lyman-limit $\lambda_\infty = 911.75$ Å.

2.1.2 The Intergalactic Medium

The Lyman-α forest is produced by the interaction of light from distant quasars with neutral Hydrogen in the intergalactic medium (IGM) along the observer's line of sight toward the quasar. As the quasar's continuous light travels toward the observer, photons with energies similar to the one from the Lyman-α transition are absorbed by neutral Hydrogen present in intervening gas clouds along the line of sight.

Recall that, as the Universe expands, light is redshifted and therefore it loses energy. Consequently, each intervening gas cloud along the line of sight of the quasar produce an absorption feature in the quasar's spectrum, at the wavelength $\lambda = (1 + z_{\rm abs})\lambda_\alpha$. This is the Lyman-$\alpha$ wavelength λ_α redshifted to the position of the absorbing gas cloud at redshift $z_{\rm abs}$. In other words, the Lyman-α forest is linked intrinsically to the evolution of the IGM through history, and how matter is distributed among it.

More precisely, the shape and distribution of the Hydrogen absorption features found in the Lyman-α forest are described by the transmitted flux fraction of the quasar radiation F, which is related to the optical depth τ of the intervening Hydrogen clouds simply by

$$F(\lambda) = e^{-\tau(\lambda)}. \tag{2.5}$$

The optical depth specifies how optically thick (or transparent) the absorbing cloud is to an incident Lyman-α photon of frequency $\nu = c/\lambda$. It is related to the interaction cross-section of the Lyman-α transition $\sigma_{\rm Ly\alpha}$ and the neutral Hydrogen number density of the absorber $n_{\rm HI}$ at a given time t by

$$\tau(\nu) = \int n_{\rm HI}(t)\sigma_{\rm Ly\alpha}(\nu)\,{\rm d}l\,, \tag{2.6}$$

where ${\rm d}l = a(t)\,{\rm d}r$ is an infinitesimal proper distance interval. Note that for convenience we have expressed the optical depth as a function of the incident photon frequency instead of its wavelength. In what follows, we will elaborate on these two quantities.

The Lyman-α cross-section

First, we will review the model of the Lyman-α cross-section. For an incident photon of frequency ν the interaction cross-section is given by

$$\sigma_{\mathrm{Ly}\alpha}(\nu) = \frac{\pi e^2}{m_e c} f_{\mathrm{Ly}\alpha} \Phi(\nu), \tag{2.7}$$

where $f_{\mathrm{Ly}\alpha} = 0.416$ is the oscillator strength of the Lyman-α, and Φ is the line profile function normalized in such a way that $\int \Phi(\nu)\, \mathrm{d}\nu = 1$.

The line profile is usually expressed as

$$\Phi(x(\nu)) = \frac{V(x, \gamma)}{\sqrt{\pi}\Delta\nu_D}, \tag{2.8}$$

where $\Delta\nu_D$ in Equation (2.8) accounts for thermal broadening introduced by the Hydrogen atoms in the absorbing cloud not being at rest. This parameter is referred to as the Doppler width defined by

$$\Delta\nu_D = \nu_\alpha \sqrt{\frac{2k_B T}{m_{\mathrm{HI}} c^2}}, \tag{2.9}$$

where k_B is the Boltzmann constant, m_{HI} is the mass of the Hydrogen atom, and T the temperature of the gas cloud. The term $V(x, \gamma)$ in Equation (2.8) is a Voigt profile defined by

$$V(x, \gamma) = \frac{\gamma}{\pi} \int_{-\infty}^{\infty} \frac{e^{-x'^2}}{(x - x')^2 + \gamma^2}\, dx'. \tag{2.10}$$

In summary, the line profile defined in Equation (2.8) accounts for both a thermal and a natural broadening caused by the fact that an excited state of any atom has a finite lifetime before decaying to the ground state. For the Voigt profile we define $x = (\nu - \nu_\alpha)/\nu_D$, where $\nu_\alpha \sim 2.47 \times 10^{15}$ Hz is the rest-frame frequency of the Lyman-α transition. The parameter γ, in this case, is the natural line width of the Lyman-α transition and accounts for the damping introduced by a spontaneous transition caused by the finite lifetime of an excited state. It is given by [44]

$$\gamma = \frac{\Gamma_\alpha}{4\pi\Delta\nu_D}, \tag{2.11}$$

where Γ_α is the spontaneous emission rate of the Lyman-α transition lines.

The neutral Hydrogen number density

Now, we will focus on the neutral Hydrogen number density $n_{\rm HI}$ of the absorber. Typically, this quantity can be derived from observable characteristics of absorption features, such as the neutral Hydrogen column density of the absorber. The column density quantifies the total number of neutral Hydrogen atoms per unit area along the line of sight, it is measured in cm^{-2} units, and defined as:

$$N_{\rm HI} = \int_0^L n_{\rm HI}(l)\,\mathrm{d}l\,, \tag{2.12}$$

where $n_{\rm HI}(l)$ is the local neutral Hydrogen number density at a given point l along the line of sight, and L is the proper length of the absorber along the line of sight. The Lyman-α forest is typically characterized by column densities below $N_{\rm HI} < 10^{17.2}$ cm^{-2} with a lower limit placed by the spectral resolution of telescopes, and an upper limit given by the fact that at this limit neutral Hydrogen clouds become optically thick and are further classified as other absorption systems to be discussed in Section 2.1.3.

In practice, the neutral Hydrogen column densities are described by the column density distribution function (CDDF) $F(N_{\rm HI}, X)$, which quantifies the average number of absorption lines per unit column density, per unit comoving absorption length $X(z)$ defined as

$$\mathrm{d}X(z) = \frac{H_0}{c}(1+z)^2\,\mathrm{d}r\,. \tag{2.13}$$

The mean value of $n_{\rm HI}$ at a given redshift z is then

$$n_{\rm HI}(z) = -\frac{(1+z)H(z)}{c}\int F(N_{\rm HI}, z) N_{\rm HI}\,\mathrm{d}N_{\rm HI} \tag{2.14}$$

An accurate measurement of $n_{\rm HI}$ is challenging. However, it can be approximated based on considerations regarding the state of the Universe including the temperature, homogeneity and/or ionized state. An example of this is the Gunn-Peterson Effect which assumes an uniform distribution of Hydrogen, for a gas in photoionization equilibrium at a temperature T_0 close to 10^4 K

$$\tau_{\rm GP}(z) = 3.038\left(\frac{1+z}{4}\right)^6\left(\frac{\Omega_b h^2}{0.0224}\right)^2\left(\frac{T}{10^4\ \mathrm{K}}\right)^{-0.7}\left(\frac{\Gamma_{\rm HI}}{10^{-12}\ \mathrm{Hz}}\right)^{-1}\left(\frac{H(z)}{100\ \mathrm{km/sMpc^{-1}}}\right)^{-1}, \tag{2.15}$$

where $\Gamma_{\rm HI}$ is the Hydrogen photoionization rate. If instead we consider a fluctuating distribution of Hydrogen, then Equation (2.15) takes the form the Fluctuating Gunn-

Peterson Approximation (FGPA).

$$\tau_{\mathrm{FGPA}}(z) = \tau_{GP}(z)(1+\delta_\rho)^\beta, \tag{2.16}$$

where $\delta_\rho = \rho/\overline{\rho} - 1$ describes the inhomogeneities in the density field and $\beta = 2 - 0.7\alpha$, which is obtained from the fact that both the density evolves as a function of ρ^2 and the temperature as a function of $T \propto T_0\rho^\alpha$ where α and T_0 are given by the history of the reionization of the Universe.

2.1.3 Astrophysical contaminants to the Lyman-α forest

Aside from the Lyman-α forest there are other absorption features present in quasar spectra. Although these are important to understand the properties of the IGM for the Lyman-α forest studies interest of this thesis, these are treated as astrophysical contaminants, which we will summarize next.

High Column Density systems

Regions with high concentrations of neutral Hydrogen, with column densities over $N_{\mathrm{HI}} > 10^{17.2}$ cm^{-2} become optically thick and thus produce a significant break on quasar spectra [45], see for example Figure 2.1. Moreover, for some systems of higher column densities, the damping introduced by the natural line width γ (Equation (2.11)) of the interaction cross-section (Equation (2.7)) dominates over the thermal broadening. In general, these objects receive the classification High Column Density (HCD) systems, and are further sub-divided in terms of their column densities as:

- Lyman-limit systems (LLS): When the absorption system has a column density between $10^{17.2}\mathrm{cm}^{-2} < N_{\mathrm{HI}} < 10^{20.3}$ cm^{-2} [45]. These are narrow line absorption systems that are optically thick to photons with the wavelengths shorter that Lyman-Limit causing a dramatic drop in the quasar spectrum. Moreover, there would exist a partial break in the quasar spectrum located at wavelength corresponding to the Lyman-Limit wavelength $\lambda_\infty = 911.75$ redshifted to the position of the LSS. These systems are usually located in the low-density regions of the circumgalactic medium (CGM), close to galaxies and galaxy halos. They are also found in the denser regions of the IGM.

- Damped Lyman-α absorber (DLA): When the neutral Hydrogen column density exceeds $N_{\mathrm{HI}} > 10^{20.3}$ cm^{-2} [46]. In such systems the natural line width

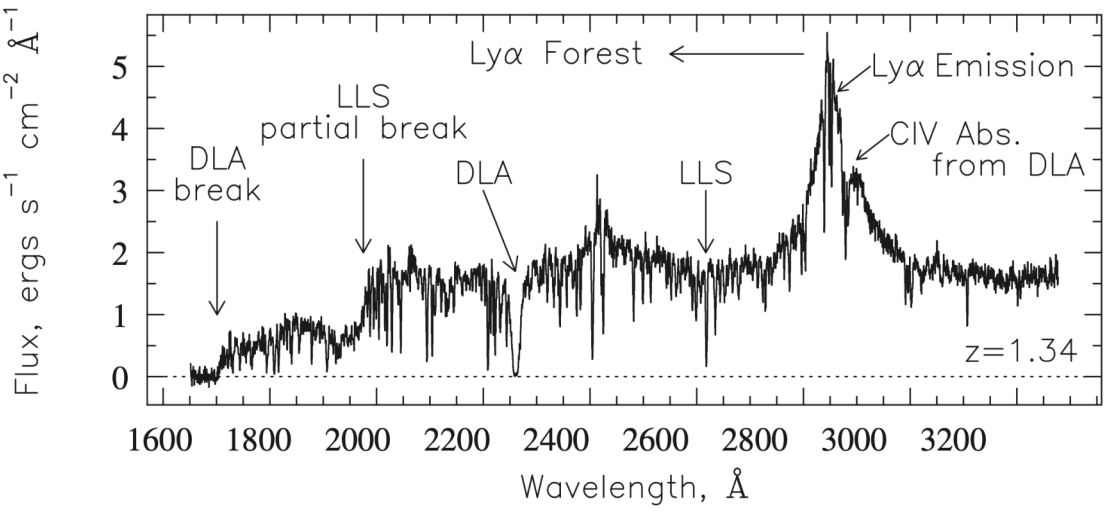

Figure 2.1: Example spectrum of a $z = 1.34$ quasar. With an LLS located at $\lambda \sim 2700$ Å that produces a partial break at $\lambda \sim 2000$. The spectrum also contains a DLA at $\lambda \sim 2300$ Å that produces a partial break at $\lambda \sim 1700$ Å. Figure taken from [44].

profile dominates over thermal broadening, producing characteristic damping wings in the absorption profile of these systems. DLAs are thought to be the seeds for forming galaxies, thus they are mostly constituted by cold Hydrogen atoms located in dark matter halos. Although a contaminant for Lyman-α forest studies, DLAs are important for the study of the properties of CGM and the abundance of other elements in addition to neutral Hydrogen in CGM and IGM.

Figure 2.1 shows an example spectrum of a $z = 1.34$ quasar that contains both an LLS and a DLA in its spectrum.

Metals

Neutral Hydrogen is not the only element present in the IGM nor the CGM. Additional to the absorption features produced by neutral Hydrogen, quasar spectra may exhibit absorption features due to other absorption lines, hereafter referred to as metals.[2]

Some examples are transitions of the Lyman series with $n > 2$, O VI(1031), O VI(1037), Si II(1190), Si II(1193), Si III(1207) with shorter wavelengths than the Lyman-α transition and are normally blended with the Lyman-α forest making

[2]Although these absorption lines are not necessarily of metal elements and some may be even in the gaseous form, we follow the convention to call metals any element heavier than Hydrogen and Helium.

them difficult to detect. Other absorption lines, such as Si II(1260) and C IV(1548), C IV(1550) have longer rest-frame wavelengths than the Lyman-α transition and thus normally appear on the red side of the Lyman-α forest making them easily identifiable. However, an absorption feature produced by a metal at redshift z_m would appear in a quasar's spectrum at the same observed wavelength as a Lyman-α absorption feature at redshift z_α if they satisfy

$$\frac{1+z_\alpha}{1+z_m} = \frac{\lambda_m}{\lambda_\alpha}, \tag{2.17}$$

where λ_m is the rest-frame wavelength of the metal transition. For example, an absorption feature produced by a C IV(1548) cloud located at $z_{\text{C IV}} \sim 1.58$ would appear at the observed wavelength $\lambda \sim 4000$ of the quasar spectrum, same as a Lyman-α absorption produced by a neutral Hydrogen gas at $z_\alpha \sim 2.29$. This is a sufficient wavelength to be observed in the Lyman-α forest of a quasar at redshift $z > z_\alpha$ by optical telescopes working in the $\lambda > 3600$ wavelength range. Refer to [47] a more complete discussion on metals on the Lyman-α forests.

Broad Absorption Line quasars

So far we have focused on absorption features in quasar spectra that are produced by material far from the vicinity of the emitting quasar. Now we will focus on a particular type of quasars defined as Broad Absorption Line (BAL) quasars. These type of quasars exhibit broad absorption troughs, caused by high velocity gas outflows thought to be caused by the accretion process of the AGN of the host quasar.

BAL quasars are further classified by the emission lines that exhibit these absorption troughs, usually blue-wards to the emission line. If the BAL quasar exhibits absorption troughs only at highly ionizing emission lines such as C IV, Si IV, and N V, these are classified as HiBALs. If they additionally present troughs due to lower ionizing emission lines such as Mg II, these receive the category of LoBALs. Furthermore, if they also exhibit absorption features caused by Fe II or Fe III they are categorized as FeLoBALs. In any case BALs, are commonly described by the quantification of the absorption troughs caused by the C IV spectral line, either by the Balnicity Index (BI, defined in Section 3 of [48]) or by the Absorption Index (AI, defined in Appendix A of [49]).

2.2 The Lyman-α forest as a matter distribution tracer

As mentioned previously in this chapter, the Lyman-α forest is constituted by absorption features due to neutral Hydrogen. These absorption features are strongly related with the distribution of matter in the Universe, and thus can be used as a tracer providing an invaluable high redshift observable that contributes to our understanding of the expansion history of our Universe and the nature of Dark Matter at these epochs. For this thesis, we will focus on the expansion history of the Universe, studies of Dark Matter models are performed mostly using the 1D Flux power Spectrum of the Lyman-α forest ($P_{\rm 1D}$) which is not covered throughout this thesis. However, the goal of this section is to present the use of the Lyman-α forest as a matter distribution tracer. Then, we will focus on the method used to measure the 3-dimensional Lyman-α forest correlation function, and its model (including nuisance parameters) used to determine the position of the BAO scale feature imprinted in the distribution of the Lyman-α absorption features. Most of the following content is based in the state of the art Lyman-α BAO analysis [50] and its validation with synthetic quasar spectra [51].

2.2.1 The 3-dimensional correlation function

The methodology to measure and model the 3D correlation function has evolved over the various studies performed over the history (to be summarized in Section 2.3). Here we will briefly detail the methodology used for Dark Energy Spectroscopic Instrument (DESI, [52–54]), in particular, for its first data release (DESI-DR1), which presents the most recent and accurate BAO scale measurements made with the Lyman-α forest up to the date this thesis is being written. DESI will be further described in Chapter 3 so here we will focus only on the state of the art computation and model of the Lyman-α 3D correlation function.

Computing of the flux-transmission field

First, we extract the flux-transmission field from the measured spectra. The methodology for doing so is fully described in [55], and briefly summarized here.

The flux-transmission field $\delta_q(\lambda)$ at a given wavelength for a particular quasar q

is defined by

$$\delta_q(\lambda) = \frac{f_q(\lambda)}{\overline{F}(\lambda)C_q(\lambda)} - 1\,, \tag{2.18}$$

where $f_q(\lambda)$ is the observed flux in the spectrum of the quasar, $\overline{F}(\lambda)$ is the mean transmitted flux, and $C_q(\lambda)$ is the unabsorbed spectrum of the quasar, also referred to as the continuum.

There are some subtleties that make it difficult to correctly measure the continuum of each quasar. An easier approach is to determine their mean flux, i.e. the $\overline{F}(\lambda)C_q(\lambda)$ product. In the current methodology this product is assumed to be a function of a universal mean flux as rest-frame $\overline{C}(\lambda_{\rm RF})$, assumed to be the same for all quasars, that is corrected by first degree polynomial of $\log\lambda$ with two parameters, the a_q and b_q, that tilt the universal mean flux, accounting for each quasar spectral diversity and the redshift dependence of $\overline{F}(\lambda)$,

$$\overline{F}(\lambda)C_q(\lambda) = \overline{C}(\lambda_{\rm RF})\left(a_q + b_q\frac{\Lambda - \Lambda_{\rm min}}{\Lambda_{\rm max} - \Lambda_{\rm min}}\right), \tag{2.19}$$

where $\Lambda = \log\lambda$. $\Lambda_{\rm max}$ and $\Lambda_{\rm min}$ define the rest-frame wavelength region where the analysis is performed. The current baseline analysis pipeline defines two regions of interest, the Lyman-α forest region A (Lyα(A)) defined between $\lambda_{\rm min} = 1040$ Å and $\lambda_{\rm min} = 1205$ Å, and the Lyman-α forest region B (Lyα(B)) defined within $\lambda_{\rm min} = 920$ Å and $\lambda_{\rm min} = 1020$ Å. Both measured in the $\lambda = [3600, 5772]$ Å observed wavelength range.

The universal mean flux $\overline{C}(\lambda_{\rm RF})$ and quasar spectral diversity parameters a_q and b_q are iteratively fitted by a log-likelihood maximization over all the forest pixels[3] of the quasar, defined by[4]

$$2\ln L = -\sum_i \frac{\left[f_q(\lambda_i) - \overline{F}C_q(\lambda_i;\overline{C},a_q,b_q)\right]^2}{\sigma_q^2(\lambda_i;\overline{C},a_q,b_q)} - \sum_i \ln\left[\sigma_q^2(\lambda_i;\overline{C},a_q,b_q)\right], \tag{2.20}$$

where $\sigma_q(\lambda)$ is the variance of the observed flux $f_q(\lambda)$, defined as

$$\sigma_q^2(\lambda) = \eta(\lambda)\sigma_{{\rm pip},q}^2(\lambda) + \sigma_{\rm LSS}^2(\lambda)\left[\overline{F}C_q(\lambda)\right]^2, \tag{2.21}$$

where $\sigma_{\rm LSS}(\lambda)$ is the intrinsic variance of the Lyman-α flux transmission fluctuations and $\sigma_{\rm pip}(\lambda)$ is the noise measured by the spectra acquisition pipeline (i.e Instrumental

[3]For this section, we will use the term 'pixel' to refer to a segment on the wavelength grid of an spectrograph. Currently set to 0.8 Åfor the Dark Energy Spectroscopic Instrument.

[4]In this equation the $\lambda_{\rm RF}$ dependence of the mean flux $\overline{C}$ is obviated.

noise, for example). The parameter $\eta(\lambda)$ accounts for inaccuracies on the noise contribution estimation.

The values of $\sigma_{\rm LSS}(\lambda)$ and $\eta(\lambda)$ are also fitted during the iteration process. Note that the iterative fitting process is done separately for the Lyα(A) and Lyα(B) regions. Refer to Section 3 of [55] for details about the iterative procedure where the five $[\overline{C}, a_q, b_q, \sigma_{\rm LSS}, \eta]$ free parameters are fitted. At this same stage spectra with low quality standard are rejected, for example if the forest not having a minimal quantity of pixels to be analyzed, failed continuum fitting, low signal-to-noise, to mention a few.

The flux-transmission field computation procedure optionally includes a previous pre-processing of the data where pixels including absorption features caused by the Milky Way galaxy (galactic absorption sky lines), BALs and/or DLAs are identified, and removed from the spectra. Extra calibration steps may also be performed, prior the computation of the flux-transmission field, by measuring the flux fluctuations on regions outside of the Lyα A and B regions where noise is expected to be the only contribution, for example the C III region defined in the $\lambda_{\rm min} = 1600$ Å and $\lambda_{\rm max} = 1850$ Å wavelength range. These steps are detailed in Sections 3.1 and 3.2 of [55].

Measurement of the correlation functions

After the flux-transmission field has been computed, the longitudinal ($r_\parallel$) and the transverse ($r_\perp$) separations of each pixel–pixel (auto-correlation), or pixel–quasar (cross-correlation) pair are defined based on their redshifts (z_i, z_j) and angular separation[5] $\Delta\theta_{i,j}$ by

$$r_\parallel = [D_C(z_i) - D_C(z_j)] \cos\left(\Delta\theta_{i,j}/2\right), \tag{2.22}$$

$$r_\perp = [D_M(z_i) + D_M(z_j)] \sin\left(\Delta\theta_{i,j}/2\right), \tag{2.23}$$

where D_C and D_M are the comoving radial and transverse distances (defined in Equations (1.37) and (1.38)), respectively. Which are equal in the flat-ΛCDM cosmology of Planck2018 [18] used as fiducial model to perform this analysis.

Then, Lyman-α auto and Lyα × QSO cross-correlation are computed by the

[5] Defined by $\Delta\theta_{i,j} = (\mathrm{DEC}_i - \mathrm{DEC}_j)^2 + \cos(\mathrm{DEC}_i - \mathrm{DEC}_j)^2(\mathrm{RA}_i - \mathrm{RA}_j)^2$, where RA and DEC refer to the right ascension and declination of the pairs.

estimators

$$\xi_A^{\mathrm{Ly}\alpha\times\mathrm{Ly}\alpha} = \frac{\sum_{i,j\in A} w_i^{\mathrm{Ly}\alpha} w_j^{\mathrm{Ly}\alpha} \delta_i \delta_j}{\sum_{i,j\in A} w_i^{\mathrm{Ly}\alpha} w_j^{\mathrm{Ly}\alpha}}, \tag{2.24}$$

$$\xi_A^{\mathrm{Ly}\alpha\times\mathrm{QSO}} = \frac{\sum_{i,j\in A} w_i^{\mathrm{Ly}\alpha} w_j^{\mathrm{QSO}} \delta_i}{\sum_{i,j\in A} w_i^{\mathrm{Ly}\alpha} w_j^{\mathrm{QSO}}}, \tag{2.25}$$

respectively. Where A is a two-dimensional bin of comoving separation $A \to (r_\parallel, r_\perp)$.[6]

The weights $w^{\mathrm{Ly}\alpha}$ of a Lyman-α forest pixel and w^{QSO} of a quasar are respectively defined as

$$w_i^{\mathrm{Ly}\alpha} = \left(\frac{1+z_i}{1+2.25}\right)^{\gamma_{\mathrm{Ly}\alpha}-1} \left[\eta(\lambda)\left(\frac{\sigma_{\mathrm{pip},q}(\lambda)}{\overline{F}C_q(\lambda)}\right)^2 + \eta_{\mathrm{LSS}}\,\sigma^2_{\mathrm{LSS}}(\lambda)\right]^{-1}, \tag{2.26}$$

$$w_i^{\mathrm{QSO}} = \left(\frac{1+z_i}{1+2.25}\right)^{\gamma_{\mathrm{QSO}}-1}, \tag{2.27}$$

where $\gamma_{\mathrm{Ly}\alpha}$ and γ_{QSO}, modulate the evolution as a function of redshift of the bias of Lyman-α and quasars, respectively. Typically set to the values found in previous Lyman-α BAO scale studies $\gamma_{\mathrm{Ly}\alpha} = 2.9$ [56] and $\gamma_{\mathrm{QSO}} = 1.44$ [57]. Equation (2.26) includes an extra parameter η_{LSS} that modulates the contribution from the intrinsic variance of the Lyman-α transmission fluctuations, here we set it to $\eta_{\mathrm{LSS}} = 7.5$ which minimizes the covariance matrix of the auto-correlation [55].

The usual is to measure the correlation function with comoving separation bins of size 4 Mpc/h, covering the $r_\perp = [0, 200]$ Mpc/h range, and $r_\parallel = [0, 200]$ Mpc/h for the auto-correlation and $r_\parallel = [-200, 200]$ Mpc/h for the cross-correlation, resulting in 2500 and 5000 comoving separation bins, respectively. However these specifications might be changed according to the purposes of the analysis.

Distortion of the flux-transmission field

The fitting procedure of the quasar spectral diversity parameters a_q and b_q during the mean flux estimation biases the mean value of the flux-transmission δ_q for each individual line of sight towards zero [58, 59], causing a distortion of the correlation functions; we will explore more on how this distortion looks in Section 6.1.1.

The distortion in the flux-transmission field is accounted for by projecting the

[6] Throughout this thesis we will indistinctly switch the notation used to describe dependency on the comoving separation bin $A \to (r_\parallel, r_\perp)$ where convenient. For example, the correlation function ξ: ξ_A, $\xi(A)$ and $\xi(r_\parallel, r_\perp)$. There will also be equations where it will be convenient to write the correlation function ξ obviating its dependence on A.

measured fluctuations δ_i, prior to the measurement of the correlation functions in Equations (2.24) and (2.25), following the linear relation:

$$\delta_i \to \sum_j \eta_{ij}\ \delta_j, \tag{2.28}$$

where η_{ij} is a projection matrix defined as

$$\eta_{ij} = \delta^K_{ij} + \frac{w^{\mathrm{Ly}\alpha}_j}{\sum_k w^{\mathrm{Ly}\alpha}_k} + \frac{w^{\mathrm{Ly}\alpha}_j (\Lambda_i - \overline{\Lambda}_q)(\Lambda_j - \overline{\Lambda}_q)}{\sum_k w^{\mathrm{Ly}\alpha}_k (\Lambda_k - \overline{\Lambda}_q)^2}\ , \tag{2.29}$$

where δ^K is the Kronecker delta, $\overline{\Lambda}_q = \overline{\log \lambda}_q$ for given quasar spectrum, and the weights $w^{\mathrm{Ly}\alpha}$ are the same as defined in Equation (2.26). See Section 4 of [59] and Appendix B of [60] for further details on the motivation of projecting the flux-transmission field.

When the flux-transmission field is projected, it results in the average value of the flux transmission field per wavelength $\overline{\delta(\lambda)}$ being allowed to deviate slightly from zero. This in principle allows the cross-correlation to have a non-zero value at large scales, which is not desired. To fix this, $\overline{\delta(\lambda)}$ is removed from the projected flux-transmission field for the cross correlation only.

$$\delta_q(\lambda) \to \delta_q(\lambda) - \overline{\delta(\lambda)}, \tag{2.30}$$

this guarantees the cross-correlation function having null value at large scales for any quasar redshift distribution.

Covariance Matrix

The method to compute the covariance matrix of the correlation functions consists in dividing the sky into sub-samples defined by HEALpix pixels [61] of `nside=16`, equivalent to regions of $(250\ \mathrm{Mpc/h})^2$ at the effective redshift $z \sim 2.3$. Then, the individual correlation functions ξ^S are computed for each sub-sample. Cross-covariance between sub-samples are neglected due to the scale of the HEALpix pixel regions being larger than the scales of interest.

The weighted covariance C_{AB} of two comoving separation bins A and B is then estimated by

$$C_{AB} = \frac{1}{W_A W_B} \sum_s W^s_A W^s_B \left[\xi^s_A \xi^s_B - \xi_A \xi_B\right], \tag{2.31}$$

where $W^s_A = \sum_{i \in A,s} w_i$, and $W_A = \sum_s W^s_A$. The weights are defined by Equa-

tions (2.26) and (2.27) according to the correlation whose covariance is to be estimated.

The covariance from Equation (2.31) is dominated by its diagonal elements, the off-diagonal elements are noisy. Therefore, they are smoothed by computing the correlation matrix, defined as

$$\mathrm{Corr}_{AB} = \frac{C_{AB}}{\sqrt{C_{AA}C_{BB}}}, \tag{2.32}$$

and replacing the off-diagonal elements of the correlation matrix, which correspond to comoving distance bins of equal $\Delta r_{\parallel} = |r_{\parallel}^{A} - r_{\parallel}^{B}|$ and $\Delta r_{\perp} = |r_{\perp}^{A} - r_{\perp}^{B}|$, by their average. The result is a smoothed covariance matrix of dimension $N_{\mathrm{bins}} \times N_{\mathrm{bins}}$. The method to compute the covariance matrix is described in more detail in Section 3.2 of [62], Section 3.3 of [50] and Section 3.3 of [51]. Previous studies have proven this method to be a sufficiently good estimate of the covariance matrix by comparing with other methods, that are slower in terms of computing time [59, 62, 63].

For sufficiently small data samples the cross-covariance between correlations is negligible[e.g 62]. However, this might not be the case for large datasets and thus the cross-covariance must be taken into account.

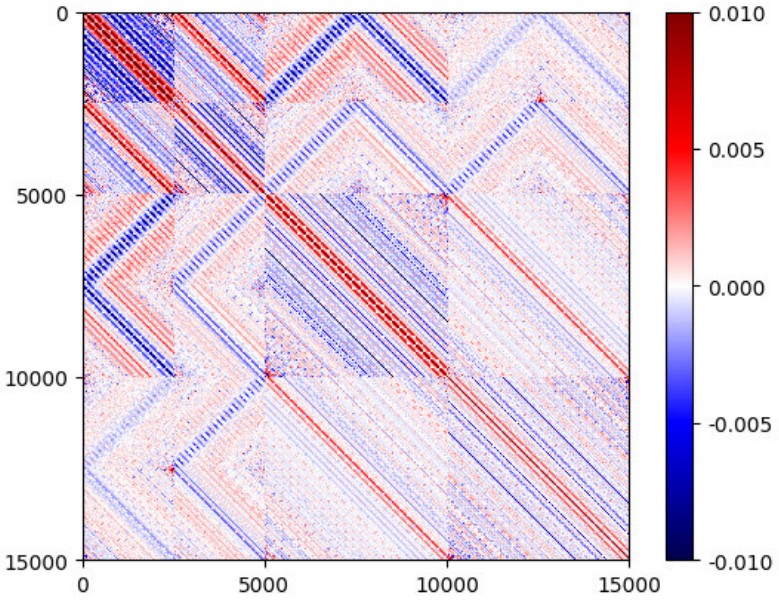

Figure 2.2: Full covariance matrix measured with the method described in this section. The first two block diagonal elements correspond to the covariance of the Lyα(A) and Lyα(B) regions auto correlations, the second two block diagonal elements are the covariance from the cross-correlation of the Lyα(A) and Lyα(B) regions. The off diagonal block elements show the cross-covariance between correlation. Figure taken from [50].

Figure 2.2 shows an example smoothed full covariance matrix measured in the most recent BAO Lyman-α analysis performed by DESI during in its first year of operations data release [50]. This matrix is of dimension 15000×15000, the diagonal is composed by two 2500×2500 block matrices, and two 5000×5000 corresponding to the covariance of the auto-correlations, and cross-correlations in the Lyman-α A and B regions.

2.2.2 Model of the 3D correlation functions

According to [64], the model of the 3D correlation functions, regardless if its the auto or cross-correlation, can be defined by the sum of a smooth component ξ^{smooth} that depends on the comoving distance separation $r_\parallel$ and $r_\perp$, and a BAO scale "peak" component ξ^{peak} that also depends on the comoving distance separation, but with two additional free parameters $\alpha_\parallel$ and $\alpha_\perp$ that are introduced to modulate the value of the $r_\parallel$ and $r_\perp$ coordinates of the BAO scale feature component, this is

$$\xi(r_\parallel, r_\perp) = \xi_{\text{smooth}}(r_\parallel, r_\perp) + \xi_{\text{peak}}(\alpha_\parallel r_\parallel, \alpha_\perp r_\perp) \tag{2.33}$$

These BAO scale parameters $\alpha_\parallel$ and $\alpha_\perp$ are related to the Hubble and comoving angular diameter distances, D_H and D_M, by

$$\alpha_\parallel = \frac{D_H(z_{\text{eff}})/r_d}{[D_H(z_{\text{eff}})/r_d]_{\text{fiducial}}}, \tag{2.34}$$

$$\alpha_\perp = \frac{D_M(z_{\text{eff}})/r_d}{[D_M(z_{\text{eff}})/r_d]_{\text{fiducial}}}, \tag{2.35}$$

where the "fiducial" subscript refers to the values according to a chosen fiducial cosmology, r_d is the scale of the sound horizon at the end of the drag epoch, and z_{eff} is the effective redshift of the measurement, computed from the mean value of the redshift of the pixel–pixel or pixel–quasar pairs.

The power spectrum

The modeling of the correlation functions begins by considering the Fourier transform of the anisotropic matter power cross-spectrum

$$P_{A\times B}(\mathbf{k}, z) = b_A b_B (1 + \beta_A \mu_k^2)(1 + \beta_B \mu_k^2) P_{\text{QL}}(\mathbf{k}, z) F_{\text{NL}}(\mathbf{k}) G(\mathbf{k}), \tag{2.36}$$

where $\mathbf{k} = (k_\parallel, k_\perp)$ is the wave-number vector of magnitude k, and $\mu_k = k_\parallel / k$. The suffix $A \times B$, on the left-hand side of the equation, refers to the tracers that are being cross-correlated, for example, Ly$\alpha \times$ Lyα or Ly$\alpha \times$ QSO.

The parameters b_X, β_X denote the linear bias and RSD parameter of such tracers, respectively. Since power spectrum and linear bias parameters evolve as a function of redshift, and the effective redshift varies in each of the ($r_\parallel$, $r_\perp$) separation bins defined to compute the correlation functions, the model takes this into account by evolving the correlations through the linear growth factor defined by the cosmology used to perform the analysis. In the case of the bias parameters their evolution is modeled differently for a Lyman-α forest pixel and for a quasar according to

$$b_{\mathrm{Ly}\alpha}(z) = b_{\mathrm{Ly}\alpha}(z_{\mathrm{eff}}) \left(\frac{1+z}{1+z_{\mathrm{eff}}} \right)^{\gamma_{\mathrm{Ly}\alpha}}, \tag{2.37}$$

$$b_{\mathrm{QSO}}(z) = b_{\mathrm{QSO}}(z_{\mathrm{eff}}) \left(\frac{1+z}{1+z_{\mathrm{eff}}} \right)^{\gamma_{\mathrm{QSO}}}, \tag{2.38}$$

respectively. Here $\gamma_{\mathrm{Ly}\alpha}$ and γ_{QSO} are the same as those defined in Equations (2.26) and (2.27), the reference values $b_{\mathrm{Ly}\alpha}(z_{\mathrm{eff}})$ and $b_{\mathrm{QSO}}(z_{\mathrm{eff}})$ are free parameters in the model. In the case of the RSD parameters β_X, they are assumed to be constant in time, the RSD value of Lyman-α pixels is a free parameter in the model. In the case of quasars, it is determined by the linear bias through $\beta_{\mathrm{QSO}} = f / b_{\mathrm{QSO}}$, where f is the linear growth rate, approximately $f \sim 0.97$ at redshift $z = 2.3$ for a Planck18 cosmology [18].

The term $P_{\mathrm{QL}}(\mathbf{k}, z)$ in Equation (2.36) is the quasi-linear power spectrum defined, in the spirit of separating the model into a smooth and a peak component, as

$$P_{\mathrm{QL}}(\mathbf{k}, z) = P_{\mathrm{smooth}}(\mathbf{k}, z) + A_{\mathrm{peak}} \exp\left[-\frac{k_\parallel^2 \Sigma_\parallel^2 + k_\perp^2 \Sigma_\perp^2}{2} \right] P_{\mathrm{peak}}(\mathbf{k}, z), \tag{2.39}$$

the smooth component P_{smooth} is derived from the linear power spectrum P_L following a side-band technique described in [64]. The linear power spectrum P_L is computed with dedicated codes, such as `CAMB`[7] [65]. The peak component P_{peak} is then obtained by substracting the smooth component from the linear power spectrum. The term A_{peak}, typically fixed to $A_{\mathrm{peak}} = 1$, modulates the amplitude of the BAO peak. The exponential next to the peak component accounts for non-linear BAO broadening effects, characterized by the $\Sigma_\parallel$ and $\Sigma_\perp$ parameters which are related to the linear growth rate f by $\Sigma_\parallel / \Sigma_\perp = 1 + f$, and are typically fixed to $\Sigma_\parallel = 6.37$ Mpc/h and

[7] https://github.com/cmbant/CAMB

$\Sigma_\perp = 3.24$ Mpc/h [e.g 62, 66].

Non-linear effects on the power spectrum at large $\mathbf{k}$ values (i.e small scales) are accounted for the term $F_{NL}(\mathbf{k})$ in Equation (2.36). The mathematical expression of this term depends on the type of correlation that is being modeled. For the auto-correlation, the small scales of the power spectrum are corrected by the model presented in [67], referred to as the Arinyo model. This model was calibrated with hydrodynamic simulations and is defined as

$$F_{NL}(\mathbf{k}) = \exp\left\{ \left[q_1 \Delta^2(\mathbf{k})\right] \left[1 - \left(\frac{k}{k_v}\right)^{a_v} \mu^{b_v}\right] - \left(\frac{k}{k_p}\right)^2 \right\}, \tag{2.40}$$

where $\Delta(\mathbf{k}) = k^2 P_L(\mathbf{k})/(2\pi^2)$, and $[q_1, k_v, a_v, b_v, k_p]$ are free parameters with best-fit values $q_1 = 0.8558, k_v = 1.11454$ h/Mpc$, a_v = 0.5378, b_v = 1.607, k_p = 19.47$ h/Mpc interpolated at the effective redshift $z_{\rm eff} \sim 2.3$ [see Table 7 of 67] that are usually fixed during the correlation functions modeling process. Note that in Equation (2.40) the first term in square brackets inside the exponential accounts for non-linear growth of structure, the second for thermal broadening and non-linear peculiar velocities smoothing, and the third term in parenthesis for suppression of the power spectrum due to gas pressure.

In the case of the cross-correlation, the most important effects are the non-linear peculiar velocities of quasars, and statistical errors on their redshift estimate. These are modeled in F_{NL} either following a Lorentzian ($F_{\rm NL,Gauss}$) or a Gaussian ($F_{\rm NL,Gauss}$) smoothing, defined respectively by

$$F^2_{\rm NL,Lorentz} = [1 + (k_\parallel \sigma_v)^2]^{-1}, \tag{2.41}$$

$$F^2_{\rm NL,Gauss} = \exp\left[-\frac{1}{2}(k_\parallel \sigma_v)^2\right], \tag{2.42}$$

where σ_v is a free parameter that determines the magnitude of the statistical redshift errors and non-linear peculiar velocities.

At last, the $G(\mathbf{k})$ term in Equation (2.36) accounts for the effects of the discrete separation grid used to measure the correlation functions it is defined as

$$G(k, \mu_k) = \mathrm{sinc}\left(\frac{k_\parallel R_\parallel}{2}\right) \mathrm{sinc}\left(\frac{k_\perp R_\perp}{2}\right), \tag{2.43}$$

where $R_\parallel$ and $R_\perp$ are the widths of the longitudinal and transverse bins ($R_\parallel =$ 4 Mpc/h and $R_\perp = 4$ Mpc/h for case discussed throughout this section), respectively.

Modeling Astrophysical contaminants

As discussed in Section 2.1.3 there are absorption features within the Lyman-α forest that are produced by other astrophysical sources, such as HCDs, LLS, DLAs, and metals. Here we will discuss how these contaminants are considered and included in the correlation functions model.

First we will review the case of HCDs (including LLS and DLAs). Although DLAs can be masked out of spectra, systems that go undetected have a contribution to the correlation functions. Moreover, HCDs with column densities below $\log N_{\rm HI} < 20.3$ cm^{-2} will always be present in spectra, since they are difficult to detect by the current identification algorithms. Thus the model takes this into account by considering the HCDs (including missed DLAs) contribution by treating the Lyman-α bias and RSD parameters as effective parameters including the effects of HCDs by

$$b'_{\rm Ly\alpha} = b_{\rm Ly\alpha} + b_{\rm HCD} F_{\rm HCD}(k_\parallel), \tag{2.44}$$

$$b'_{\rm Ly\alpha}\beta'_{\rm Ly\alpha} = b_{\rm Ly\alpha}\beta_{\rm Ly\alpha} + b_{\rm HCD}\beta_{\rm HCD} F_{\rm HCD}(k_\parallel), \tag{2.45}$$

where $b_{\rm HCD}$ and $\beta_{\rm HCD}$ are the bias and RSD parameters associated to HCDs, the RSD parameter is set free in the model, while the bias parameter is set to a constant value of $b_{\rm HCD} = 2$ based on previous studies [e.g 68]. The factor $F_{\rm HCD} = \exp\left(-L_{\rm HCD} k_\parallel\right)$ modulates the contribution of HCDs to the correlation functions and depends on the typical length scale $L_{\rm HCD}$ of HCDs present on the analyses spectra [69]. This last parameter is set free in the model, however up to the date of writing this thesis this parameter is still under study, and some variations of the analysis might fix it to known values. There also exists the possibility to use a Voigt profile, instead of an exponential, to model $F_{\rm HCD}$, see Appendix A of [50] for more information about other HCD models.

Next, let us consider the case of metals. As discussed in Section 2.1.3, the Lyman-α forest might contain traces of absorption features due to other spectral lines, which we have referred to as metals. Metals contribute to the auto-correlation function in three ways. First, through their cross-correlation with absorption features due to Lyman-α ($\xi^{\rm Ly\alpha \times m}$). Second, through their own auto-correlation ($\xi^{\rm m \times m}$). And third, by their cross-correlation with other metals $\xi^{\rm m_i \times m_j}$, with $i \neq j$. In the case of the cross-correlation these only contribute through their cross-correlation with quasars.

Thus the complete correlation function model ($\xi^{\rm Total}$) would have the form of

$$\xi^{\rm Ly\alpha\times Ly\alpha}_{\rm Total} = \xi^{\rm Ly\alpha\times Ly\alpha} + \sum_m \xi^{\rm Ly\alpha\times m} + \sum_{i\neq j} \xi^{\rm m_i\times m_j}. \tag{2.46}$$

$$\xi^{\rm Ly\alpha\times QSO}_{\rm Total} = \xi^{\rm Ly\alpha\times QSO} + \sum_m \xi^{\rm QSO\times m}. \tag{2.47}$$

The power spectrum of the aforementioned correlation functions would be then similar to Equation (2.36) by taking B a metal, i.e each metal has its own bias b_m and RSD β_m parameters, the bias of each metal are set as free parameters in the model, while the RSD parameters are fixed to $\beta_m = 0.5$ based on previous studies [e.g 58, 59, 62]. The evolution as a function of redshift of the metal biases b_m is assumed to follow a similar relation to those defined in Equations (2.37) and (2.38) with their own γ_m value, typically set to 1 for all metals.

Since the method to extract the flux-transmission field (described in Section 2.2.1) assumes that all the absorption features present in the analyzed regions are due to Lyman-α only. This causes that the redshift, and therefore the comoving separation bins, of the absorption features caused by metals be incorrectly estimated, thus having an important spurious contribution at the level of their cross-correlation with Lyman-α (or quasars), and with other metals. Then, the BAO scale feature in the correlation functions would appear, with an amplitude that depends on the absorption strength of the metal relative to Lyman-α, at an incorrect scale defined by

$$r_\parallel = (1+z) D_H(z) \frac{\lambda_m - \lambda_{\rm Ly\alpha}}{\lambda_{\rm Ly\alpha}}, \tag{2.48}$$

at small $r_\perp$ values. In this equation, z is the redshift of the absorption feature caused by the metal, λ_m and $\lambda_{\rm Ly\alpha}$ are the rest-frame wavelength of metal transition and Lyman-α, respectively.

Given the regions of interest of the analysis ($|r_\parallel|$ scales between 0 and 200 Mpc/h), the metals that contribute the most, in terms of their cross-correlation with Lyman-α or quasars, are located at rest-frame wavelengths in the vicinity of the Lyman-α transition line. The most important of them, and therefore considered in the model, being Si II(1260), Si III(1207), Si II(1193), and Si II(1190) at the scales of $r_\parallel \sim 105$, -22, -53 and -59 Mpc/h, at the effective redshift $z \sim 2.33$ assuming a Planck18 cosmology [18], respectively.

Metals with a particular strong absorption, such as C IV(1548) and C IV(1550), have a contribution in terms of their own auto-correlation, although at the moment of writing this thesis, the signal from the C IV auto-correlation has not been sig-

nificantly detected. The cross-correlation between metals are negligible due to their small contribution in terms of the correlation functions.

As mentioned above, the comoving separation bins of the metals are incorrectly calculated, due to the fact that only Lyman-α absorption is assumed. This is taken into account for the modeling process through "metal matrices" that "move" the metal correlation functions from their correct comoving separation bins $(\tilde{r}_\parallel, \tilde{r}_\perp)$ to the "incorrect" separation bins $(r_\parallel, r_\perp)$ by

$$\xi^{\mathrm{m}}(A) = \sum_B M_{AB}\xi^{\mathrm{m}}(B), \tag{2.49}$$

where $A \to (r_\parallel, r_\perp)$ and $B \to (\tilde{r}_\parallel, \tilde{r}_\perp)$, and M_{AB} is the metal distortion matrix defined as

$$M_{AB} = \frac{1}{W_A} \sum_{(i,j)\in A,(i,j)\in B} w_i w_j, \tag{2.50}$$

where the weights w are defined by Equations (2.26) and (2.27) according to the correlation being modeled, and $W_A = \sum_{i\in A} w_i$.

Quasar proximity effects

The energy of quasars is high enough to ionize their IGM environment. Thus, light from a distant Lyman-α quasar that passes through the surrounding region of another quasar may encounter less neutral Hydrogen than usual (due to ionization). These effects are accounted for in the cross-correlation model, in line with following [62, 70], by the quasar proximity model defined by

$$\xi^{\mathrm{TP}} = \xi_0^{\mathrm{TP}} \left(\frac{1\ \mathrm{Mpc/h}}{r} \right)^2 \exp\left(\frac{-r}{\lambda_{\mathrm{UV}}} \right), \tag{2.51}$$

where ξ_0^{TP} is a free parameter that modulates the amplitude of this contribution to the cross-correlation, and λ_{UV} is the mean free path of photons with ultra-violet (UV) wavelengths, fixed to $\lambda_{\mathrm{UV}} = 300$ Mpc/h.

Systematic quasar redshift measurement errors

The σ_v parameter in Equations (2.41) and (2.42) account for systematic redshift errors. However, it might be the case that there exists a systematic error on the estimated redshift of quasars. This offset has a major impact on the determination of the position of a quasar. This effect is seen as an offset on the position of the cross-

correlation peak on small scales along the line of sight ($r_\perp = 0$). This is accounted for by introducing a free parameter $\Delta r_\parallel$ that captures this shift.

Instrumental spurious correlations

There might be spurious correlations introduced by the data acquisition pipeline of any spectroscopic survey that mainly affect the auto-correlation function. For example, errors in the calibration of the spectrograph, correlated noise between measurements, spectrographs or fibers, correlations in the background sky model substracted from the spectra, to mention some.

This is taken into account in the modeling process by introducing a spurious correlation ξ_{inst} contribution to the auto-correlation that accounts for these instrumental effects. The exact model of this contribution is inherent to the instrument used to measure the spectra used in the calculation of the flux-transmission field and should therefore be characterized for each instrument independently[8]. Some examples are explained in Section 4.5 in [62] for eBOSS, Section 4.3.2 in [66] for the early data release of DESI, and Section 4.5 from [50] for the first data release of DESI. The model for DESI is defined by

$$\xi^{\text{inst}}(r_\parallel, r_\perp) = A_{\text{inst}} \delta^K(r_\parallel) f(r_\perp) \tag{2.52}$$

where A_{inst} a free parameter that modulates the amplitude of the instrumental pipeline contributions to the correlation functions, and $f(r_\perp)$ is a function that decreases with $r_\perp$. See Section 4.5 of [50] and Section 4.1.3 of [71] for further details of this model.

Broadband corrections

As seen throughout this section, the model of the correlation functions is quite complicated with the systematic effects considered. In case any effect has not been captured by the aforementioned considerations, there is the option to include a Broadband polynomial correction introduced by

$$B(r, \mu) = \sum_{j=0}^{j_{\max}} \sum_{i=0}^{i_{\max}} a_{ij} \frac{L_j(\mu)}{r^i}, \tag{2.53}$$

[8] Additionally, the model of the same instrument may evolve as the spurious correlations induced by their spectroscopic pipeline are characterized.

where j is even (usually set up to a $j_{\max} = 6$), and $L_j(\mu)$ are the Legendre polynomials of order j, and $i_{\max} = 2$, and a_{ij} are free parameters, thus introducing 12 extra free parameters for each correlation function if broadband polynomials are considered.

The distortion matrix

As seen before the complete model of the correlation functions can take the form

$$\xi^{\mathrm{Ly\alpha\times Ly\alpha}}_{\mathrm{Total}} = \xi^{\mathrm{Ly\alpha\times Ly\alpha}} + \sum_m \xi^{\mathrm{Ly\alpha\times m}} + \sum_{i\neq j} \xi^{\mathrm{m_i\times m_j}} + \xi^{\mathrm{inst}} \tag{2.54}$$

$$\xi^{\mathrm{Ly\alpha\times QSO}}_{\mathrm{Total}} = \xi^{\mathrm{Ly\alpha\times QSO}} + \sum_m \xi^{\mathrm{QSO\times m}} + \xi^{\mathrm{TP}}, \tag{2.55}$$

if no broadband corrections are considered. The last thing to consider into the modeling of the correlation functions is to include the distortion caused by the continuum fitting procedure described by Equations (2.28) and (2.29). This is introduced in the model by the "Distortion Matrix", constructed via the projection matrix η_{ij} defined in Equation (2.28) for the auto and cross-correlation by

$$D^{\mathrm{Ly\alpha\times Ly\alpha}}_{AB} = \frac{1}{W_A} \sum_{i,j\in A} w_i^{\mathrm{Ly\alpha}} w_j^{\mathrm{Ly\alpha}} \sum_{i',j'\in B} \eta_{ii'}\eta_{jj'}, \tag{2.56}$$

$$D^{\mathrm{Ly\alpha\times QSO}}_{AB} = \frac{1}{W_A} \sum_{i,j\in A} w_i^{\mathrm{Ly\alpha}} w_j^{\mathrm{QSO}} \sum_{i',j'\in B} \eta_{ii'}, \tag{2.57}$$

respectively. With weights w defined by Equations (2.26) and (2.27),A and B define two comoving separation bins , and $W_A = \sum_{i\in A} w_i$.

The modeled correlation function would be then

$$\hat{\xi}(A) = \sum_B D_{AB}\xi_{\mathrm{Total}}(B). \tag{2.58}$$

2.3 Review of BAO scale constraints with the Lyman-α forest.

Given the high redshifts at which quasars can be observed, the Lyman-α forest absorption features present in their spectra provide a continuous, high redshift tracer. From these features, a 3-dimensional auto-correlation function and cross-correlation with the positions of nearby quasars can be computed. In this section, we will briefly summarize the efforts made on measuring the BAO scale signal in the Lyman-α

forest 3D correlation function throughout history, quoting the state of the art results presented by DESI [50].

During the early 2000´s [72] studied the requirements of an spectroscopic Lyman-α survey, in terms of total effective area covered, number of targets, spectral resolution, and signal-to-noise ratio, to be able to effectively measure the BAO scale in the Lyman-α 3-dimensional correlation function. This study consisted in performing a Fisher matrix formalism with the specifications of a given galaxy spectroscopic survey and using as a base the Lyman-α forest 3-dimensional power spectrum P_F inferred through Hydrodynamic-Particle Mesh (Hydro-PM) simulations [73].

The first measurement of the Lyman-α 3D correlation function was presented in 2011 by [74], using the spectra of 14k high redshift ($z > 2.1$) quasars from the first year of operations of the Baryon Oscillation Spectroscopic Survey (BOSS, [75]), part of the Sloan Digital Sky Survey (SDSS). Although they did not present a detection of the BAO scale since their measurement comprised only the r region up to 60 Mpc/h, below the expected scale of the BAO feature in the Lyman-α 3D-correlation function.

The first ever detection of the BAO scale at high redshift in the Lyman-α auto-correlation function was presented in [76] which used the Lyman-α forest found in the spectra of 48.6k quasars in the redshift range $2.1 \leq z \leq 3.5$ from the Lyman-α sample of the ninth SDSS data release (DR9, [77]) of BOSS. This study detected the BAO peak signal located at $r = 105$ Mpc/h at the effective redshift of $z = 2.3$ with a significance ranging from 2.3 to 4.2σ depending on the method used [76]. This was later confirmed by [78], which measured a BAO scale in concordance with the measurements of [76] at redshift $z = 2.4$ with a significance level between 3 and 5σ, measuring the auto-correlation of the Lyman-α forest of 58.2k $2.1 \leq z \leq 3.5$ quasar targets from same DR9 dataset with different cuts applied, and following a different, independently developed, analysis pipeline from the one presented in [76], also including an alternative method to fit the correlation functions and measure the BAO scale position, presented in [64]. This data release also presented the first study of the Ly$\alpha \times$ QSO cross-correlation covering the range of $0 \leq r \leq 80$ Mpc/h [70] , testing the robustness of the method prior BAO detection.

Later, during the eleventh data release of SDSS (DR11, [79]) the precision on the measurement of the Lyman-α auto-correlation was improved with the forests of 137.5k quasars in the redshift range $2.1 \leq z \leq 3.5$ measuring the values $D_H/r_d = 9.18 \pm 0.28$ and $D_M/r_d = 37.68 \pm 2.17$ at redshift $z = 2.34$ [63]. The first detection of the BAO signal in the Lyman-α cross-correlation was done with the same data release

using the spectra and positions of 164k quasars of redshift $2 \leq z \leq 3.5$ measuring $D_H/r_d = 9.0 \pm 0.3$ and $D_M/r_d = 36.3 \pm 1.3$ at the effective redshift $z = 2.36$ [80].

The SDSS target number density was improved by the extended BOSS (eBOSS, [81]) project. The Lyman-α auto-correlation study using the SDSS twelfth data release (DR12, also presented in [79]) contained 157.7k quasars in the redshift range $2.1 \leq z \leq 3.5$ and measured $D_H/r_d = 9.07 \pm 0.31$ and $D_M/r_d = 37.77 \pm 2.13$ at redshift $z = 2.33$ [59]. Using this same data release the cross-correlation measurement was performed in 168.8k forests and the position of 234.3k quasars yielding $D_H/r_d = 8.94 \pm 0.22$ and $D_M/r_d = 36.6 \pm 1.2$ at redshift $z = 2.4$ [58]. The SDSS fourteenth data release (DR14, [82]) permitted to expand the data sample of the auto-correlation to 179.9k quasars forests achieving measurement of $D_H/r_d = 8.86 \pm 0.29$ and $D_M/r_d = 37.41 \pm 1.86$ [83]. In the case of the cross-correlation, the data sample grew to 266.6k quasars in the redshift range $1.77 < z < 3.5$ and achieved a measurement of $D_H/r_d = 9.20 \pm 0.36$ and $D_M/r_d = 36.3 \pm 1.8$ at redshift $z = 2.35$ [84] which were combined with the results from the auto-correlation of [83] to yield $D_H/r_d = 9.00 \pm 0.22$ and $D_M/r_d = 37.0 \pm 1.3$ at $z = 2.34$. These studies were the first to include the Lyα(B) region in their measurements. The eBOSS survey concluded in 2019 and its data was presented during the sixteenth data release of SDSS (DR16, [85]), with this data release [62] presented the measurements of the Lyman-α auto-correlation using the forest of 210k quasars in the Lyα(A) region and 69.6k in the Lyα(B) region along with their cross-correlation with the position of 341.4 $z > 1.77$ yielding a $D_H/r_d = 8.99 \pm 0.19$ and $D_M/r_d = 37.5 \pm 1.1$ result at redshift $z = 2.334$ when combining all the Lyman-α forest auto and cross-correlation results in the A and B regions.

The DESI survey, to be described in Chapter 3, allowed the next jump in precision of the measurement of the BAO signal in the Lyman-α 3D correlation functions. In 2023, [66] presented the measurement of the auto-correlation of using the forest data of 88.5k $z \geq 2$ Lyman-α quasars, and its cross-correlation with 147.8k $z \geq 1.77$ quasars from the early data release of DESI (DESI-EDR) [86], although the BAO scale signal was detected with 3.8σ significance, its position was kept blinded during this stage of the DESI survey.

The latest BAO scale measurements with the Lyman-α forest, up to the date this thesis is being written, were presented by DESI early on 2024. Providing the most precise measurement of BAO in the Lyman-α forest with a sample of 428.4k forest in the Lyα(A) region and 137.4k in the Lyα(B) region, along their cross-correlation with 709.5k $z \geq 1.77$ quasars from the first DESI Data Release (DESI-

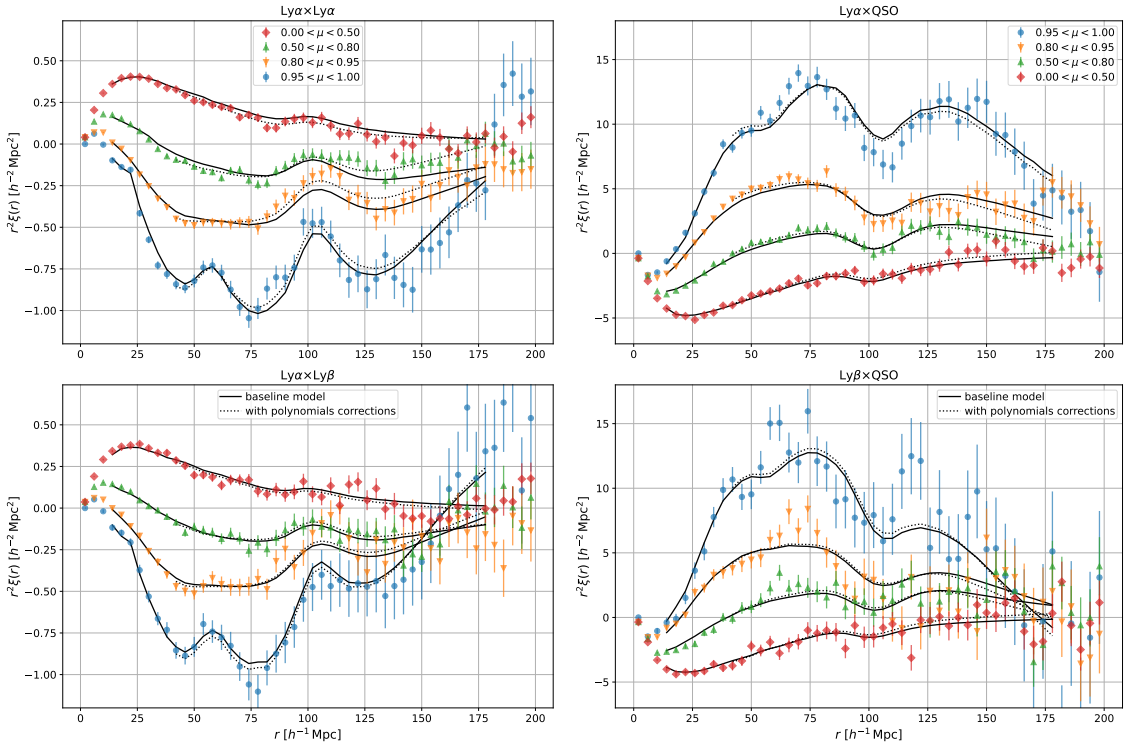

Figure 2.3: Lyman-α auto (left panels) and cross (right panels) correlation functions in the Lyα(A) (top panels) and Lyα(B) regions measured from the DESI-DR1 Lyman-α sample. The best-fit model of the correlation functions is shown as a black solid line, the dotted black line corresponds to the best-fit model including Broadband polynomial corrections (equation (2.53)). Figure taken from [50].

DR1, [87]). The correlation functions and BAO scale measurements from DESI-DR1 were performed following the method and model described Sections 2.2.1 and 2.2.2 and are shown in Figure 2.3 and their respective BAO scale parameters measurements in Figure 2.4, these results are traduced into measurement of $D_H/r_d = 8.52 \pm 0.17$ and $D_M/r_d = 39.71 \pm 0.95$ at redshift $z = 2.33$ when the auto and cross-correlations from the Lyα forest A and B regions are combined. It is worth to mention that these result used Lyman-α mocks presented in [51] and in Section 5.4 of this thesis as a validation step of the analysis pipeline.

DESI will conduct a five-year long survey period expected to improve the accuracy of the Lyman-α BAO scale measurements below the percent level, see [52, 54, 88] and Section 6.2.2 of this thesis for forecasts of the accuracy of the completed DESI survey measurements.

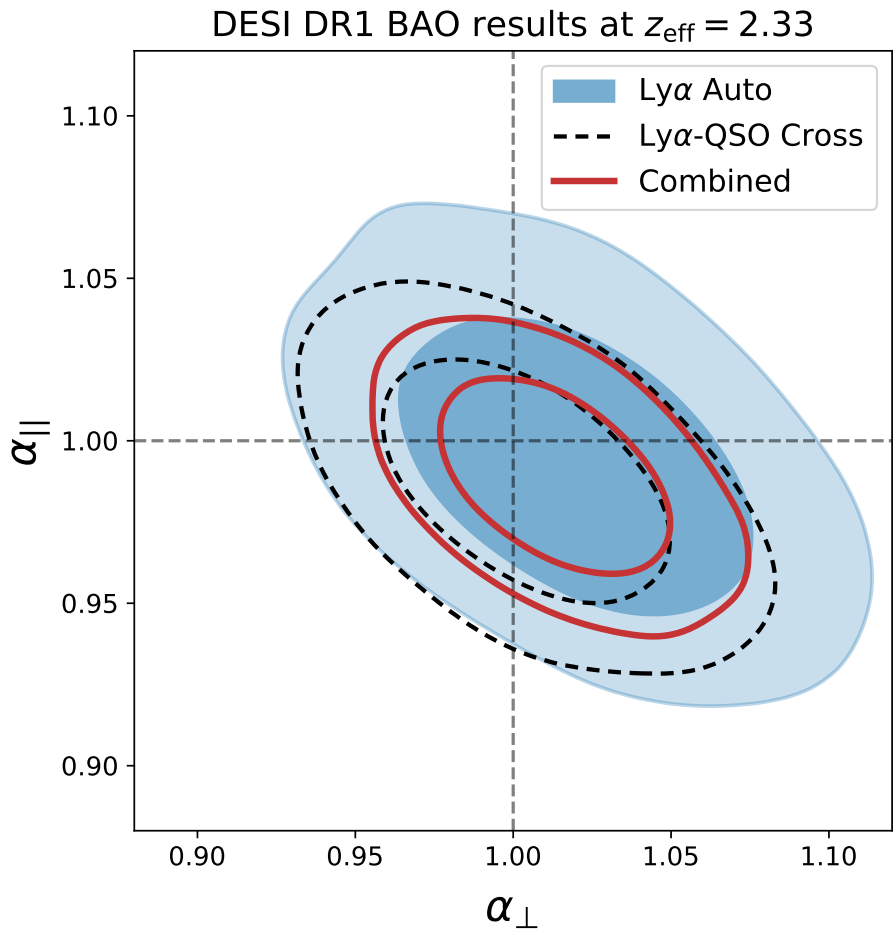

Figure 2.4: BAO scale parameters $\alpha_{\parallel}$ and $\alpha_{\perp}$ measured from the Lyman-α auto-correlation (blue contours), the cross-correlation (black dashed contours) and their combination (red contours). The contours correspond to the 68% and 95% credible regions. Figure taken from [50].

3

The Dark Energy Spectroscopic Instrument

Spectroscopic galaxy surveys are key in probing cosmological models, providing invaluable insights into the large-scale structure and evolution of the Universe. Over the years, a series of galaxy surveys have significantly contributed to our understanding of the Universe. The Center of Astrophysics (CfA) conducted a spectroscopic survey, spanning from 1977 to 1982, which measured the spectra of approximately 2.4k galaxies [89]. This survey was later expanded by the CfA-II survey, to include spectra from nearly 13k galaxies in the early 1990s [90].

The evolution of galaxy surveys has gone through an exponential increase in the number of observed galaxies. The Two-degree-Field Galaxy Redshift Survey (2dFGRS) marked an order of magnitude increase on the number of studied galaxies by collecting spectra from approximately 221k galaxies between 1997 and 2002 [91]. The Sloan Digital Sky Survey (SDSS) further increased the number of observed galaxies in a four-stage project that spanned from 2003 to 2020. The SDSS stages comprise various initiatives, including SDSS-I, SDSS-II, the Baryon Oscillation Spectroscopic Survey (BOSS) [75] (part of SDSS-III) and the extended Baryon Oscillation Spectroscopic Survey (eBOSS) [81] (part of SDSS-IV), which altogether measured spectra from nearly 4 million galaxies [92–94]. SDSS started its SDSS-V stage period in 2020 and plans on performing more galaxy spectroscopic surveys.

The most recent effort on spectroscopic galaxy surveys is the Dark Energy Spectroscopic Instrument (DESI) with the scientific goal of studying the role of Dark Energy in the expansion history of the Universe through the measurement of the BAO imprint on galaxy clustering. Additionally, it will measure the growth of structure through redshift-space distortions combined with BAO measurements. These goals will be achieved by observing approximately 40 million galaxy spectra in a 14,000 $\deg^2$ area during a five-year period, reaching a precision on its measurements of sub-percent level, see Section 6 of [95] and Section 5 of [88] for examples of the forecasted uncertainties precision level of the DESI survey.

The goal of this chapter is to describe the main features of DESI, including

the instrument itself, target selection strategies, data acquisition pipelines, and the survey status up to the date this thesis is being written. The contents are mainly based on key DESI publications, which will be cited accordingly for further details on the corresponding sections. Some of the figures presented in this chapter are taken from the cited references and the DESI figures database with all the corresponding credits given.

3.1 The instrument

The major components of DESI include a telescope, an optical corrector, a focal plane assembly with 5000 positioners attached to a fiber system and ten 3-band spectrographs. Figure 3.1 shows a 3D model of DESI installed in the telescope with various of its main components labeled.

The goal of this section is to describe the instrumental characteristics of DESI. The text is mainly based on the Key Paper on the Overview of the DESI Instrumentation [96] and the dedicated references of each component with credits to the DESI Collaboration and the members involved in writing these references.

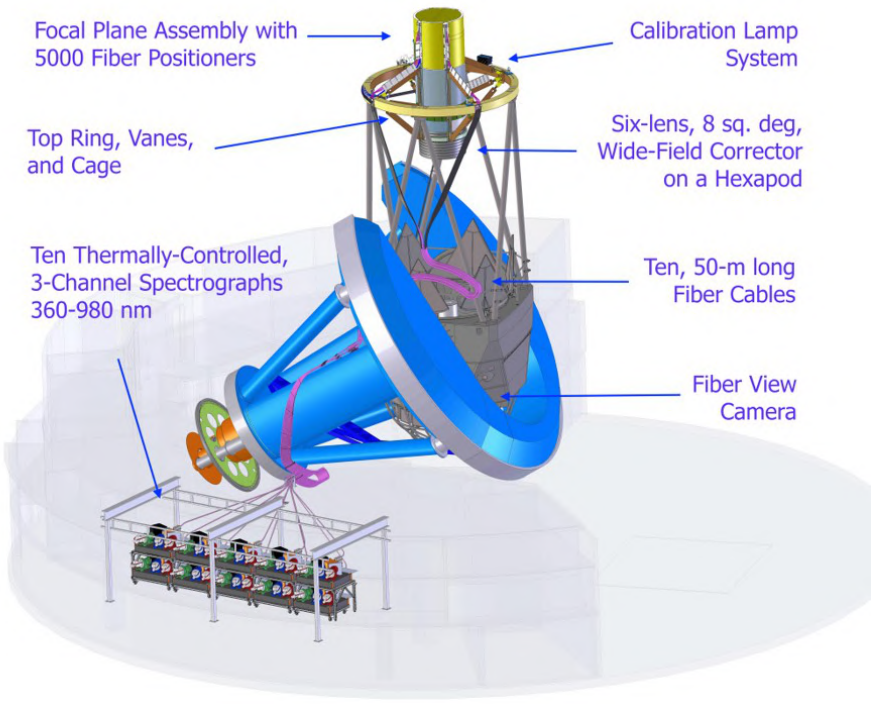

Figure 3.1: Three-dimensional Model of DESI installed in the host telescope. All of the main components are labeled. Figure taken from [96].

3.1.1 Telescope

DESI is installed at the Mayall 4 meter telescope located at the Kitt Peak National Observatory in Arizona (USA) and is operated by the National Optical-Infrarred Astronomy Research Laboratory (NOIRLab). Figure 3.2 shows the exterior and interior of the Mayall telescope.

Figure 3.2: Photographs of the exterior (left) and interior (right) of the Mayall telescope where DESI is installed. Credits to Marilyn Sargent and the DESI collaboration.

The Mayall telescope underwent major upgrades to support the installation requirements of DESI. These updates include the modernization of the telescope control system (TCS) with the goal of minimizing the time between exposures and accurate guiding of the telescope during observations. The Mayall facility was also improved by increasing the building electric wiring capacity, enhancing the cooling system, and bandwidth upgrades on the Ethernet system. Refer to section 10 of [96] for further details on the improvements made to the Mayall telescope and facility.

3.1.2 Optical corrector and support systems

The instrument includes a wide-field optical corrector operating in the $360--980$ nm wavelength range and changes the focal ratio of the telescope's primary mirror from $f/2.8$ to $f/3.9$, also correcting for optical aberrations over a 3.2-degree field of view. The corrector consists of six one meter diameter lenses: four fused silica lenses, two of them with an aspheric surface, and two additional lenses that function as an Atmospheric Dispersion Corrector (ADC) made from borosilicate glass that can be rotated to correct for the dispersion introduced by the atmosphere when the telescope is pointed towards angles off the zenith up to 60 degrees. The left panel of Figure 3.3 shows a diagram of the optical corrector lens array.

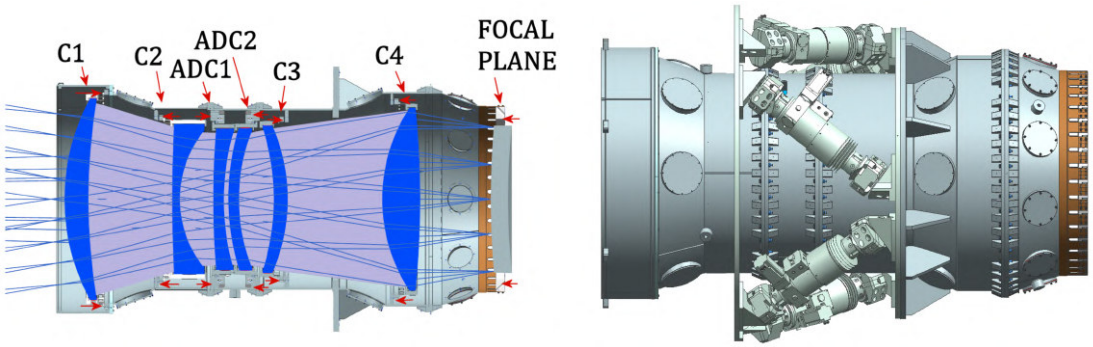

Figure 3.3: Left panel: DESI optical corrector interior diagram highlighting the four correcting lenses array (C1-C4) and two ADC lenses (ADC1, ADC2). Right panel: Model of the corrector barrel as viewed from the exterior. Figures taken from [96].

The corrector is held by a supporting system that includes a steel barrel, shown in the right panel of Figure 3.4, that holds cells where the lenses of the corrector are placed. Additionally, this system includes a mechanism dedicated to rotate the ADC lenses and an hexapod mechanism that ensures that all of the lenses from the corrector are aligned with the primary mirror of the telescope, also visible in the right panel of Figure 3.4. The supporting system also comprises a prime focus cage, a top ring and four sets of three vanes with adjustment features to center the prime focus cage to the optical axis of the primary mirror, shown in Figure 3.4, that are placed at the top end of the Mayall telescope.

Refer to Sections 3 and 4 of [96] and the dedicated paper [97] for more technical details of the corrector and supporting system.

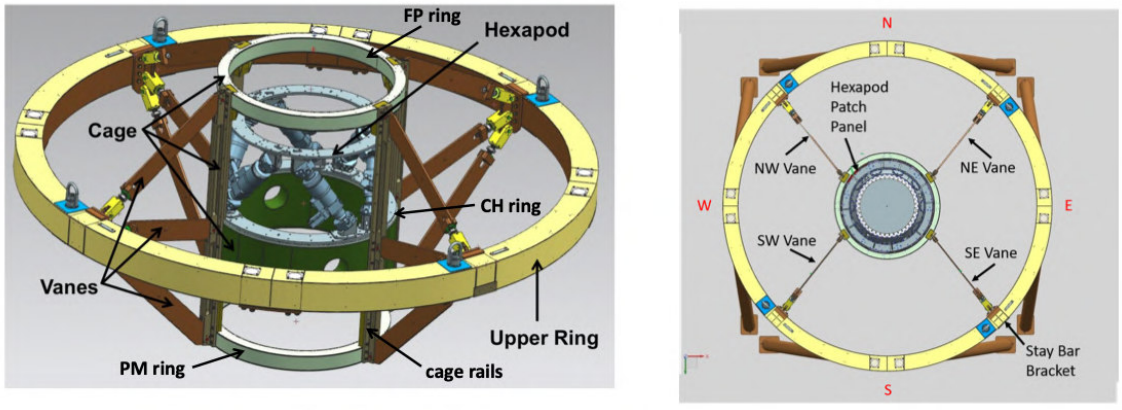

Figure 3.4: Diagram of the prime focus cage, top ring and vanes as part of the corrector support system placed at the top end of the Mayall telescope as seen from aside (left panel) and above (right panel). Figure taken from [97].

3.1.3 Focal plane and fiber systems

The focal plane system (FPS) of DESI includes a Focal Plane Assembly (FPA), which is mounted at the rear end of the optical corrector and consists of 10 identical subdivisions, called petals. Each petal includes 500 fibers of 107 μm diameter and 47.5 meters long that connect to one of the spectrographs on the Instrument and two additional fibers that connect to a sky monitor system. Each fiber is attached to robotic positioner with two rotational degrees of freedom, mounted in a hexagonal grid with separations of 10.4 mm between positioners, that automatically re-targets the fiber in less than 2 minutes for each observation with a precision of less than 10 μm. An anti-collision software plans the movement sequence of each positioner to avoid crashes.

Additionally, each petal includes a Guide/Focus/Alignment (GFA) system placed at their outer edge. Of the ten total GFAs, six are set as guide cameras, calibrated with standard reference stars, and the four remaining as wavefront cameras used to maintain the optical alignment between the optical corrector and the primary mirror.

The DESI FPS also includes a Focal View Camera (FVC) placed at the center of the primary mirror that monitors the correct positioning of the fibers through 123 illuminated point sources (12 per petal)[1] referred to as fiducials, whose positions are well determined on the focal surface. The fibers are highlighted through a back-illumination system and their relative positions on the focal plane are mapped using the illuminated fiducials as a reference. This allows to perform corrections on the fiber positions with a typical accuracy of 50 μm RMS.

All the components of the FPS, including electrical systems are stored within an insulated Focal Plane Enclosure (FPE) with a cooling system that regulates the temperature. Figure 3.5 shows one of the petals of the Focal Plane system and the complete layout of the petal, fibers, positioners and electronics. Refer to Section 5 of [96] and the dedicated paper of the focal plane [98] for further details.

3.1.4 Spectrograph system

DESI includes ten identical spectrographs connected to the focal plane through the fiber system and mounted on a support system. Each spectrograph has three wavelength channels: Blue (B), Red (R), and Near Infrared (NIR or Z). That respectively work on the 3600 – 5930 Å 5600 – 7720 Å and 7470 – 9800 Å wavelength range with

[1] One petal contains 13 and another 14 allowing to distinguish the petals when viewed from the FVC.

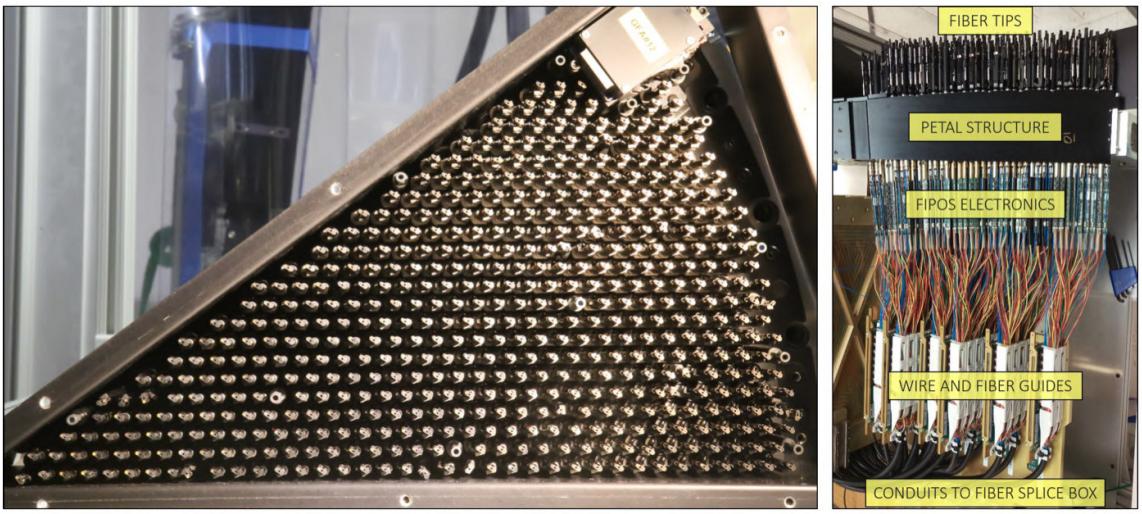

Figure 3.5: Left panel: Front view of one of the focal plane petals fully assembled. Right panel: Side view of a petal assembly including fibers and electronics. Figure taken from [98].

2000 – 3200, 3200 – 4100, 4100 – 5000 resolution power.

The optical design is as follows: light enters the spectrograph through a fiber slit placed directly in a slot on a spherical dichroic designed for the NIR channel. Then, it is directed to a collimator mirror and bounced back to the dichroic. Light beams with wavelengths corresponding to the NIR channel pass through this dichroic directly to a grating and then enter a camera to be finally collected by a Charged-Coupled Device (CCD) detector. Light at shorter wavelengths are reflected off the NIR dichroic to a second spherical dichroic that lets red light through and reflects blue light. Then, both blue and red light beams follow a process similar to the NIR channel where both beams pass through a grating and collected by a CCD inside a camera designed for the corresponding channel. Each channel has a vacuum cryostat placed directly after the CCDs that controls the environment temperature and reduces the readout noise.

The top panel of Figure 3.6 shows a ray diagram of one DESI spectrograph, with its main components labeled. The ten spectrographs of DESI are mounted in two rows of five in an structure, referred to as the Rack, and maintained into an environmental enclosure referred to as the Shack, both seen on the bottom panel of Figure 3.6. The spectrograph system also includes a calibration and sky monitor systems. The calibration system is installed in the top ring of the telescope and consists in a dome screen and calibration lamps attached to four boxes with six drawers, one with a set of LED, the second with two halogen lamps, and the remaining two

with a combination of Cadmium (Cd), Xenon (Xe), Neon (Ne), Krypton (Kr), Mercury (Mg) and Argon (Ar) calibration lamps. The sky monitor system is described in 7.7 of [96]. Refer to Section 7 of this same reference for more technical details on the spectrographs.

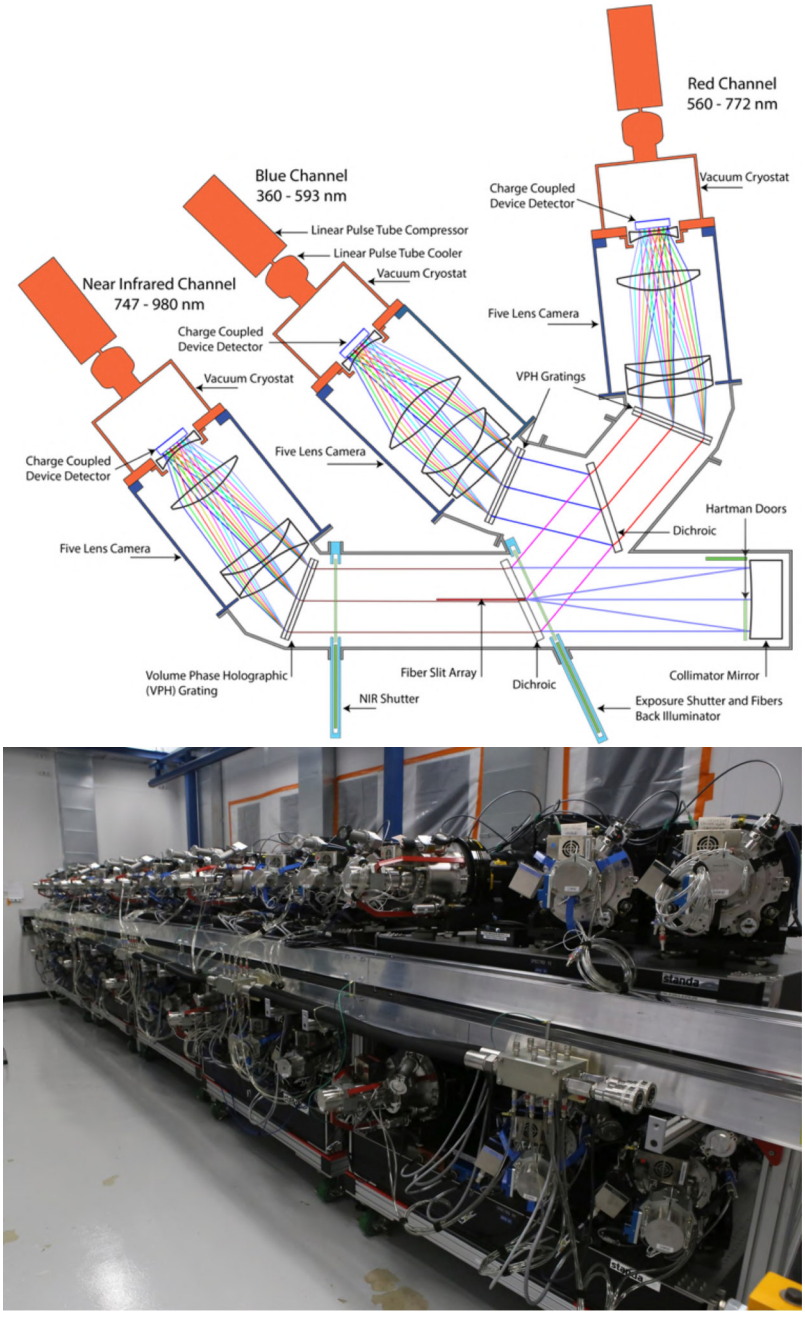

Figure 3.6: Top: Light ray diagram of a DESI spectrograph with its main components labeled. Bottom: The ten DESI spectrographs mounted into the supporting structure (Rack) inside an environmental enclosure (Shack) at the Mayall telescope. Figures taken from [96].

3.2 Target Selection

Prior to the start of its operations, the possible targets to be spectroscopically confirmed by DESI were selected via an imaging survey and color-based selection algorithms. The goal of this section is to briefly describe the process involved into the selection of these targets, including the imaging surveys, and to describe the overall characteristics of the various target samples of the DESI imaging surveys and the expected number of spectra to be measured for each target type. This section is based mostly on the DESI Legacy Surveys webpage[2], their overview paper [99] and the dedicated target selection description and validation papers of the different DESI target types [100–104] with all credits due in the corresponding sections.

3.2.1 The DESI Legacy Imaging Surveys

An imaging survey, referred to as the DESI Legacy Imaging Surveys (LS), was performed from August 2014 until March 2019 in order to select the possible targets to be confirmed by the DESI spectroscopic survey. The Legacy Surveys consisted into three sub-projects: the Dark Energy Camera Legacy Survey (DECaLS), the Bejing-Arizona Sky Survey (BASS) and the Mayall z-band Legacy Survey (MzLS) that altogether cover the 14k deg^2 of the DESI footprint, defined in the region between declination $-18° < \mathrm{DEC} < 84°$, and Galactic latitude $|b| > 15°$, with three optical bands (g, r, and z). Figure 3.7 shows the footprint of the DESI Legacy Surveys on top of the regions covered by other spectroscopic surveys like SDSS, 2dF and BOSS. The three optical bands are complemented by two infrared bands (W1 and W2) of the Wide-field Infrared Survey Explorer (WISE) satellite [105].

DECaLS: This survey was conducted during the whole LS campaign, using the Dark Energy Camera (DECam; [106]) installed on the Blanco 4-meter diameter telescope at the Cerro Tololo International American Observatory located in the Coquimbo Region (Chile). DECam is designed to work in the wavelength range of 4000 – 10000 Å, optimal for the g, r and z photometric bands, respectively, with nominal exposures under ideal conditions of 166, 134 and 200 seconds reaching depths of $g = 23.95$, $r = 23.54$ and z=22.50 with 5σ accuracy. The footprint of this survey covers 5770 deg^2 of the North Galactic Cap (NGC) and 3580 deg^2 of the South Galactic Cap (SGC) in regions with declination below $\mathrm{DEC} \leq 32°$ and $\mathrm{DEC} \leq 34°$,

[2] https://www.legacysurvey.org

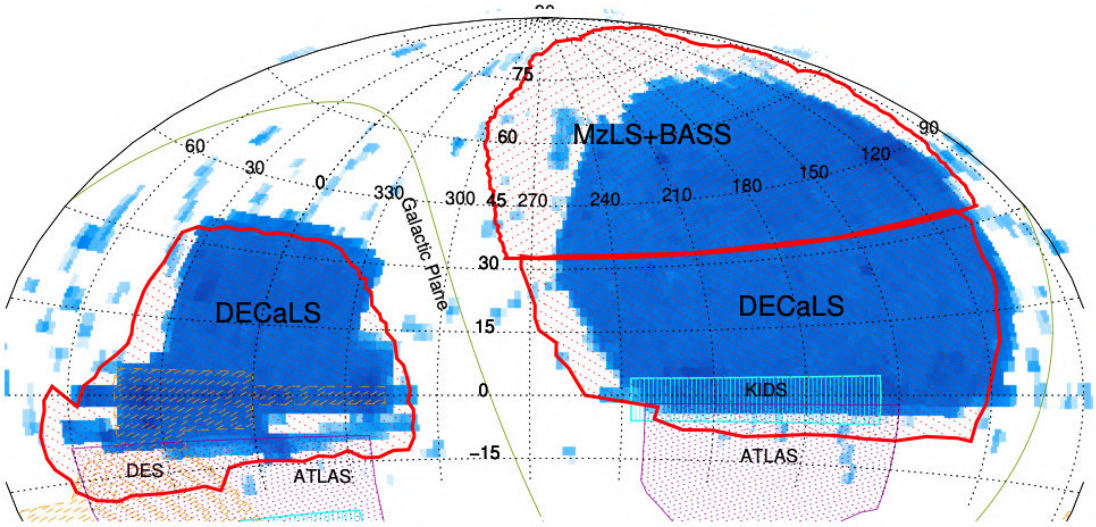

Figure 3.7: Footprint of the BASS, MzLS and DECaLS imaging surveys (red solid lines). Blue colors show the regions covered by the SDSS, 2dF and BOSS spectroscopic surveys. Other imaging surveys are displayed as shaded regions. Figure taken from [99].

respectively. This survey also includes datasets from the Dark Energy Survey (DES; [107]) covering 5000 deg^2 contiguous area on the SGC taken with the same camera.

BASS: Covered 5100 deg^2 of the NGC at declination DEC $\geq 32°$, complementing the region covered by DECaLS, with the 90Prime camera installed at the Bok 2.3-meter telescope at Kitt Peak, Arizona (USA), observing in the wavelength range of 5680 – 7110 Å ideal for the g and r bands, reaching depths of $g = 23.65$ and $r = 23.08$ with 5σ accuracy. The survey was performed from January 2015 until the end of the LS surveys. Refer to the BASS project overview reference [108] for further details.

MzLS: Performed from February 2016 until February 2018, using the MOSAIC-3 camera installed at the prime focus of the Mayall telescope before DESI was installed. It imaged 5100 deg^2 of DEC $\geq 32°$ NGC, similar to BASS. The MOSAIC-3 camera imaged the sky in the 8500 to 10,000 Åwavelength range, covering the z band and complementing BASS, reaching a 5σ depth of $z = 22.60$.

WISE: The WISE telescope conducted an all-sky imaging with a 4-band mission, W1, W2, W3 and W4, centered at 3.4, 4.6, 12 and 22 μm. The three sub-programs of the DESI Legacy Surveys are complemented by the first two bands. The first WISE observational campaign was performed from January until February, 2010,

with two additional extensions completed in February, 2011. These observations were later extended by the NEOWISE [109] initiative starting in December 2013 and still running in 2024.

3.2.2 Target Samples

DESI will measure the spectra of four types of galaxies covering a vast range of redshifts: the bright galaxies, luminous red galaxies, emisision line galaxies and quasars, with additional spectra of stars within our galaxy. The targets of each sample were previously selected from the ninth DESI Legacy Surveys data release (LS-DR9) following the specifications of each dedicated target group, to be briefly summarized during the next part of this section, and that follow a common target selection pipeline, fully described in [110].

The Milky Way Survey (MWS). The specifications of DESI allow to measure the spectra of stars located in our galaxy. Therefore, the target selection of DESI includes the Milky Way Survey (MWS). The primary MWS program targets were selected from the LS-DR9 catalog and combined with astrometric measurements of the Gaia telescope, reaching approximately 30.4 million targets available to be confirmed with spectroscopy. From which, the MWS program expects to measure the spectra of approximately 6.6 million stars in the outer thick disk and stellar halo of the galaxy divided into three main samples that cover the $16 < r < 19$ r-band magnitude range: 3.7 million blue stars characterized by the $g - r < 0.7$ criteria (blue sample), 805k redder stars with $g - r > 0.7$ that fulfill an astrometry criteria based on the Gaia telescope (red sample) and 2.1 million red stars that do not match this astrometry criteria (broad sample). The MWS main program also expects to measure the spectra of 66.3k White Dwarfs (WD), 20.8k faint nearby stars, 8.9k RR Lyrae variables, 17.9k blue horizontal branch stars (BHBs), and stars similar to the blue, red and broad samples but in the fainter $19 < r < 20$ magnitude range (71k red and 486k blue) which combined with the main star sample aforementioned comprise approximately 7.2 million unique MWS targets measured spectra. There is also a secondary targets program including fainter BHBs, white dwarf binary candidates, dwarf satellites and globular clusters that will be targeted if there are fibers of the spectrograph available. Refer to [100] for an overview of the MWS program.

Bright Galaxy Sample (BGS). DESI expects to measure the spectra of over approximately 13.8 million low redshift ($z < 0.6$) bright galaxies. To reach this goal,

possible BGS targets were selected from LS-DR9 complimented with photometry measurements from catalogs of Gaia, Tycho-2, OpenNGC and the Siena Galaxy Atlas to perform star-galaxy separation and spatial masking. The selected BGS sample is subdivided into bright galaxies with magnitudes $r < 19.5$ (BGS Bright sample) with a density 854 targets/deg^2, fainter galaxies with r-band magnitudes in the range $19.5 < r < 20.175$ (BGS Faint sample) with 526 targets/deg^2 selected by imposing an additional color cut criteria based on $r_{\text{fiber}} - \text{color}$[3], and low redshift quasars (BGS AGN sample) with a selection criteria based on optical and WISE colors and a density of 4 targets/deg^2. Refer Section 3 of [101] for further details on the selection criteria imposed to the BGS sample. For details about the BGS program strategy design and validation refer to to Sections 4 and 5 of this same reference.

Luminous Red Galaxies (LRG). Luminous red galaxies (LRGs) are primarily populated by old stars and are easily identified by their characteristic red color and a prominent break on their spectra at 4000 Å. DESI expects to measure the spectra of approximately 7.46 million LRGs in the $0.4 < z < 1.0$ redshift range. The sample of possible LRG targets available includes 624 targets /deg^2, selected with LS-DR9 after a cut in the z-band fber magnitude z_{fiber} and additional color cuts based on the grz and W1 photometric bands with different cut values applied to the north and south regions of the DESI footprint. Refer to Section 2 of [102] for further details on the LRG target selection criteria and to Sections 3 and 4 for a validation of this methodology.

Emission Line Galaxies (ELG). Emission line galaxies (ELGs) are star-forming galaxies with a young population of stars that produce characteristic emission lines in their spectra and whose redshift is easily measured using the position of these lines, in particular an O II doublet at 3726 Å rest-frame. The ELG program comprises the largest dataset of DESI, and consequently the tightest cosmological constraints, expecting to measure the spectra of approximately 15.7 million ELGs in the $0.6 < z < 1.6$ redshift range. The possible ELG targets were selected from the LS-DR9 following a quality cut, a $g_{\text{fiber}} < 24.1$ g-band fiber magnitude cut and a $(g - r)$ vs. $(r - z)$ selection box that primarily sets the redshift range of the targets. The ELG selected sample is further divided into the low priority (ELG LOP) subsample with 1941 targets/deg^2 and the very low priority (ELG VLO) subsample

[3]In this section some selection criteria will be referring to fiber magnitudes. These are defined as the predicted magnitude within a DESI fiber for the corresponding photometric band.

with 463 targets/deg^2. These two subsamples are complimented by the high priority subsample (ELG HIP), a random selection of 10% of the ELG LOP and VLO targets of high higher priority in order to assess equality in the assignment of the focal plane fibers to the ELG and LRG targets during observations. Refer to Section 3 of [103] for more details on the target selection criteria of these subsamples and to the rest of this reference for a complete description of the properties of the ELG sample and a validation of the target selection procedure.

Quasars (QSO). The highest redshift range covered by DESI is constituted by the quasars (QSO) program. It expects to measure the spectra of approximately 2 million quasars $0.9 < z < 2.1$ to be used as point tracers and 840k Lyman-α quasars at redshift $z > 2.1$ to perform measurements of the Lyman-α forest correlation functions, as explained in Section 2.2.1. To achieve this goal, the selected QSO sample contains about 310 targets/deg^2 in the $16.5 < r < 23$ r-band magnitude range. During the Survey Validation phase of DESI (see Section 2.2.2) two quasar selection methods were tested, the first of them based on color cuts on the grz optical, and the W1, W2 infra-red bands. The second, and final, quasar target target selection strategy selects objects with $W1 < 22.3$ and $W2 < 22.3$ to avoid stellar contamination, and then applies a machine-learning algorithm based on the Random-Forest (RF) technique that allows to increment the sample size by 15% for tracer quasars and by 21% for Lyman-α quasars. The RF algorithm was trained using two samples, one with known quasars and another with a set of unresolved objects that do not exhibit quasar characteristics in their light curve, allowing the algorithm to discriminate the latter. See [104] for more details on the quasar target selection procedure and its validation.

3.3 Data acquisition

Achieving the goals set by DESI requires a good planning of both the survey and its daily operations to acquire data. The latter is mainly controlled by the DESI survey operations pipeline which includes the planning of the observational campaign of the night, the assignment of the instrument fibers to perform the targets exposures, and post-processing tasks including the spectroscopic extraction of data from the CCD, the measurement of the redshift of each observed object and their classification, among other steps. Figure 3.8 shows a flow chart of the main steps of the DESI operations pipeline. The goal of this section is to briefly describe this pipeline and is

mainly based in the DESI survey operations paper [111] and the dedicated reference for the spectroscopic data extraction pipeline [112] with corresponding credits given too the authors of these references.

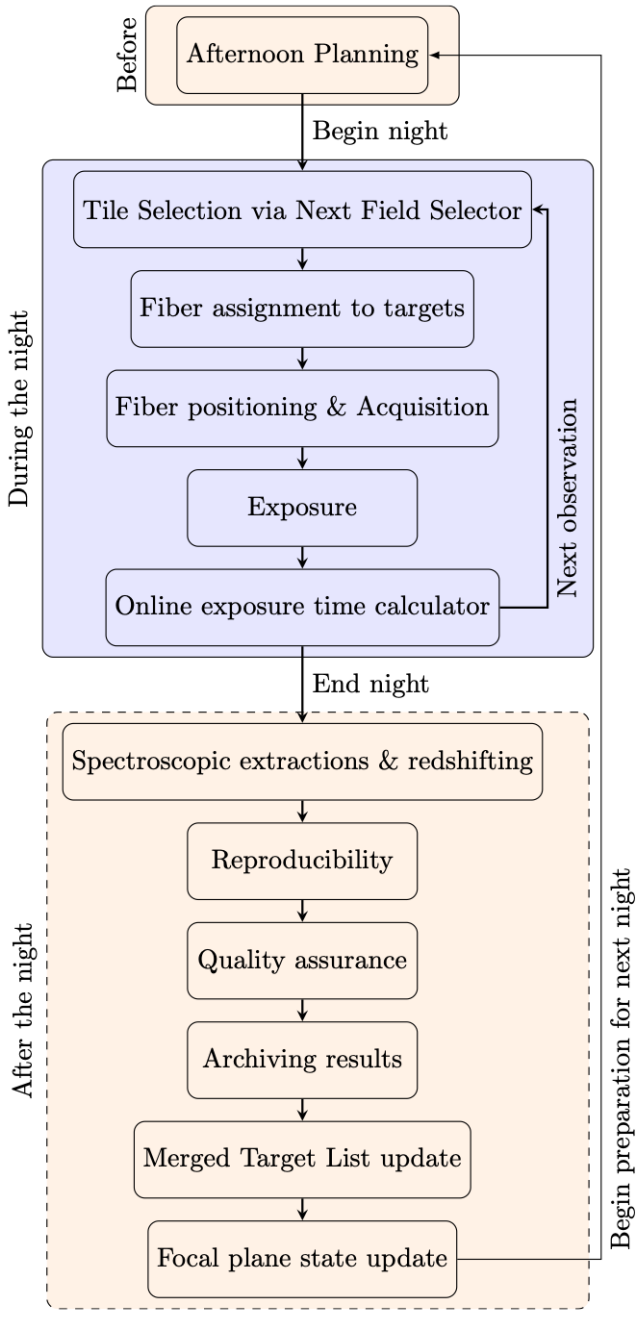

Figure 3.8: Flow chart of the DESI daily operations steps. The tasks inside the dashed box can be optionally postponed if the systems are not available. Figure taken from [111].

3.3.1 Survey Strategy

The DESI survey strategy includes two main observation programs. The dark-time program, aimed to observe the LRG, ELG and quasar samples when the seeing, airmass, transparency and sky background conditions are optimal to perform observations at a survey speed above 0.4.[4] And the bright-time program, dedicated to measure the spectra of the BGS and MWS programs, when conditions are not good enough to perform dark-time observations and the survey speed is above 0.08. If neither the conditions to perform bright or dark time observations are satisfied, there exists a backup program dedicated to observe the brightest stars in the MWS sample.

The dark and bright-time programs will respectively cover the DESI footprint with 9929 and 5676 tiles, defined by the coordinates of the center of the tile, which corresponds to sky-coordinates where the DESI the focal plane center is pointed, and the sky-coordinate positions of each fiber assigned to a target. The tiles are distributed in seven dark-time and four bright-time passes of 1427 non overlapping tiles, where each pass is rotated in order to fill the gaps between tiles allowing a complete coverage of the sky. Figure 3.9 shows an example of a DESI tile centered at RA = 0, DEC = 0. The DESI survey follows a "depth-first" strategy that aims to fully cover a region before moving to another part of the sky. Refer to section 4 of [111] for further details on the survey strategy.

3.3.2 Afternoon planning

The first step in the daily operations pipeline involves planning the observation schedule for each afternoon. Where completed, pending and unobserved tiles are identified, and priorities are assigned to each uncompleted or unobserved tile based on their declination, if the tile has been started or not on previous nights, and the number of finished tiles overlapping with such tile.

The tiles to be observed at night are selected by the Next Field Selector (NFS). Two minutes before finishing each observation the NFS selects a program, based on the current observational conditions, and assigns a score to each tile of that program based on their priority and other factors including the time that it takes to move DESI from its current position to that tile, the hour angle of the tile and the hour

[4]The survey speed measures how fast DESI accumulates signal-to-noise if the instrument is pointed at the zenith with no dust-extinction (reduction of light due to dust) at the observational conditions of that night. Reaching a value of zero at clouded nights and a maximum of 2.5 at clear sky conditions.

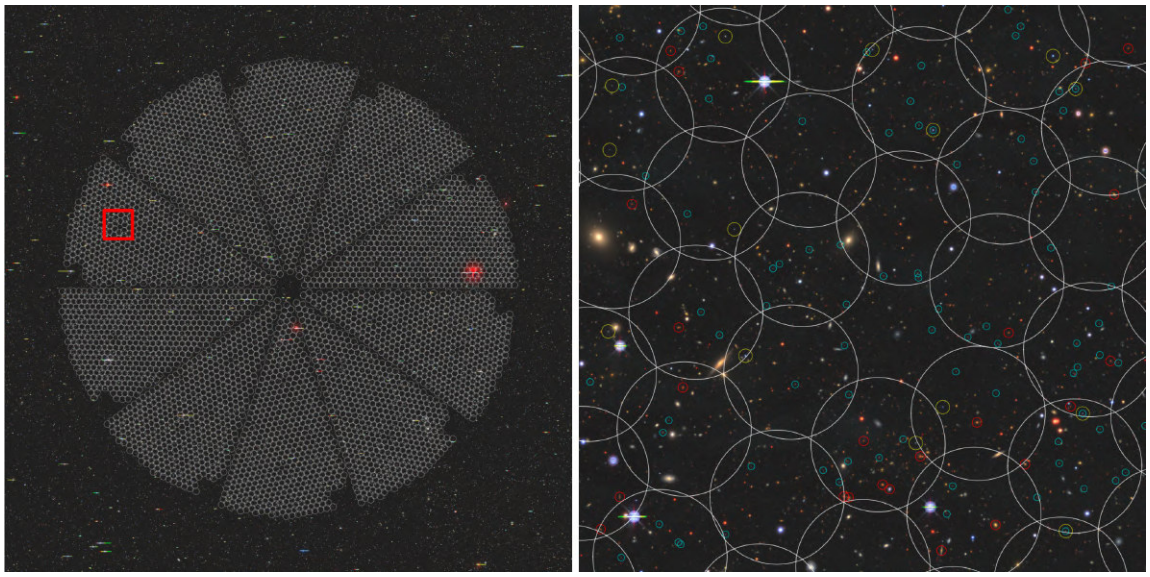

Figure 3.9: Left panel: Example of a DESI tile centered at the RA = 0, DEC = 0 coordinates. The small white circles display the regions reachable by each individual fiber. Right panel: Zoom of the region highlighted by a red square in the left panel. Small colored circles display the type of targets: red for LRGs, cyan for ELGs, and yellow for quasars. Figure taken from [113].

angle of the mid-point of the next observation. The NFS selects the tile with the highest score to be observed next. Refer to Sections 5.4 and 5.5 of [111] for more details on the selection criteria of the tiles to be observed during the night.

3.3.3 Fiber assignment and exposures

After a tile has been selected by the NFS, the `fiberassign`[5] package assigns each available DESI fiber to a reachable target based on the focal plane state and a Merged Target List (MTL), a catalog that contains relevant information about each potential target and its priority to be observed.

The priorities in the MTL are set based on the target type and their observing status. The initial priorities of unobserved targets of the dark-time program place quasars first, then LRGs and ELGs of the HIP sample, both with the same priority, and ELGs of the LOP and VLO samples with the lowest dark-time priority. All unobserved targets of the bright-time program have lower initial priority than those of the dark-time program targets. White Dwarfs have the highest bright-time priority, followed by BGS, and stars of the MWS samples at the lowest bright-time priority. Some rare events of secondary observation programs of DESI, such as gravitational

[5] https://github.com/desihub/fiberassign

lenses, have the highest priority among other targets. The targets designed for the backup program have the lowest priority of all the unobserved targets. Refer to Section 6 of [111] for more details on the MTL.

Once each available target has been assigned to a target, an exposure time calculator (ETC) algorithm assigns an exposure time to the tile, based on the current observational conditions, measured from the sky monitor of the spectrograph system, and the total accumulated effective exposure time of the tile. This algorithm is also responsible for splitting long observations into multiple exposures if the conditions are not optimal to perform the complete exposure. Refer to Section 5.7 of [111] and to the dedicated reference [114] for more details on the ETC.

3.3.4 Spectroscopic data processing pipeline

After every night of observation, if systems are available, the DESI spectroscopic pipeline processes the data of the observed targets. This pipeline is responsible of extracting the flux-calibrated spectra from the CCDs raw images obtained during the observation campaign of each night. The complete details of this pipeline are given in [112].

The first step in this pipeline involves converting the CCD Analog-Digital Counts (ADC) into an electron per pixel count, and performing a calibration and preprocessing of the CCD raw image data. This requires a series of calibration exposures performed every afternoon prior the start of observations.[6] Some of these calibration exposures are then used to extract a bias image and the dark current from the observed images, perform a flat-field correction accounting for non-uniformities in the CCD pixel response, and mask cosmic rays and possible bad pixels or columns in the CCDs. The arc calibration exposures are used to adjust the spectral trace coordinates, calibrate the wavelength, and measure a 2-dimensional Point Spread Function (PSF) of each CCD.

Once the CCD images are calibrated and the PSF is determined, the pipeline extracts the spectra following the "spectro-perfectionism" algorithm [115] which performs a 2-dimensional forward modeling of the CCD pixel data using the measured PSF model. The extracted spectra are then calibrated by performing a flat field

[6]The calibration exposures consist in a set of 25 exposures of 0 seconds with the spectrographs shutters closed, denominated zero; a 300 seconds exposure with the shutters still closed, denominated dark; one set of 3 exposures of 120 seconds on the LED array for each of the four calibration system boxes, denominated flat; 3 exposures of 5 seconds on the Hg, Ar, Cd, Ne, Xe and Kr calibration lamps, and 5 exposures of 30 seconds on the Cd and Xe calibrations lamps, denominated respectively as short arc and long arc.

correction, using the flat calibration exposures that account for variations between the fibers or the spectrographs, and by subtracting a sky spectrum model measured from dedicated fibers that observe empty space locations in the sky.

The flux from the sky-subtracted spectra is calibrated by comparing the measured spectra of standard stars, observed during the same night of observations, with their expected flux from photometry. Allowing to perform a conversion of the flux from electron counts into physical units (10^{-17}ergs/s cm^{-2} Å^{-1}). Finally, if the target has multiple exposures, the flux-calibrated spectra are co-added allowing one to improve its signal-to-noise ratio.

The last step on the data processing pipeline is to pass the measured spectra through the `redrock` package,[7] [116, 117] which classifies them into target types and measures their redshifts. The measurements from `redrock` are then complimented by two additional algorithms, also called afterburners, the neural network classifier `QuasarNet` [118] and the `Mg II` line finder algorithms, both dedicated to identify quasars.

After completing the DESI data processing pipeline, the information of each observed target in the MTL is updated including the redshift, target type, total exposure time, and priorities. Targets that accumulate one observation of total effective exposure time are assigned low priority. Confirmed $z > 2.1$ quasars are the exception to this rule since they receive a higher priority than LRGs and HIP ELGs, but less than unobserved quasars, until they accumulate four effective observations.

3.4 Survey status overview

In this section we will review the relevant information about the DESI survey progress through time, including the Early Data Release (EDR), the works towards the first DESI data release of the first year of operations (DESI-DR1 or DESI-Y1) and a quick overview of the survey progress as today (early 2024, third year of operations).

3.4.1 Survey Validation: DESI Early Data Release

The journey of the DESI survey has gone through ups and downs throughout its progress. The completed instrument officially saw its first-light in October 2019. After this, the instrument went through a commissioning phase to ensure that everything worked fine until March 2020. However, the validation of survey operations

[7] https://github.com/desihub/redrock

could not be performed until December 2020 due to the start of the COVID-19 pandemic. The Survey Validation (SV) phase of DESI officially started on December 14, 2020 and culminated in May 13, 2021. The data obtained during SV was made public in June 2023, referred to as the DESI Early Data Release (EDR), see [86, 113] for the complete details of the Survey Validation phase and DESI-EDR.

The DESI-EDR contained a total of 466.5k MWS, 428.7k BGS, 227.3k LRGs, 437.7k ELGs and 137.1k quasars gathered through three subphases. The first of them, known as Target Selection Validation or SV1, was dedicated to the test, optimization, and validation of the target selection procedure of each target sample, including extensions and variations of the pipelines, see [110] for more details. During SV1 a total of 175 high exposure tiles (mostly 4 times greater than the value designed for the main survey) where observed during 73 nights, achieving a total of 137.5 observed effective time hours. Some of these tiles were observed with even deeper exposure time[8] and were used to asses the efficiency of the redshift measurement and classification algorithms by comparing with the results of visually inspected spectra, see [119, 120] for further details on Visual Inspection (VI).

The second subphase of SV, called Operations Development or SV2, was performed on few tiles with the purpose to test and validate the end-to-end survey operations procedure, described in Section 3.3.1. This includes a complete enhancement and establishment of the procedure to generate and update the MTL used to plan observations. During the SV2 phase, a total of 39 tiles were observed accumulating 6.4 hours of effective exposure time in 10 nights.

The last subphase, referred to as the One Percent Survey or SV3, started in April 5, 2021 until May 13, 2021[9] and covered a total 20 fields, referred to as rosettes, of 10 to 11 overlapping tiles for the bright-time program and 12 to 13 tiles for the dark-time program. The goal of SV3 was to perform observations following a procedure as close as possible to the main survey design with high fiber assignment completeness and good redshift estimation in zones overlapping with other surveys. This is done to characterize the efficiency of the automated data acquisition procedure and to determine the overall expected characteristics of the main DESI survey. SV3 provided a total of 939k spectra collected across a total of 488 tiles achieving a total of 102.2 hours of effective exposure time.

[8]The deeper effective exposure times in hours for each target class are 0.85 for BGS, 0.95 for MWS, 2.6 for LRGs, 4.2 for ELGs and 6.7 for QSOs.

[9]Some tiles observed after this date, during the main phase of the survey to enhance the completeness of the dataset. The last tile of the SV phase, included in the DESI-EDR sample was finalized in June, 2021.

3.4.2 Main survey

First year of operations

DESI officially started its nominally five-year long main survey in May 14, 2021. From that date until June 10, 2021, DESI compiled the spectra of nearly 18 million targets that will be part of the first official DESI data release, referred to indistinctly as DESI-DR1 or DESI-Y1, in April 4 2024. During this period, DESI operated without a problem with only a programmed interruption during from July 12 until September 12, 2021 to perform hardware upgrade on the focal plane and spectrographs, and an overall maintenance of the instrument.

DESI-DR1 will contain a total of 3.64 million MWS, 6.28 million BGS, 2.83 million LRGs, 3.93 million ELGs, and 1.34 million quasar targets were observed across 2,744 dark-time and 2,257 bright-time tiles, accumulating 148.8 and 782.9 hours of effective exposure time, respectively. Figure 3.10 shows the tiles that cover the DESI-DR1 footprint, with the colors showing the completeness of each covered region. As noted, the dark-time and bright-time programs reached 29% and 41% completeness in just one year. The latter was reason enough to increment the number of passes of the bright-time survey to five. The backup program observed 327 additional tiles.

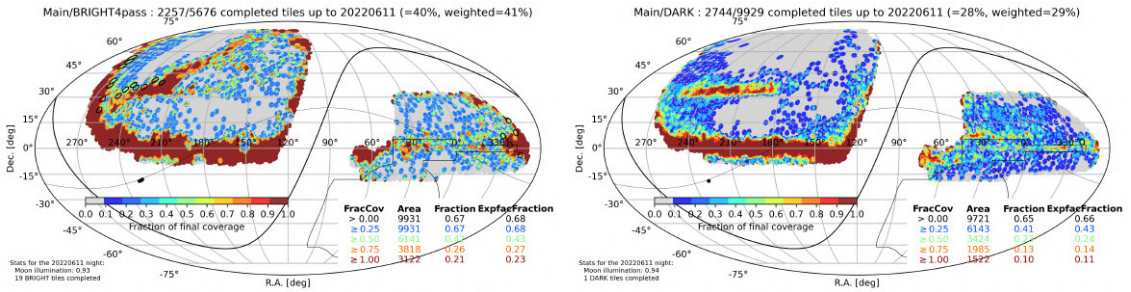

Figure 3.10: Sky map with the tiles covered by the DESI main survey, from May 2021 to June 2022, in the bright-time (left) and dark-time (right) programs. Colors display the observational completeness of each region. Credits to Anand Raichoor.

Current status

From the date marked as the end of DESI-Y1 until today DESI has ceased operations in two additional occasions the first of them occurred in June 2022 when a lightning strike in the surrounding area of the Kitt Peak National Observatory produced a wildfire, known as the Contreras Fire, that compromised the operations.[10] Fortunately, no significant damage was done to the observatory or DESI and operations could be resumed in September 2022. The next shutdown, was a programmed maintenance shutdown from June 22 until August 4, 2023 to re-aluminize the primary mirror of the telescope.

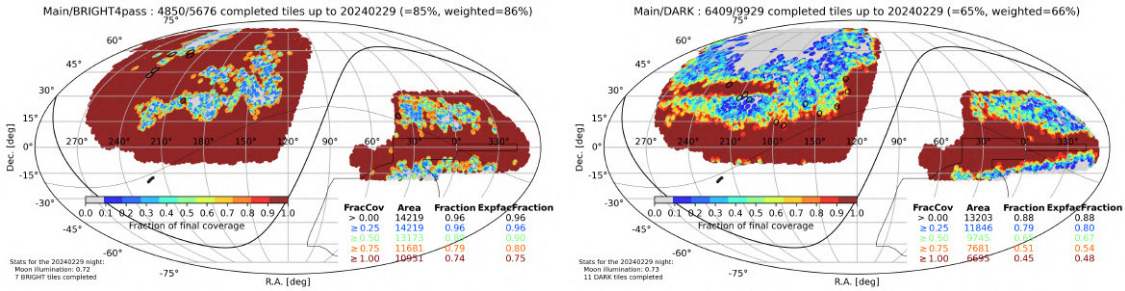

Figure 3.11: Similar to Figure 3.10 but for the current status of the main DESI survey at the moment of writing this thesis (February 29, 2024). Credits to Anand Raichoor.

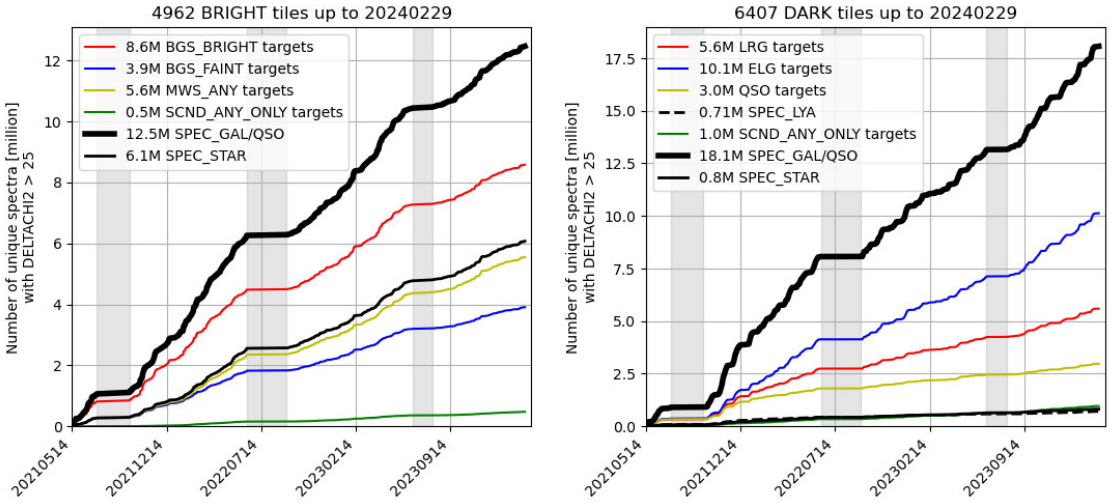

Figure 3.12: Progression of the number of measured spectra by DESI with time. Grey bands show three operation interruptions that the instrument have went through: two programmed maintenance shutdowns during the Summers of 2021 and 2023, and another shutdown in 2022 caused by a wildfire on the observatory area. Credit: Anand Rachoor.

[10] Refer to the DESI blog for more information about the wildfire.

At the moment of writing this thesis (February 2024, close to the third anniversary of the start of the main survey), DESI has reached the milestone of 40 million observed spectra. This includes 5.6 million MWS, 12.5 million BGS, 5.6 million LRG, 10.1 million ELG, and 3 million quasars gathered in 4962 bright-time and 6407 dark-time tiles. Figure 3.11 shows the survey footprint as February 29, 2024 of the bright-time and dark-time programs. Figure 3.12 shows evolution of the number of spectra measured by DESI through time including the three shutdowns aforementioned displayed as gray bands. Even with these shutdowns, the bright-time survey is 86% complete (70% for the five pass program) and the dark-time program is 66% completed, if the survey keeps progressing as it has been doing until now, the nominal DESI five-year survey is predicted to end 6.2 months ahead of schedule.

4

Generating synthetic Lyman-α spectra

Various Lyman-α BAO analyses have used synthetically generated Lyman-α spectra. These mock spectra have played an important role in validating analysis procedures and characterizing of systematics.

The BAO analysis through the Lyman-α auto-correlation function of the Baryon Oscillation Spectroscopic Survey (BOSS; [74]) early data and subsequent BOSS Data Release 9 (DR9; [64, 76, 78]) included a validation step with Lyman-α mocks following the method presented in [121], which relies on the generation of a transmitted flux field through generating a Gaussian random field and performing a log-normal transformation, resulting in a set of correlated lines of sight.

Later, the BOSS DR11 [63] and DR12 [59] Lyman-α auto-correlation analysis made use of mocks developed following the method described in [122], which in principle follow the same methodology as the previous mocks with some improvements, and mimic the statistical properties of the BOSS survey. These mocks also include contamination from HCDs and metals.

While the aforementioned mocks were useful for the Lyman-α auto-correlation analysis validation, the position and redshift of the quasars were the same as in the BOSS survey and therefore uncorrelated with the absorption field, causing the cross-correlation to vanish and making these mocks unsuitable for the purposes of Lyα $\times$ QSO cross-correlation measurements of BOSS DR11 [80]. In this regard, the mocks generated for BOSS DR12 cross-correlation analysis [58] included mock spectra that were correlated with the position of quasars following the technique described in [123], which placed quasars at the peaks of the generated Gaussian random field.

Finally, the extended version of BOSS (eBOSS) DR16 Lyman-α measurement [62] included the use of two new sets of mocks as a validation step for both auto and cross correlations generated following an early version of the methodology described in [88].

The current Lyman-α mock generation procedure makes use of the `desisim`[1] and `specsim`[2] repositories, in particular a script called `quickquasars`,[3] which compiles multiple of the modules and functions of these packages and follows a method similar to the aforementioned BOSS and eBOSS mocks.

The first step to generate Lyman-α synthetic spectra with the `quickquasars` code is to create noiseless spectra. This is done in three main steps and work the same regardless of the survey to be simulated. First, we use previously computed transmitted flux fraction datasets (referred to as raw mocks). Then, we multiply this value by a generated unabsorbed spectrum template (referred to as the continuum). Finally, we add instrumental noise and observational features to our mock spectra. As an additional feature to these three main steps, our mock generation method can include astrophysical contaminants (HCDs, BALs and metals) and also emulate a survey given its specifications (footprint, quasar number density, and magnitude distributions).

Most of these steps are described in [88]. Therefore, the goal of this chapter is to provide a brief summary of these steps while referencing the corresponding sections of the paper where a more extended description is given. Specific Lyman-α mocks for the DESI survey and their generation method will be described in Chapter 5.

Figure 4.1 shows an example spectrum through the various stages in our mock generation method. First, as the Lyman-α transmitted flux fraction from raw mocks. Then contaminated with HCDs, BALs and metals. Then, the total transmitted flux fraction is multiplied by a continuum template, generated with `SIMQSO`, and at last instrumental noise is added, emulating 4000 seconds of exposure time with the DESI spectrograph.

4.1 Transmitted flux fraction (raw mocks)

Including the Lyman-α forest in the simulated datasets rely on an input set of Lyman-α transmitted flux fractions F, related to the optical of the simulated density field by τ by $F = e^{-\tau}$. We refer as raw mocks to the set of transmitted flux fractions without instrumental noise, continuum template, nor astrophysical contaminants added.

Raw mocks can be constructed in a variety of ways depending on the science case of use. For example, using one of the following methods: log-normal, Hydrodynamic,

[1] https://github.com/desihub/desisim
[2] https://github.com/desihub/specsim
[3] https://github.com/desihub/desisim/blob/main/py/desisim/scripts/quickquasars.py

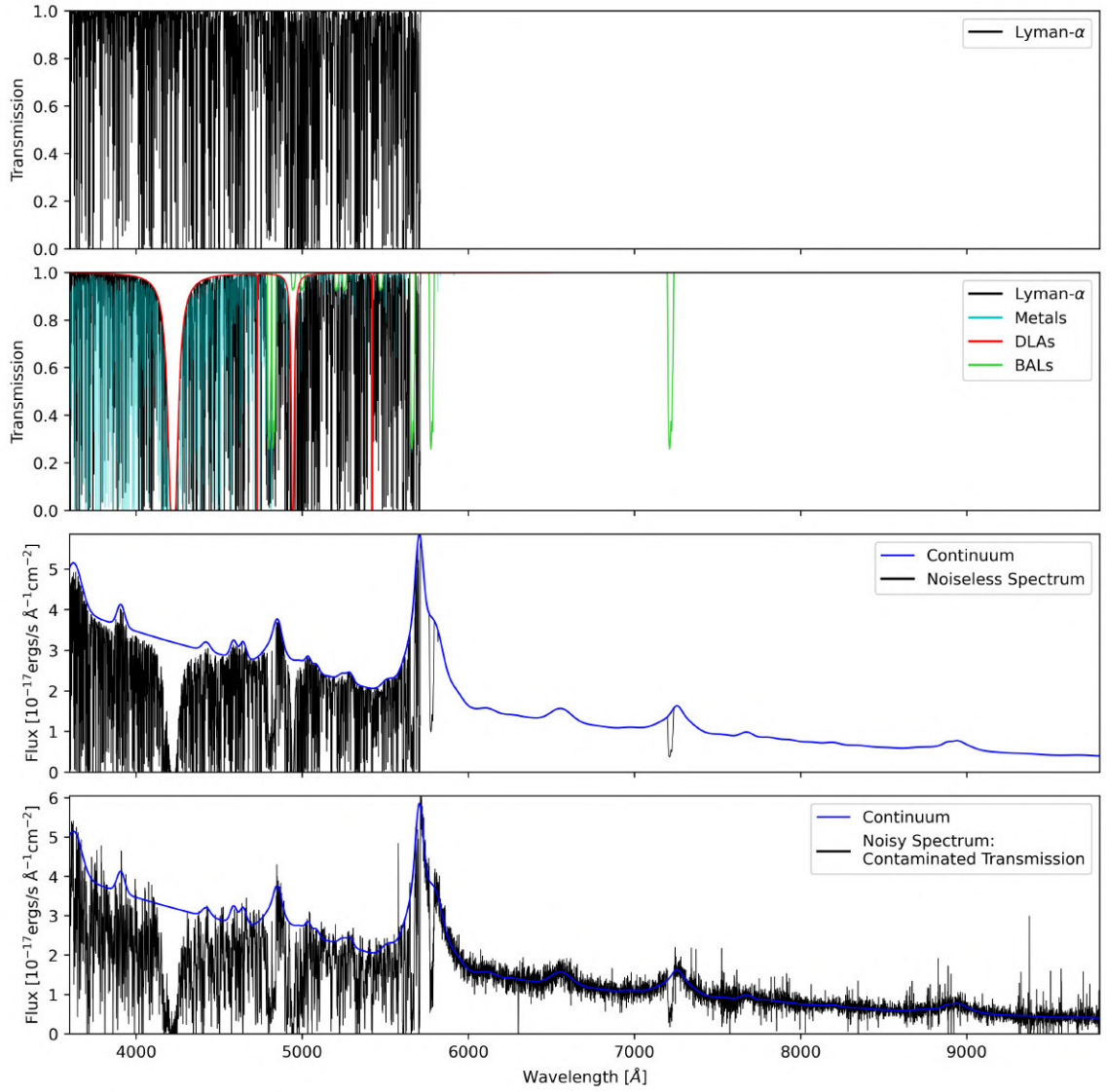

Figure 4.1: Simulated spectrum of a quasar at redshift $z = 3.69$ through the various stages in `quickquasars`. The first row shows the Lyman-α transmission from the raw mocks. The second row shows the transmission after contaminants have been added: DLA features (highlighted in red) of column densities $\log N_{\rm HI} = 21.26, 17.29, 20.01, 17.34$ at wavelength $\lambda = 4224, 4731, 4945, 5424$ Å, respectively. BAL features (highlighted in green) with Absorption Index $\mathrm{AI_{CIV}} = 557.108$, and metal absorption features (highlighted in cyan). The third row shows the noiseless spectrum, i.e the total transmission with contaminants multiplied by the continuum template (blue solid line). The last row shows the quasar spectrum with instrumental noise added simulating 4000 seconds of exposure time.

N-body, Lagrangian Perturbation theory (LPT), to mention some. The simulations presented as a validation step on various of the aforementioned Lyman-α analysis made use of the log-normal method due to their low computational cost. This method consists in generating Gaussian random fields from an input matter power spectrum, then a flux field is computed through a log-normal transformation of the Gaussian field. Finally, this flux transmission field is converted into an optical depth τ field through a Fluctuating Gunn-Peterson Approximation (FGPA).

There are three main raw mock datasets used to produce DESI Lyman-α simulations, referred as the `LyaCoLoRe` [124], `Saclay` [125] and `Ohio`[126, 127] mocks, all based on the aforementioned method for different purposes and some subtle differences. These mocks are further described in their respective references and briefly discussed in Section 2.1 of [88].

4.2 Unabsorbed quasar spectra (continuum)

Our mock generation method includes the generation of templates of the unabsorbed quasar spectrum, also referred to as continuum, and this is further detailed in Section 2.2 of [88]. The default method for DESI mocks is based on the `simqso` package[4] [128, 129]. The `simqso` continuum template generation method is constructed using two main components. First, a broken power-law model model is defined by two elements, the slopes, and the break points. On the one hand, the slopes are randomly sampled following a Gaussian distribution of mean m and dispersion σ_m emulating quasar spectral diversity. On the other hand, break points set the rest-frame wavelength λ of the slopes. For DESI mocks we use the values defined in Table 4.1. The last three rows are based on measurements made in BOSS DR9, while the first two were obtained from a PCA analysis with the goal of better reproducing the mean continuum and spectral diversity observed in the eBOSS DR16 data [130].

Table 4.1: Broken power law model parameters set for the `simqso` continuum generation method.

Slope (m)	**Dispersion (σ_m)**	**Wavelength region** [Å]
-1.50	0.7	$\lambda < 1100$
-0.50	0.7	$1100 < \lambda < 5700$
-0.37	0.3	$5700 < \lambda < 9730$
-1.70	0.3	$9730 < \lambda < 22300$
-1.03	0.3	$\lambda > 22300$

[4] https://github.com/imcgreer/simqso

The second component of the `simqso` continuum generation method are the emission lines. These are a set of emission lines modeled by a Gaussian profile of width σ and amplitude given by the Equivalent Width (EW) of the emission lines, centered at the corresponding observed wavelength of the emission line feature. The value of the equivalent width is randomly generated with a Gaussian distribution centered on the mean value of the EW and a given scatter dispersion. The model used for the DESI mocks is shown in Table A.1 of Appendix A in [88].

There exists an alternative method referred as `QSO`, which was used in previous Lyman-α mocks, for example [122]. This method consists in generating continuum templates based on the eigenvalues and eigenvectors of a PCA analysis on observed spectra. While this method is outdated and is not currently being used for DESI mocks at the point of writing this thesis, it is worth re-implementing it based on DESI spectra for future generation of mocks.

4.3 Astrophysical contaminants

The astrophysical contaminants discussed during Section 2.1.3 can be included into mock spectra using various methods. Here these methods will be summarized, see section 2.3 of [88] for further details.

4.3.1 High Column Density systems

In our synthetic spectra methodology, HCDs can be included into mock spectra by two different methods. On the one hand, they can be added randomly into spectra. This is done by sampling the column density of a random number of HCDs and their redshift given a distribution function. In our case, we follow the column density distribution from the `pyigm`[5] [131] package, defined as the number of HCDs per unit column density N_{HI} per unit absorption distance [132, 133]. This distribution is shown in Figure 4.2. In this case, the generated HCDs would not be not correlated with the density field that was used to produce the raw mocks.

On the other hand, HCDs can be produced during the raw mock generation procedure. In this case HCDs are correlated with the density field unlike the aforementioned method. This method consists in placing HCDs in peaks of the density field above a given threshold and then assigning them a column density given the `pyigm` distribution function. The complete procedure for including HCDs into raw

[5] https://github.com/pyigm/pyigm

mock transmissions is described in [134].

Regardless of the method used to include HCDs on the mock spectra, they are introduced into the mock spectra using a Voigt profile characterized by their column density

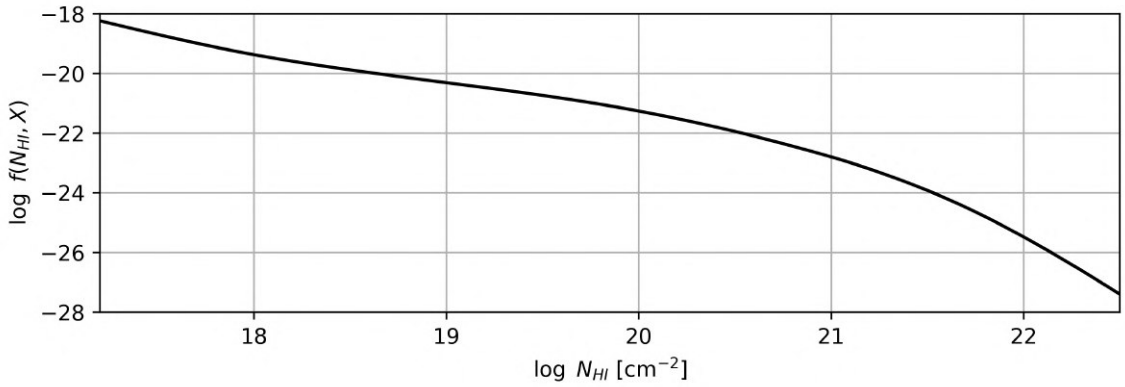

Figure 4.2: HCDs column density distribution function from the `pyigm` package.

4.3.2 Broad Absorption Lines

BALs are randomly included into the quasar synthetic spectra for a given fraction of spectra, 16% for current DESI mocks. In this method random templates are selected from the set of BAL templates presented in [135] which in order include BAL features associated with O VI(1031), O VI(1037), Lyman-α, N V(1240), Si IV(1398), and C IV(1549). DESI mocks emulating the first year of operations include features due to Si IV(1062), Si IV(1074), P V(1118), P V(1128), and C III* (1175), these mocks and updates will be discussed during Section 5.4.

4.3.3 Metals

We have two methods to include metals. The first of them being the `quickquasars` method, where the optical depth of the Lyman-α absorption is re-scaled by a factor C_m according to the relative absorption strength of each metal relative to Lyman-α. Then, the transmitted flux of each metal is computed by $F = \exp(-C_m\tau_m)$ and moved to their corresponding observed wavelength via $\lambda_{obs,m} = (1 + z_{abs})\lambda_m$, where $z_{abs} = \lambda/\lambda_{Ly\alpha} - 1$ is the redshift of the absorption. This method causes metals to have the same redshift-space distortions parameter as Lyman-α.

The second method involves generating the transmitted flux of metal absorptions simultaneously with the Lyman-α transmission during the raw mocks generation pro-

cess. At the moment, only the `LyaCoLoRe` mocks include this feature. In this method metal transmissions are generated in a similar way than the Lyman-α transmission however their optical depth is defined by a relative absorption A_m in such way that $\tau_m = A_m \tau_{Ly\alpha}$. Then, RSDs are applied to the flux-transmitted fraction of each metal. In this method, the absorption features are read by `quickquasars` and then added to the spectra. Notice that, in principle, the re-scaling factor C_m and the relative absorption coefficient A_m are not necessarily the same.

4.4 Adding Noise

The last step in our general mock generation procedure is to include noise, this is detailed in Sections 3.2.1, 3.2.2 and 3.2.4 of [88]. The noise is simulated by the `specsim` package following various models. First, noise due to observational conditions is introduced following an atmosphere modeled by applying extinction and adding a sky background. The atmosphere model also depends on the observing conditions, accounting for seeing, air mass, moon illumination fraction, moon altitude, and moon separation from the tile being simulated. Then, noise introduced by the instrument is modeled by the specific instrument characteristics such as primary mirror area, corrector optics, cameras, throughput, exposure time, to mention some. At last, we add noise accounting for shot-noise, thermal dark currents from the CCD and readout electronics contributions.

5 DESI Lyman-α Mocks

The `quickquasars` mock generation method introduced in [88] includes the flexibility to mirror survey specific characteristics such as footprint, quasar number density distribution as a function of positions, redshift-magnitude distributions, and number of exposures.

The goal of this chapter is to provide general aspects of constructing DESI mocks and describe how mocks for specific stages of the DESI survey were constructed. Lyman-α mocks emulating the results of the complete DESI survey (DESI-Y5) and mocks emulating the Early Data Release (EDR) plus 2 months of observations of the main DESI survey (M2), known as the EDR+M2 sample, were presented in [88]. This chapter will briefly describe these mocks, although we will focus on the enhancements made to the methodology for generating mocks emulating the first year of operations of DESI (DESI-Y1). This means we will review a more detailed explanation of the caveats in the method used to construct the EDR+M2 mocks and how these issues were addressed for the making of the DESI-Y1 mocks.

5.1 General attributes

There are some aspects of the mocks that are common regardless of which DESI survey stage is to be simulated. For example, the redshift-magnitude distributions. The method to emulate the distributions by `quickquasars` is described in Section 3.2 of [88]. It consists of computing the expected quasar number density distribution for DESI-Y5 tabulated as a function of redshift and r-band magnitude. Then, randomly downsample the available quasars from the raw mocks by redshift bin, following the number density distribution marginalized over magnitudes. Finally, a random r-band magnitude is assigned to quasars according to their redshift following the redshift-magnitude distribution. This method was effectively validated in section 4.1 of [88].

By construction, we follow the number density distribution as expected by the DESI quasar target selection pipeline [104]. However, there is flexibility to adjust the distributions as the DESI survey progresses. The distribution used to construct all the mocks presented in this thesis is given in Figure 5.1.

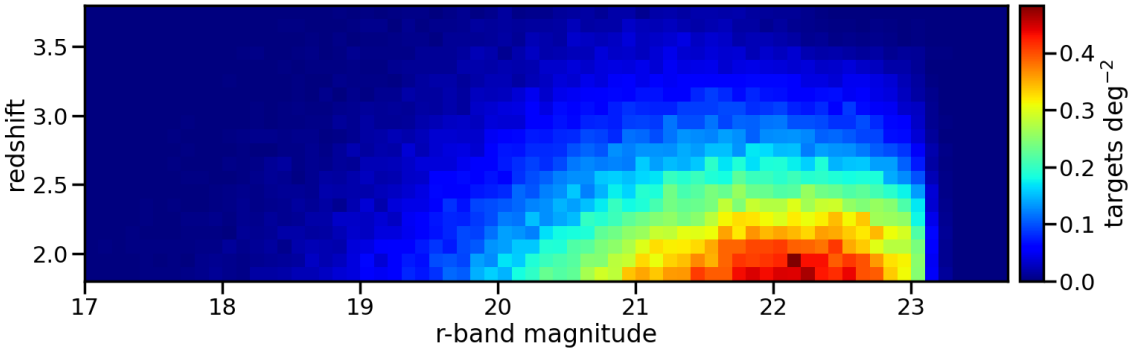

Figure 5.1: Quasar number density distribution as a function of redshift and magnitudes as expected by the DESI quasar target selection pipeline. Figure already presented in [88] as Figure 1.

5.2 Overview of the DESI-Y5 mocks

The simplest to make are the DESI-Y5 mocks since they do not require applying a downsampling on the number density of quasars as a function of their position. The only requirement is to define the expected DESI region of observations. When applying the expected redshift distribution to DESI-Y5 mocks, the result is approximately 1.4 million Quasars from which 929k are Lyman-α. For this kind of mocks we assign all quasars an exposure time of 4000 seconds which reflects the effective exposure time expected for a Lyman-α quasar in the DESI survey. The resulting footprint is shown in Figure 5.2.

5.3 Overview of the EDR+M2 mocks

DESI EDR+M2 mocks were presented in section 4 of [88] presenting a qualitative comparison of these mocks with the actual EDR+M2 observed data. The results were overall favorable, leaving some studies for future mocks releases. One of these improvements was the possibility to modify the method used to downsample mock targets to mirror the footprint and spatial number density distribution. The method presented in [88] consists of dividing the EDR+M2 footprint into HEALpix pix-

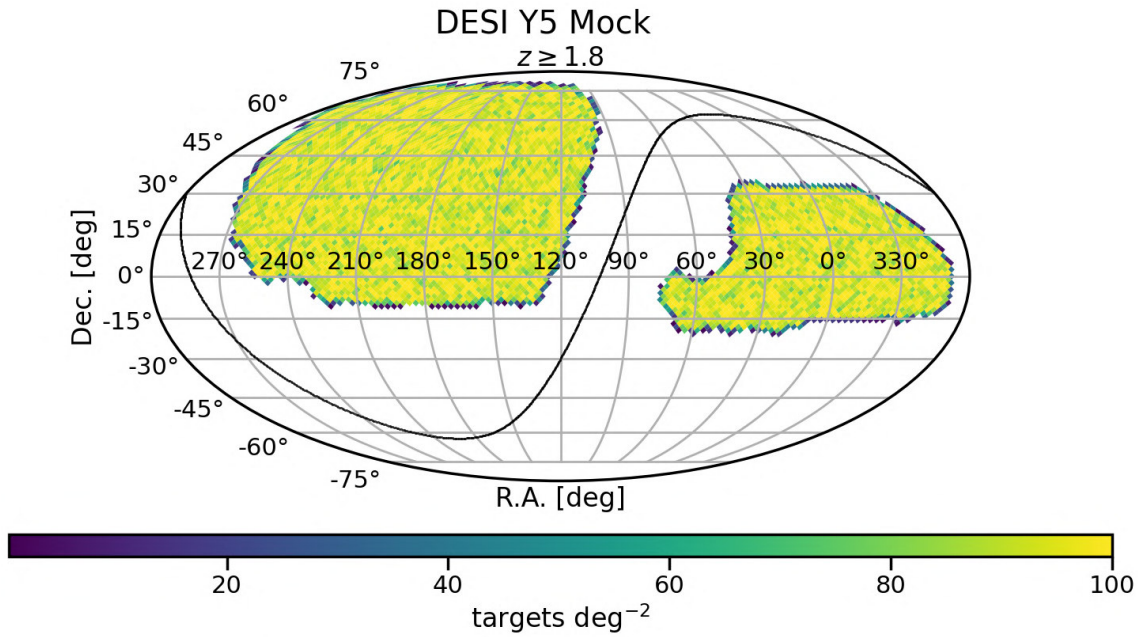

Figure 5.2: DESI-Y5 footprint and Lyman-α quasar density as obtained by applying the redshift distribution expected by the QSO target selection pipeline to our mocks. Figure already presented in [88] as Figure 15.

els [61] of `nside=16`, the native `nside` of the raw mocks. Then, randomly select the exact number of targets in each HEALpix pixel than in data. The exposure time of each Lyman-α target was randomly assigned according to a number of exposures probability function computed by HEALpix pixel.

While this method was found to be good enough for the purposes of the EDR+M2 mocks, there are some caveats in it. First, while the footprint and object density matches at the level of `nside=16` HEALpix pixels, when looking the detail of the objects footprint, mocks do not reflect the actual geometry of observed data. This is clearly seen in Figure 5.3 when looking into the scatter plot of the mock quasar positions, the data footprint on the upper panel is well reproduced. However, if we compare with the actual geometry of data (red dots in the bottom panel), then the resulting mocks are not a good representation.

Second, Section 3.1 of [88] discusses the possibility of this method having an effect on the QSO auto-correlation function at large scales depending on the HEALpix pixel size used. Although evidence of this was not presented in the paper, in this thesis, we will provide more details on the results of studying the QSO 2-point correlation function in mocks. For the following results the 2-point correlation functions were

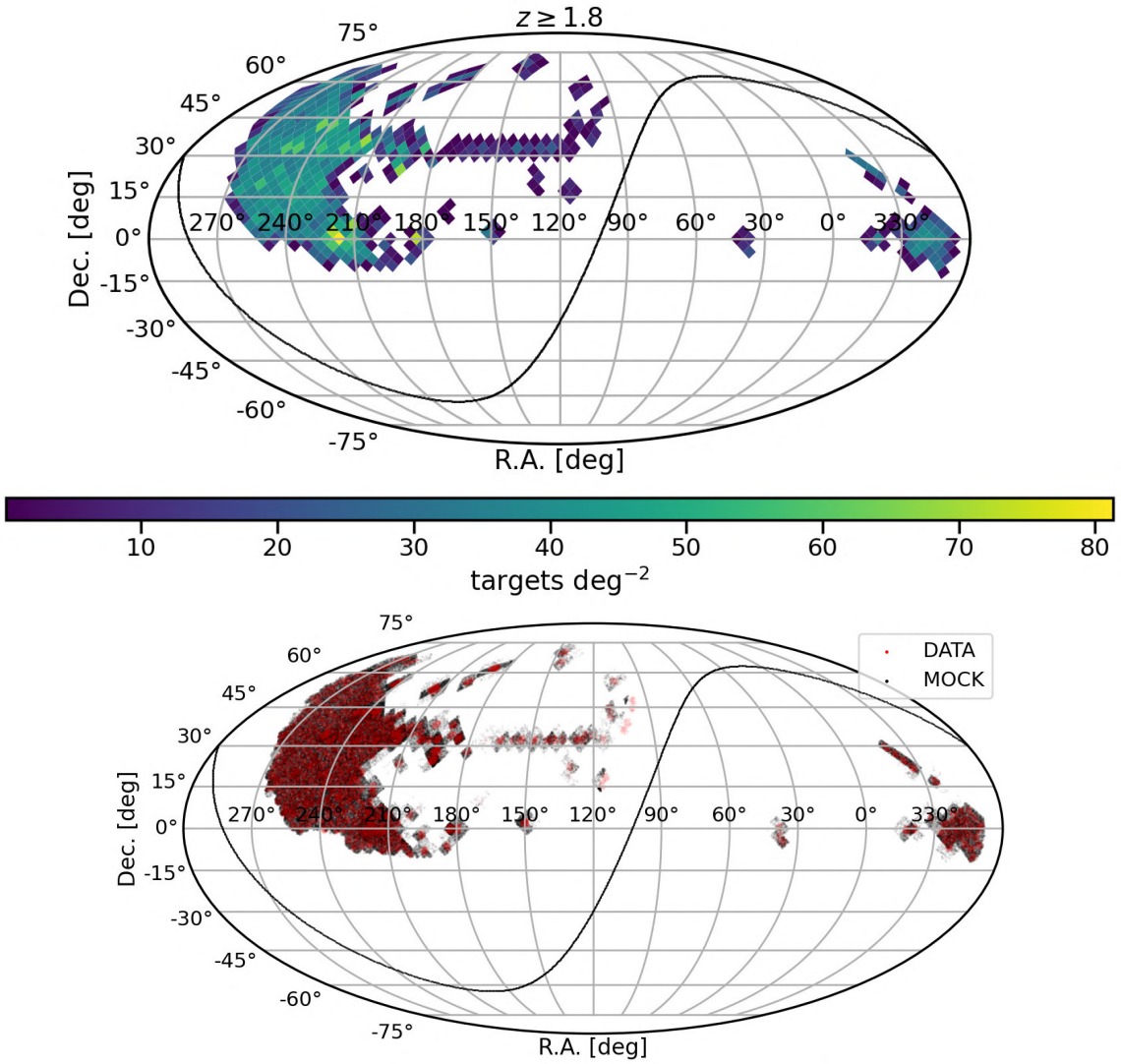

Figure 5.3: Top: EDR+M2 data footprint and quasar number density divided into HEALpix pixels of `nside=16`. Figure already presented in [88] as Figure 4. Bottom: Scatter plot of the quasar positions of data (red) compared to a mock realization (black).

computed using the Landy-Szalay estimator [136] defined by

$$\xi(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)}, \tag{5.1}$$

where DD, RR and DR denote number of pairs within the data, random sample and both samples, respectively.

For this study, we generated a randomly distributed sample of quasars of 10 times the size of data sample. Figure 5.4 shows the comparison of mock realizations using different `nside` (pixel sizes) values for doing the downsampling of targets to match the EDR+M2 data footprint and number density. The figure includes the QSO auto-correlation function that results from applying the redshift distributions to the raw mocks, but without selecting any particular footprint or downsampling. Additionally, we explored the results when uniformly downsampling to match the global number of targets of the EDR+M2 sample. From this figure it is clear that when a uniform downsample is applied, the 2-point correlation function of the complete mock is recovered. However, applying a downsample by HEALpix pixel produces fall in the correlation function at large scales that increases with the size of the HEALpix pixel, which makes this method less suitable for our mocks purposes as stated in [88].

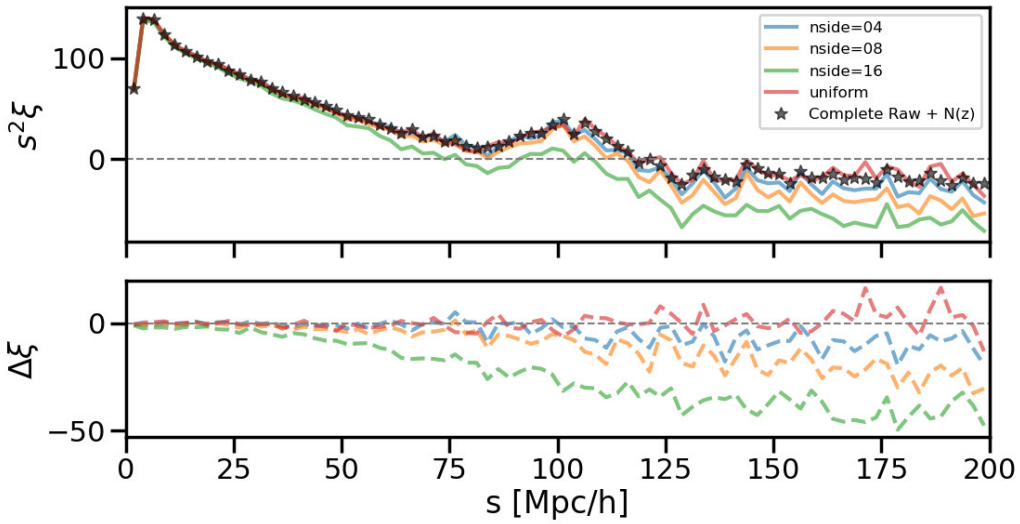

Figure 5.4: Top: QSO 2-point auto correlation as computed by Equation (5.1) for five mock catalogs: a complete raw mock with number density distribution as function of redshift $N(z)$ following the expected DESI redshift distribution (black starts), three mocks following different `nsides` to perform the downsampling on mocks to mirror the EDR+M2 footprint and spatial number density, and one mock where the EDR+M2 footprint is emulated but the number density is uniformly downsampled. Bottom: Differences with respect to complete raw mock with $n(z)$ distribution applied.

5.4 Introducing the DESI-Y1 Lyman-α mocks

The method used to mimic the footprint and number density distribution on the EDR+M2 mocks based on HEALpix pixels was presented as a first approximation method, even thought this method alters the shape of the QSO auto-correlation function depending on the `nside` value used as discussed in Section 5.3. Therefore, a better and more suitable method was needed to perform this task.

In this regard, various methods to properly mirror the spatial distribution of quasars of data were explored to produce the DESI-Y1 mocks. This includes a method to mirror the spatial distribution as a function of the number of tiles (N_{passes}) that have covered a certain region of the sky in observed data. First, we divide the sky region into HEALpix pixels of `nside=1024`, then, for each HEALpix pixel, we count how many tiles have observed that region. The resulting map is shown in Figure 5.5, for each region a maximum amount of seven tiles can overlap in a region as a consequence of the seven pass program discussed in Section 3.3.1. However, the maximum number of observations a Lyman-α quasar can have is four.

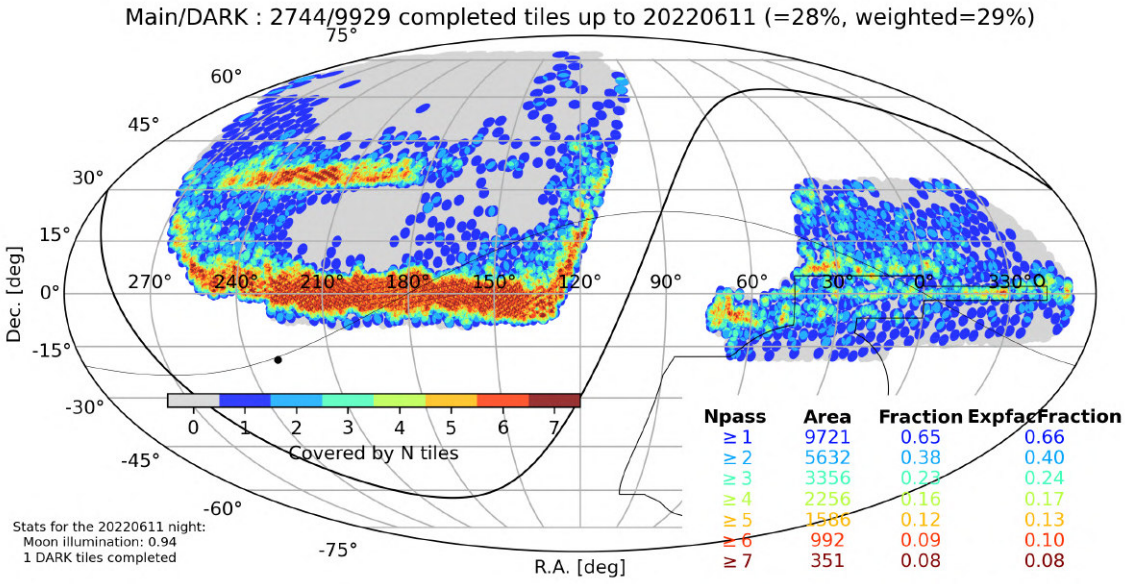

Figure 5.5: Sky map of the DESI-Y1 data sample dark-time tiles footprint. Color displays the number of tiles that overlap in a given region (N_{pass}). Credits: Anand Raichoor.

After the N_{passes} map has been generated we count how many of the observed QSOs N_{data} are in regions observed by 1,2,...,7 tiles and similar for mocks N_{mock}. Then, we use the fraction $N_{\text{data}}/N_{\text{mock}}$ as a function of N_{passes} to perform a down-sampling. The last step of this method is assigning number of observations to our

mock spectra; for this, we compute exposure time probability distribution function as a function of $N_{\rm passes}$ on data. This is done by computing the effective exposure time in the data defined by the spectroscopic data processing pipeline by $T_{\rm eff} = 12.15$ seconds $\times$ TSNR$^2_{\rm LRG}$, where TSNR$_{\rm LRG}$ is the signal to noise ratio of the LRG template [112]. Then, we use this distribution to randomly assign exposures to our mocks as a multiple of 1000 seconds.

Figure 5.6 shows a scatter plot of the positions of the observed (top) and mock (bottom) targets, while the colors show the exposure time of each target. From this figure, there are two clear things: first, with this method we mirror the actual data geometry, unlike with EDR+M2 mocks, and second, the exposure time distribution as a function of $N_{\rm passes}$ is qualitatively reproduced. Thus, at this point, this method gives a better result than the one presented for EDR+M2 mocks.

Regarding the effect of this method on the QSO auto-correlation, Figure 5.7 shows the QSO auto-correlation of the complete raw mock catalog, the result of applying the redshift distribution sub-sampling, and the result when applying the tile coverage (NPASS) sub-sampling method. Notice that the shape of the correlation function is preserved in all cases contrary to the by HEALpix downsampling method. Therefore, we decided to use this method to generate the DESI-Y1 Lyman-α mocks, since it does not show any obvious drawback compared to the previous implementation.

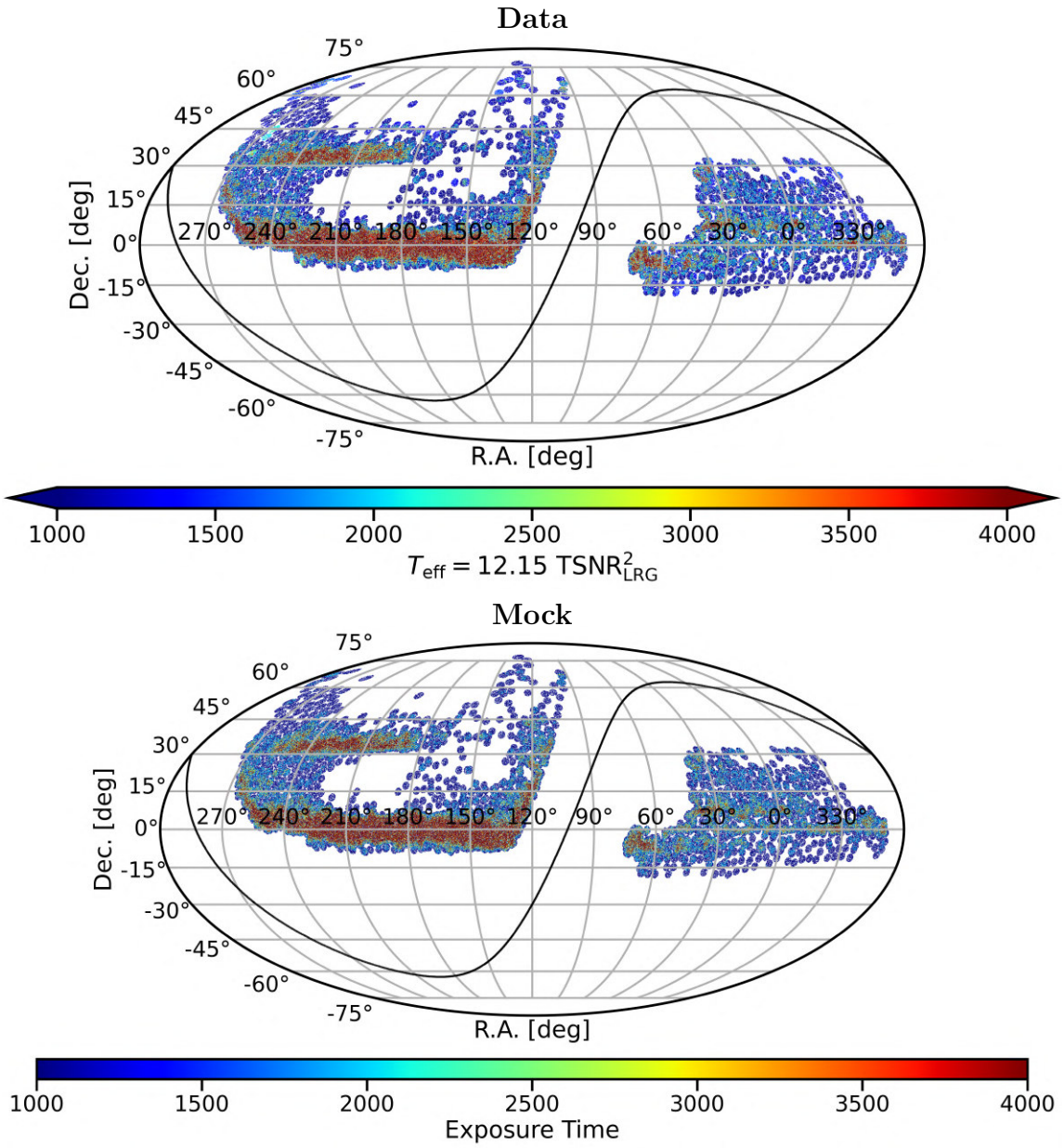

Figure 5.6: Top: DESI-Y1 footprint as observed in data, colors display the exposure time as computed by the effective exposure time $T_{\rm eff} = 12.15$ seconds $\times$ TSNR$^2_{\rm LRG}$. The effective exposure time range has been restricted to match the possible exposure times of our mocks. Bottom: Similar to the top panel, but for mocks, colors display the assigned exposure time as a multiple of 1000 seconds. Figure already presented in [51] as the left panel of Figure 1.

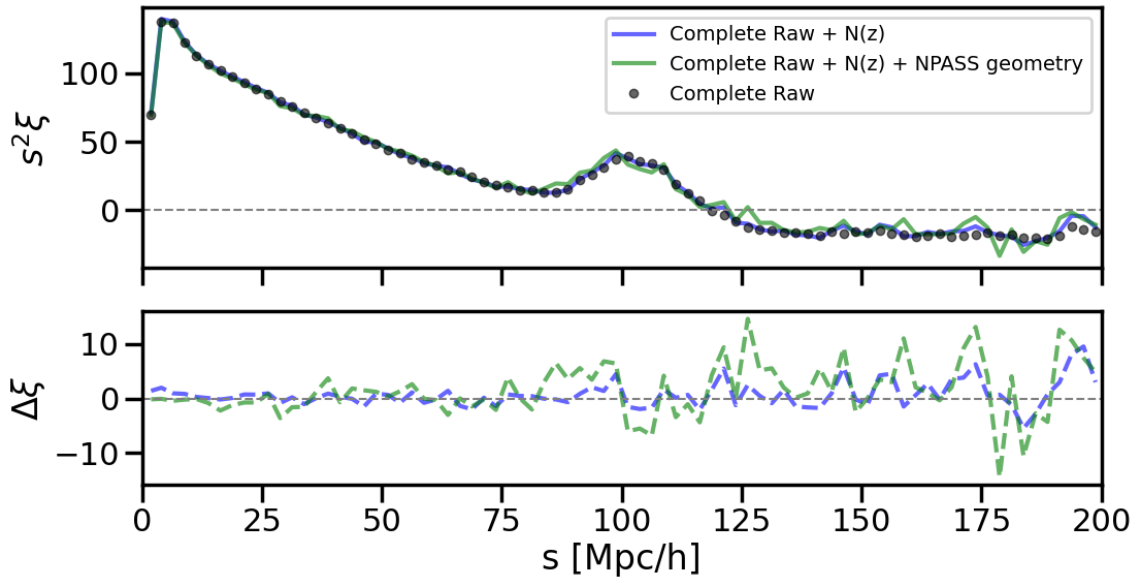

Figure 5.7: Similar to Figure 5.4 but for the downsampling method as a function of number of tiles covering an area (NPASS). Displayed are the QSO auto correlation of the complete raw catalogs without restrictions, when applying the redshift distribution $N(z)$ downsampling and when applying the NPASS geometry downsampling is applied along with the redshift distribution.

5.4.1 Qualitative comparison with observed data

We generated 100 `LyaCoLoRe` and 50 `Saclay` fully contaminated mocks with the usual specifications. This is, we include correlated HCDs, add BALs to 16% of the generated spectra, include metals and add redshift errors following a Gaussian distribution with $\sigma_v = 400$ km/s. For metals we used the `quickquasars` method for both `LyaCoLoRe` and `Saclay` mocks, with small modifications on the C_m coefficients used, this will be further discussed during the last part of this subsection. The main goal of this subsection is to qualitatively compare the results of these mock datasets with respect to observed data and provide an update of the results and conclusions presented for the EDR+M2 mocks in section 4 of [88].

Demographics

The official DESI-Y1 Lyman-α sample resulted in a 683k quasar sample at redshift $z > 1.8$, of which 448.9k are Lyman-α quasars with redshifts above $z > 2.1$ [50], including 6.8k Lyman-α quasars with $z > 3.8$ exceeding the redshift available for our mocks. Similar to the EDR+M2 mocks, we ignore the quasars that fall on regions outside the area covered by both the `LyaCoLoRe` and `Saclay` mocks. To generate the proper number of mock quasars, we also restrict the observed data redshift to the range of the raw mocks $1.8 \leq z \leq 3.8$ for `LyaCoLoRe` and $1.8 \leq z \leq 3.6$ for `Saclay`. This yields to a 630k `LyaCoLoRe` mock quasars sample, of which 409k are Lyman-α and 620k `Saclay` mock quasars, with 401k Lyman-α, these values are consistent across mocks, with small dispersion.

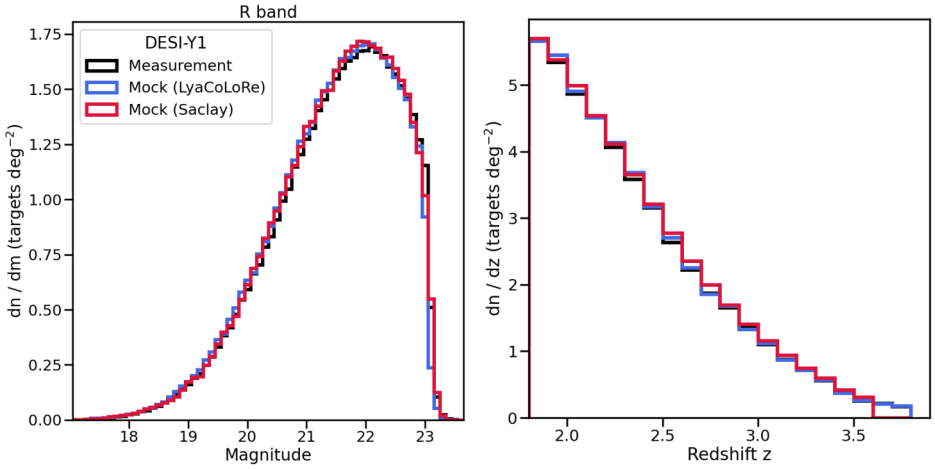

Figure 5.8: R-band magnitude (left) and redshift (right) distributions of the DESI-Y1 observed data compared to one realization of the `LyaCoLoRe` (blue) and one of the `Saclay` (red) mocks.

Figure 5.8 shows the comparison of the r-band magnitude (left) and redshift (right) distributions of one mock realization of both `LyaCoLoRe` and `Saclay` DESI-Y1 mocks and observed data. The observed data redshift has been restricted to $1.8 \leq z \leq 3.8$ to match the redshift range of the `LyaCoLoRe` mocks. From this figure we notice that the distributions are qualitatively reproduced and we conclude that our sub-sampling method does not affect the results nor the conclusions obtained in [88]. Note that the observed data magnitude distribution differs from the one presented in the EDR+M2 case; this is due to the fact that we are analyzing only in the main DESI-Y1 survey. In other words, in this case, we are not including the fainter objects that were studied in EDR.

Figure 5.9 shows a comparison of the median signal-to-noise ratio (SNR) of a 100k randomly selected quasar sample of observed data and one realization of each type of mocks in the Lyman-α forest region A subdivided into redshift and r-band magnitude bins of with 0.5 and 0.1, respectively. As with the EDR+M2 mocks we included in our DESI instrument model a throughput that includes a dip feature at $\lambda \approx 440$ nm (see Section 3.2.2 and Figure 3 of [88] for further details). The results are similar to what was found on EDR+M2 mocks. This is, the SNR of mock low redshift bright quasars seems to be overestimated when compared to the results on observed data. This issue is yet to be studied and goes beyond the scope of this thesis, leaving it as an improvement for future mock realizations of DESI releases such as the third year of DESI observations (DESI-Y3.

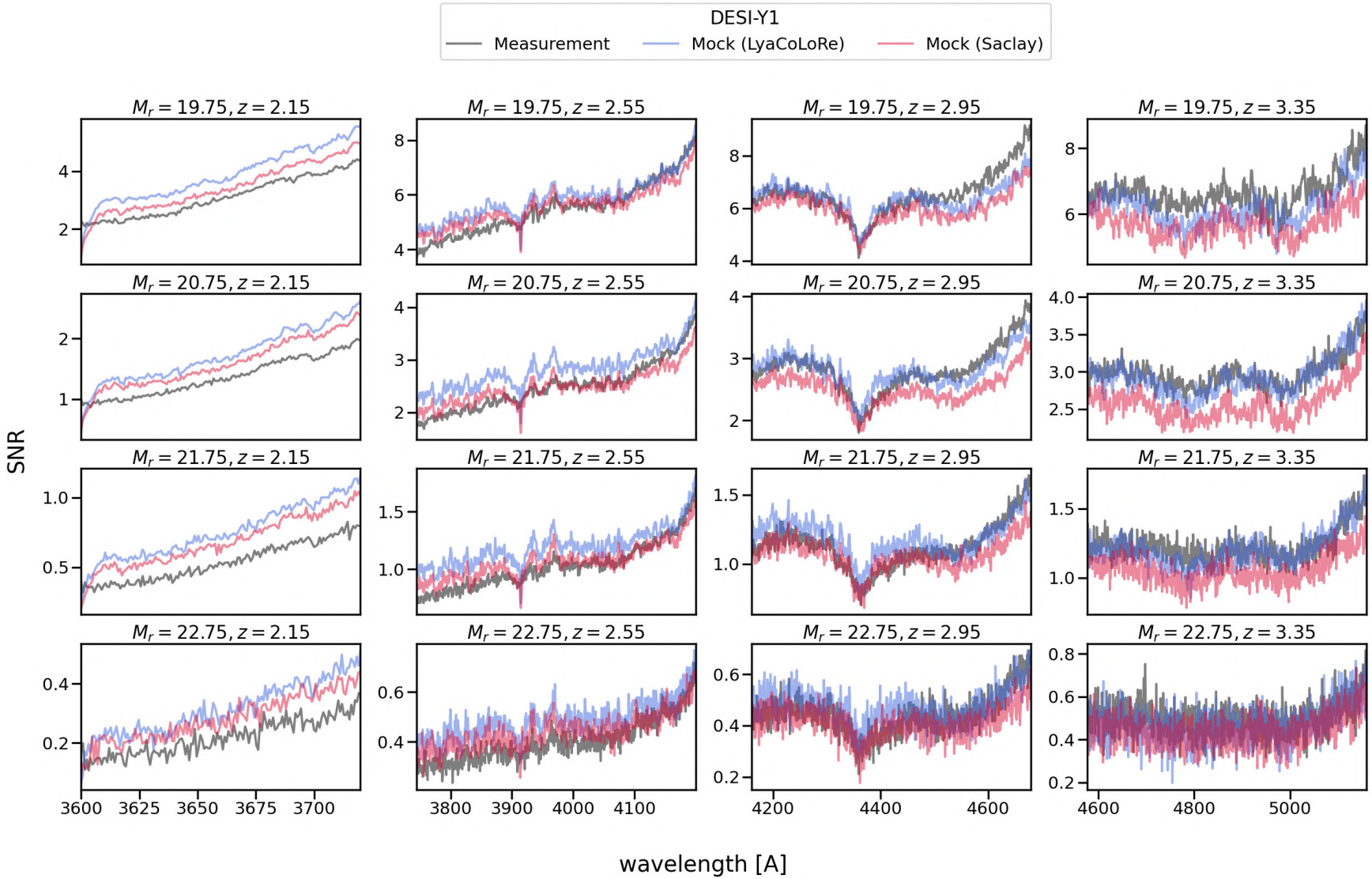

Figure 5.9: Median value of the signal to noise ratio (SNR) in the Lyα(A) region of a 100k QSOs randomly selected from the observed and mock samples. Each panel shows the results of dividing the sample into redshift bins of width 0.5 and magnitude bins of width 0.1. The numbers above each panel represent the central values of the each bin.

Astrophysical Contaminants

The HCD value added catalog of DESI-Y1 is presented in [137], this catalog contains HCDs detected by two algorithms, one based on a Convolutional Neural Network (CNN, [138, 139]), and another on a Gaussian Process (GP, [140]) model. The observed data Lyman-α BAO analysis restricted the sample of this catalog to those DLA systems with $\log N_{\rm HI} > 20.3$ cm^{-2} confirmed by the two detection algorithms with at least 50% probability and signal-to-noise ratio $S/N > 3$ to ensure purity on the sample, this yields in 18.3k DLA sample that are masked in the observed data analysis. For mocks, we included HCDs correlated with the density field used to generate the raw transmissions and that follow the `pyigm` [131] column density distribution, this is we includes system with densities $\log N_{HI} > 17.2$ cm^{-2}. This results in 58k mock DLA systems with column density $\log N_{\rm HI} > 20.3$ cm^{-2} for `LyaCoLoRe` mocks and 53k for `Saclay` mocks. The difference on the number of DLAs on the samples are due to the strict restrictions imposed to the observed data catalog.

In the case of BALs, the DESI-DR1 value added quasar catalog contains 114k BAL quasars with AI > 0 found by the BAL finder algorithm [135, 141]. For mocks, we find 100k BAL quasars with AI > 0, the differences are simply for the 1% difference on the fraction of BAL quasars on data (16.7% with respect to the total sample) and the fraction of AI¿0 BAL quasars introduced in mocks (15.8% of the total sample). The BAL templates used to include these systems into mock spectra were updated to include features associated with Si IV(1062), Si IV(1074), P V(1118), P V(1128), and C III* (1175) additional to the O VI(1031), O VI(1037), Lyman-α, N V(1240), Si IV(1398), and C IV(1549) features from the previous templates version.

Figure 5.10 shows the comparison of the DLA (left panel) and BAL (right panel) distributions of our mocks and the observed data catalogs used to perform the DESI-DR1 Lyman-α analysis [50]. In the DLA column density distribution panel we show the results of one randomly selected mock realization of both `LyaCoLoRe` and `Saclay` mocks. In the case of the right panel displaying the log AI we only display the results of one mock realization. In this case, there is no need for a distinction between `LyaCoLoRe` and `Saclay` given that they use the same templates to include BALs in mock spectra and therefore follow the same distribution.

We obtain similar results as in the EDR+M2 mocks. The DLA normalized distributions present good agreement, however the number of objects on the observed data case contains 3.2 times less DLAs than mocks, this is due to the restrictions imposed to the observed DLA catalog as discussed earlier on this section. In the

case of BALs, we find a discrepancy between the AI distribution shape, as discussed in [141] and [88] this might be due to the differences between the SNR of DESI and the eBOSS DR14 dataset used to generate the templates. A possible improvement for future DESI mocks is to use DESI-DR1 data to produce new BAL templates to be included on spectra, this is left as a future work.

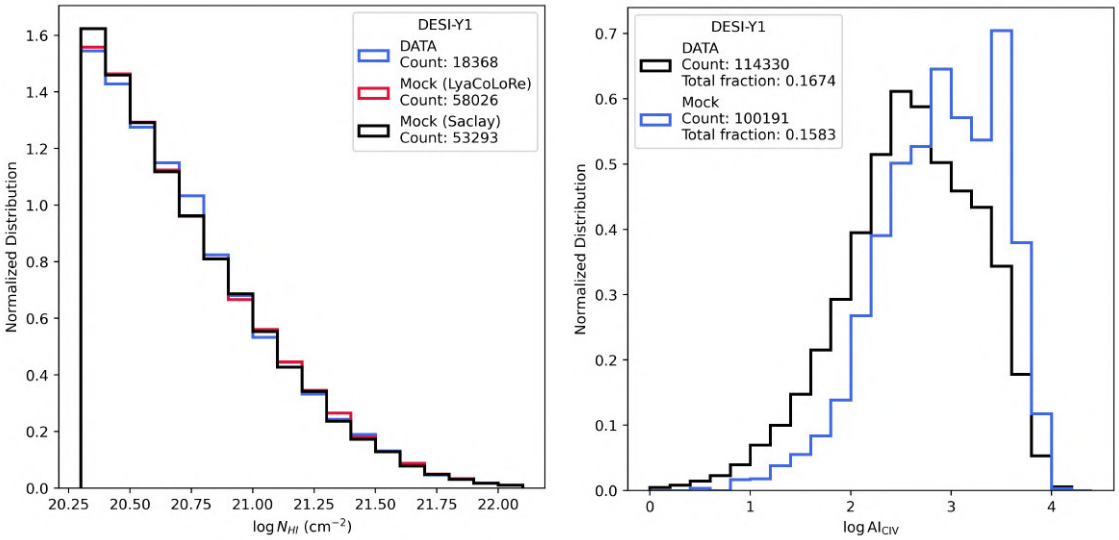

Figure 5.10: Left: Distribution of the DLA column density N_{HI} of the DESI-Y1 observed data as obtained by the DLA finder algorithms with 50% probability and $S/N > 3$, and one realization of both `LyaCoLoRe` and `Saclay` mocks. Right: Distribution of the DESI-Y1 $\mathrm{AI_{CIV}} > 0$ BAL quasars found by the BAL finder algorithm [135, 141] and one mock realization. All the distributions are normalized so their cumulative sum is equal to 1.

In the case of metals, we found that the coefficient used for the EDR+M2 mocks resulted in metal biases b_m that were statistically inconsistent with those of the DESI-Y1 observed data for both available methods to include metals into mock spectra. As stated in [88], in the `quickquasars` method the C_m coefficients and the biases follow a linear relation. Therefore, we used this method to find proper coefficients in order to produce mocks which result consistent with observed data. This tuning procedure consisted in finding the linear relation of each metal by using the results obtained by the previous coefficients and the results of re-scaling the coefficients by the ratio between the obtained biases and the results of observed data. Once the linear relations were obtained we interpolated in order to find the coefficients required to reproduce the DESI-Y1 results. The obtained values for `LyaCoLoRe` and `Saclay` mocks are reported in Table 5.1, these were used to produce the 100 `LyaCoLoRe` and 50 `Saclay` mocks used for the DESI-Y1 Lyman-α analysis validation [50]. The results for one mock realization of each type will be discussed during Section 5.4.2,

the bias values found for all the generated mocks are statistically consistent with observed data and therefore, the conclusions in such section are valid for the rest of the mocks.

Table 5.1: Coefficient C_m values of Si III(1207), Si II(1190), Si II(1193), and Si II(1260) used to produce the DESI-Y1 Lyman-α `LyaCoLoRe` and `Saclay` mocks.

	`LyaCoLoRe`	`Saclay`
Transition	$\mathbf{C_m}$	$\mathbf{C_m}$
Si II(1190)	1.4×10^{-3}	6.8×10^{-4}
Si II(1193)	0.7×10^{-3}	5.3×10^{-4}
Si II(1260)	1.3×10^{-3}	5.7×10^{-4}
Si III(1207)	3.5×10^{-3}	1.6×10^{-3}

Correlation Functions

To measure the correlation functions we follow the same methodology described in Section 2.2.1 and use the `picca` package `picca`[1][142].

We compute the flux-transmission field δ_q in the Lyman-α region A (Lyα(A)) defined in the rest-frame wavelength range 1040 Å $< \lambda <$ 1205 Å in the spectra of quasars in the 3600 Å $< \lambda <$ 5772 Å observed wavelength range. In this analysis, we also include the Lyman-α region B (Lyα(B)) defined in the rest-frame wavelength range 920 Å $< \lambda <$ 1020 Å in the same observed wavelength ranges as the Lyα(A) region analysis. We also require a minimum of 150 pixels in the forest of 0.8 Å width pixels for them to be accepted as part of the analysis. All of the aforementioned specifications are similar to the observed data analysis for a direct comparison. Additionally, we mask all HCDs with $\log N_{\rm HI} > 20.3$ and all the BAL features in our mock spectra. The remaining HCDs with $\log N_{\rm HI} < 20.3$ are assumed to be undetected in our mocks and kept unmasked.

The observed data analysis results in the analysis of 428k accepted forests for the Lyα(A) region and 137k for the B region. In the case of mocks, the `LyaCoLoRe` mock datasets result in an accepted forests mean value of 403k for the Lyman-α A region and 129k for the B region. The `Saclay` mocks result in an accepted forests mean value of 394k for the Lyα(A) region and 124k for Lyα(B) region. There is small dispersion across the values of the 100 `LyaCoLoRe` and 50 `Saclay` mocks. Differences with respect to observed data could be due to the signal to noise ratio of the mocks in the case of the Lyα(A) region and due to the observed data having more objects

[1] https://github.com/igmhub/picca/

at higher redshift reaching redshift $z = 4.55$, while our mocks only reach $z = 3.8$ for `LyaCoLoRe` and $z = 3.6$ for `Saclay`.

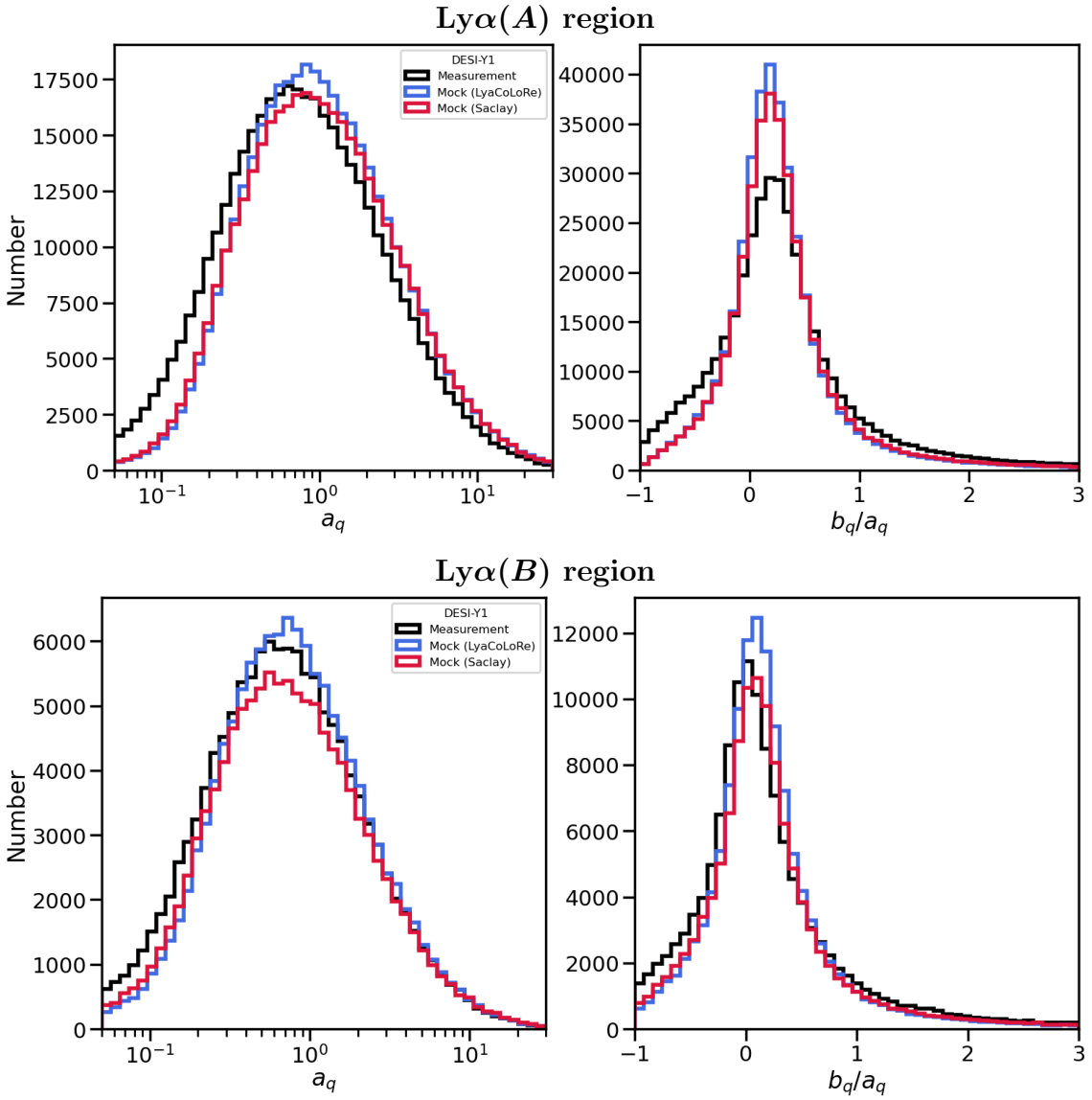

Figure 5.11: Quasar spectral diversity parameters a_q and b_q/a_q distributions. Top and bottom panels show the results for the Lyman-α absorption within the Lyα(A) (1040 Å $< \lambda <$ 1200 Å) and Lyα(B) (920 Å $< \lambda <$ 1020 Å) regions.

Figure 5.11 shows the comparison of the distributions of the quasar spectral diversity parameters a_q and b_q/a_q as measured from data and one mock realization of each type on the Lyα(A) (top panel) and Lyα(B) (bottom panel) regions. We have restricted to those objects whose amplitude parameter is $a_q > 0$. There is a discrepancy in the shape of the b_q/a_q distribution on the Lyα(A) region similar to what was found in [88], which is expected given that we did not update the

continuum template generation method for these mocks, on the Lyα(B) region the difference is smaller. Studying this issue and refining the template methods requires modifying the model slopes and dispersion values of the `SIMQSO` method discussed in Section 4.2 and go beyond the scope of this thesis. We might retake this issue for future mocks, however an alternative is to use the `QSO` method with updated PCA templates generated from the observed DESI-Y1 spectra.

Figure 5.12 shows the comparison between observed data and the results of 100 `LyaCoLoRe` and 50 `Saclay` mock realizations mean flux at rest-frame $\bar{C}_q(\lambda_{\rm RF})$ (left panels) and large-scale structure variance $\sigma_{\rm LSS}$ (right panels) of the Lyman-α absorption within the Lyα(A) (top panels) and Lyα(B) (bottom panels) regions. The mean flux of the Lyman-α qualitatively reproduce the results from observed data, similar to what was found in [88]. In the case of the Lyα(B) region, the mean flux also presents an overall qualitative agreement with a discrepancy at $\lambda_{\rm R.F.} < 940$ Å caused by Lyman-limit systems present in observed data which are not modeled by the `SIMQSO` continuum template generation method, introducing these systems into our template model is left as a future study. This issue is not expected to be present in the `QSO` method which will be studied for future mocks. In the case $\sigma_{\rm LSS}$ we also obtain a slight difference on mocks with respect to observed data for both analyzed regions, this is discussed in [88] and might be due to the DLA and BAL masking procedure, which assumes a 100% efficiency for mocks, while observed data depends on the performance of the DLA and BAL finder algorithms. The effects of these systems of the measured $\sigma_{\rm LSS}$ values are studied with DESI-Y5 mocks on Section 6.1. The $\sigma_{\rm LSS}$ of mocks on the Lyα(B) region drops to zero as we reach the mocks redshift upper limit ($z = 3.8$ for `LyaCoLoRe` and $z = 3.6$ for `Saclay` and therefore there are no objects whose Lyα(B) region is above the corresponding wavelengths.

Given that we have two analysis regions, we compute two Lyman-α auto correlations, Ly$\alpha(A) \times$ Ly$\alpha(A)$ and Ly$\alpha(A) \times$ Ly$\alpha(B)$, by the estimators defined in Equation (2.24) with weights defined by the values $\gamma_{\rm Ly\alpha} = 2.9$ and $\sigma_{\rm mod} = 7.5$ in Equation (2.26) for the Lyman-α flux-transmission field pixels. Similarly, we compute two Lyman-α – QSO cross-correlations, Ly$\alpha(A) \times$ QSO and Ly$\alpha(B) \times$ QSO from Equation (2.25) with weights defined by $\gamma_{\rm QSO} = 1.44$ in Equation (2.27) for the QSO weights. We chose these values to match the ones used in the observed data analysis.

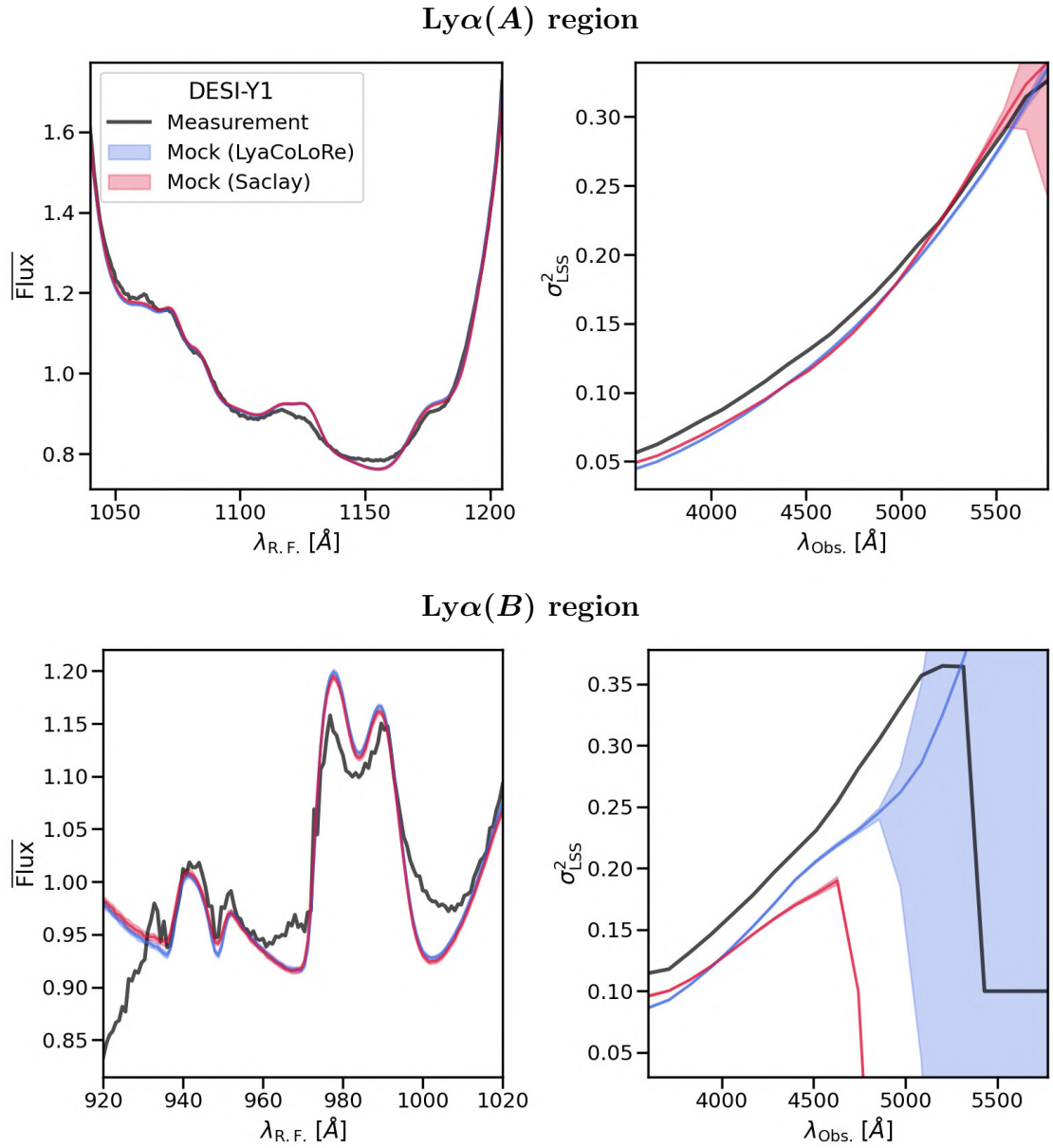

Figure 5.12: Comparison of the results on observed data and 50 realizations of `LyaCoLoRe` and `Saclay` mocks on the mean flux at rest frame $\overline{F}(\lambda)C_q(\lambda_{\rm RF})$ (left panels) and the large-scale structure variance $\sigma_{\rm LSS}$ (right panel) obtained for the Lyα(A) region (top panels) defined in the range 1040 Å $< \lambda <$ 1200 Å and the Lyα(B) region (bottom panels) defined in the range 920 Å $< \lambda <$ 1020 Å. Solid line shows median value of our mock realizations while colored band shows the 1-σ percentiles.

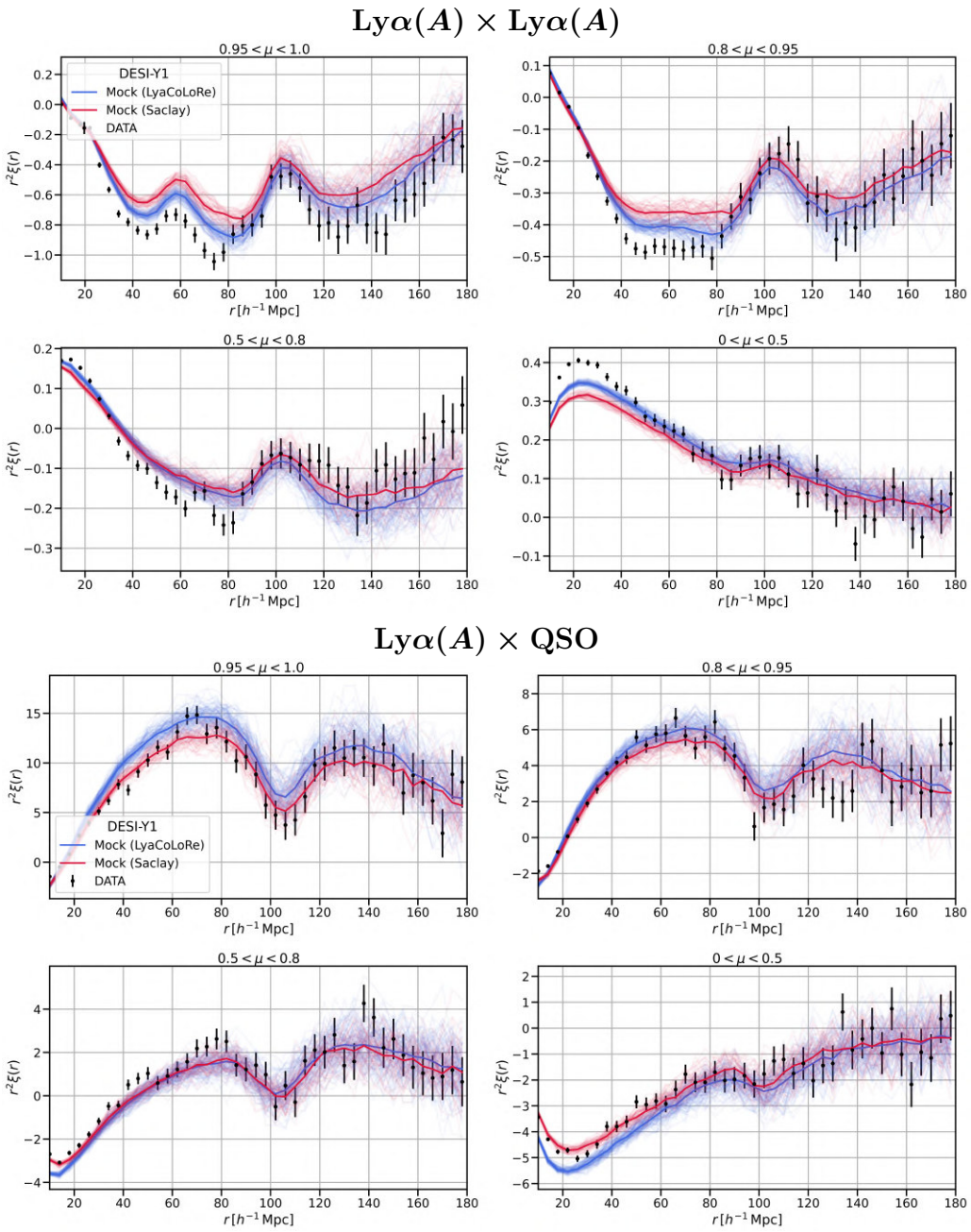

Figure 5.13: Lyman-α auto (top) and Lyman-α – QSO cross (bottom) correlation functions computed using Lyman-α absorption features in the A region. The correlations are multiplied by r^2 and are presented as averages in four ranges of $\mu = r_{\parallel}/r$ (four panels). The transparent red and blue lines show individual correlations of 50 mocks for each `LyaCoLoRe` and `Saclay` mocks, while the solid line shows the median values. Black dots are the results on observed data as obtained by [50].

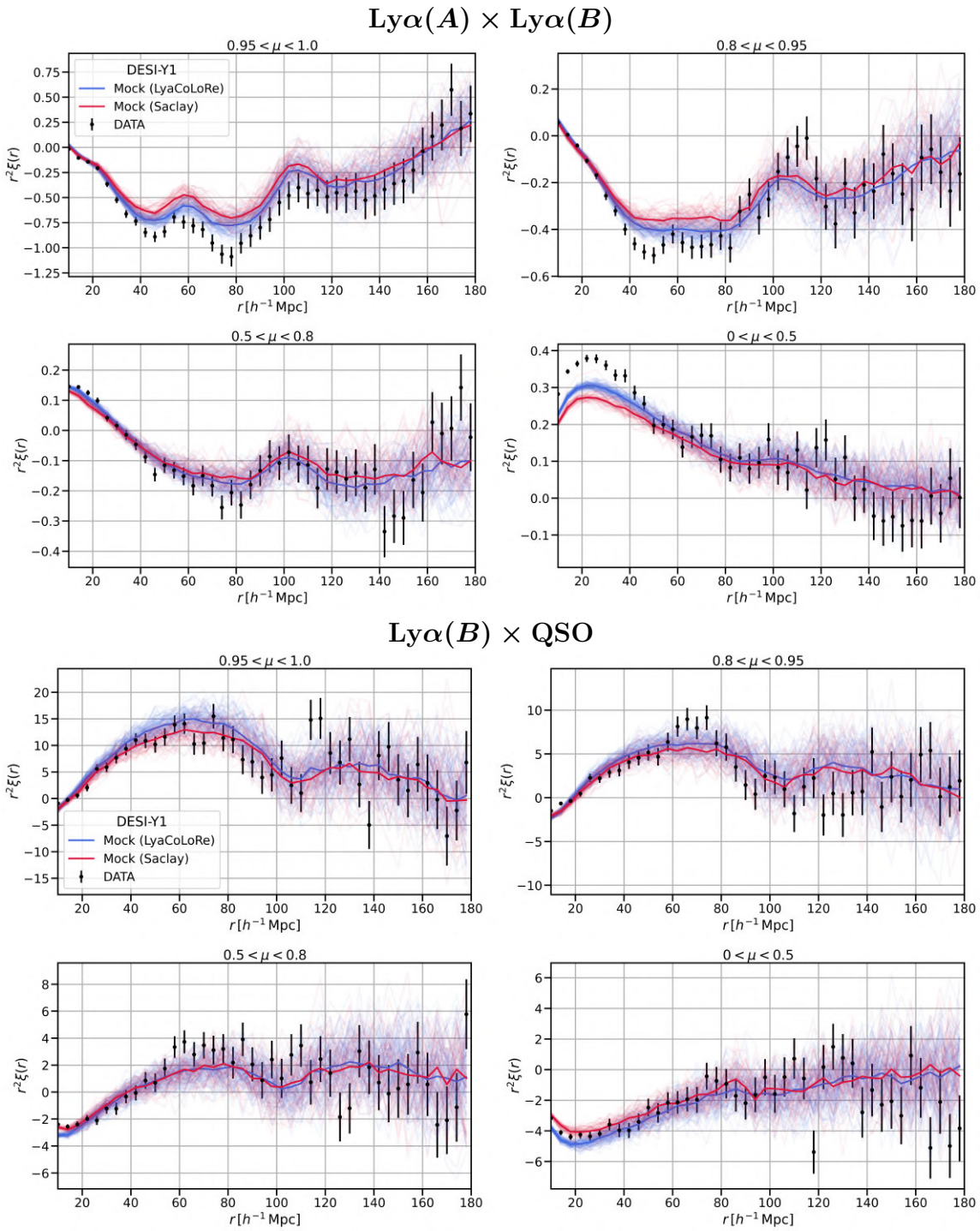

Figure 5.14: Similar to Figure 5.13 but using Lyman-α absorption features in the B region.

Figures 5.13 and 5.14 show the Lyman-α auto (top panels) and Lyman-α – QSO cross correlations (bottom panels) averaged in four $\mu = r_{\parallel}/r$ regions computed using Lyman-α absorption features in the Lyman-α A and B regions, respectively. This figure displays the results on observed data and individual results of 100 `LyaCoLoRe` and 50 `Saclay` mocks as transparent lines and their median value as solid line. The dispersion on mocks is enough to be statistically consistent with the results of the observed data with some discrepancies. As in the EDR+M2 mocks there are some discrepancies on the correlation functions obtained on mocks and observed data. As discussed in Section 4.2 of [88], these discrepancies might be caused by several reasons. The first of them, BAL and DLA masking which in the observed case, relies on the efficiency of the BAL and DLA finder algorithms and the catalog used to mask them, while mocks mask all these features. Second, the instrumental noise model included on mock spectra that alters the shape of the correlations. Other possible contributions are the Lyman-α and quasar biases used to produce the raw mock transmissions, the relative absorption coefficients used to include metals on mocks, to mention some. The effects of some of these possible sources of discrepancy on the shape of the correlations are further studied in Section 6.1 with DESI-Y5 mocks.

5.4.2 Best-fit Model

Additionally to the qualitative comparison carried out throughout the previous subsection, an additional task to do is to compare the results of the best-fit model on the observed data with the results obtained by the correlations measured in one mock realization of each `LyaCoLoRe` and `Saclay` mocks DESI-Y1 mocks when performing a joint-fit combining the auto and cross correlations of the A and B Lyman-α regions. This section does not attempt to perform an analysis pipeline validation which should require a large number of mocks, but rather attempts to serve as a guide on the what parameters should be set free to perform a model fitting on mocks and compare the results with observed data. The uses of mocks as a validation step are further discussed in Section 6.3.1. A complete validation of the DESI-Y1 Lyman-α BAO analysis with mocks is presented in [51].

The fits in this section are performed using `vega`[2] package [143] as in the case of observed data. We follow the same methodology as observed data that makes use of the complete covariance matrix of all the Lyα(A) and Lyα(B) auto and cross correlations. We exclude QSO radiation effects and instrumental systematics from

[2] https://github.com/andreicuceu/vega

the mocks model since these were not included in our simulations.

The correlation function model of the mocks also includes a Gaussian anisotropic smoothing that accounts for the low resolution of the raw mocks given by

$$P_{\text{smoothed}}(k,z) = \exp\left[-(\sigma_{\parallel}^2 k_{\parallel}^2 + \sigma_{\perp}^2 k_{\perp}^2)\right] P(k,z), \tag{5.2}$$

where $\sigma_{\parallel}$ and $\sigma_{\perp}$ describe the smoothing along and across the line of sight, respectively. This smoothing is done on both the Lyman-α and metal correlation models. For the fits performed in this section these values are fixed to the values found when performing a fit on the stack of a large set of DESI-Y1 mocks of each type: $\sigma_{\parallel} = 1.83$ Mpc/h, $\sigma_{\perp} = 2.12$ Mpc/h found on a stack 100 `LyaCoLoRe` mocks, and $\sigma_{\parallel} = 2.01$ Mpc/h, $\sigma_{\perp} = 1.91$ Mpc/h found in a stack of 50 `Saclay`, see [51] for further details.

Table 5.2 shows the best-fit model parameters on one randomly selected mock realization of `LyaCoLoRe` and another of `Saclay` compared to data (last column). The second column displays the priors set for mocks, where U and $\mathcal{N}$ denote flat and Gaussian priors, respectively. With the same free parameters and priors, those parameters not present in this table are fixed to the values mentioned during the description of the model in Section 2.2.2. We fit the model on the region between $r = 10$ Mpc/h and $r = 180$ Mpc/h of the correlation functions, and set the results obtained by Planck in 2015 [144] as fiducial cosmology , this in order to match the cosmology used to generate the raw mocks that are input to our mock spectra.

First, we will focus on the statistical consistency by comparing the resulting central values of the parameters and their confidence intervals. Note that the following conclusions are for these particular mocks only and are not general; a complete study requires using a large set of mocks and is left for future work.

- The BAO $\alpha_{\parallel}$ and $\alpha_{\perp}$ parameters results on mocks are statistically consistent with the expected values $\alpha_{\parallel} = 1$ and $\alpha_{\perp} = 1$.

- We obtain results slightly different with respect to the Lyman-α forest bias and RSD parameters $b_{Ly\alpha}$, $\beta_{Ly\alpha}$ expected values of the raw mocks: $b_{Ly\alpha} = -0.121$ and $\beta_{Ly\alpha} = 1.624$ for `LyaCoLoRe` [124] and $b_{Ly\alpha} = -0.1171$ and $\beta_{Ly\alpha} = 1.669$ [125] for `Saclay`. The value of the quasar bias b_{QSO} is consistent with the expected value $b_{QSO} = 3.7$ in the case of `LyaCoLoRe`, while the `Saclay` mock presents a discrepancy, however it is consistent with the DESI-DR1 value. The reason of the aforementioned discrepancies with respect to the expected values and the values obtained on DESI-DR1 require further study and are left for

Table 5.2: Best-fit model obtained when combining the results on the auto and cross correlations of the Lyα(A) and Lyα(B) regions for one mock realization of each `LyaCoLoRe` (third column) and `Saclay` (fourth column) mocks compared to the results of the best fit model on observed data (fifth column) as reported by [50]. The second column displays the priors used to perform the fits on mocks. Here $U(\min, \max)$ denotes flat priors defined in the [min,max] interval and $\mathcal{N}(\mu, \sigma)$ represents Gaussian priors.

Parameter	**Prior**	`LyaCoLoRe`	`Saclay`	**Observed Data**
$\alpha_{\parallel}$	$U(0.01, 2.0)$	1.004 ± 0.018	1.004 ± 0.017	0.988 ± 0.020
$\alpha_{\perp}$	$U(0.01, 2.0)$	1.012 ± 0.021	0.996 ± 0.019	1.013 ± 0.023
$b_{Ly\alpha}$	$U(-2.0, 0.0)$	-0.125 ± 0.012	-0.130 ± 0.002	-0.109 ± 0.005
$\beta_{Ly\alpha}$	$U(0.0, 5.0)$	1.495 ± 0.104	1.466 ± 0.042	1.726 ± 0.085
b_{QSO}	$U(0.0, 10.0)$	3.670 ± 0.024	3.403 ± 0.026	3.411 ± 0.048
$\sigma_{v,\mathrm{QSO}}(h^{-1}\mathrm{Mpc})$	$U(0.0, 15.0)$	4.563 ± 0.180	6.287 ± 0.141	3.665 ± 0.142
$\Delta r_{\parallel,QSO}(h^{-1}\mathrm{Mpc})$	$U(-3.0, 3.0)$	-0.232 ± 0.049	-0.102 ± 0.058	0.067 ± 0.059
b_{HCD}	$U(-0.2, 0.0)$	-0.027 ± 0.010	-0.017 ± 0.005	-0.056 ± 0.004
β_{HCD}	$\mathcal{N}(0.5, 0.09)$	0.497 ± 0.089	0.514 ± 0.089	0.629 ± 0.080
$L_{0,\mathrm{HCD}}$	$U(0.0, 40.0)$	6.005 ± 3.655	33.915 ± 10.355	6.541 ± 0.894
$10^3 b_{\mathrm{SiII}(1190)}$	$U(-20.0, 0.0)$	-5.731 ± 0.646	-6.514 ± 0.669	-4.500 ± 0.639
$10^3 b_{\mathrm{SiII}(1193)}$	$U(-20.0, 0.0)$	-1.987 ± 0.588	-1.447 ± 0.680	-3.052 ± 0.624
$10^3 b_{\mathrm{SiIII}(1207)}$	$U(-20.0, 0.0)$	-7.682 ± 0.787	-7.560 ± 0.753	-9.765 ± 0.677
$10^3 b_{\mathrm{SiII}(1260)}$	$U(-20.0, 0.0)$	-3.232 ± 0.563	-4.153 ± 0.578	-4.001 ± 0.625

future mock realizations. Nevertheless the values obtained on the observed data case indicate that these values might require a re-tuning for future generations of raw mock realizations.

- Since we introduced a systematic error of $\sigma_v = 400$ km/s on our mocks, we would expect to measure a statistical redshift error parameter close to $\sigma_{v,\mathrm{QSO}} = 4$ Mpc/h, which is the case for `LyaCoLoRe`, but not for `Saclay`, however these results have contributions from a $\sigma_{\mathrm{FoG}} = 150$ km/s redshift shift introduced in our mocks prior the continuum template generation to emulate the Fingers of Good effect, as discussed in Section 4 of [145]. These values could be improved by introducing a more realistic redshift error distribution model.

- In the case of the systematic redshift error parameter $\Delta r_{\parallel,QSO}$, we would expect to measure a value close to zero since we did not introduce this kind of errors in mocks. While the `Saclay` mock present a result consistent with zero and other studies using these kind of mocks [146] as well with observed data, the `LyaCoLoRe` mock has a non-zero consistent result, this is a known issue with the `LyaCoLoRe` mocks which have a dipole on the QSO correlation functions that cause this problem when an inhomogeneus footprint is introduced. This will be fixed for future generations of these mocks.

- The HCD bias b_{HCD}, RSD β_{HCD} and typical length scale $L_{0,\mathrm{HCD}}$ present values that disagree with those of the observed data case. Although, the recovered value of the RSD parameter agrees with the expected $\beta_{HCD} = 0.5$ set as an input on the raw mocks generation method for HCDs, which suggests that these values should be tuned. Studying these discrepancies with observed data requires HCD dedicated mocks and are left as future projects.

- The biases measured for metals are in general statistically consistent with observed data which suggests that the tuning procedure described in Section 5.4.1 was correctly performed. The only exception to this is the Si (1193) metal bias which is consistent between the `LyaCoLoRe` and the `Saclay` mock, suggesting that this metal should be further tuned, however this task is complicated given this metal is highly correlated with the Si (1190) line.

At last, focusing only on the amplitude of the uncertainties obtained, the results are overall consistent between mocks and observed data, which might indicate that the covariance matrix of mocks is representative of data. However, to fully confirm this, it is simpler to compare the uncertainties through a forecast, this will be discussed in Section 6.2.1.

6

Examples of the usefulness of Lyman-α mocks.

Simulations provide a controlled environment in which the cosmological and physical properties of the data are known accurately. Therefore, applying an analysis pipeline on mock datasets should result in the recovery of these properties and in an accurate measurement of the related parameters. This means that mocks serve as an invaluable tool for characterizing systematic errors, performing analysis pipeline validations, and testing the accuracy of some algorithms like the redshift estimation algorithm, and the BAL and DLA finder algorithms, to mention a few.

The main objective of this chapter is to illustrate the usefulness of Lyman-α mocks through examples, serving as a guide for conducting various studies and tests. To do so, we will explore the effect of systematics on the correlation functions measurement procedure, review the use of mocks as a validation step on some previous BAO studies, focusing particularly on the validation of the DESI-DR1 Lyman-α analysis (reported in [51]). Furthermore, we will review the use of mocks as a forecast tool and update the DESI-Y5 BAO uncertainty forecast presented in [88] using the most recent analysis pipeline, validating these results by performing a forecast with DESI-Y1 mocks and compare it with the observed results. At last, we will review other studies where DESI Lyman-α mocks have been useful.

6.1 Characterization of the effects of systematics

In this section, we are interested in the use of mocks for studying the effects of systematics on the shape of the correlation functions and the BAO scale measurements. However, we will not attempt a complete characterization of each individual systematic, which requires the generation of a large set of mocks dedicated to each of them individually and the complete analysis on each realization or the stack of all of them. This is computing time and memory consuming and goes beyond the scope of this thesis.

Instead, this section attempts to serve as an example for those unfamiliar with

the effects of the systematics on the Lyman-α forest correlation functions, and also as a starting point if a particular characterization is to be performed on future mocks. For this reason, we will simply explore their effects on the correlation functions of one specific `LyaCoLoRe` realization of the complete DESI survey mocks presented in Chapter 5. That means that the conclusions obtained in this section are for that specific mock only, and therefore, should not be considered as a comprehensive characterization of systematics. The reason to use `LyaCoLoRe` and not `Saclay` is based primarily on their wider redshift range, which provides more extensive data coverage. For simplicity, we will focus only on the Lyα(A) region. The effects of systematics on the Lyα(B) region are similar.

Some systematics have already been characterized with mocks or observed data; in such cases, the corresponding references and relevant sections will be cited.

6.1.1 Continua Fitting

The first systematic we will focus on is the continua fitting procedure. As discussed in Section 2.2.1, during the estimation of the flux-transmission field δ_q, the continuum parameters a_q and b_q of each individual spectrum are estimated in an iterative fitting procedure, this continuum fitting procedure biases the measured flux-transmission field towards zero, and thus produces a distortion of the correlation functions. There are two natural questions in this context. First, how precise is the method in recovering the actual quasar continua. Second, how significant is this distortion on the correlation functions. In other words, what are the differences between the distorted and the undistorted correlations.

We address the first question in Figure 6.1, which shows a comparison of the true and fitted continua a randomly selected mock quasar spectrum with a redshift of $z = 3.12$. The overall shape of the true continuum is recovered by the baseline fitting procedure of `picca`. It is important to remember that the algorithm fits the mean flux $C_q(\lambda)\overline{F}(z)$, and therefore the result is shifted with respect to the true continuum, which is part of the causes of the correlation functions distortion that we have mentioned.

Regarding the second question, we study this by measuring the correlation functions in the usual way (fitting the mean flux), by measuring the correlations using the continuum templates generated on mocks (referred to as the true continuum) and the correlation functions from the raw transmissions, without noise, contaminants, or continuum templates added. Figure 6.2 shows the results of the auto (top) and cross (bottom) correlation functions of the aforementioned cases. The differences

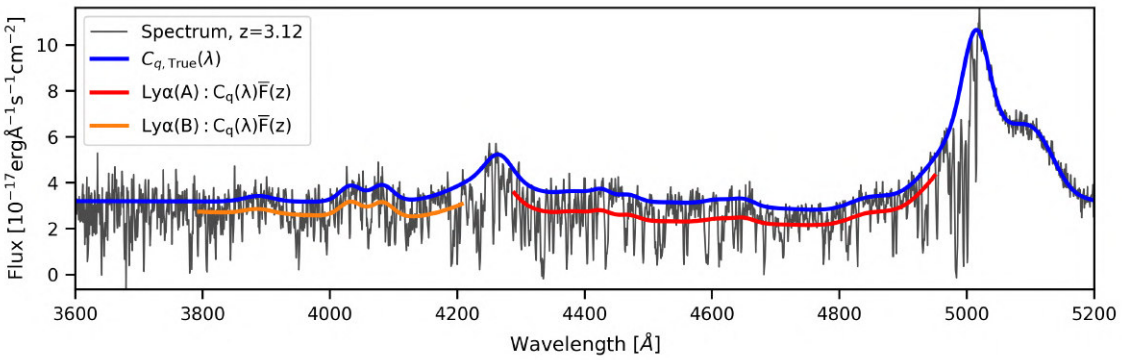

Figure 6.1: Spectrum of a redshift $z = 3.12$ mock quasar. We display the true continuum $C_{q,\mathrm{True}}$ as generated by `quickquasars` (blue line) and the mean flux ($C_q(\lambda)\overline{F}(z)$) as obtained by `picca` in the Lyα(A) (red line) and Lyα(B) (orange line) regions.

between the results of the true continuum analysis and those of the raw transmission analysis are due to the instrumental noise added to our mock spectra and will be discussed in more detail during Section 6.1.4.

The distortion of the shape of the correlation functions produced by the continuum fitting, seen as black dots in Figure 6.2, is accounted for by the distortion matrix defined in Equations (2.56) and (2.57). The inclusion of the distortion matrix on the model of the correlation functions was studied in the DR12 [58, 59] and DR16 [62] analyses, and was found not to induce biases on the BAO scale parameters measurements.

6.1.2 Astrophysical contaminants

Next, we focus on systematics from astrophysical contaminants: HCDs, BALs, and metals. First, we will focus on their effect on the mean flux and σ_{LSS} measured from `picca` and thus on the measurement of the correlation functions and the estimation of the covariance matrix.

Figure 6.3 displays the results obtained when the analysis is performed on contaminated mocks with respect to the uncontaminated case. For BALs and HCDs, the results with and without masking these systems prior measuring the flux-transmission field are shown as solid and dashed lines, respectively. HCDs also include the case where only mask DLAs (dotted lines) with column density $\log N_{\mathrm{HI}} > 20.3$ cm^{-2} as in the observed data case. In the case of the mean flux, the contaminant that most affects the results are the BALs when not masked, the rest of the contaminants have a negligible impact (below the 1% level) on the mean flux.

In contrast, the results on σ_{LSS} are mostly affected by HCDs, with a percentage

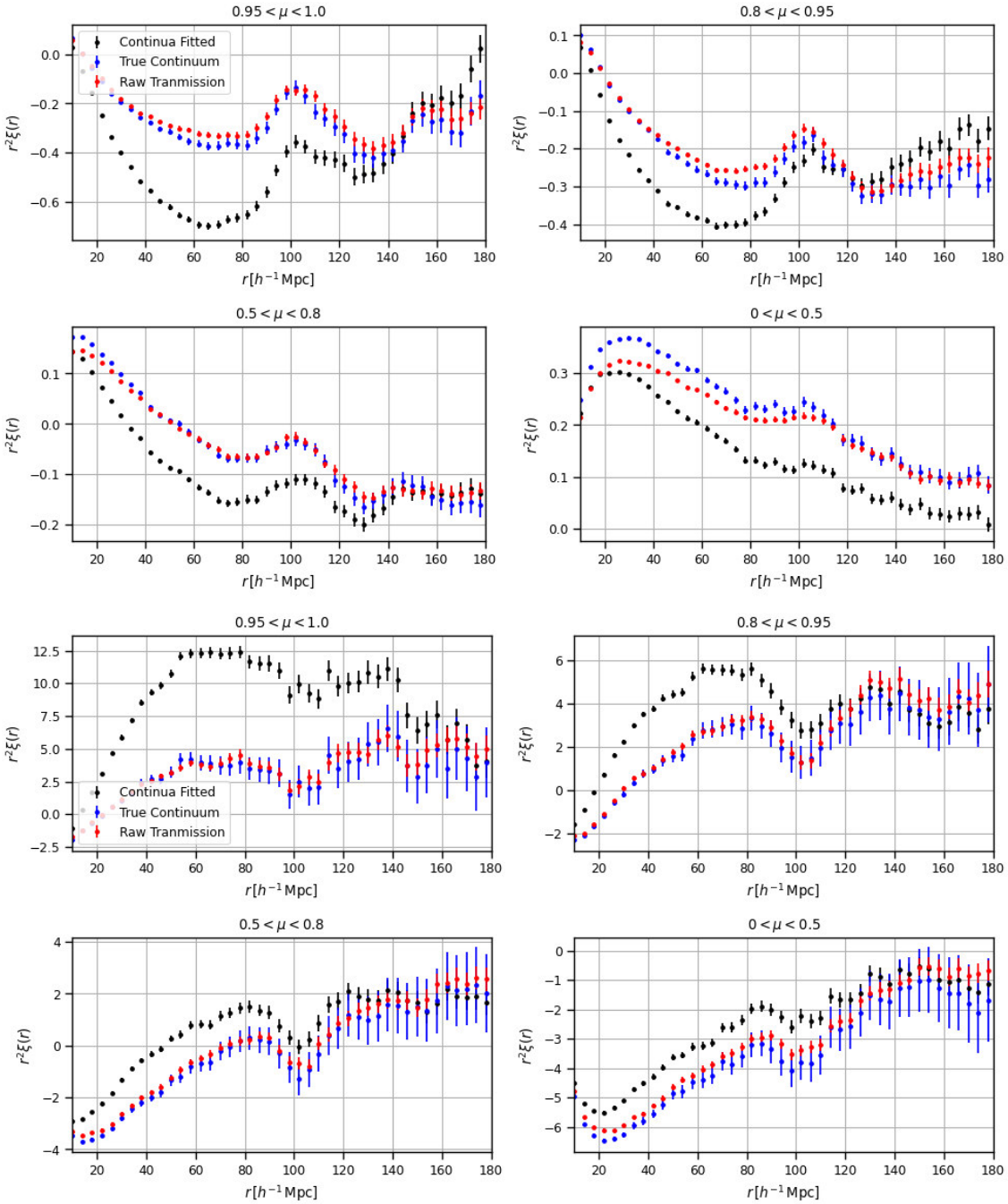

Figure 6.2: Lyman-α auto (top) and cross (bottom) correlation presented as the averages on four $\mu = r_\parallel/r$ ranges measured on a DESI-Y5 mock when performing the continua fitting procedure in `picca` (black dots), when skipping the continua fitting procedure and using the true continuum templates from the mocks (blue dots) and the results as extracted directly from the raw transmissions without a continuum template and instrumental noise added (red dots).

difference of up to 15% when not masking any HCD and close to 10% when only BALs are masked followed by BALs when not masked, which produces a difference of 5% with respect to the uncontaminated case. On both quantities, the results from the uncontaminated case are recovered when masking all BALs and HCDs, this is not precisely realistic given that the results on observed data rely on the efficiency of the BAL and DLA finder algorithms. In any case, HCDs with $\log N_{\rm HI} < 20.3$ cm^{-2} will always be present in the spectra and therefore they must be accounted for on the correlations model. Finally, we note that metals have a negligible impact on both quantities.

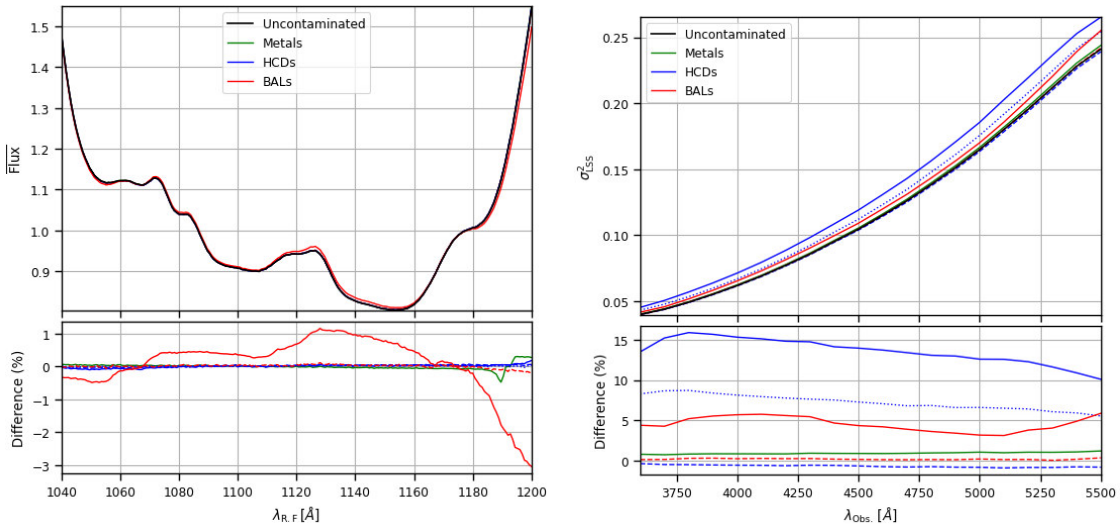

Figure 6.3: Mean flux and $\sigma_{\rm LSS}$ as measured by the pipeline analysis on DESI-Y5 mocks including different contaminants. Bottom panels show the percent difference of the contaminated mocks with respect to the uncontaminated case. The red solid and dashed lines respectively show the result with and without BAL masking on spectra. Blue solid lines show the result when HCDs are not masked, dotted lines when only DLAs with $\log N_{\rm HI} > 20.3$ cm^{-2} are masked and dashed lines when all HCDs are masked.

Now, we will focus on the effect of contaminants on the shape of the auto and cross correlation functions. We will do this for each contaminant individually for clarity.

First, we will start with BALs. Figure 6.4 shows the results of the auto (top) and cross (bottom) correlation functions displayed as the average on four $\mu = r_\parallel/r$ regions with (blue dots) and without (red dots) BAL masking compared to the uncontaminated case (black dots). In this case, the effect of not masking BALs has small effect on the shape of the correlations. Additionally, we recover the measurements from the uncontaminated case with small differences. A more detailed study of the effect of BALs and the test of several masking strategies related to the value of

the Absorption Index (AI) on the continuum and the auto-correlation was performed in [147] using observed data from eBOSS DR14 and the BAL catalog presented in [135].

In the case of HCDs, we are interested in three scenarios: when all HCDs are masked, when only DLAs with column densities $\log N_{\rm HI} > 20.3$ cm^{-2} are masked, and when none of the HCDs found in the spectra are masked. Figure 6.5 shows the auto and cross correlation functions presented as averages in four $\mu = r_\parallel/r$ regions for the three aforementioned cases. Not masking any HCDs or just masking of DLAs produces a modification of the shape and amplitude of the correlation functions clearer seen along the line of sight ($0.95 < \mu < 1.0$) and at small scales across the line of sight ($0 < \mu < 0.5$). Masking all HCDs recovers the results of the uncontaminated case as expected.

Finally, as discussed in Section 2.2.2, metals and neutral hydrogen at the same physical position introduce spurious correlations located at a scale defined by the relative difference of the metal transition wavelength with respect to the one from Lyman-α (defined in Equation (2.48)). Figure 6.6 shows the auto and cross correlation functions, presented as a function of $r_\parallel$ and $r_\perp$, measured on a uncontaminated mock (ξ_0) and a mock contaminated ($\xi_{\rm Metals}$) with Si II(1260), Si III(1207), Si II(1193), and Si II(1190). The differences $\xi_{\rm Metals} - \xi_0$, shown on the right panels, highlight the contribution of the metals to the correlations. Mostly noticeable along the line of sight (small $r_\perp$ values) as blue bumps on the cross-correlation at the scales $r_\parallel = 104$, -21, -54 and -61 Mpc/h, respectively caused by Si II(1260), Si III(1207), Si II(1193), and Si II(1190); and as red bumps on the auto correlation at the absolute value of these same scales.

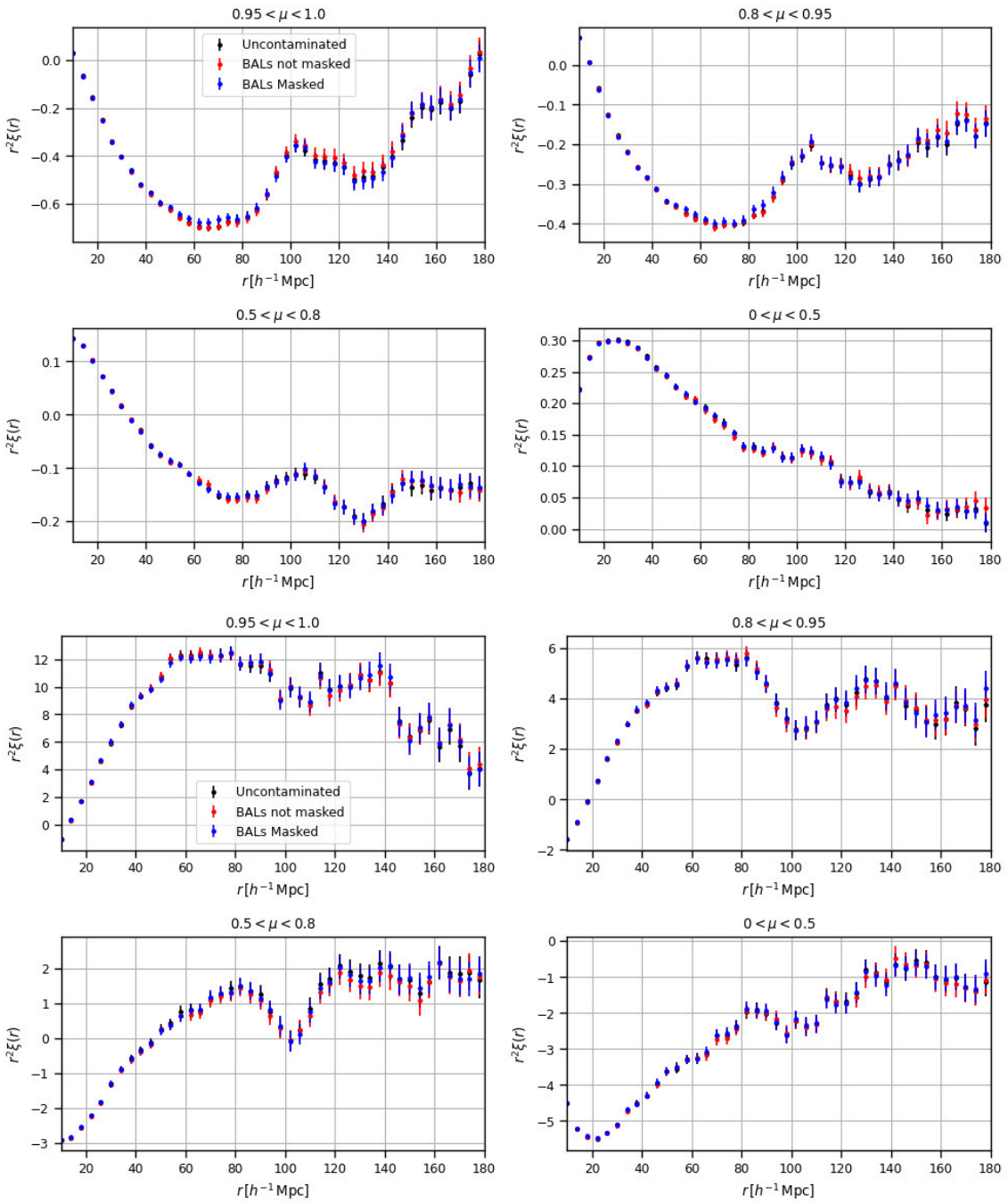

Figure 6.4: Lyman-α auto (top) and cross (bottom) correlation presented as the averages on four $\mu = r_{\parallel}/r$ ranges measured on a DESI-Y5 mock contaminated with BALs only when BALs are masked (blue dots) and when they are left on spectra (red dots) compared to the uncontaminated case (black dots).

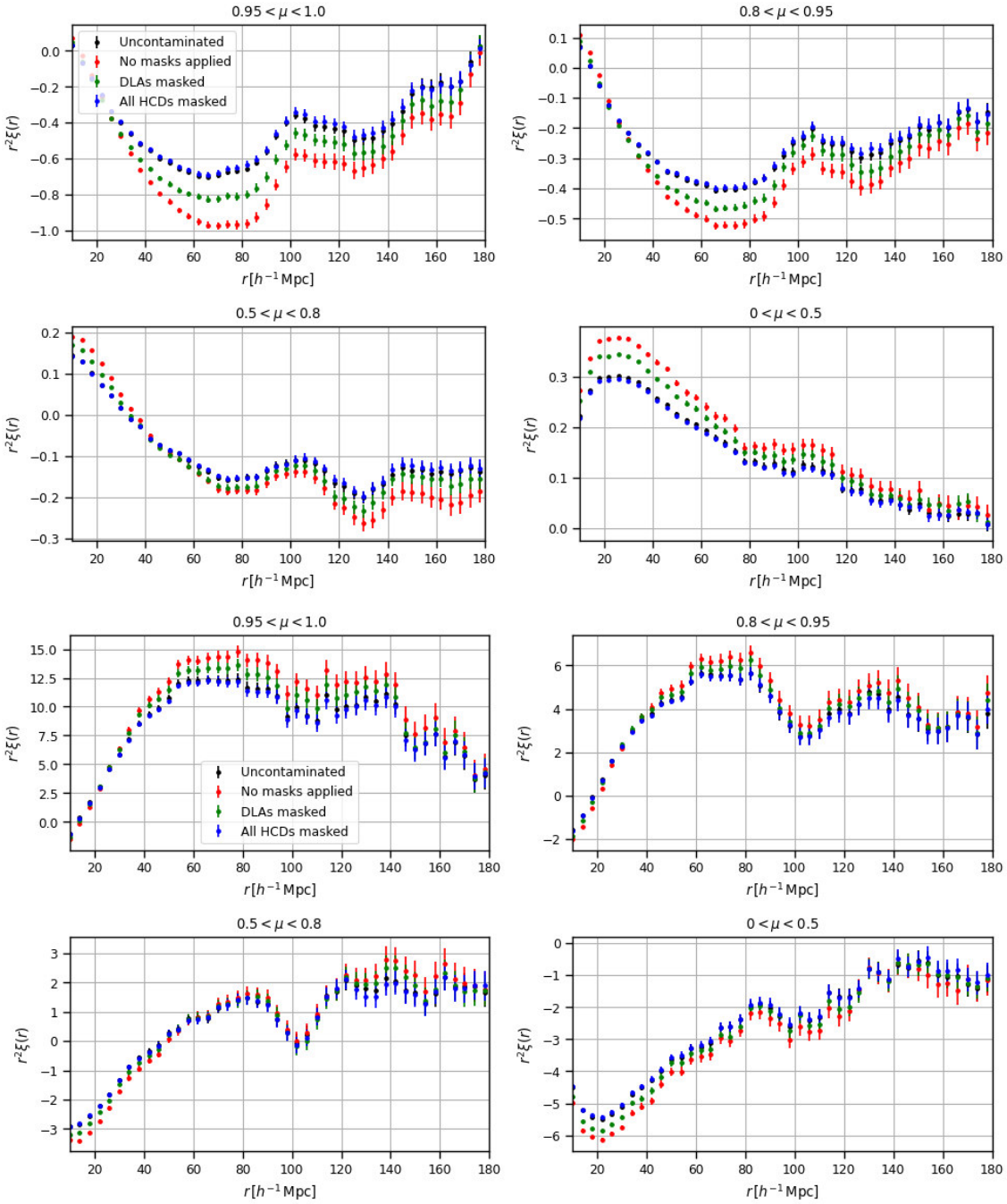

Figure 6.5: Lyman-α auto (top) and cross (bottom) correlation measured on a DESI-Y5 mock contaminated mock with DLAs when: all HCDs are masked (blue dots), only BALs with $\log N_{\rm HI} > 20.3$ cm^{-2} are masked (green dots) and when none of the HCDs introduced to spectra are masked (red dots) compared to the uncontaminated case (black dots).

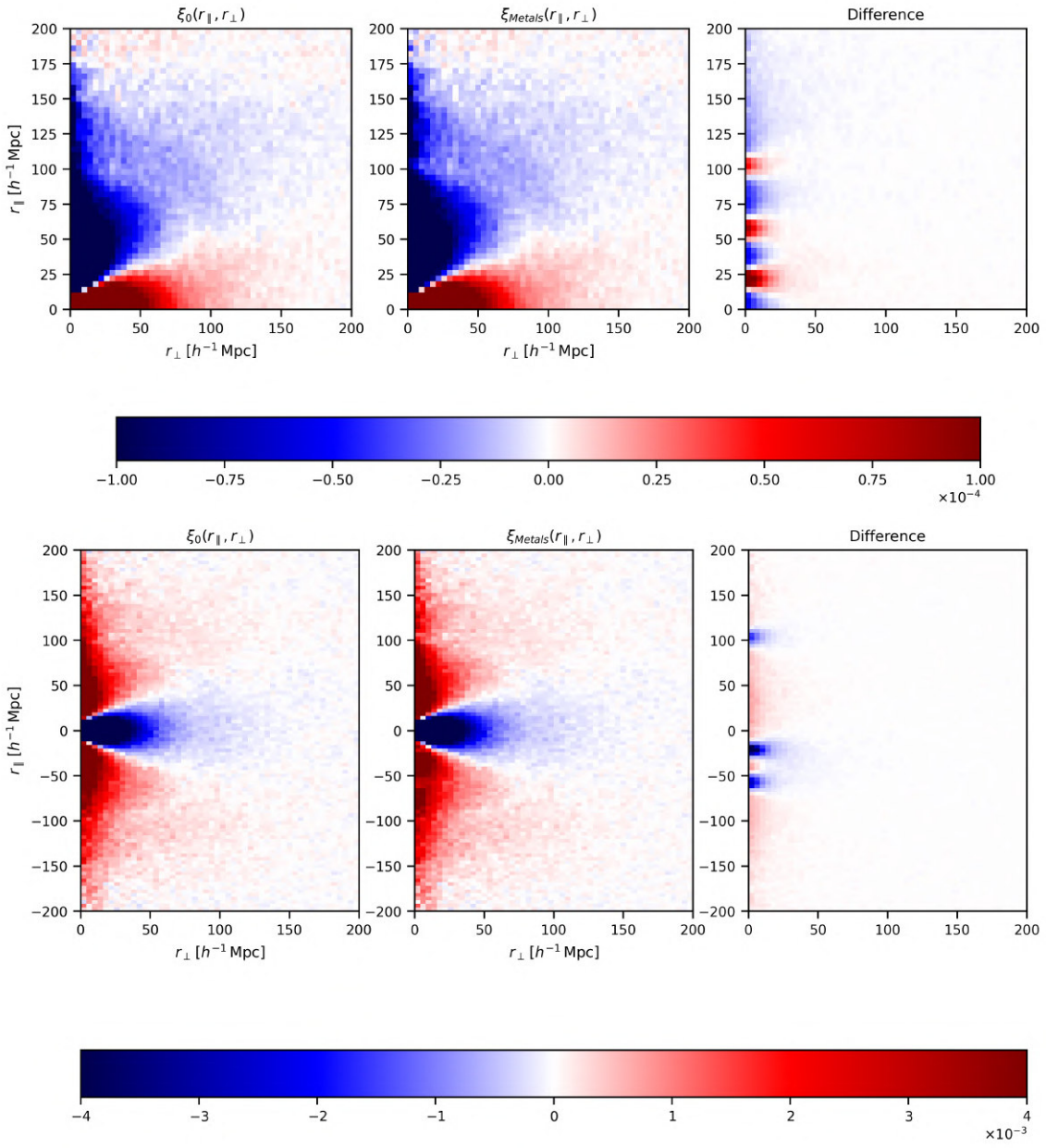

Figure 6.6: Auto (top) and cross (bottom) correlation functions measured a uncontaminated mock (ξ_0) and a mock contaminated (ξ_{Metals}) with Si II(1260), Si III(1207), Si II(1193), and Si II(1190) presented as a function of $r_\parallel$ and $r_\perp$. The right panels display the differences of the correlation functions computed as $\xi_{\mathrm{Metals}} - \xi_0$.

6.1.3 Statistical redshift errors

Now we will focus on statistical errors on the redshift estimation of each quasar, which produce an smoothing of the spectral lines in the estimated mean flux region with respect to the value they would have if all redshifts were correctly estimated. Figure 6.7 shows the mean flux measured on an uncontaminated mock (black lines) and the same mock but when a random redshift error following a Gaussian distribution with dispersion $\sigma_v = 400$ km/s is introduced to the QSO catalog before measuring the flux transmission field (green lines) and the percentage. In this figure we note that the differences (bottom panel) on the mean flux are higher at the position of emission lines, for example C III*(1175), P V(1128), P V (1118), O III(1084), Fe II(1082), S IV (1073) and S IV (1063). In this case the differences are below the 1% level, however they scale as the value of σ_v of the redshift errors distribution.

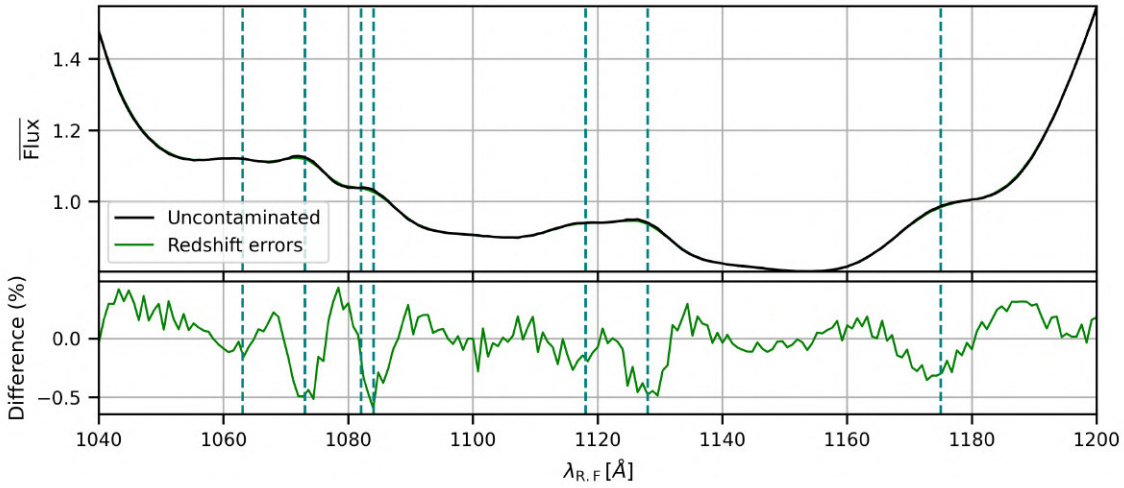

Figure 6.7: Mean flux measured on a DESI-Y5 uncontaminated mock (black lines) and the same mock when random redshift errors following a Gaussian distribution of $\sigma_v = 400$ km/s are added to the quasar catalog (green lines). Bottom panels display the differences with respect to the case without redshift errors. The dashed vertical lines show the position of the S IV (1063), S IV (1073), Fe II(1082) O III(1084), P V(1118), P V(1128), and C III*(1175) emission lines (from left to right).

Even though the differences shown in Figure 6.7 are below the 1% level, these are enough to introduce spurious correlations to the 3D correlations. This is clearly seen in Figure 6.8 which shows a comparison of the auto and cross-correlation functions presented as a function of $r_\perp$ and $r_\parallel$ measured on an uncontaminated mock (ξ_0) and the correlations measured on the same mock when Gaussian redshift errors of $\sigma_v = 400$ km/s are added (ξ_z) to the QSO catalog. The right panels highlight the differences ($\xi_z - \xi_0$) clearer seen as red and blue bumps along the line of sight.

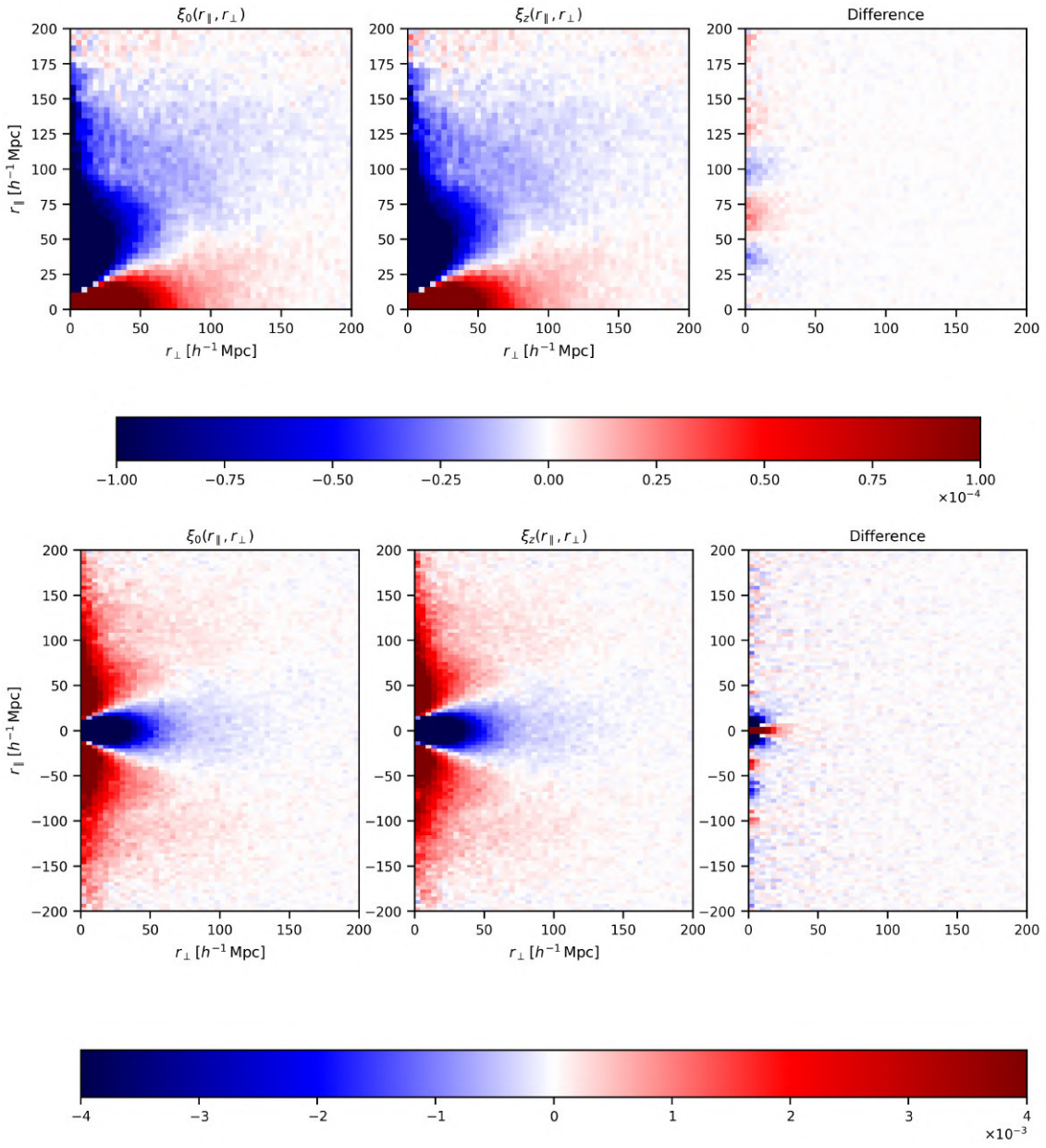

Figure 6.8: Auto (top) and cross (bottom) correlation functions measured on an uncontaminated mock (ξ_0) and the correlations measured on the same mock when Gaussian redshift errors of $\sigma_v = 400$ km/s are added (ξ_z). The right panels display the difference computed as $\xi_z - \xi_0$.

Refer to [145] for a more exhaustive characterization of the effect of redshift errors on the correlation function. In this reference these effects are studied through multiple stacks of 10 mocks where the value of σ_v is varied. Refer to Figures 4 and 5 of this reference for a clear example of how the amplitude of σ_v affects the correlation functions and the mean flux measurements.

Notice that statistical redshift errors have an impact similar to metals on the correlation functions, with contributions mostly on the line-of-sight, and almost at the same scales. This might cause some difficulties to measure the metal biases on the observed data case that is fully contaminated.

6.1.4 Instrumental Noise

The last type of systematic that we will focus on this subsection is the instrumental noise. As mentioned in Chapter 4 we introduce noise to our mocks emulating the observing conditions and response of the instrument. Even though this noise follows ideal observational conditions and therefore might not completely reflect the reality of observed data, it is important to illustrate the effect of noise on the correlation functions measurements. Figure 6.9 shows the mean flux and σ_{LSS} of our baseline mock with added noise (referred to as uncontaminated in the previous sections) and the same mock but without instrumental noise, the differences are seen in the mean flux shape, as a consequence of the better resolution on spectra, and in the magnitude of σ_{LSS}.

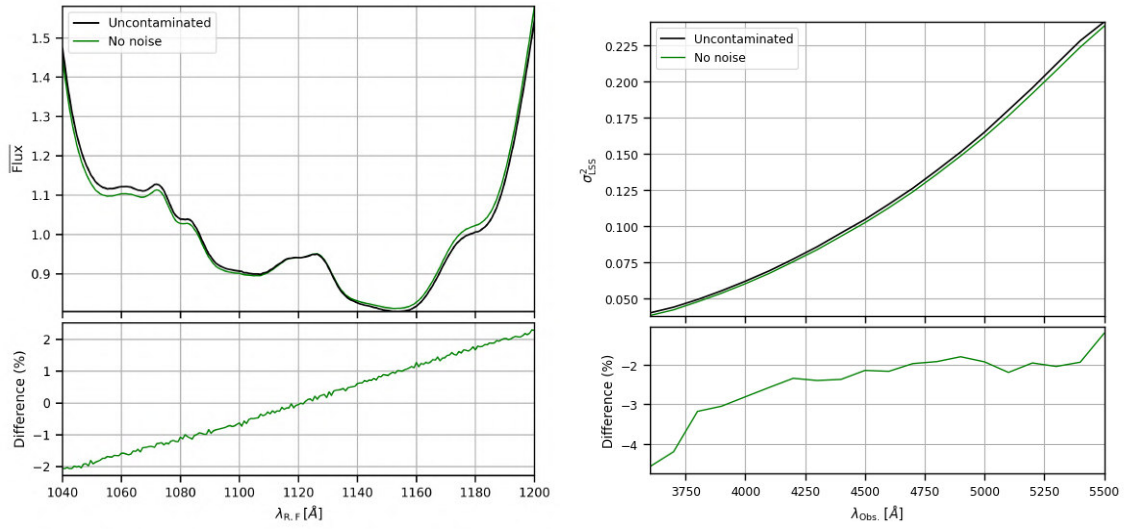

Figure 6.9: Similar to Figure 6.3 with our baseline (uncontaminated) mock including instrumental noise (black lines) compared too the same moock without instrumental noise added (green lines).

The results of the correlation functions averaged on four $\mu = r_{\parallel}/r$ ranges as function of r are shown in Figure 6.10. On the one hand, notice that instrumental noise has a direct impact on the size of the error bars, as naturally expected from what was obtained in $\sigma_{\rm LSS}$. On the other hand, the shape of the correlation functions is altered, most noticeable at small scales across the line of sight. This explains the differences presented in the analysis using the true continuum compared to the raw analysis in Figure 6.2 as the mocks used to measure the correlation functions include instrumental noise.

6.2 Forecasts

Spectroscopic survey forecasts can be done through simple techniques such as a Fisher Formalism, which are mostly done by assuming ideal scenarios. For the purposes of forecasting this provides a best-case scenario uncertainty that is usually underestimated compared to the actual observations, although they serve as a lower limit. In this regard, Lyman-α mock spectra datasets provide an alternative for predicting the constraining power with the flexibility to include additional, more realistic, considerations such as contaminants or instrumental noise.

An example of forecasts using Lyman-α mock datasets was performed in Section 5 of [88], where the predicted BAO scale parameters uncertainties for DESI-Y5 were found to be $\sigma_{\alpha_{\parallel}}/\alpha_{\parallel} = 0.65\%$ and $\sigma_{\alpha_{\perp}}/\alpha_{\perp} = 0.73\%$, if the same observing strategy and expected distributions as the EDR+M2 sample are followed

These forecasts were validated by comparing the prediction of the uncertainties on EDR+M2 mocks against the actual measurements on observed data. With the main caveat being that the EDR+M2 observed data analysis did not provide a measurement on the BAO scale parameters, and thereby the validation of this forecast technique relied on the comparison of other parameters of the model (see Table 2 of [88]). For this reason, the goal of this section is to perform a more complete version of this forecast validation and to update the DESI-Y5 forecasts following the most recent analysis pipeline.

We will follow the same procedure as in the forecast presented in [88]. The only difference being that in this case we will use the full covariance measured in one realization of our fully contaminated `LyaCoLore` DESI-Y1 mocks. This is, we use the forecast mode of `vega` which generates a simulated noiseless correlation function given a fiducial model. Then, use the full covariance matrix result of the mock to construct a Gaussian likelihood. In this case, for simplicity, we perform a Maximum

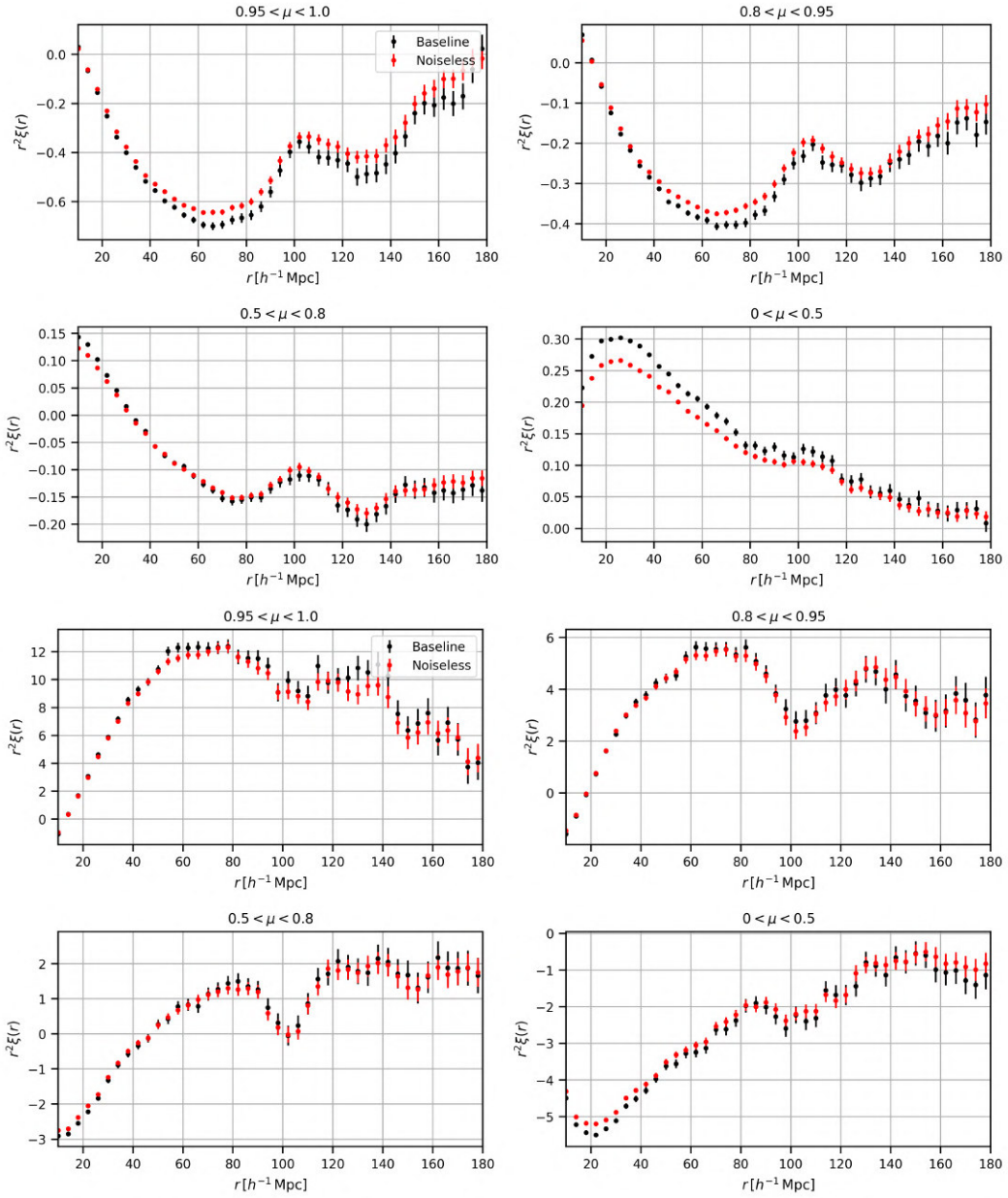

Figure 6.10: Lyman-α auto (top) and cross (bottom) correlation presented as the averages on four $\mu = r_{\parallel}/r$ ranges measured on a DESI-Y5 mock contaminated with BALs only when BALs are masked (blue dots) and when they are left on spectra (red dots) compared to the uncontaminated case (black dots).

Likelihood Estimation (MLE) through a minimization instead of sampling the posterior distribution of each free parameter as was done in [88]. This is performed directly by `vega` using the `iminuit` minimizer package[1] [148, 149]. The main difference is that this method assumes Gaussian distributions of the uncertainties and approximates them using the second derivative of the likelihood around the central value point.

The BAO uncertainties might not be exactly Gaussian and the Posterior Sampling method is therefore a more appropriate approach. However, the BAO scale parameters of the DESI-Y1 mock datasets were found to closely follow a Gaussian distribution [51], and so do the DESI-Y5 mocks. Thus, the MLE approach provides a close estimate of the uncertainties with far less computing resources consumption. See [150] for a more detailed explanation of the differences of these methods.

6.2.1 Validation: DESI-Y1 mocks forecast vs observations

To perform these forecasts, we use the full covariance matrix of a fully contaminated `LyaCoLoRe` realization presented in Section 5.4 based only on their higher redshift range and higher statistics on the Lyα(B) region than the `Saclay` mocks. We use a similar correlation function model and priors as the observed data analysis, see [50] for further details. The fiducial model is set to the central values found by the combined fit from the observed data except for $\alpha_{\parallel}$ and $\alpha_{\perp}$ which are set to 1.

Table 6.1 shows the results of the forecast performed using a DESI-Y1 mock full covariance. Most forecast errors agree with the observed quantity below the 15% level, the exceptions are the quasar radiation proximity ξ_0^{TP}, the instrumental systematics A_{inst}, and the metal biases $b_{\mathrm{SiII(1190)}}$, $b_{\mathrm{SiII(1193)}}$. The first two are expected since the mocks do not include these effects and therefore the covariance matrix do not reflect these effects. The latter two might indicate that the coefficients used to include these metals in spectra might require further tuning to better represent what is seen in data. However, this task is complicated since these metals are affected by the spurious correlations introduced by redshift estimation errors (see Section 6.1.3), and thus this study is left as a future project.

For the purposes of this forecast we are interested in the BAO scale parameters $\alpha_{\parallel}$ and $\alpha_{\perp}$ the values obtained are respectively in $\sim 5\%$ and $\sim 12\%$ underestimated with the observed uncertainty. This is considered to be a good approximation of the uncertainties given the characteristics of our mocks, e.g we lack of BAO broadening

[1] https://github.com/scikit-hep/iminuit

and QSO radiation proximity effects), thus validating this approach and allowing updating the DESI-Y5 forecast of [88], which will be done during the next subsection.

Table 6.1: Forecast uncertainty comparison of one mock against observed data. The second column displays the central value used for our fiducial model. Mock uncertainties σ_{Forecast} are reported as obtained by the `iminuit` minimizer computed from the joint fit of the Lyα(A) and Lyα(B) regions auto and cross correlation. The uncertainties of the observed data (σ_{Obs}) are taken from the results of the observed data [50]. The last column displays the relative difference $\Delta = 1 - \sigma_{\text{Forecast}}/\sigma_{\text{obs}}$ displayed as percentage. The signs in the difference column indicate an underestimation ($-$) or an overestimation ($+$) of the forecasted uncertainties with respect to the observed case.

Parameter	**Central Value**	$\boldsymbol{\sigma_{\text{obs}}}$	$\boldsymbol{\sigma_{\text{Forecast}}}$	**Difference(%)**
$\alpha_{\parallel}$	1.000	0.020	0.019	-4.923%
$\alpha_{\perp}$	1.000	0.023	0.021	-11.702%
$b_{Ly\alpha}$	-0.109	0.005	0.005	-2.480%
$\beta_{Ly\alpha}$	1.726	0.085	0.086	$+1.148\%$
b_{QSO}	3.411	0.048	0.042	-12.061%
$\sigma_{v,\text{QSO}}(h^{-1}\text{Mpc})$	3.665	0.142	0.126	-11.448%
$\Delta r_{\parallel,QSO}(h^{-1}\text{Mpc})$	0.067	0.059	0.058	-0.991%
ξ_0^{TP}	0.399	0.051	0.041	-18.968%
b_{HCD}	-0.056	0.004	0.004	$+1.619\%$
β_{HCD}	0.629	0.080	0.078	-2.523%
$L_{0,\text{HCD}}$	6.541	0.894	0.814	-8.904%
$10^3 b_{\text{SiII(1190)}}$	-4.500	0.639	0.506	-20.726%
$10^3 b_{\text{SiII(1193)}}$	-3.052	0.624	0.495	-20.664%
$10^3 b_{\text{SiIII(1207)}}$	-9.765	0.677	0.601	-11.301%
$10^3 b_{\text{SiII(1260)}}$	-4.001	0.625	0.548	-12.358%
$10^3 b_{\text{CIV}(eff)}$	-24.347	1.499	1.497	-0.167%
$10^4 A_{\text{inst}}$	3.538	0.165	0.117	-29.179%

6.2.2 DESI-Y5 forecast

The favorable results of the forecast of the BAO scale parameters of the previous section allow an update of the full DESI survey forecasts presented in [88]. The forecast of this thesis presents three major differences with respect to the one presented in [88]. First, the forecasts of [88] do not include contaminants, therefore the uncertainties of the BAO scale parameters might be underestimated. For this thesis we will explore the results on a fully contaminated full DESI survey mock, i.e the most realistic forecast possible with out mock generation procedure. Second, we use the full covariance obtained from the correlation functions of this mock, similar to the DESI-DR1 data analysis. Third, the purpose of the forecast of [88] was to

compare the results with a Fisher forecast formalism forecast presented in [52], thus the fiducial model set for the forecast was the same as in the Fisher formalism case. In this case we will set the results from the DESI-DR1 Lyman-α analysis [50] as our fiducial model, i.e this forecast is similar to the one performed on the previous section with the only difference being the covariance matrix.

The characteristics of a full DESI survey mock realization were described in Section 5.2. These include 1.4 million Quasars from which approximately 929k are Lyman-α across the 14,000 square degrees footprint of DESI, and follow the redshift and magnitude distributions as expected by the main DESI survey program. All of the Lyman-α quasars are set to be observed for a total of 4000 seconds. As with the DESI-Y1 mocks for this forecast we include HCDs correlated with the density field used to generate the Lyman-α transmission, add BAL features to 16% of the spectra, include metal contamination, and add random Gaussian errors on the redshift with $\sigma_v = 400$ km/s

We focus only on the parameters of the BAO scale $\alpha_\parallel$ and $\alpha_\perp$. The uncertainties predicted by our forecast are:

$$\sigma_{\alpha_\parallel}/\alpha_\parallel = \sigma_{D_H}/D_H = 0.96\%, \tag{6.1}$$

$$\sigma_{\alpha_\perp}/\alpha_\perp = \sigma_{D_M}/D_M = 1.01\%, \tag{6.2}$$

which according to the definitions of the BAO scale parameters (Equations (2.34) and (2.35)) taking $\alpha_\parallel = \alpha_\perp = 1$, and assuming a Planck2018 [18] cosmology, correspond to

$$D_H(z_{\text{eff}} = 2.33)/r_d = 8.629 \pm 0.083, \tag{6.3}$$

$$D_M(z_{\text{eff}} = 2.33)/r_d = 39.151 \pm 0.396, \tag{6.4}$$

note that these values would correspond to DESI fully confirming the measurements from Planck18, i.e the real central values might vary in reality. Also note that these predicted uncertainties are naturally larger than the presented in [52, 88] since we consider a fully contaminated, more realistic, case. Even in this case, we expect DESI to measure the BAO scale through the Lyman-α forest with a 1% level precision, inclusively if we consider that these forecasts might be underestimated at the same level as the results from the Y1 forecast comparison from Table 6.1.

6.3 Analysis pipeline validation: DESI-DR1

As mentioned at the beginning of Chapter 4, Lyman-α mocks have been used as a test of the robustness of BAO scale measurement pipeline of previous analyses such a BOSS and eBOSS. In this regard, the DESI-DR1 Lyman-α forest analysis [50] includes a validation step, detailed in [51], that makes use of the 100 `LyaCoLoRe` and 50 `Saclay` fully contaminated mocks presented in Section 5.4, and analyzes them with a procedure as close as possible to the analysis pipeline used in observed data. The main goal of this section is to summarize the results of this validation and therefore is based on this references with some figures taken directly from it, with the corresponding credits given.

6.3.1 BAO scale measurement

The study of [51] focuses on two types of analyses of the robustness of the BAO scale measurement procedure. The first, by computing the posterior distribution of the stack of all the realizations performed for each type of raw mock. This allows to perform a BAO measurement on a dataset with 100 times more statistical power than the DESI-DR1 quasar sample in the case of `LyaCoLoRe` and 50 times in the case of `Saclay`. The posterior distribution is sampled with the `Polychord` nested sampler included in `vega`. The main goal of this analysis was to determine if there was a bias introduced by the analysis pipeline on the BAO scale parameters measurement. This is facilitated by mocks since we use the same cosmology to generate them and to analyzed them, i.e we should recover a result consistent with $\alpha_{\parallel} = 1$ and $\alpha_{\perp} = 1$ if no systematic shifts are introduced by the analysis pipeline. A goal set by the DESI-DR1 Lyman-α analysis was to recover these values with an uncertainty threshold set to 1/3 of the DESI-DR1 Lyman-α BAO measurement uncertainty.

The left panel of Figure 6.11 shows the posterior distribution measured on these stack correlations and the combination of both results. The credible regions of the stack of both types of raw mocks contain the expected $\alpha_{\parallel} = 1$ and $\alpha_{\perp} = 1$ values in their 95% credible region. Another thing to note is that the 95% credible regions of both realizations are fully contained in the 1/3 DESI-DR1 uncertainty region. Both these results confirm an unbiased measurement of the BAO scale position with the DESI-DR1 Lyman-α pipeline.

The second type of analysis was aimed to study on the statistics of the population of the BAO scale measurements on each individual mock. Such study serves for two purposes. First, to determine if the individual measurements of the BAO scale

parameters of the mocks scatter around the expected $\alpha_{\parallel} = 1$ and $\alpha_{\perp} = 1$ value, i.e another test to asses that no biases are introduced. The right panel of Figure 6.11 shows the individual results of the 100 `LyaCoLoRe` and 50 `Saclay` mock. All points scatter around the expected value, this is yet another probe that the analysis pipeline does not introduce a bias on the BAO scale parameters measurement.

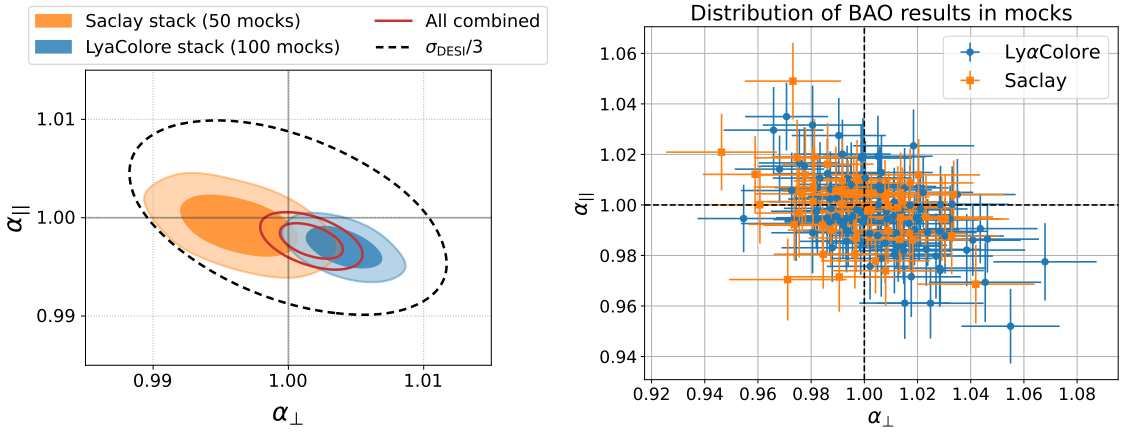

Figure 6.11: Left panel: BAO measurement posterior distribution from the stack of 100 `LyaCoLoRe` (blue contours) and 50 `Saclay` (orange contours) mocks, and the two results combined (red contour). The colored regions highlight the 68% and 95% credible regions. The panel also includes the contour of 1/3 of the uncertainty of the DESI-DR1 Lyman-α BAO measurement. Right panel: BAO scale parameters individually measured from the same mocks as the left panel. Figures taken from [51]. Credits to Andrei Cuceu.

The second purpose is to evaluate that the uncertainties are correctly estimated and are representative of the measurements, and to study the shape of the uncertainty distribution. This is done by comparing the variance of the measured BAO scale parameters of each individual mock to the uncertainties that are estimated. In principle, if the uncertainties are Gaussian and the covariance matrix is correctly estimated the two values should be consistent. Figure 6.12 shows the distribution of the uncertainties obtained on the 150 DESI-Y1 mock presented in this thesis. The red vertical line and bands show the mean value, and the 68% and 95% confidence region of the $\alpha_{\parallel}$ and $\alpha_{\perp}$ values measured in each mock. The error distributions agree with the variance on the measured BAO parameters both these results and a more robust test also performed with the individual measurements on mocks are fully detailed in Section 4.2 of [51]. In contrast, this figure also shows the results obtained on the observed data case, which are outside both the distribution of the error on mocks and the scatter of the best-fit parameters, which might be explained to the lack of some effects on mocks like BAO broadening, this is fully studied and discussed in Appendix A of this same reference.

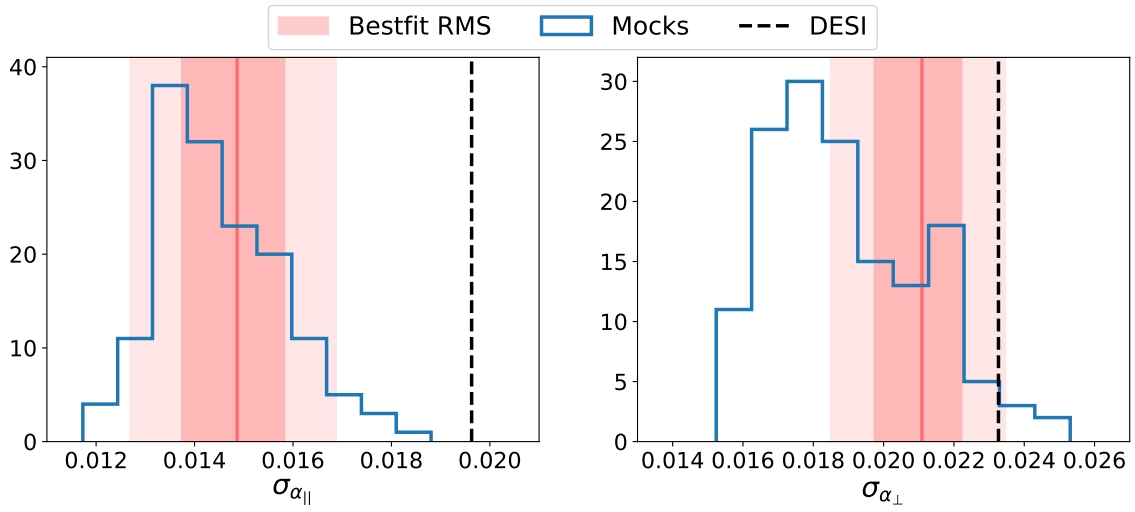

Figure 6.12: Distribution of the uncertainties on BAO parameters of 150 (100 `LyaCoLoRe` and 50 `Saclay`) DESI-Y1 mocks. Red vertical regions show the mean (solid line) and the 68% and 95% confidence intervals of the individual $\alpha_\parallel$ and $\alpha_\perp$ best-fit values. The vertical dashed black line shows the uncertainties obtained in the observed data analysis [50]. Figure taken from [51]. Credits to Andrei Cuceu.

6.3.2 Tests on covariance matrix

The subsampling method used to obtain the covariance matrix for each individual correlation was studied in various of the eBOSS BAO analyses with mocks for example DR12 [58, 59] and DR16 [62] by generating multiple mocks and comparing the covariance measured on one mock by the subsampling method with a covariance obtained from considering the mock-to-mock variation of the correlation functions. This is, if we have the correlation of N mocks $\xi_m(I)$ defined in a given bin $I \to (r_\parallel, r_\perp)$ and the correlation of the stack of all mocks $\xi_{\rm stack}(I)$ the mock-to-mock covariance the covariance matrix, in the bins A and B, $C^{\rm M2M}(A,B)$ is defined as

$$C^{\rm M2M}(A,B) = \frac{1}{N}\sum_{m=0}^{N}[\xi_m(A) - \xi_{\rm stack}(A)][\xi_m(B) - \xi_{\rm stack}(B)], \tag{6.5}$$

this covariance is smoothed in the same way as the covariance matrix obtained from the sub-sampling method (Equation (2.31)).

The aforementioned eBOSS studies showed that the covariance obtained by the subsampling method is similar to the obtained from the mock-to-mock variation. For the DESI-DR1 Lyman-α analysis we studied if this was still the case. Figure 6.13 shows the comparison of the correlation matrix, defined as $\mathrm{Corr}(A,B) = C(A,B)/\sqrt{C(A,A)C(B,B)}$, in the $\Delta r_\perp = 0,4$ Mpc/h separation bins of the Lyα(A)

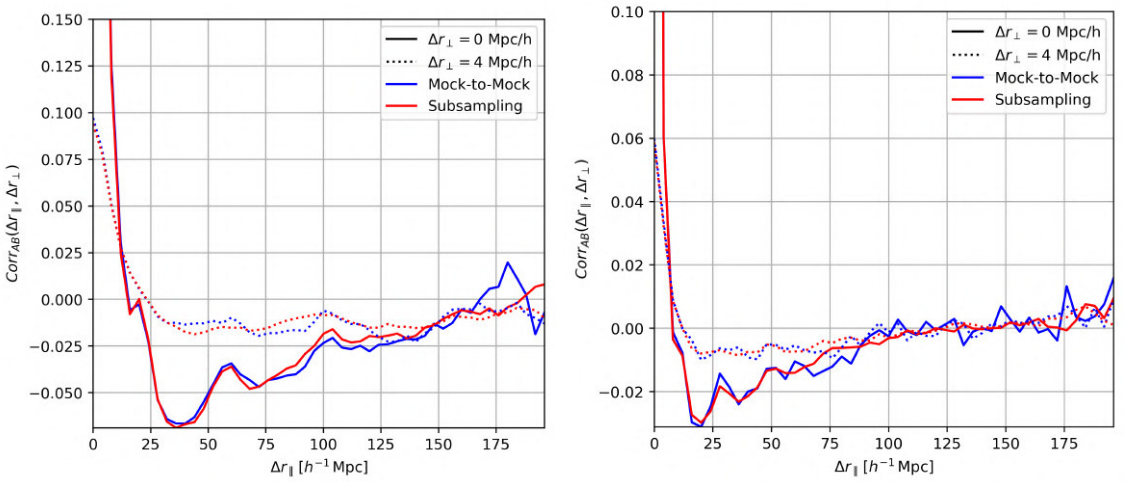

Figure 6.13: Correlation matrix of the Lyα(A) region auto (left panel) and cross (right panel) correlations obtained from the `picca` subsampling method (red color) and the obtained from the mock-to-mock covariance (blue color) of the $\Delta r_\perp = 0,4$ Mpc/h separation bins (solid and dashed lines, respectively).

region auto and cross correlations obtained by both methods. As observed, the results are in concordance, which validates the appropriateness of the subsampling method for the DESI-DR1 sample. The results of the auto and cross correlations on the Lyα(B) region are found to be similar.

For the eBOSS [58, 59, 62, 83, 84] and the DESI-EDR [66] the Lyman-α datasets the cross-covariance across correlations was negligible (see Figure 11 of [62], for example), and thus performed their analysis with the covariance matrix of each individual correlation functions, ignoring cross-covariance. However, during the preparations for the DESI-DR1 Lyman-α we noticed that the cross-covariance could not longer be ignored on the analysis since doing so resulted in 10% underestimation of the BAO uncertainties (discussed in Section 3.3 of [51]). Thus, the decision of using the full covariance matrix shown in Figure 2.2.

The DESI-DR1 BAO analysis validated the full covariance estimation through the comparison of the full covariance obtained from the subsampling method with the results obtained from the stack of the 100 `LyaCoLoRe` and the stack of 50 `Saclay` mocks presented in this thesis. Then, compare the BAO results of each individual mock when using the subsampled covariance matrix against the result of using the covariance from the stack of mocks, showing the results to be consistent with either method up to a tenth of the DESI-DR1 Lyman-α BAO uncertainty. See Section 4.3 of [51] for a more detailed explanation of these results.

6.4 Other examples of Lyman-α mocks uses

This section, will summarize other relevant studies in which Lyman-α mocks have been used that have not been previously mentioned throughout this chapter. In particular, we will focus on those research works where DESI mocks produced with `quickquasars` have been used. This section is based on the corresponding references with all corresponding credits given.

Some of these studies used early versions of the DESI-Y1 mocks, developed prior to the start of DESI operations, and presented in [151], hereafter referred to as PRIOR-DESI-Y1 mocks. These mocks follow a different approach from the DESI-Y1 mocks presented in this thesis and were developed as a prediction tool of the DESI observations. They were produced by performing a survey simulation which includes stochastic weather realizations and follow a color-cut targeting method, as it was the expected method to be used on DESI prior Survey Validation (see Section 3.2). The result is a realistic prediction of the DESI-Y1 nominal footprint, object number density and exposure time distributions, which was mirrored with `quickquasars` following an early version of the method used to generate the EDR+M2 mocks.

6.4.1 Test of DLA identification algorithm performance

Identifying DLA systems is an important part of any spectroscopic survey. On the one hand, as stated in Sections 2.1.3 and 2.2.1, these systems act as a contaminant for Lyman-α studies on the 3D and 1D Lyman-α correlation functions and the 1D flux power spectrum. On the other hand, DLAs are tracers of the underlying matter density field and therefore can also be used to perform clustering studies, for example, through the DLA–Lyman-α forest cross-correlation [60, 68, 134] among other techniques.

Examples of DLA catalogs were produced for SDSS data releases [152–155] following various approaches, including visual inspection. More recently, the large amount of spectra analyses in modern surveys such as BOSS and eBOSS raised the difficulty of producing these catalogs, relying on automatic detection algorithms such as Gaussian Process (GP) [156–158] and Convolutional Neural Networks (CNN) [138, 159].

The DLA catalog presented in eBOSS DR16 had a total of approximately 57k DLA candidates with columns densities above $\log N_{\rm HI} > 20.3$ cm^{-2} [159] that were classified with a CNN algorithm, known as the DLA finder. In this study, the efficiency of the DLA finder was tested by running the algorithm on Lyman-α synthetic

spectra datasets that emulate eBOSS DR16 produced with `quickquasars` [62]. Finding a purity greater than 90% for DLAs with column densities 20.1 $\mathrm{cm}^{-2} \geq \log N_{\mathrm{HI}} \geq$ 22 cm^{-2} in forest with mean flux above $\bar{f}_\lambda > 2 \times 10^{-19}\,\mathrm{Wm}^{-2}\mathrm{nm}^{-1}$.

The DLA finder was later updated to work on the DESI survey spectra [139]. In this study the CNN was trained and validated on PRIOR-DESI-Y1 resulting on an accuracy greater than 99% on quasars with signal to noise above S/N > 5 and 97% on quasars with signal to noise S/N ≈ 1. These results are further discussed in Section 4 of [139] and shown in Figure 6.14 where the left panel displays the purity and the right panel the completeness of the DLA finder as a function of signal to noise and column density, when used on the spectra of a PRIOR-DESI-Y1 mock which include only DLAs as a contaminant.

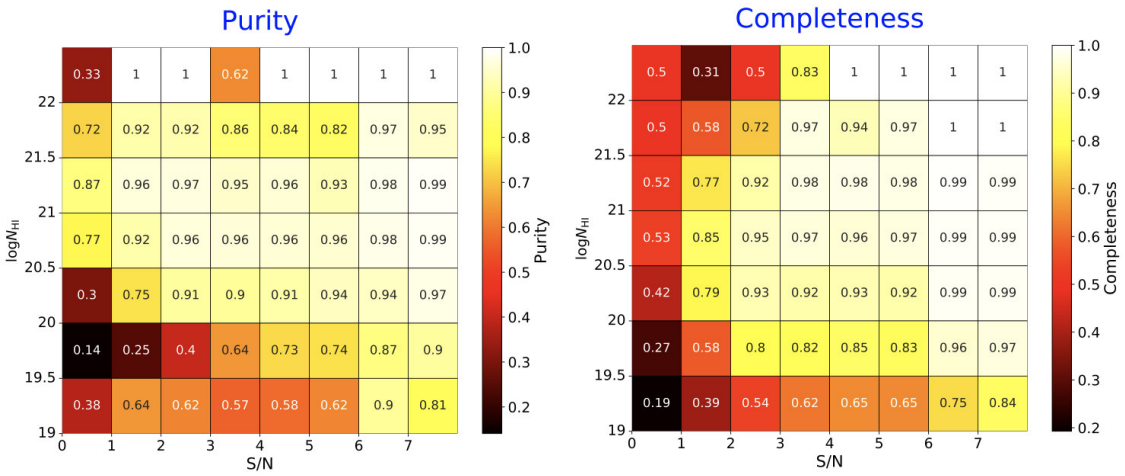

Figure 6.14: Purity (left) and completeness (right) of the DLA finder CNN algorithm found on a PRIOR-DESI-Y1 mock. Figures taken from [139].

6.4.2 Test of BAL identification algorithm performance

Another crucial part of Lyman-α forest BAO studies is identifying BAL features in quasar spectra. In this regard, BOSS DR12 included a catalog of approximately 29.5k BAL quasars identified by visual inspection [160]. Later, given the large amount of data, this task was performed with automatic detection machine learning algorithms. With this kind of algorithms, the eBOSS DR14 [161] and DR16 [162] respectively included a catalog of 21.8k and 99.8k BAL quasars. The algorithm used for these releases, namely the BAL finder, was presented in [135] and consists of a CNN algorithm trained in the C IV (1549) region of BAL quasars identified by visual inspection, finding an accuracy of 98% when used in the DR12 data.

This algorithm was recently updated for DESI to use a PCA approach instead of a CNN, and used to identify almost 30k BAL quasars on the EDR+M2 sample [141]. The performance of the algorithm was also tested on DESI EDR+M2 and DESI-Y5 mock datasets, including BAL features on 16% of the sample, and resulted on a 50% purity and 68% completeness on AI > 0 systems (see sections 2.3 and 3 of [141] for more details).

6.4.3 Effect of BALs on the redshift estimation pipeline

BAL features on quasar spectra can complicate the redshift estimation process as they distort the shape of the main emission line features used to determine the quasar redshift. As stated during Section 6.3 errors on the measurement of quasars redshift lead to spurious correlations on the 3D correlation function analysis. Therefore, characterizing the effect of BAL features on the redshift measurement procedure comprises an important task in the spectroscopic pipeline procedure.

This was studied in [163] by using the synthetic spectra of PRIOR-DESI-Y1 mocks. The procedure was to run the DESI object classifier and redshift estimator, `redrock` [116, 117], on two types of mock datasets, one uncontaminated and another with contamination only due to BAL features on 16% of the spectra. This study found that the presence of BAL features in quasar spectra affects the performance of the `redrock` algorithm. One the one hand, by raising the rate of misidentified quasars from 1.6% on uncontaminated mocks to 3.6% when BAL features are included, this rate decreases to 2.0% if the BAL features are masked. On the other hand, catastrophic redshift failures, defined as those estimated redshifts with more than 15000 km/s difference with respect to the true redshift, were reduced in 80% of the cases. Suggesting that BAL features found in quasar spectra should be masked and the quasar redshift estimation should be re-done after this procedure. Refer to [163] for further details and discussion about this study.

Conclusions and future work

As stated during the introductory section, the main goal of this thesis was twofold. First, to highlight the importance of the Lyman-α forest as a matter tracer and its role in studying the expansion history of the Universe. Second, and the main purpose of this thesis, to describe and showcase the use of synthetic Lyman-α forest datasets as a tool for several studies, including the validation of the BAO scale measurement analysis pipeline, characterization of systematics, algorithm performance tester, to mention a few.

To achieve these goals, Chapter 2 described the theoretical framework of the Lyman-α forest and the state-of-the-art analysis pipeline. Then, Chapter 4 briefly described the key aspects for generating synthetic Lyman-α quasar spectra, including my contributions to the methodology. The current methodology relies on a code called `quickquasars`, to which I have heavily contributed on its development. `Quickquasars` takes as an input a set of Lyman-α transmission field datasets, based in log-normal simulations, and converts them into a realistic quasar spectra by including a continuum template, instrumental noise, and astrophysical contaminants such as metals, Damped Lyman-α Absorbers (DLAs) and Broad Absorption Line systems (BALs).

An important feature of the methodology I developed during my PhD is its flexibility and adaptability to generate synthetic datasets that mimic the characteristics of any Lyman-α spectroscopic survey in terms of its observed footprint, object density distribution, signal-to-noise ratio, redshift distribution, and magnitude distribution. This is an improvement in the realism of synthetic Lyman-α datasets compared to those used in previous experiments such as the Baryon Oscillation Spectroscopic Survey (BOSS) and the extended BOSS (eBOSS). Chapter 5 exemplifies this by presenting dedicated mocks for the most recent, and largest, spectroscopic survey up to the date this thesis was written: the Dark Energy Spectroscopic Instrument (DESI, described in Chapter 3).

Part of the work and efforts presented in this thesis resulted in a first-author article [88], that fully details the aforementioned methodology, and presented Lyman-α mocks emulating the DESI Early Data Release (DESI-EDR). Since the purpose of

this thesis was not to repeat this description, Chapter 4 and the first three sections of Chapter 5 only provided a summary of this article.

Furthermore, Section 5.4 includes the description and enhancements done to the methodology in order to be able to emulate the spatial distribution of the DESI Lyman-α survey sample at an early stage, compared to the method described in [88]. This method was then applied to generate 150 Lyman-α mocks emulating the first year of operations of DESI (DESI-DR1) dataset, one of the key products of this thesis. These mocks, briefly described in [51], served as a key component in the validation of the DESI-DR1 Lyman-α analysis [50], which presented the most precise measurement of the BAO scale up to the date this thesis was written, yielding $D_H(z_{\rm eff} = 2.33)/r_d = 8.52 \pm 0.17$ and $D_M(z_{\rm eff} = 2.33)/r_d = 39.71 \pm 0.95$.

As an additional feature, this thesis provided a more detailed comparison of the DESI-DR1 mock with observed data than the one presented [51] in terms of survey statistics, the Lyman-α forest correlation functions, and BAO scale measurements. The results were favorable, however, they reveal key aspects to improve for future realizations of mocks, the most important of them being the requirement of improving (or moving forward from) the log-normal simulations approach, which may benefit from a better representation of small-scale clustering. Additionally, in [51], we noticed the need to introduce the non-linear BAO broadening which could benefit from more realistic transmission fields based on N-body simulations and/or mocks based on 2nd order Lagrangian Perturbation theory (2LPT), as an input for the `quickquasars` code. Another key improvement, is to find a more realistic method to include metals in the synthetic spectra. Other possible improvements would be to include further contaminants in the mocks, such as correlated sky residuals, UV background fluctuations, contamination by C IV, and update the BAL and DLA distributions used to inject these features into our mocks. These aspects will be studied for future mocks.

To further highlight the usefulness of synthetic Lyman-α datasets, Chapter 6 exemplifies their use for characterizing the effect of contaminants on correlation functions, validate analysis pipelines, forecast and validate the performance of different algorithms. This chapter included the study of the effects of certain contaminants on the correlation functions on variations of one mock realization, while it does not serve as a contaminant characterization itself, since this would require making a large batch of mocks dedicated for each contaminant individually, it serves as a guide to clearly see these effects on one mock realization. This section did not present a study on the effect of the contaminants on the measurement of the BAO scale parameters;

this will be left as a future study.

Then, Section 6.2.2 took advantage of the synthetic Lyman-α datasets generation method and provided a realistic prediction of the BAO scale uncertainties to be obtained by the completed DESI Lyman-α survey. This forecast follows the state-of-the-art pipeline, and uses a covariance measured in mocks that realistically mimic the one expected for the completed DESI Lyman-α survey, resulting in a constraining power of $\sigma_{D_H}/D_H = 0.96\%$ and $\sigma_{D_M}/D_M = 1.01\%$. This forecast method was validated during Section 6.2.1 by comparing the forecasted uncertainty on a DESI-DR1 mock against the results on observed data, revealing that the predicted values are in agreement with the observed quantities at a 4.9% level for $\alpha_{\parallel}$ and 11.7% level for $\alpha_{\perp}$.

The increasing number and quality of the spectra measured by contemporary spectroscopic surveys, such as DESI, and achieving their scientific goals require further improvement and validation on the analysis pipeline. In this regard, the synthetic datasets to be generated for the future realizations must be of a quality comparable to the survey projects themselves. This could benefit from the methodology I have developed to realistically emulate the characteristics of synthetic Lyman-α spectroscopic datasets. This method could be modified to generate synthetic datasets for other surveys besides DESI (including future surveys) and to include the generation of synthetic spectra of other types of objects such as the Lyman-Break galaxies (LBGs). My plan is to study on these enhancements and continue contributing with my little grain of sand to this wonderful task that is studying the Universe.

Bibliography

1. Cambridge Dictionary. *Cosmology* Accessed: 2024-05-31. https://dictionary.cambridge.org/dictionary/english/cosmology.
2. Carroll, S. M. Lecture notes on general relativity. arXiv: gr-qc/9712019 (Dec. 1997).
3. Ryden, B. *Introduction to Cosmology* ISBN: 9781292039718. https://books.google.com.mx/books?id=DtZCnwEACAAJ (Pearson, 2014).
4. Dodelson, S. & Schmidt, F. *Modern Cosmology* doi:10.1016/C2017-0-01943-2 (2020).
5. Einstein, A. Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften,* 831–839 (Jan. 1915).
6. Dyson, F. W., Eddington, A. S. & Davidson, C. A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919. *Phil. Trans. Roy. Soc. Lond. A* **220,** 291–333. doi:10.1098/rsta.1920.0009 (1920).
7. Walsh, D., Carswell, R. F. & Weymann, R. J. 0957+561 A, B: twin quasistellar objects or gravitational lens? Nature **279,** 381–384. doi:10.1038/279381a0 (May 1979).
8. Abbott, B. P. *et al.* Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.* **116,** 061102. doi:10.1103/PhysRevLett.116.061102. arXiv: 1602.03837 [gr-qc] (2016).
9. Event Horizon Telescope Collaboration *et al.* First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole. ApJ **875,** L1. doi:10.3847/2041-8213/ab0ec7. arXiv: 1906.11238 [astro-ph.GA] (Apr. 2019).
10. Einstein, A. The Field Equations of Gravitation. *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)* **1915,** 844–847 (1915).

11. Aghanim, N. *et al.* Planck 2018 results. I. Overview and the cosmological legacy of Planck. *Astron. Astrophys.* **641,** A1. doi:10.1051/0004-6361/201833880. arXiv: 1807.06205 [astro-ph.CO] (2020).

12. Hubble, E. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science* **15,** 168–173. doi:10.1073/pnas.15.3.168 (Mar. 1929).

13. Einstein, A. Cosmological Considerations in the General Theory of Relativity. *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)* **1917,** 142–152 (1917).

14. Perlmutter, S. *et al.* Measurements of Ω and Λ from 42 High Redshift Supernovae. *Astrophys. J.* **517,** 565–586. doi:10.1086/307221. arXiv: astro-ph/9812133 (1999).

15. Riess, A. G. *et al.* Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116,** 1009–1038. doi:10.1086/300499. arXiv: astro-ph/9805201 (1998).

16. eBOSS Collaboration *et al.* Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Cosmological implications from two decades of spectroscopic surveys at the Apache Point Observatory. *Phys. Rev. D* **103,** 083533. doi:10.1103/PhysRevD.103.083533. arXiv: 2007.08991 [astro-ph.CO] (2021).

17. DESI Collaboration *et al.* DESI 2024 VI: Cosmological Constraints from the Measurements of Baryon Acoustic Oscillations. *arXiv e-prints,* arXiv:2404.03002. doi:10.48550/arXiv.2404.03002. arXiv: 2404.03002 [astro-ph.CO] (2024).

18. Aghanim, N. *et al.* Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.* **641.** [Erratum: Astron.Astrophys. 652, C4 (2021)], A6. doi:10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO] (2020).

19. Fixsen, D. J. The Temperature of the Cosmic Microwave Background. ApJ **707,** 916–920. doi:10.1088/0004-637X/707/2/916. arXiv: 0911.1955 [astro-ph.CO] (Dec. 2009).

20. Fukuda, Y. *et al.* Evidence for oscillation of atmospheric neutrinos. *Phys. Rev. Lett.* **81,** 1562–1567. doi:10.1103/PhysRevLett.81.1562. arXiv: hep-ex/9807003 (1998).

21. Zwicky, F. Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta* **6,** 110–127 (Jan. 1933).

22. Rubin, V. C. & Ford W. Kent, J. Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. ApJ **159,** 379. doi:10.1086/150317 (Feb. 1970).

23. Clowe, D. *et al.* A direct empirical proof of the existence of dark matter. *Astrophys. J. Lett.* **648,** L109–L113. doi:10.1086/508162. arXiv: astro-ph/0608407 (2006).

24. Arbey, A. & Mahmoudi, F. Dark matter and the early Universe: a review. *Prog. Part. Nucl. Phys.* **119,** 103865. doi:10.1016/j.ppnp.2021.103865. arXiv: 2104.11488 [hep-ph] (2021).

25. Chow, T. *Gravity, Black Holes, and the Very Early Universe: An Introduction to General Relativity and Cosmology* ISBN: 9780387736297. https://books.google.com.mx/books?id=_Ah8tgAACAAJ (Springer New York, 2007).

26. Weinberg, S. *The First Three Minutes: A Modern View Of The Origin Of The Universe* ISBN: 9780465024377. https://books.google.com.mx/books?id=Yw47_BsQlJ0C (Basic Books, 1993).

27. Guth, A. H. Inflationary universe: A possible solution to the horizon and flatness problems. Phys. Rev. D **23,** 347–356. doi:10.1103/PhysRevD.23.347 (Jan. 1981).

28. Linde, A. D. A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. *Physics Letters B* **108,** 389–393. doi:10.1016/0370-2693(82)91219-9 (Feb. 1982).

29. Albrecht, A. & Steinhardt, P. J. Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking. Phys. Rev. Lett. **48,** 1220–1223. doi:10.1103/PhysRevLett.48.1220 (Apr. 1982).

30. De Boer, W. Grand unified theories and supersymmetry in particle physics and cosmology. *Prog. Part. Nucl. Phys.* **33,** 201–302. doi:10.1016/0146-6410(94)90045-0. arXiv: hep-ph/9402266 (1994).

31. Alpher, R. A. & Herman, R. C. On the Relative Abundance of the Elements. *Physical Review* **74,** 1737–1742. doi:10.1103/PhysRev.74.1737 (Dec. 1948).

32. Penzias, A. A. & Wilson, R. W. A Measurement of Excess Antenna Temperature at 4080 Mc/s. ApJ **142,** 419–421. doi:10.1086/148307 (July 1965).

33. Smoot, G. F. COBE observations and results. *AIP Conf. Proc.* **476** (eds Maiani, L., Melchiorri, F. & Vittorio, N.) 1–10. doi:10.1063/1.59326. arXiv: astro-ph/9902027 (1999).

34. Bennett, C. L. *et al.* Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results. ApJS **208,** 20. doi:10.1088/0067-0049/208/2/20. arXiv: 1212.5225 [astro-ph.CO] (Oct. 2013).

35. Swetz, D. S. *et al.* Overview of the Atacama Cosmology Telescope: Receiver, Instrumentation, and Telescope Systems. ApJS **194,** 41. doi:10.1088/0067-0049/194/2/41. arXiv: 1007.0290 [astro-ph.IM] (June 2011).

36. Carlstrom, J. E. *et al.* The 10 Meter South Pole Telescope. PASP **123,** 568. doi:10.1086/659879. arXiv: 0907.4445 [astro-ph.IM] (May 2011).

37. Einstein, A. Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen. *Jahrbuch der Radioaktivität und Elektronik* **4,** 411–462 (Jan. 1908).

38. Webb, S. *Measuring the Universe: The Cosmological Distance Ladder* ISBN: 9781852331061. https://books.google.com.mx/books?id=ntZwxttZF-sC (Springer, 1999).

39. Eisenstein, D. J., Seo, H.-j. & White, M. J. On the Robustness of the Acoustic Scale in the Low-Redshift Clustering of Matter. *Astrophys. J.* **664,** 660–674. doi:10.1086/518755. arXiv: astro-ph/0604361 (2007).

40. Hu, W. & White, M. Acoustic Signatures in the Cosmic Microwave Background. ApJ **471,** 30. doi:10.1086/177951. arXiv: astro-ph/9602019 [astro-ph] (Nov. 1996).

41. Kaiser, N. Clustering in real space and in redshift space. MNRAS **227,** 1–21. doi:10.1093/mnras/227.1.1 (July 1987).

42. Eisenstein, D. J. *et al.* Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. ApJ **633,** 560–574. doi:10.1086/466512. arXiv: astro-ph/0501171 [astro-ph] (Nov. 2005).

43. McQuinn, M. The Evolution of the Intergalactic Medium. *Ann. Rev. Astron. Astrophys.* **54,** 313–362. doi:10.1146/annurev-astro-082214-122355. arXiv: 1512.00086 [astro-ph.CO] (2016).

44. Mo, H., van den Bosch, F. & White, S. *Galaxy Formation and Evolution* ISBN: 9780521857932. https://books.google.com.mx/books?vid=9780521857932 (Cambridge University Press, 2010).

45. Prochaska, J. X., O'Meara, J. M. & Worseck, G. A Definitive Survey for Lyman Limit Systems at z ~3.5 with the Sloan Digital Sky Survey. ApJ **718,** 392–416. doi:10.1088/0004-637X/718/1/392. arXiv: 0912.0292 [astro-ph.CO] (July 2010).

46. Wolfe, A. M., Gawiser, E. & Prochaska, J. X. Damped lyman alpha systems. *Ann. Rev. Astron. Astrophys.* **43,** 861–918. doi:10.1146/annurev.astro.42.053102.133950. arXiv: astro-ph/0509481 (2005).

47. Pieri, M. M. *et al.* The Composite Spectrum of Strong Lyα Forest Absorbers. ApJ **724,** L69–L73. doi:10.1088/2041-8205/724/1/L69. arXiv: 1001.5282 [astro-ph.CO] (Nov. 2010).

48. Weymann, R. J., Morris, S. L., Foltz, C. B. & Hewett, P. C. Comparisons of the Emission-Line and Continuum Properties of Broad Absorption Line and Normal Quasi-stellar Objects. ApJ **373,** 23. doi:10.1086/170020 (May 1991).

49. Hall, P. B. *et al.* Unusual Broad Absorption Line Quasars from the Sloan Digital Sky Survey. ApJS **141,** 267–309. doi:10.1086/340546. arXiv: astro-ph/0203252 [astro-ph] (Aug. 2002).

50. DESI Collaboration *et al.* DESI 2024 IV: Baryon Acoustic Oscillations from the Lyman Alpha Forest. *arXiv e-prints,* arXiv:2404.03001. doi:10.48550/arXiv.2404.03001. arXiv: 2404.03001 [astro-ph.CO] (2024).

51. Cuceu, A. *et al.* Validation of the DESI 2024 Lyα forest BAO analysis using synthetic datasets. *arXiv e-prints,* arXiv:2404.03004. doi:10.48550/arXiv.2404.03004. arXiv: 2404.03004 [astro-ph.CO] (2024).

52. Aghamousa, A. *et al.* The DESI Experiment Part I: Science,Targeting, and Survey Design. arXiv: 1611.00036 [astro-ph.IM] (Oct. 2016).

53. Aghamousa, A. *et al.* The DESI Experiment Part II: Instrument Design. arXiv: 1611.00037 [astro-ph.IM] (Oct. 2016).

54. DESI Collaboration *et al.* Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument. AJ **164,** 207. doi:10.3847/1538-3881/ac882b. arXiv: 2205.10939 [astro-ph.IM] (Nov. 2022).

55. Ramirez-Perez, C. *et al.* The Lyman-α forest catalog from the Dark Energy Spectroscopic Instrument Early Data Release. arXiv: 2306.06312 [astro-ph.CO] (June 2023).

56. McDonald, P. *et al.* The Lyman-alpha forest power spectrum from the Sloan Digital Sky Survey. *Astrophys. J. Suppl.* **163,** 80–109. doi:10.1086/444361. arXiv: astro-ph/0405013 (2006).

57. Du Mas des Bourboux, H. *et al.* The Extended Baryon Oscillation Spectroscopic Survey: Measuring the Cross-correlation between the MgII Flux Transmission Field and Quasars and Galaxies at $z = 0.59$. *Astrophys. J.* **878,** 47. doi:10.3847/1538-4357/ab1d49. arXiv: 1901.01950 [astro-ph.CO] (2019).

58. du Mas des Bourboux, H. *et al.* Baryon acoustic oscillations from the complete SDSS-III Lyα-quasar cross-correlation function at z = 2.4. A&A **608,** A130. doi:10.1051/0004-6361/201731731. arXiv: 1708.02225 [astro-ph.CO] (Dec. 2017).

59. Bautista, J. E. *et al.* Measurement of baryon acoustic oscillation correlations at z = 2.3 with SDSS DR12 Lyα-Forests. A&A **603,** A12. doi:10.1051/0004-6361/201730533. arXiv: 1702.00176 [astro-ph.CO] (June 2017).

60. Pérez-Ràfols, I. *et al.* The SDSS-DR12 large-scale cross-correlation of damped Lyman alpha systems with the Lyman alpha forest. *Mon. Not. Roy. Astron. Soc.* **473,** 3019–3038. doi:10.1093/mnras/stx2525. arXiv: 1709.00889 [astro-ph.CO] (2018).

61. Górski, K. M. *et al.* HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. ApJ **622,** 759–771. doi:10.1086/427976. arXiv: astro-ph/0409513 [astro-ph] (Apr. 2005).

62. du Mas des Bourboux, H. *et al.* The Completed SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations with Lyα Forests. ApJ **901,** 153. doi:10.3847/1538-4357/abb085. arXiv: 2007.08995 [astro-ph.CO] (Oct. 2020).

63. Delubac, T. *et al.* Baryon acoustic oscillations in the Lyα forest of BOSS DR11 quasars. A&A **574,** A59. doi:10.1051/0004-6361/201423969. arXiv: 1404.1801 [astro-ph.CO] (Feb. 2015).

64. Kirkby, D. *et al.* Fitting methods for baryon acoustic oscillations in the Lyman-α forest fluctuations in BOSS data release 9. J. Cosmology Astropart. Phys. **2013,** 024. doi:10.1088/1475-7516/2013/03/024. arXiv: 1301.3456 [astro-ph.CO] (Mar. 2013).

65. Lewis, A., Challinor, A. & Lasenby, A. Efficient computation of CMB anisotropies in closed FRW models. ApJ **538,** 473–476. doi:10.1086/309179. arXiv: astro-ph/9911177 [astro-ph] (2000).

66. Gordon, C. *et al.* 3D Correlations in the Lyman-α Forest from Early DESI Data. arXiv: 2308.10950 [astro-ph.CO] (Aug. 2023).

67. Arinyo-i-Prats, A., Miralda-Escudé, J., Viel, M. & Cen, R. The Non-Linear Power Spectrum of the Lyman Alpha Forest. *JCAP* **12,** 017. doi:10.1088/1475-7516/2015/12/017. arXiv: 1506.04519 [astro-ph.CO] (2015).

68. Pérez-Ràfols, I. *et al.* The cross-correlation of galaxies in absorption with the Lyman α forest. *Mon. Not. Roy. Astron. Soc.* **524,** 1464–1477. doi:10.1093/mnras/stad1994. arXiv: 2210.02973 [astro-ph.CO] (2023).

69. Rogers, K. K. *et al.* Correlations in the three-dimensional Lyman-alpha forest contaminated by high column density absorbers. MNRAS **476,** 3716–3728. doi:10.1093/mnras/sty603. arXiv: 1711.06275 [astro-ph.CO] (May 2018).

70. Font-Ribera, A. *et al.* The large-scale quasar-Lyman α forest cross-correlation from BOSS. J. Cosmology Astropart. Phys. **2013,** 018. doi:10.1088/1475-7516/2013/05/018. arXiv: 1303.1937 [astro-ph.CO] (May 2013).

71. Guy, J. *et al.* Characterization of contaminants in the Lyman-alpha forest auto-correlation with DESI. *arXiv e-prints,* arXiv:2404.03003. doi:10.48550/arXiv.2404.03003. arXiv: 2404.03003 [astro-ph.CO] (2024).

72. McDonald, P. & Eisenstein, D. Dark energy and curvature from a future baryonic acoustic oscillation survey using the Lyman-alpha forest. *Phys. Rev. D* **76,** 063009. doi:10.1103/PhysRevD.76.063009. arXiv: astro-ph/0607122 (2007).

73. McDonald, P. Toward a Measurement of the Cosmological Geometry at z ~2: Predicting Lyα Forest Correlation in Three Dimensions and the Potential of Future Data Sets. ApJ **585,** 34–51. doi:10.1086/345945. arXiv: astro-ph/0108064 [astro-ph] (Mar. 2003).

74. Slosar, A. *et al.* The Lyman-α forest in three dimensions: measurements of large scale flux correlations from BOSS 1st-year data. J. Cosmology Astropart. Phys. **2011,** 001. doi:10.1088/1475-7516/2011/09/001. arXiv: 1104.5244 [astro-ph.CO] (Sept. 2011).

75. Dawson, K. S. *et al.* The Baryon Oscillation Spectroscopic Survey of SDSS-III. AJ **145,** 10. doi:10.1088/0004-6256/145/1/10. arXiv: 1208.0022 [astro-ph.CO] (Jan. 2013).

76. Busca, N. G. *et al.* Baryon acoustic oscillations in the Lyα forest of BOSS quasars. A&A **552,** A96. doi:10.1051/0004-6361/201220724. arXiv: 1211.2616 [astro-ph.CO] (Apr. 2013).

77. Lee, K.-G. *et al.* The BOSS Lyα Forest Sample from SDSS Data Release 9. AJ **145,** 69. doi:10.1088/0004-6256/145/3/69. arXiv: 1211.5146 [astro-ph.CO] (Mar. 2013).

78. Slosar, A. *et al.* Measurement of baryon acoustic oscillations in the Lyman-α forest fluctuations in BOSS data release 9. J. Cosmology Astropart. Phys. **2013,** 026. doi:10.1088/1475-7516/2013/04/026. arXiv: 1301.3459 [astro-ph.CO] (Apr. 2013).

79. Alam, S. *et al.* The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. ApJS **219,** 12. doi:10.1088/0067-0049/219/1/12. arXiv: 1501.00963 [astro-ph.IM] (July 2015).

80. Font-Ribera, A. *et al.* Quasar-Lyman α forest cross-correlation from BOSS DR11: Baryon Acoustic Oscillations. J. Cosmology Astropart. Phys. **2014,** 027. doi:10.1088/1475-7516/2014/05/027. arXiv: 1311.1767 [astro-ph.CO] (May 2014).

81. Dawson, K. S. *et al.* The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. AJ **151,** 44. doi:10.3847/0004-6256/151/2/44. arXiv: 1508.04473 [astro-ph.CO] (Feb. 2016).

82. Abolfathi, B. *et al.* The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment. ApJS **235,** 42. doi:10.3847/1538-4365/aa9e8a. arXiv: 1707.09322 [astro-ph.GA] (Apr. 2018).

83. de Sainte Agathe, V. *et al.* Baryon acoustic oscillations at z = 2.34 from the correlations of Lyα absorption in eBOSS DR14. A&A **629,** A85. doi:10.1051/0004-6361/201935638. arXiv: 1904.03400 [astro-ph.CO] (Sept. 2019).

84. Blomqvist, M. *et al.* Baryon acoustic oscillations from the cross-correlation of Lyα absorption and quasars in eBOSS DR14. A&A **629,** A86. doi:10.1051/0004-6361/201935641. arXiv: 1904.03430 [astro-ph.CO] (Sept. 2019).

85. Ahumada, R. *et al.* The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. ApJS **249,** 3. doi:10.3847/1538-4365/ab929e. arXiv: 1912.02905 [astro-ph.GA] (July 2020).

86. Adame, G. *et al.* The Early Data Release of the Dark Energy Spectroscopic Instrument. doi:10.5281/zenodo.7964161. arXiv: 2306.06308 [astro-ph.CO] (June 2023).

87. DESI Collaboration. DESI 2024 I: Data Release 1 of the Dark Energy Spectroscopic Instrument. *in preparation* (2025).

88. Herrera-Alcantar, H. K. *et al.* Synthetic spectra for Lyman-α forest analysis in the Dark Energy Spectroscopic Instrument. *arXiv e-prints,* arXiv:2401.00303. doi:10.48550/arXiv.2401.00303. arXiv: 2401.00303 [astro-ph.CO] (Dec. 2023).

89. Davis, M., Huchra, J., Latham, D. W. & Tonry, J. A survey of galaxy redshifts. II. The large scale space distribution. ApJ **253,** 423–445. doi:10.1086/159646 (Feb. 1982).

90. Falco, E. E. *et al.* The Updated Zwicky Catalog (UZC). PASP **111,** 438–452. doi:10.1086/316343. arXiv: astro-ph/9904265 [astro-ph] (Apr. 1999).

91. Colless, M. *et al.* The 2dF Galaxy Redshift Survey: spectra and redshifts. MNRAS **328,** 1039–1063. doi:10.1046/j.1365-8711.2001.04902.x. arXiv: astro-ph/0106498 [astro-ph] (Dec. 2001).

92. York, D. G. *et al.* The Sloan Digital Sky Survey: Technical Summary. AJ **120,** 1579–1587. doi:10.1086/301513. arXiv: astro-ph/0006396 (2000).

93. Eisenstein, D. J., Weinberg, D. H., *et al.* SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems. AJ **142,** 72. doi:10.1088/0004-6256/142/3/72. arXiv: 1101.1529 [astro-ph.IM] (Sept. 2011).

94. Blanton, M. R. *et al.* Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. AJ **154,** 28. doi:10.3847/1538-3881/aa7567. arXiv: 1703.00052 [astro-ph.GA] (July 2017).

95. Adame, G. *et al.* Validation of the Scientific Program for the Dark Energy Spectroscopic Instrument. doi:10.5281/zenodo.7858207. arXiv: 2306.06307 [astro-ph.CO] (June 2023).

96. Abareshi, B. *et al.* Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument. *Astron. J.* **164,** 207. doi:10.3847/1538-3881/ac882b. arXiv: 2205.10939 [astro-ph.IM] (2022).

97. Miller, T. N. *et al.* The Optical Corrector for the Dark Energy Spectroscopic Instrument. arXiv: 2306.06310 [astro-ph.IM] (June 2023).

98. Silber, J. H. *et al.* The Robotic Multiobject Focal Plane System of the Dark Energy Spectroscopic Instrument (DESI). *Astron. J.* **165,** 9. doi:10.3847/1538-3881/ac9ab1. arXiv: 2205.09014 [astro-ph.IM] (2023).

99. Dey, A. *et al.* Overview of the DESI Legacy Imaging Surveys. AJ **157,** 168. doi:10.3847/1538-3881/ab089d. arXiv: 1804.08657 [astro-ph.IM] (May 2019).

100. Cooper, A. P. *et al.* Overview of the DESI Milky Way Survey. ApJ **947,** 37. doi:10.3847/1538-4357/acb3c0. arXiv: 2208.08514 [astro-ph.GA] (Apr. 2023).

101. Hahn, C. *et al.* The DESI Bright Galaxy Survey: Final Target Selection, Design, and Validation. AJ **165,** 253. doi:10.3847/1538-3881/accff8. arXiv: 2208.08512 [astro-ph.CO] (June 2023).

102. Zhou, R. *et al.* Target Selection and Validation of DESI Luminous Red Galaxies. AJ **165,** 58. doi:10.3847/1538-3881/aca5fb. arXiv: 2208.08515 [astro-ph.CO] (Feb. 2023).

103. Raichoor, A. *et al.* Target Selection and Validation of DESI Emission Line Galaxies. AJ **165,** 126. doi:10.3847/1538-3881/acb213. arXiv: 2208.08513 [astro-ph.CO] (Mar. 2023).

104. Chaussidon, E. *et al.* Target Selection and Validation of DESI Quasars. ApJ **944,** 107. doi:10.3847/1538-4357/acb3c2. arXiv: 2208.08511 [astro-ph.CO] (Feb. 2023).

105. Wright, E. L. *et al.* The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. AJ **140,** 1868–1881. doi:10.1088/0004-6256/140/6/1868. arXiv: 1008.0031 [astro-ph.IM] (Dec. 2010).

106. Flaugher, B. *et al.* The Dark Energy Camera. AJ **150,** 150. doi:10.1088/0004-6256/150/5/150. arXiv: 1504.02900 [astro-ph.IM] (Nov. 2015).

107. The Dark Energy Survey Collaboration. The Dark Energy Survey. *arXiv e-prints.* arXiv: astro-ph/0510346 [astro-ph] (Oct. 2005).

108. Zou, H. *et al.* Project Overview of the Beijing-Arizona Sky Survey. PASP **129,** 064101. doi:10.1088/1538-3873/aa65ba. arXiv: 1702.03653 [astro-ph.GA] (June 2017).

109. Mainzer, A. *et al.* Initial Performance of the NEOWISE Reactivation Mission. ApJ **792,** 30. doi:10.1088/0004-637X/792/1/30. arXiv: 1406.6025 [astro-ph.EP] (Sept. 2014).

110. Myers, A. D. *et al.* The Target-selection Pipeline for the Dark Energy Spectroscopic Instrument. AJ **165,** 50. doi:10.3847/1538-3881/aca5f9. arXiv: 2208.08518 [astro-ph.IM] (Feb. 2023).

111. Schlafly, E. F. *et al.* Survey Operations for the Dark Energy Spectroscopic Instrument. AJ **166,** 259. doi:10.3847/1538-3881/ad0832. arXiv: 2306.06309 [astro-ph.CO] (Dec. 2023).

112. Guy, J. *et al.* The Spectroscopic Data Processing Pipeline for the Dark Energy Spectroscopic Instrument. AJ **165,** 144. doi:10.3847/1538-3881/acb212. arXiv: 2209.14482 [astro-ph.IM] (Apr. 2023).

113. DESI Collaboration *et al.* Validation of the Scientific Program for the Dark Energy Spectroscopic Instrument. *arXiv e-prints,* arXiv:2306.06307. doi:10.48550/arXiv.2306.06307. arXiv: 2306.06307 [astro-ph.CO] (June 2023).

114. Kirkby et al. in preparation (2024).

115. Bolton, A. S. & Schlegel, D. J. Spectro-Perfectionism: An Algorithmic Framework for Photon Noise-Limited Extraction of Optical Fiber Spectroscopy. PASP **122,** 248. doi:10.1086/651008. arXiv: 0911.2689 [astro-ph.IM] (Feb. 2010).

116. Bailey et al. , *in preparation* (2024).

117. Brodzeller, A. *et al.* Performance of the Quasar Spectral Templates for the Dark Energy Spectroscopic Instrument. *arXiv e-prints,* arXiv:2305.10426. doi:10.48550/arXiv.2305.10426. arXiv: 2305.10426 [astro-ph.IM] (May 2023).

118. Busca, N. & Balland, C. QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks. arXiv: 1808.09955 [astro-ph.IM] (Aug. 2018).

119. Alexander, D. M. *et al.* The DESI Survey Validation: Results from Visual Inspection of the Quasar Survey Spectra. *Astron. J.* **165,** 124. doi:10.3847/1538-3881/acacfc. arXiv: 2208.08517 [astro-ph.GA] (2023).

120. Lan, T.-W. *et al.* The DESI Survey Validation: Results from Visual Inspection of Bright Galaxies, Luminous Red Galaxies, and Emission-line Galaxies. *Astrophys. J.* **943,** 68. doi:10.3847/1538-4357/aca5fa. arXiv: 2208.08516 [astro-ph.CO] (2023).

121. Font-Ribera, A., McDonald, P. & Miralda-Escudé, J. Generating mock data sets for large-scale Lyman-α forest correlation measurements. J. Cosmology Astropart. Phys. **2012,** 001. doi:10.1088/1475-7516/2012/01/001. arXiv: 1108.5606 [astro-ph.CO] (Jan. 2012).

122. Bautista, J. E. *et al.* Mock Quasar-Lyman-α forest data-sets for the SDSS-III Baryon Oscillation Spectroscopic Survey. J. Cosmology Astropart. Phys. **2015,** 060. doi:10.1088/1475-7516/2015/05/060. arXiv: 1412.0658 [astro-ph.CO] (May 2015).

123. Le Goff, J. M. *et al.* Simulations of BAO reconstruction with a quasar Ly-α survey. A&A **534,** A135. doi:10.1051/0004-6361/201117736. arXiv: 1107.4233 [astro-ph.CO] (Oct. 2011).

124. Farr, J. *et al.* LyaCoLoRe: synthetic datasets for current and future Lyman-α forest BAO surveys. J. Cosmology Astropart. Phys. **2020,** 068. doi:10.1088/1475-7516/2020/03/068. arXiv: 1912.02763 [astro-ph.CO] (Mar. 2020).

125. Etourneau, T. *et al.* Mock data sets for the Eboss and DESI Lyman-α forest surveys. *arXiv e-prints,* arXiv:2310.18996. doi:10.48550/arXiv.2310.18996. arXiv: 2310.18996 [astro-ph.CO] (Oct. 2023).

126. Karaçaylı, N. G. *et al.* Optimal 1D Lyα Forest Power Spectrum Estimation – III. DESI early data. arXiv: 2306.06316 [astro-ph.CO] (June 2023).

127. Ravoux, C. *et al.* The Dark Energy Spectroscopic Instrument: One-dimensional power spectrum from first Lyman-α forest samples with Fast Fourier Transform. arXiv: 2306.06311 [astro-ph.CO] (June 2023).

128. Ross, N. P. *et al.* The SDSS-III Baryon Oscillation Spectroscopic Survey: The Quasar Luminosity Function from Data Release Nine. ApJ **773,** 14. doi:10.1088/0004-637X/773/1/14. arXiv: 1210.6389 [astro-ph.CO] (Aug. 2013).

129. McGreer, I., Moustakas, J. & Schindler, J. *simqso: Simulated quasar spectra generator* Astrophysics Source Code Library, record ascl:2106.008. June 2021. ascl: 2106.008.

130. De la Cruz, R. *Realismo en el espectro continuo de los cúasares en las regiones de Ly-α y Ly-β simulados en DESI* MA thesis (Universidad de Guanajuato, 2020).

131. Prochaska, J. X., Madau, P., O'Meara, J. M. & Fumagalli, M. Towards a unified description of the intergalactic medium at redshift z $\approx$ 2.5. *Mon. Not. Roy. Astron. Soc.* **438,** 476–486. doi:10.1093/mnras/stt2218. arXiv: 1310.0052 [astro-ph.CO] (2014).

132. Carswell, R. F. *et al.* The absorption line profiles in Q 1101-264. ApJ **278,** 486–498. doi:10.1086/161815 (Mar. 1984).

133. Rudie, G. C., Steidel, C. C., Shapley, A. E. & Pettini, M. The Column Density Distribution and Continuum Opacity of the Intergalactic and Circumgalactic Medium at Redshift <z> = 2.4. ApJ **769,** 146. doi:10.1088/0004-637X/769/2/146. arXiv: 1304.6719 [astro-ph.CO] (June 2013).

134. Font-Ribera, A. & Miralda-Escudé, J. The effect of high column density systems on the measurement of the Lyman-α forest correlation function. J. Cosmology Astropart. Phys. **2012,** 028. doi:10.1088/1475-7516/2012/07/028. arXiv: 1205.2018 [astro-ph.CO] (July 2012).

135. Guo, Z. & Martini, P. Classification of Broad Absorption Line Quasars with a Convolutional Neural Network. *Astrophys. J.* **879,** 72. doi:10.3847/1538-4357/ab2590. arXiv: 1901.04506 [astro-ph.GA] (2019).

136. Landy, S. D. & Szalay, A. S. Bias and Variance of Angular Correlation Functions. ApJ **412,** 64. doi:10.1086/172900 (July 1993).

137. J. Zou et al. The DESI Damped Lyα System Survey: Data Release 1. *in preparation* (2024).

138. Parks, D., Prochaska, J. X., Dong, S. & Cai, Z. Deep learning of quasar spectra to discover and characterize damped Lyα systems. MNRAS **476,** 1151–1168. doi:10.1093/mnras/sty196. arXiv: 1709.04962 [astro-ph.GA] (May 2018).

139. Wang, B. *et al.* Deep Learning of Dark Energy Spectroscopic Instrument Mock Spectra to Find Damped Lyα Systems. ApJS **259,** 28. doi:10.3847/1538-4365/ac4504. arXiv: 2201.00827 [astro-ph.GA] (2022).

140. Ho, M.-F., Bird, S. & Garnett, R. Detecting multiple DLAs per spectrum in SDSS DR12 with Gaussian processes. MNRAS **496,** 5436–5454. doi:10.1093/mnras/staa1806. arXiv: 2003.11036 [astro-ph.CO] (Aug. 2020).

141. Filbert, S. *et al.* Broad Absorption Line Quasars in the Dark Energy Spectroscopic Instrument Early Data Release. arXiv: 2309.03434 [astro-ph.CO] (Sept. 2023).

142. du Mas des Bourboux, H. *et al. picca: Package for Igm Cosmological-Correlations Analyses* Astrophysics Source Code Library, record ascl:2106.018. June 2021. ascl: 2106.018.

143. Cuceu, A. *et al.* The Alcock–Paczyński effect from Lyman-α forest correlations: analysis validation with synthetic data. *Mon. Not. Roy. Astron. Soc.* **523,** 3773–3790. doi:10.1093/mnras/stad1546. arXiv: 2209.12931 [astro-ph.CO] (2023).

144. Ade, P. A. R. *et al.* Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594,** A13. doi:10.1051/0004-6361/201525830. arXiv: 1502.01589 [astro-ph.CO] (2016).

145. Youles, S. *et al.* The effect of quasar redshift errors on Lyman-α forest correlation functions. *Mon. Not. Roy. Astron. Soc.* **516,** 421–433. doi:10.1093/mnras/stac2102. arXiv: 2205.06648 [astro-ph.CO] (2022).

146. Bault, A. *et al. Impact of Systematic Redshift Errors on the Cross-correlation of the Lyman-α Forest with Quasars at Small Scales Using DESI Early Data* Feb. 2024. doi:10.48550/arXiv.2402.18009. arXiv: 2402.18009 [astro-ph.CO].

147. Ennesser, L., Martini, P., Font-Ribera, A. & Pérez-Ràfols, I. The impact and mitigation of broad-absorption-line quasars in Lyman α forest correlations. *Mon. Not. Roy. Astron. Soc.* **511,** 3514–3523. doi:10.1093/mnras/stac301. arXiv: 2111.09439 [astro-ph.CO] (2022).

148. Dembinski, H. & et al., P. O. scikit-hep/iminuit. doi:10.5281/zenodo.3949207. https://doi.org/10.5281/zenodo.3949207 (Dec. 2020).

149. James, F. & Roos, M. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput. Phys. Commun.* **10,** 343–367. doi:10.1016/0010-4655(75)90039-9 (1975).

150. Cuceu, A., Font-Ribera, A. & Joachimi, B. Bayesian methods for fitting Baryon Acoustic Oscillations in the Lyman-α forest. *JCAP* **07,** 035. doi:10.1088/1475-7516/2020/07/035. arXiv: 2004.02761 [astro-ph.CO] (2020).

151. Herrera-Alcantar, H. K. *The Dark Energy Spectroscopic Instrument: Simulations of the first year Lyman-α quasars observations* M.S. thesis (Universidad de Guanajuato, 2020). http://repositorio.ugto.mx/handle/20.500.12059/5236.

152. Prochaska, J. X. & Herbert-Fort, S. The SDSS damped Ly-alpha survey: Data Release 1. *Publ. Astron. Soc. Pac.* **116,** 622. doi:10.1086/421985. arXiv: astro-ph/0403391 (2004).

153. Prochaska, J. X., Herbert-Fort, S. & Wolfe, A. M. The sdss damped lya survey: data release 3. *Astrophys. J.* **635,** 123–142. doi:10.1086/497287. arXiv: astro-ph/0508361 (2005).

154. Noterdaeme, P., Petitjean, P., Ledoux, C. & Srianand, R. Evolution of the cosmological mass density of neutral gas from Sloan Digital Sky Survey II - Data Release 7. A&A **505,** 1087–1098. doi:10.1051/0004-6361/200912768. arXiv: 0908.1574 [astro-ph.CO] (Oct. 2009).

155. Noterdaeme, P. *et al.* Column density distribution and cosmological mass density of neutral gas: Sloan Digital Sky Survey-III Data Release 9. A&A **547,** L1. doi:10.1051/0004-6361/201220259. arXiv: 1210.1213 [astro-ph.CO] (Nov. 2012).

156. Garnett, R., Ho, S., Bird, S. & Schneider, J. Detecting damped Ly α absorbers with Gaussian processes. *Mon. Not. Roy. Astron. Soc.* **472,** 1850–1865. doi:10.1093/mnras/stx1958. arXiv: 1605.04460 [astro-ph.CO] (2017).

157. Ho, M.-F., Bird, S. & Garnett, R. Detecting Multiple DLAs per Spectrum in SDSS DR12 with Gaussian Processes. *Mon. Not. Roy. Astron. Soc.* **496,** 5436–5454. doi:10.1093/mnras/staa1806. arXiv: 2003.11036 [astro-ph.CO] (2020).

158. Ho, M.-F., Bird, S. & Garnett, R. Damped Lyman-α absorbers from Sloan digital sky survey DR16Q with Gaussian processes. MNRAS **507,** 704–719. doi:10.1093/mnras/stab2169. arXiv: 2103.10964 [astro-ph.GA] (Oct. 2021).

159. Chabanier, S. *et al.* The Completed Sloan Digital Sky Survey IV Extended Baryon Oscillation Spectroscopic Survey: The Damped Lyα Systems Catalog. *Astrophys. J. Supp.* **258,** 18. doi:10.3847/1538-4365/ac366e. arXiv: 2107.09612 [astro-ph.CO] (2022).

160. Pâris, I. *et al.* The Sloan Digital Sky Survey Quasar Catalog: Twelfth data release. A&A **597,** A79. doi:10.1051/0004-6361/201527999. arXiv: 1608.06483 [astro-ph.GA] (Jan. 2017).

161. Pâris, I. *et al.* The Sloan Digital Sky Survey Quasar Catalog: Fourteenth data release. A&A **613,** A51. doi:10.1051/0004-6361/201732445. arXiv: 1712.05029 [astro-ph.GA] (May 2018).

162. Lyke, B. W. *et al.* The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release. ApJS **250,** 8. doi:10.3847/1538-4365/aba623. arXiv: 2007.09001 [astro-ph.GA] (Sept. 2020).

163. García, L. Á. *et al.* Analysis of the impact of broad absorption lines on quasar redshift measurements with synthetic observations. arXiv: 2304.05855 [astro-ph.CO] (Apr. 2023).

List of Figures

List of Tables

UNIVERSIDAD DE GUANAJUATO

León, Gto. a 10 de junio de 2024.

Dr. Delepine, David Yves Ghislain
División de Ciencia e Ingenierías
DIRECTOR

Estimado Dr. Delepine:

Por medio de la presente me permito informarle que he leído la tesis doctoral intitulada **"Synthetic Lyman-α quasar spectra datasets as a tool for cosmological studies"** , escrita por el estudiante Hiram Kalid Herrera Alcantar para llenar uno de los requisitos para obtener el grado de Doctor en Física por la Universidad de Guanajuato.

Después de la lectura de esta tesis, discutí con el alumno su contenido y le dí algunas sugerencias para mejorarla, mismas que el alumno incluyó en la versión final.

Considero que la tesis doctoral es de calidad y es original y qué por lo tanto está lista para ser defendida de manera oral por el estudiante.

Atentamente

Dra. Ana Laura Benavides Obregón
Profesora Invitada

División de Ciencias e Ingenierías, Campus León
Loma del Bosque 103, Col. Lomas del Campestre, León, Gto, México, C.P. 37150
Tel. 01 (477) 788-5100 Ext. 8422 Fax. Ext. 8410
www.ugto.mx

Universidad de Guanajuato

CAMPUS LEÓN
DIVISIÓN DE CIENCIAS E INGENIERÍAS
DEPARTAMENTO DE FISICA

Asunto: Revisión de tesis de doctorado
León, Gto., Junio de 2024

DR. DAVID Y. G. DELEPINE
DIRECTOR DE LA DIVISIÓN DE CIENCIAS E INGENIERÍAS
PRESENTE

Por este medio le comunico que he revisado la tesis **Synthetic Lyman-α quasar spectra datasets as a tool for cosmological studies**, escrita por el M. en F. **Hiram Kalid Herrera Alcantar**, para efecto de presentarla para la obtención del grado de Doctorado en Física en la División de Ciencias e Ingenierías de la Universidad de Guanajuato.

El texto de la tesis se encuentra completo y se presentan resultados interesantes y bien explicados sobre el uso de bases de datos sintéticas relacionadas con la fisica de cuasares de Lyman-alpha. En mi opinión, las hipótesis de trabajo y el análisis teórico se conectan bien con los resultados presentados, incluso dentro del contexto numérico. Igualmente, he podido ver que el texto fue modificado por el autor para reflejar las sugerencias y comentarios que le fueron expresados durante la revisión. En mi opinión la tesis cumple con los elementos necesarios para ser defendida ante el comité sinodal asignado en la fecha próxima que sea acordada de manera conjunta.

Agradeciendo su amable atención, aprovecho la ocasión para enviarle un cordial saludo.

ATENTAMENTE
“LA VERDAD OS HARÁ LIBRES”

DR. LUIS ARTURO UREÑA LÓPEZ
PROFESOR TITULAR C

Instituto de Física

Instituto de Fisica, UNAM.
Edificio Moshinsky, Oficina 226
mmaganav@fisica.unam.mx
mmarianav@gmail.com
Telefono: +52 (55) 56-22-50-00
Extensión: 2614

Dr. David Yves Ghislain Delepine
Director de la División de Ciencias e Ingenierías Campus León, Universidad de Guanajuato
PR E S E N T E

Por este medio, me permito informarle que he leído y revisado la tesis titulada *"Synthetic Lyman-α quasar spectra datasets as a tool for cosmological studies,"* que realizó el estudiante **Hiram Kalid Herrera Alcantar** como requisito para obtener el grado de Doctor en Física.

Considero que el trabajo de tesis realizado por Hiram reúne los requisitos necesarios de calidad e interés académico para que sea defendido en un examen de grado, razón por la cual extiendo mi aval para que así se proceda.

Sin más que agregar, agradezco su atención y aprovecho la ocasión para enviarle un cordial saludo.

ATENTAMENTE

Dra. Mariana Vargas Magaña
Instituto de Fisica, UNAM

Alberto Diez Tejedor
Profesor Asociado C
Departamento de Física
DCI-León

León, Guanajuato; a 8 de junio de 2024

Dr. David Yves Ghislain Delepine
Director de la División de Ciencias e Ingenierías
Campus León, Universidad de Guanajuato
P R E S E N T E

Estimado Dr. Delepine,

Por este medio, me permito informarle que he leído y revisado la tesis titulada **"Synthetic Lyman-α quasar spectra datasets as a tool for cosmological studies,"** que realizó el estudiante **Hiram Kalid Herrera Alcantar** como requisito para obtener el grado de Doctor en Física.

Considero que el trabajo de tesis realizado por Hiram reúne los requisitos necesarios de calidad e interés académico para que sea defendido en un examen de grado, razón por la cual extiendo mi aval para que así se proceda.

Sin más que agregar, agradezco su atención y aprovecho la ocasión para enviarle un cordial saludo.

ATENTAMENTE
"LA VERDAD OS HARÁ LIBRES"

Dr. Alberto Diez Tejedor
Departamento de Física
DCI, Campus León

UNIVERSIDAD DE GUANAJUATO

León, Guanajuato, 5 de junio de 2024

Dr. David Delepine

Director

División de Ciencias e Ingenierías

PRESENTE

Por medio de la presente me permito informar que he leído la tesis titulada **"Synthetic Lyman-α quasar spectra datasets as a tool for cosmological studies"**, que para obtener el grado de Doctorado en Física ha sido elaborada por el **M. F. Hiram Kalid Herrera Alcantar**. En mi opinión, la tesis cumple con los requisitos de calidad correspondientes al grado académico al que se aspira. Las correcciones sugeridas por mi parte han sido atendidas, por lo cual recomiendo se proceda a la defensa de la tesis.

Sin más por el momento quedo a sus órdenes para cualquier aclaración.

Atentamente

Dr. Carlos Alberto Vaquera Araujo
Investigador por México Conahcyt
Departamento de Física
DCI, Campus León
vaquera@fisica.ugto.mx

Universidad de Guanajuato

CAMPUS LEÓN
DIVISIÓN DE CIENCIAS E INGENIERÍAS
DEPARTAMENTO DE FISICA

25 de Junio, 2024

DR. DAVID Y. G. DELEPINE
DIRECTOR DE LA DIVISIÓN DE CIENCIAS E INGENIERÍAS
PRESENTE

Por este medio le comunico que he revisado la tesis titulada **Synthetic Lyman-α quasar spectra datasets as a tool for cosmological studies**, escrita por el estudiante **Hiram Kalid Herrera Alcantar**, para efecto de presentarla y obtener el grado de Doctorado en Física. El texto de la tesis se encuentra completo.

En mi opinión, la tesis cumple con los elementos necesarios para ser defendida ante el comité sinodal designado en fecha próxima que sea acordada de manera conjunta.

Agradeciendo su amable atención, aprovecho la ocasión para enviarle un cordial saludo.

Alma X. González Morales
Departamento de Física
División de Ciencias e Ingenierías
Universidad de Guanajuato

Gustavo Niz Quevedo
Departamento de Física
División de Ciencias e Ingenierías

León, Gto., 29 de Mayo de 2024.

Dr. David Yves Ghislain Delepine
Director de la División de Ciencias e Ingenierías
Universidad de Guanajuato

Estimado Dr. David Yves Ghislain Delepine

Por medio de la presente le informo que he recibido, leído y revisado la tesis de Doctorado titulada "Synthetic Lyman-α quasar spectra datasets as a tool for cosmological studies" del alumno Hiram Kalid Herrera Alcantar, bajo la supervisión de la Dra. Alma X. González-Morales.

El trabajo cumple con los estándares requeridos para la obtención del grado, y apoyo la defensa del mismo en la fecha convenida.

Me pongo a su disposición para cualquier duda sobre la revisión de dicho trabajo de tesis.

Atentamente,

Gustavo Niz